Chapter 2
Health Data

Health data are diverse with multiple modalities. This chapter will introduce different types of health data, including structured health data (e.g., diagnosis codes, procedure codes) and unstructured data (e.g., clinical notes, medical images). We will also present the popular health data standards for representing those data.

2.1 The Growth of Electronic Health Records

Over the past decade, more and more health service providers worldwide have adopted electronic health record (EHR) systems to manage data about patients and records of health care services. For example, Fig. 2.1 shows the increase of national basic and certified EHR adoption rate over time according to the American Hospital Association Annual Survey. Here the basic EHR adoption curve corresponds to the EHR systems having basic EHR functions such as patient demographics, physician notes, lab results, medications, diagnosis, clinical and drug safety guidelines. A certified EHR just has to cover essential EHR technology that meets the technological capability, functionality, and security requirements adopted by the Department of Health and Human Services. From Fig. 2.1, we can see nearly all reported hospitals (96%) possessed a certified EHR technology by 2015. In 2015, 84.8% of hospitals adopted at least a Basic EHR system; this represents a ninefold increase since 2008. Thanks to the wide deployment of EHR systems, many healthcare institutions have collected diverse health data. This chapter provides an overview of health data: what different data types are available and how the data are collected, and by whom. All these data are potential inputs for training deep learning models for supporting diverse healthcare tasks.

C. Xiao, J. Sun, *Introduction to Deep Learning for Healthcare*,
https://doi.org/10.1007/978-3-030-82184-5_2

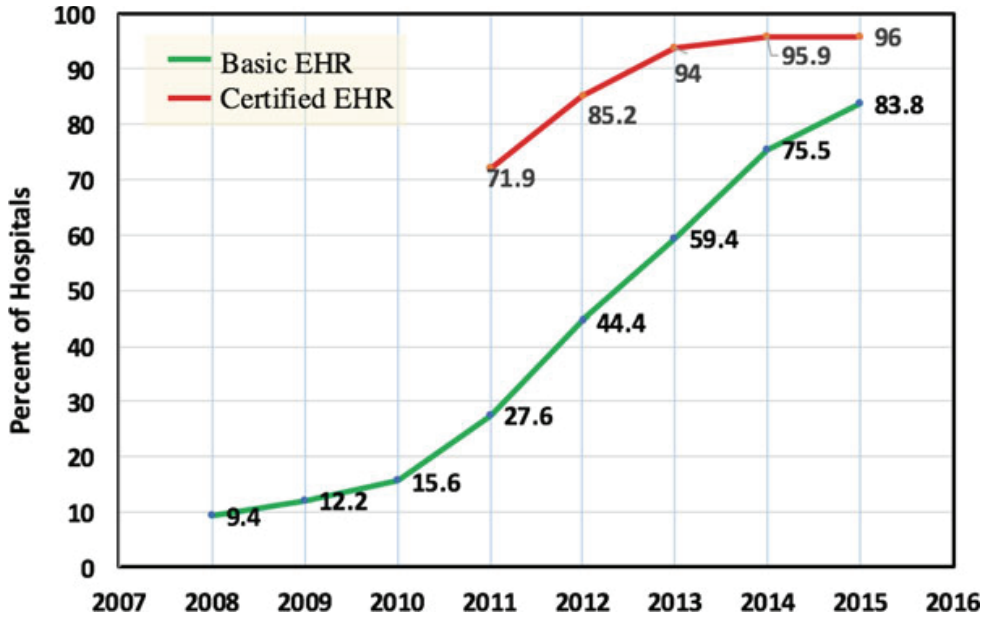

Fig. 2.1 Percentage of EHR system adoption over time. Basic EHR means EHR systems with a set of required functionalities such as patient demographics, physician notes, lab results, medications, diagnosis, clinical and drug safety guidelines. A certified EHR system means the hospital has the essential EHR technology certified by Department of Health and Human Services. Source: American Hospital Association Annual Survey

2.2 Health Data

Shifting from the traditional paper-based records to electronic records has generated a massive collection of health data, which created opportunities for enhanced patient care, data-driven care delivery, and accelerated healthcare research. According to the definition from Centers for Medicare & Medicaid Services (CMS)—a federal institution that administers government-owned health insurance services, EHR is "an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative, clinical data relevant to that person's care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports".[1] From the modeling perspective, EHR can be viewed as a longitudinal record of comprehensive medical services provided to patients and documentation of patient medical history. There are several important observations of EHR data:

1. EHR data are mostly managed by providers (although there is an ongoing movement to enable patients to augment additional data into their EHRs);
2. Each provider manages their own EHR systems; as a result, partial information about the same patient may scatter across EHR systems from multiple providers;

[1] https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html.

