Data Standards and Terminology Including Biomedical Ontologies

Spiros Denaxas and Christian Stoeckert

Abstract

Electronic health records are routinely collected as part of care and have variable data types, quality and structure. As a result, there is a need for standardization of clinical data from health records if these are to be used in software applications for data mining and/or machine learning and artificial intelligence approaches. Clinical terminologies and classification systems are available that can serve as standards to enable the harmonization of disparate data sources. In this chapter, we discuss different types of biomedical semantic standards including medically-relevant ontologies, their uses, and their limitations. We also discuss the application of semantic standards in order to provide features for use in machine learning particularly with respect to phenotypes. Finally, we discuss potential areas of improvement for the future such as covering genotypes and steps needed.

Keywords

EHR · Semantic standards · Ontologies · Clinical terminologies

1 Introduction

Medicine is inherently a data driven practice. The widespread adoption of electronic health record (EHR) systems in the US and Europe has rapidly increased the amounts of health related data that are electronically generated and captured during routine interactions of patients with the healthcare system [1]. Patient interactions with the healthcare system, for example an outpatient visit or a hospital admission, generate a substantial amount of data and metadata. These data are organized, recorded and curated using different healthcare standards and clinical terminologies. Healthcare standards enable the storage and exchange of health information across healthcare providers while clinical terminologies enable the systematic and standardized recording of healthcare information.

Before raw EHR data can be used as input features into analytical AI pipelines, a significant amount of preprocessing and harmonization must occur. For example, multiple EHR sources utilizing different clinical terminologies to record information need to be aligned to a common format. With unstructured data, such

S. Denaxas (✉)
Institute of Health Informatics, University College London, London, UK
e-mail: s.denaxas@ucl.ac.uk

Data Science Centre, British Heart Foundation, London, UK

C. Stoeckert
University Of Pennsylvania, Philadelphia, USA
e-mail: stoeckrt@pennmedicine.upenn.edu

F. W. Asselbergs et al. (eds.), *Clinical Applications of Artificial Intelligence in Real-World Data*,
https://doi.org/10.1007/978-3-031-36678-9_3

as information recorded in clinical text, Natural Language Processing (NLP) approaches can be deployed to extract clinically-meaningful markers and transform them into input features for the pipeline (this process is often referred as entity extraction). Finally, depending on the purpose of each dataset, different biases might exist in the data which need to be accounted for. For example, administrative hospitalization EHR might be influenced by local coding guidelines which in turn affect the observed data recording patterns and need to be accounted for prior to analyses.

The outcome of such a data preprocessing pipeline would be features extracted from complex, multidimensional EHR that can be used as input features to AI analytical approaches. Extracting clinically important markers from complex EHR (e.g. disease status, biomarkers, prescriptions, procedures, symptoms etc.) is often referred to as phenotyping [2]. The main objective therefore of this chapter is to provide a succinct overview of the main clinical terminologies used to record EHR data, their characteristics, and outline different approaches for creating and evaluating EHR-derived phenotypes. The methods outlined here will cover a set of phenotyping methodologies ranging from rule-based deterministic algorithms, to aggregated coding systems and finally to more complex learnt representations).

1.1 The Need for Standards and Their Application

Standards in the context of this chapter are defined as common representations of data. They may be approved by a governing body (e.g., ISO dates [3]) or they may simply represent established formats (Variant Call Format (VCF) files of genomic variants [4]). For clinical terminologies, standards may be mandated by the government, institution (e.g., National Institutes of Health [5]), or professional societies. Terminologies may be developed by communities adhering to common principles (e.g., OBO Foundry [6]).

Standards are needed in healthcare to effectively find, store and analyze data. If different representations are used for syntax and semantics, there is no guarantee that the data used for analysis is complete or can be correctly combined across sources. If data is not standardized, it can prevent information sharing and reuse of clinical data [7]. Often data can come from different systems even within the same institution and mappings to a common standard is needed. The challenge however is that there may be competing standards (PCORNet [8], FHIR [9], OMOP [10] and others).