3. The main purpose of EHR data is to support accurate and efficient billing services, which creates challenges for other secondary use of EHR data such as research.

2.2.1 The Life Cycle of Health Data

Let us first introduce the key players in the healthcare industry and the life cycle of health data from the perspectives of those key healthcare players.

Key Healthcare Players There are diverse healthcare institutions that generate and manage health data.

- **Providers** are hospitals, clinics, and health systems, which provide healthcare services to patients. Providers use electronic health records (EHR) systems to capture everything that happened during patient encounters, such as diagnosis codes, medication prescription, lab and imaging results, and clinical notes. Providers interact with other players such as payers, pharmacies, and labs.
- **Payers** are entities that provide health insurance. They can be private insurance companies such as United Healthcare and Anthem. Or they can be public insurance programs owned by government entities such as MEDICARE and MEDICAID. Payers reimburse the full or partial cost associated with healthcare services to providers. Payers interact with providers and pharmacies via claims. More specifically, providers and pharmacies submit claims to corresponding payers from which they have medical insurance. Claims are usually structured data with diagnosis, procedure, and medication codes, and associated cost information.
- **Pharmacies** prepare and dispense medications often based on medication prescriptions. Pharmacies know what and when patients actually fill medications. Pharmacies produce pharmacy claims, which are also sent to payers of the patients for reimbursement.
- **Pharmaceutical companies** discover, develop, produce, and market drugs. They conduct clinical trials to validate new drugs. Pharmaceuticals generate experimental data for drug discovery and clinical trials data.
- **Contracted Research Organizations (CROs)** provide outsourced contract services to pharmaceutical companies such as pre-clinical and clinical research and clinical trial management. Depending on the services line, CROs produce various datasets that support pharmaceutical companies to get their drugs to market.
- **Government agencies** play multiple roles in the healthcare ecosystem. Food and Drug Administration (FDA) is the most important regulator to approve new drugs and monitor existing drugs. Centers for Disease Control and Prevention (CDC) is a public health institution focusing on monitoring, controlling, and preventing diseases. For example, government agencies worldwide have been collecting the reports from Spontaneous Reporting Systems (SRS) submitted by pharmaceutical companies, healthcare professionals, and consumers to facilitate

post-market drug surveillance. These SRSs have served as a cornerstone for post-marketing drug surveillance, and the FDA Adverse Event Reporting System (FAERS) is one of the most prominent SRSs.

- **Patients** are at the center of healthcare, who interact with all the other players. For example, the healthcare benefits of patients are usually covered by their payers. Patients are also increasingly empowered to produce and manage their own health-related data. Most EHR systems provide patient portals (e.g., Web-based or Apps) for patients to assess their own EHR data and interact with their healthcare providers. For example, with wearable sensors such as wristbands and smartphones, more people can have activity monitoring data such as movement and heart rates, which are essential to monitoring individual health.
- **Researchers** are an important group of individuals that try to push the frontier of medical and healthcare research. Healthcare researchers can have a diverse background, including medicine, biology, chemistry, engineering, and data science. On one end, they can be conducting basic research in biology and chemistry that can help discover new drugs in the future or understand the basics of disease mechanisms. On the other end, they can be analyzing EHR data to produce translational insights that immediately change clinical practice. Researchers produce medical literature and clinical guidelines, which are important data in itself.

Life Cycle of Health Data When a patient comes to a clinic or a hospital, an electronic health record will be created about this clinical encounter or visit. This record will be documented by doctors or nurses (or generally healthcare providers) to describe what happened during this visit. A **clinical note** will be created to describe the visit in narrative text. Then various **medical codes** will be assigned to this visit, including diagnosis codes, procedure codes, and medication prescriptions. **Lab and imaging tests** may be ordered, which are done either at the clinic or sent to an external lab. The **lab report** contains both structured data and unstructured text. After the clinical visit, a **medical claim** containing most structured medical codes will be filed to the payer, usually by the provider on behalf of the patient. Then the payer will verify and reimburse the associated cost to the providers or patients. Meanwhile, the patient may take the medication prescription to a pharmacy to fill their medication. The corresponding medication will be dispensed to the patient. Then pharmacy may file a **pharmacy claim** to the payer to obtain reimbursement of the medication. In parallel, to invent new drugs, pharmaceutical companies (pharma) often work with providers to recruit patients to participate in clinical trials. Many **clinical trials** are managed by external CRO for pharma. Once patients are enrolled in the trials, various measurements related to the drug's efficacy and safety will be collected as part of the trial results. The complete results and their analysis will be submitted to the FDA for approval. The new drugs can only be widely distributed once three phases of trials are conducted with positive results and the corresponding FDA's approval is acquired. Different kinds of health data and the associated players are illustrated in Fig. 2.2.

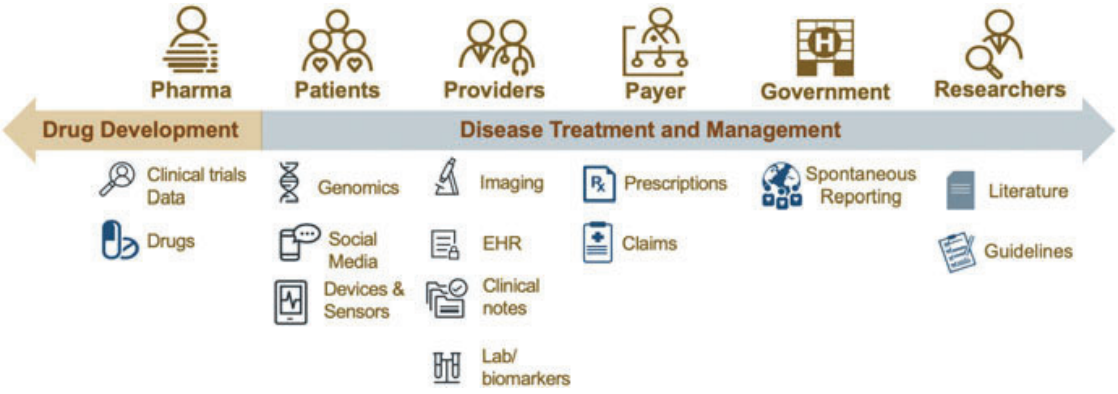

Fig. 2.2 Important players and the life cycle of health data

2.2.2 Structured Health Data

Structured data are common in healthcare, which are often represented as medical codes.

Various **medical codes** are used in both EHR and claim data, which usually follow common data standards. For example, diagnosis codes follow international disease classification (ICD); procedures use current procedure terminology (CPT) codes. The number of unique codes from each data standard is large and growing. For example, ICD version 9 (ICD-9) has over 13,000 diagnosis codes, while ICD-10 has over 68,000 diagnosis codes. Each encounter is only associated with a few codes. The resulting data are high-dimensional but sparse. A simple and direct way to represent such data is to create a *one-hot vector* for a medical code and a *multi-hot vector* for the patient with multiple codes as shown in Fig. 2.3. For example, to represent diagnosis information in an encounter, one can create a 68,000-dimensional binary vector where each element is a binary indicator of a corresponding diagnosis code. If only one dimension is one and zeros otherwise, we call it *one-hot vector*. If multiple ones are present, it is a *multi-hot vector*. As we will show in later chapters, such multi-hot vectors can be improved with various deep learning approaches to construct appropriate lower-dimensional representation.

Most medical codes follow certain standards such as ICD for diagnosis, CPT for procedures, and NDC for drugs, which will be explained later in Health data standards. Most of these medical codes are organized in hierarchies defined by medical ontologies such as CCS codes. These hierarchies are instrumental in constructing more meaningful and low-dimensional input features for the deep learning models. For example, instead of directly treating each ICD-10 code as a feature, we can group them into a few hundred CCS codes,[2] higher disease categories and treat each CCS code as a feature.