To understand the need and application of standards consider the how, when and why data are generated during routine clinical interactions. Data can be generated by physicians and healthcare professionals entering data directly in the EHR for patient care. Data can also be generated through clinical coding for billing and reimbursement purposes can subsequently be used for research FInally, data may be processed and curated through clinical audits for registries, quality of care, and planning. Each of these may use different systems with different representations that need to be harmonized before analyses. Furthermore, different stakeholders and systems may attempt to record the same information but choose different levels of granularity. For example a healthcare professional might record detailed information on presenting signs, symptoms and diagnoses while a clinical coder might distill this information into a small number of terminology concepts. A coding system therefore should be able to account for these differences and enable their harmonization.

In this section, we will provide working definitions of key concepts in data standardization to guide understanding of the different options and complexity of choosing and applying a standard. An excellent review of different semantic representations is provided elsewhere [11]. Here we highlight commonly used and mentioned types of semantic standards and provide details of different levels of standardization and what they offer.

Semantic standards can be understood at three levels of abstraction of increasing

complexity. The first is as entities (terms) that make up classes (general concepts) and instances (individual members) of those classes. For example, 'heart failure' is a class whereas 'the first heart failure diagnosis of a patient' is an instance of that class. Most usage of terminologies and ontologies is at this first level where terms are used as annotations. A second level is the organization of the entities into structures such as hierarchies or assertions and statements including axioms and logical definitions. Hierarchies can be simple taxonomies ('heart failure' is-a 'disorder of cardiac function') or can be poly-hierarchies to accommodate a term having more than one parent. The structure of assertions/statements can be in the form of triples: subject-predicate-object such as: 'heart failure' 'occurs in' 'heart structure'. These structures provide the ability to connect concepts in a defined manner. The third level is the representational model adhering to open versus closed worlds and languages such as Resource Description Framework (RDF) [12], Web Ontology Language (OWL), Simple Knowledge Organization System (SKOS) [13] and schema languages as part of the Semantic Web [14]. These can be employed in messaging systems such as FHIR and Common Data Models (CDM) like Observational Medical Outcomes Partnership (OMOP). The products of semantic standards can be browsed in repositories such as the NCBO BioPortal [15] or used in knowledge bases linking classes or terms (TBox) to instances or assertions (ABox) about data [16].

Clinical classification systems, medical ontologies, and clinical terminologies make use of these different levels of abstraction. In this context, ontologies are distinguished by formal relations between entities and use of logical definitions or axioms. The W3C provides approved standards such as OWL and a query language (SPARQL) which enables ontologies based on these standards to be programmatically accessed and searched [12]. Clinically relevant ontologies include the Disease Ontology [17], the Drug Ontology [18], and the Ontology for Biomedical Investigations [19] which can be used to link to diagnoses, medications, and lab tests respectively in EHR. Those ontologies, which are part of the OBO Foundry, not only provide hierarchies for capturing related data at different levels of granularity but also have formal links to other external ontologies (e.g., for chemicals in CHEBI [20]) that can be used to connect them and build more complex knowledge structures (e.g., classes of drugs containing chemicals that are used as an antineoplastic agent).

Multiple ontologies or terminologies may be needed to annotate or instantiate data. When this is done, care should be taken to avoid conflicts or redundancies, i.e. the chosen terminologies should be semantically interoperable. This however is not guaranteed if different sources of terms are used as they can have different contexts and thus different meanings. With the OBO Foundry, the objective is that adhering ontologies are semantically consistent with respect to meaning of terms and use of relations.

2 Controlled Clinical Terminologies and Clinical Classifications Systems

EHR provide the infrastructure for healthcare professionals to record information that is relevant for the care of a patient. This information can include symptoms, medical history information on the patient or their direct family, laboratory or anthropometric measurements, prescriptions, diagnoses, and surgical procedures. The data recorded within the EHR allow healthcare professionals to assess and treat a patient but are also widely used for a number of other purposes (often referred to as secondary uses) such as reimbursement, planning, billing, auditing and research. Although clinical terminologies and clinical classification systems are often used interchangeably, they serve two distinct purposes [21]. The former were created to enable healthcare professionals to record information that is pertinent to clinical care. The latter are a tool which enables the aggregation and statistical analyses of health information (Table 1).