[2] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

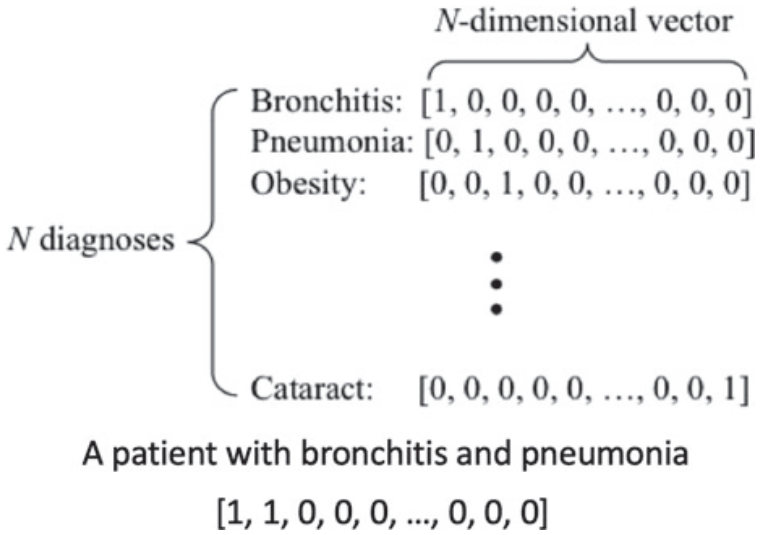

Fig. 2.3 Examples of one-hot vectors of medical codes, a multi-hot vector for a patient. Here 1 or 0 indicates the presence or absence of a particular diagnosis

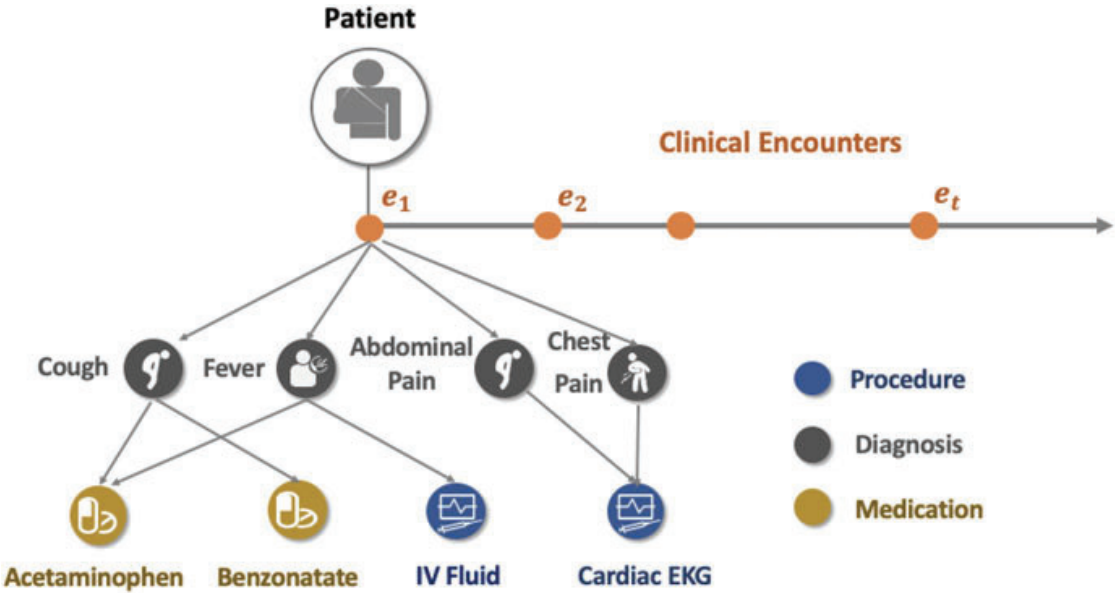

Fig. 2.4 In EHR data, medical codes are structured hierarchically with heterogeneous relations, e.g., medication *Acetaminophen* and procedure *IV fluid* are correlated, while both occur due to the same diagnosis *Fever*

All medical codes are interconnected in a hierarchical way within an encounter, as depicted by Fig. 2.4. In particular, EHR data can be seen as a collection of individual patient records, where each patient has a sequence of encounters over time. Within an encounter, multiple interrelated medical codes are assigned. Once a diagnosis code is assigned, the corresponding medication orders and procedure orders are then created. Each medication order contains information such as the medication code, start date, end date, and instructions. Procedure orders contain the procedure code and possible lab results. As shown by Fig. 2.4, procedures such

Table 2.1 Structured data in EHR

Codes	Standard	Example
Diagnoses	International Classification of Disease (ICD)	S06.0x1A
Procedures	Current Procedural Terminology (CPT)	70010
Labs	Logical Observation Identifiers Names & Codes (LOINC)	5792-7
Medication	RxNorm, ATC	
Demographics	NA	Gender, age
Vital signs	NA	Numeric
Behavioral status	NA	Smoking status

Table 2.2 Unstructured components of EHRs

Component	Description	Format
Discharge summary	Summary of an encounter	Text
Progress notes	Timestamped detailed description	Text
Radiology report	Summary of imaging test	Image+text
Electroencephalogram (EEG)	Brain activity monitoring	Time series
Electrocardiogram (ECG)	Heart monitoring	Time series

as *Cardiac EKG* come with several results (QRS duration, Q-T interval, notes), but *IV Fluid* does not. Note that some diagnoses might not be associated with any medication or procedure orders.

Other Structured Data In addition to medical codes, there are other structured data such as patient demographics (e.g., age and gender), vital signs (e.g., blood pressure), and social history (e.g., smoking status). These are also stored as structured fields in the EHR database. Various data standards are used for documenting different medical codes, as summarized in Tables 2.1 and 2.2.

Challenges Several challenges exist in analyzing structured data:

1. *Sparsity:* Raw data such as medical codes are often high dimensional and extremely sparse. For example, with 68,000 possible ICD-10 codes, each patient encounter may have only a few codes present;
2. *Temporality:* Health data often have an important time aspect that needs to be modeled. For example, EHR data of a patient may contain multiple visits over time. Discovering temporal relations is crucial for making an accurate assessment or prediction of any health outcome.

2.2.3 Unstructured Clinical Notes

Clinical notes are recorded for various purposes, either for the documentation of an encounter in discharge summaries, describing the reason and use of prescriptions in

medication notes, interpreting the results of medical images with radiology reports, or analyzing the results from lab tests pathology reports. The reports include different sets of functional components, thus yield various difficulties in understanding. There are growing interests in using machine learning for these unstructured clinical notes, especially deep learning methods for automated indexing, abstracting, and understanding. Also, some works focus on automated classification of clinical free text to standardized medical codes, which are important for generating appropriate claims for that visit [114]. For example, a progress note is one type of clinical note commonly used in describing clinical encounters. A progress note is usually structured in four sections with the acronym SOAP:

- *Subjective* part describes what the patient tells you;
- *Objective* part presents the objective findings such as lab test results and imaging rests;
- *Assessment* part provides the diagnosis of the encounter;
- *Plan* part describes the treatment plan for the patient.

There are many different types of notes, including Admission notes, Emergency department (ED) notes, Progress notes, Nursing notes, Radiology reports, ECG reports, Echocardiogram reports, Physician notes, Discharge summary, and Social work notes. Each type can be written in a very different format with different lengths and quality.

Challenges Several key technical challenges exist in mining clinical text data:

1. *High dimension*—the number of unique words in clinical text corpus is large. Many words are acronyms, which are important to understand using their context.
2. *External knowledge*—in addition to clinical text in EHR data, there is a large amount of medical knowledge encoded in the text, such as medical literature and clinical guidelines. It is important to be able to incorporate that knowledge in modeling EHR data.
3. *Privacy*—Besides technical challenges, it is challenging to access clinical text due to privacy concerns as the data are very sensitive and affect individual privacy. As a result, very limited clinical notes are openly accessible for method development. And the volume of the shared data is also limited due to largely privacy concerns.