Controlled clinical terminologies (also referred to as controlled clinical ontologies, controlled medical ontologies, controlled medical vocabularies) are the basic building blocks used by healthcare professionals to record information within an EHR system. The main purpose of clinical terminologies is to enable the consistent and systematic recording of clinical data and metadata which in turn are used for direct patient care. As a result, controlled clinical terminologies often encapsulate a wide and diverse set of domains and healthcare-related actions.

The US Bureau of Labor Statistics defines classification systems as "ways of grouping and organizing data so that they may be compared with other data" [22]. In the context of medicine, clinical classification systems enable the aggregation and analysis of data related to health can healthcare on a national or international level. One of the most commonly used classification systems worldwide is the ICD-10 which is maintained by the World Health Organization (WHO) [23]. Clinical classification systems are also used for other secondary purposes, one of the most common being reimbursement where clinical data get transformed and aggregated into a clinical classification system. The process by which raw data are transformed into ICD codes is defined as *coding*. The WHO defines coding as "the translation of diagnoses, procedures, comorbidities and complications that occur over the course of a patient's encounter from medical terminology to an internationally coded syntax" [24].

2.1 SNOMED-CT

SNOMED Clinical Terms (SNOMED-CT) is a controlled clinical terminology providing a set of hierarchically-organized, machine-readable codes, terms, synonyms and definitions used to record information related to health and healthcare within EHR information systems [25]. SNOMED-CT is maintained and distributed by the International Health Terminology Standards Development Organisation (IHTSDO). SNOMED-CT was created in 1965 as the Systematized Nomenclature of Pathology (SNOP) which in turn evolved in the SNOMED Reference Terminology (SNOMED-RT) and finally merged with the NHS Clinical Terms Version 3 (Read codes Version 3, CTV3) [26] to create SNOMED-CT in 2002. Similarly to ICD, different countries can maintain their own versions of SNOMED-CT that are tailored to their local healthcare system or needs; in the UK for example, the National Health Service (NHS) maintains a UK version of SNOMED-CT [27] that is used.

SNOMED-CT consists of three components [28] which are explained below (Tables 2 and 3):

Table 1 Comparison between ICD-10 (statistical classification system) and SNOMED (clinical terminology

	ICD-10	SNOMED-CT
Type	Clinical classification system	Controlled clinical terminology
N concepts	10^4	10^5
Relationships	A concept has a single parent	A concept can have multiple hierarchical relationships and multiple parents
Age related diagnoses	Information on age is encapsulated within the term	The term used is the same across all ages and the age of onset is derived by the date of diagnosis and the age of the patient
Fidelity	Information organized in mutually exclusive categories with generic "not otherwise specified" or "not elsewhere classified" terms used to record information if required	NOS/NEC are not used in SNOMED-CT

1. **Concept**: Every SNOMED-CT concept represents a unique clinical meaning and has a unique numerical identifier which is persistent across the ontology and can be used to reference the concept. The January 2021 version of SNOMED CT contains approximately 350,000 concepts.
2. **Description**: Each SNOMED-CT concept has a unique description, the Fully Specified Name (FSN), which offers an unambiguous description of the concept's meaning. Additionally, a concept can have one or more synonym terms (*Synonyms*) which are associated with the concept.
3. **Relationship**: SNOMED-CT offers several types of relationships between concepts in order to enable logical computable definitions of complex concepts. The terminology contains approx 1.4 million relationship entries defining these. All concepts are organized in an acyclic hierarchy using the "is-a" relationship and concepts can have multiple parents (as opposed to most statistical classification systems that only support a single parent child relationship). Additionally, SNOMED-CT offers more than 60 other relationship types for example finding site, causative agent and associate morphology.