2.2.4 Continuous Signals

With an increasing number of new medical and wellness sensors being created, there will be more continuous signals as part of health data. Most commonly collected continuous signals in clinics include electrocardiogram (ECG) and electroencephalogram (EEG). ECG measures the electrical activities by placing electrodes on the chest, arms, and legs. In contrast, EEG measures the electrical activities of a brain via electrodes that are placed on the scalp. Both ECG and EEG are routine clin-

ical measurements captured by multiple sensors (multi-channel) in high frequency (e.g., 200 Hz). Human doctors currently conduct most interpretations of ECG and EEG recordings (sometimes with machines' help). Beyond clinical signals like ECG and EEG, there are also consumer-grade wearables that monitor movement and heart rate. Several such sensors, including an accelerometer, gyroscope, and heart rate sensor, are already built into a typical smartphone.

Challenges Two significant challenges involved in analyzing continuous signals:

1. *Noise*—Sensor data often contain a significant amount of noises, making many simple models fail. For example, ECG and EEG signals can easily interfere with physical movement and power lines.
2. *Lack of labels*—To create direct clinical values, the continuous signals have to be mapped to meaningful phenotypes such as disease diagnoses. However, learning such a mapping requires sufficient labeled data that map continuous signals to phenotypes, which can be difficult to obtain. Because it may require significant time from human experts to produce the labels. Sometimes the data generated by new types of sensors are not well studied before, making it difficult for anyone to produce accurate labels.

2.2.5 Medical Imaging Data

Medical imaging is about creating a visual representation of the human body for diagnostic purposes. Various imaging technologies have been introduced, such as X-ray radiography, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (e.g., echocardiography). The resulting data are 2D images or 3D representations from multiple 2D images and videos. Medical imaging data are stored and managed by a separate system called *Picture archiving and communication system* (PACS). The images themselves are stored and transmitted in DICOM format (Digital Imaging and Communications in Medicine). Given the raw imaging data, radiologists read and mark the images and write a text report (radiology reports) to summarize the findings. The radiology reports are often copied into the EHR systems so that clinicians and patients can access the findings of the imaging tests. Thanks to the digitization of the radiology field, a large number of high-resolution images and their corresponding phenotypes (labels) are available for modeling. Thus there is tremendous excitement in using deep learning for radiology tasks [65, 101].

Challenges Data quality and lack of reliable and detailed labels are still challenges in analyzing such data. As the raw input images are very large (high-dimensional), it demands a sufficient sample size to train accurate and generalizable models.

2.2.6 Biomedical Data for In Silico Drug Discovery

In silico drug discovery is about using computational methods to create and select molecule compounds for new drugs. The data used *in silico* modeling include molecule compound graphs or sequences, protein target sequences, genome sequences, and disease-related knowledge graphs. For a molecule compound, a molecular graph represents the structural formula of the compound in terms of nodes (atoms) and edges (chemical bonds). Many computational models are not directly conducted on their original graph form but using specialized encodings:

- **SMILES**: A special string encoding called simplified molecular-input line-entry system (SMILES) is used to represent the chemical compounds.
- **ECPF/FCFP**: One can describe the compound using the functional connectivity fingerprints, including the Extended Connectivity Fingerprint (ECPF), which is a list of integer identifiers or a fixed-size bit string, where each identifier and bit corresponds to a neighbor substructure; and similarly the Functional-Class Fingerprint (FCFP), which is a variant of ECFP, integrates the functional features to the ECFP fingerprint.
- **Functional genes** are characterized by a unique gene identifier in a gene database such as GeneBank. There are also text descriptions of gene functions.

More generally, various knowledge graphs are constructed to represent the relations among entities within and/or across different data types. For example, the human disease network is a taxonomy of diseases themselves, the disease-drug network, disease-gene network. These networks describe the association between disease and drugs, as well as diseases and genes.

Challenges Several significant challenges in analyzing data associated with *in silico* studies.

1. *Incorporating domain knowledge*—Data for in-silico studies are mainly chemical data and various knowledge graphs. All the representations have a precise meaning in their domains. It is crucial to understand and incorporate their domain knowledge.
2. *Interpretable models*—Since the purpose of such data is to support drug discovery, it is important to provide more convincing evidence and interpretable explanation of each prediction.

2.3 Health Data Standards

Next, we overview a set of commonly used standards in healthcare data.

- **ICD** stands for International Classification of Diseases, which is a set of codes that represents diseases, symptoms, and clinical procedures [12]. ICD codes follow a hierarchical structure where related codes can be grouped into a higher

level category. ICD codes are a widely used international standard maintained by World Health Organization (WHO). The latest version is ICD-11 as of 2020. Most of the world is currently using ICD-10. For example, "I50" corresponds to the ICD-10 category for heart failure, I50.2 is Systolic (congestive) heart failure, and I50.21 is Acute systolic (congestive) heart failure. ICD codes are used to represent disease diagnosis in EHR and claims data. Most EHR data will have ICD codes either in ICD-9 or ICD-10 format. An ICD9 code has up to 5 digits. The first digit is either alphabetic or numeric, and the remaining digits are numeric. For example, the ICD-9 code for *Diabetes mellitus without mention of complications* is 250.0x. The first three digits of an ICD-9 code corresponding to the disease category. And the last 1 or 2 digits reflect the subcategories of the disease. Besides numeric codes, ICD-9 codes can have an initial letter of V or E. For example, V85.x is an ICD-9 code for body mass index (BMI). In particular, V85.0 corresponds to BMI<19, V85.1 BMI between 19 and 25, and V85.2x indicates BMI>25. ICD-10 codes are more granular than ICD-9 codes. Each ICD-10 code has up to 7 digits. The first digit is always a letter; the second digit is always numeric; the third to seventh digits are alphanumeric. For example, E10.9 is the ICD-10 code for *Type 1 diabetes mellitus without complications*.

- **CPT** corresponds to Current Procedural Terminology, which is a standard created and copyrighted by American Medical Association. CPT codes represent medical services and procedures that doctors can document and bill for payment. CPT codes also follow a hierarchical structure. For example, CPT codes between 99201 and 99215 correspond to Office/other outpatient services, while a high-level category 99201–99499 corresponds to codes for evaluation and management. Like ICD codes, CPT codes are commonly present in structured EHR data.
- **NDC** codes are 10- or 11-digit national drug codes, which are managed by Food and Drug Administration (FDA). It consists of three segments: labeler, product, and package. For example, 0777-3105-02 is an NDC code where 0777 corresponds to labeler Dista Products Company, 3105 maps to the product Prozac, and 02 indicates the package of 100 capsules in 1 bottle. The same drugs with different packages will have different codes. From an analytic modeling perspective, NDC codes are probably too specific to be used directly as features.
- **LOINC** is a terminology standard for lab tests. LOINC stands for Logical Observation Identifiers Names and Codes (LOINC). Like other standards, LOINC has LOINC codes and associated descriptions of the code. To support lab tests, LOINC description follows a specific format with six parts: (1) COMPONENT (ANALYTE): The substance being measured or observed; (2) PROPERTY: The characteristic of the analyte; (3) TIME: The interval of time of the observation; (4) SYSTEM (SPECIMEN): The specimen upon which the observation was made; (5) SCALE: How the observation is quantified: quantitative, ordinal, nominal; (6) METHOD: how the observation was made (which is an optional part). For example, LOINC code 806-0 is the lab test of the manual count of white blood cells in the cerebral spinal fluid specimen. The different parts of the description are Component:Leukocytes, Property:NCnc (Number concentration),

Time:Pt(Point in time), System:CSF (Cerebral spinal fluid), Scale:Qn (Quantitative), Method:Manual count. LOINC codes demonstrate even structured data can encode multiple aspects of information.

- **SNOMED CT** is a comprehensive ontology of all medical terminologies. SNOMED CT stands for Systematized Nomenclature Of Medicine Clinical Terms. The core components of SNOMED include concept codes (or SNOMED ID), concept description, and the relationships between concepts. For example, 22298006 is the SNOMED code for a heart attack; there are various heart attack descriptions, including Myocardial infarction, Infarction of heart, Cardiac infarction, and Heart attack, Myocardial infarction (disorder), and Myocardial infarct. There are many associated relationships to heart attacks, such as a parent relationship to Ischemic heart disease (disorder), a child relationship to Acute myocardial infarction (disorder), an associated-morphology relation to infarct, and a finding-site relation to myocardium structure. Computationally SNOMED CT provides a large knowledge graph that connects many clinical terminologies, which can be extremely useful to combine with EHR data for predictive model building.

In addition to the data standards, various mapping software packages can process different types of healthcare data.

- **CCS** codes are a hierarchical categorization of ICD and CPT codes maintained by the Healthcare Cost and Utilization Project (HCUP). The purpose of CCS codes is to aggregate detailed ICD and CPT codes into clinically meaningful groups to support better statistical analysis. CCS codes have much fewer categories than the original ICD and CPT codes. For example, there are about a few hundred CCS codes, while ICD and CPT have tens of thousands of codes. From a machine learning modeling perspective, CCS codes can often be more informative than raw ICD and CPT codes.
- **RxNorm** is a terminology system for drugs and the associated software for mapping various mentions of drugs to normalized drug names. RxNorm group synonyms of drug expressions into drug concept. Each concept is assigned with a normalized name. In addition to drug name normalization, RxNorm also creates relations for each drug. For example, The drug "Naproxen 250 MG Oral Tablet" has a dose relation to "Oral Tablet", an ingredient relation of "Naproxen" and an *is-a* relation to "Naproxen Oral Tablet."
- **UMLS** standards for Unified Medical Language System, which integrates many biomedical terminologies. UMLS has three knowledge sources: (1) **Metathesaurus** integrates many terminologies including ICD, CPT, LOINC, SNOMED, and RxNorm, normalizes concepts and provides concept unique identifiers (CUIs) for each concept; (2) **Semantic Network** specifies all the relations among concepts; (3) **Lexical Tools** normalizes strings, handles lexical variants and provides basic natural language capability for biomedical text.

2.4 Exercises

1. What are the most useful health data for predicting patient outcome (e.g., mortality)?
2. What are the most accessible health data? And why?
3. What are the most difficult health data (to access and to model)?
4. What are the important health data that are not described in this chapter?
5. Which of the following is NOT true about electronic health records (EHR)?

   (a) EHR data from a single hospital consists of complete clinical history from each patient.
   (b) Outpatient EHR data are viewed as point events
   (c) EHR data contain longitudinal patient records.
   (d) Inpatient EHR data are viewed as interval events.

6. Which of the following is not true about clinical notes?

   (a) They can provide a detailed description of patient status.
   (b) Most EHR systems provide clinical notes functionality.
   (c) Clinical notes can contain sensitive protected health information.
   (d) Because of its unstructured format, it is easy for computer algorithms to process the notes

7. Which of the following are the limitations of claims data?

   (a) Coding errors can commonly occur in the claims data.
   (b) Since claims data are for billing purposes, they do not accurately reflect patient status.
   (c) Claims data are rare and difficult to find.
   (d) Claims data of a patient are often incomplete because they can go to different hospitals.

8. Which of the following is not true?

   (a) EHR are richer than claims.
   (b) EHR captures the medication prescription information but does not capture whether the prescription are filled.
   (c) Continual signals are rarely collected in hospitals.
   (d) Continuous signals provide objective assessments of patients.

9. Which of the following are not imaging data?

   (a) X-rays
   (b) Computed tomography
   (c) Electrocardiogram
   (d) Magnetic resonance imaging

10. What is true about medical literature data?

    (a) They are difficult to parse because of the natural language format.

(b) They are noisy and often low in quality.
(c) They often contain sensitive patient identifiers.
(d) They are in a machine-friendly format.

11. Which of the following is a medical ontology for medications?

(a) CPT codes
(b) RxNorm
(c) SNOMED codes
(d) MESH terms

12. Which of the following is not clinical trial data?

(a) Trial protocols
(b) Trial eligibility criteria
(c) Data in clinical trial management systems
(d) Electronic health records

13. Which of the following is not true about drug data?

(a) Drugs are often represented in molecule structures.
(b) Drug data are standard.
(c) Drug data are often encoded in 3D molecule structures.
(d) ChEMBL is a large bioactivity database.