Table 2 Example SNOMED-CT concept core components

Fully specified name	Heart failure (disorder)
SCTID	84,114,007
Synonyms	Heart failure Myocardial failure Weak heart Cardiac failure Heart failure (disorder) HF—Heart failure Cardiac insufficiency
Parents	Disorder of cardiac function (disorder)
Finding site (relationship)	Heart structure

Subsets of SNOMED-CT components (e.g. of concepts, their descriptions and relationships between concepts) can be represented using a standardized approach enabled by *Reference Sets*. Reference Sets are commonly used to provide a subset of the terminology that has been curated to serve a particular process and to enable the standardized recording of clinical data at the point of care (for example, in an emergency department [29]).

Precoordination and Postcoordination of Concepts

Complex clinical information can often be represented by combinations of multiple concepts or modifiers for example "chronic migraine", "major depression with psychotic symptoms", "recurrent deep vein thrombosis" or "accidental burning or scalding caused by boiling water". The concepts can contain information on the chronicity, morphology, severity or other aspect

Table 3 Selected top level SNOMED hierarchy concepts and examples (based on the SNOMED-CT UK hierarchy [30])

Name	Example
Body structure	83,419,000 Femoral vein structure (body structure)
Clinical finding	1,362,251,000,000,108 Recurrent bleeding from nose (finding)
Environment or geographical location	285,201,006 Hospital environment (environment)
Event	419,620,001 Death (event)
Procedure	414,089,002 Emergency percutaneous coronary intervention (procedure)
Qualifier value	90,734,009 Chronic (qualifier value)
Situation with explicit context	406,140,001 Discussion about care plan with family (situation)
Social concept	236,324,005 Factory worker (occupation)
Specimen	258,583,001 Bone marrow clot sample (specimen)
Staging and scales	1,077,341,000,000,105 Diagnosing Advanced Dementia Mandate Tool (assessment scale)
Substance	447,208,001 Alcaftadine (substance)

```
284196006 | burn of skin | :
   116676008 | associated morphology | = 80247002 | third degree burn injury |
 , 272741003 | laterality | = 7771000 | left |
 , 246075003 | causative agent | = 47448006 | hot water |
 , 363698007 | finding site | = 83738005 | index finger structure
```

Fig. 1 Example of the SNOMED-CT compositional syntax used to create a postcoordinated concept which can be used to record a third degree burn caused by hot water of the left index finger (*Source* WIkipedia [33])

of the information being recorded. Clinical terminologies have traditionally tried to enable the recording of such information by creating and providing terms for them, a process often referred to as *precoordination*. The core SNOMED-CT ontology contains approx 350.000 precoordinated concepts as they are available upfront for use. The use of precoordinated concepts greatly improves the storage and manipulation of information as it effectively reduces the dimensionality of the data (i.e. the use of one concept versus the use of multiple concepts to record the same data point).

The approach of offering precoodinated concepts for any possible combination of clinically meaningful concepts however does not scale given the complex, highly heterogeneous, and multidisciplinary nature of health and healthcare. For example, it would be unreasonable to expect a precoordinated term for "third degree burn of left index finger caused by hot water". To enable the recording of complex concepts in a machine readable manner, SNOMED-CT offers a compositional grammar (Fig. 1) [31] that can be used to combine multiple concepts together into clinical expressions that are more accurate as opposed to only using a single concept. The created concepts are referred as "postcoordinated" as they are not available upfront in the ontology but have been created a posteriori. Postcoordination however introduces considerable challenges, both in terms of data recording by clinicians, storage and retrieval of information and significantly increases the complexity of the underlying data [32].

2.2 International Classification of Disease (ICD)

The 10th edition of the International Classification of Disease (ICD), commonly referred to as ICD-10, is maintained and published by the WHO and is the most commonly used statistical classification system worldwide. The 11th edition of ICD (ICD-11) officially came was adopted by the 72nd World Health Assembly in 2019 and came into effect on 1st January 2022 [34]. While the WHO maintains the core ICD system, individual countries often develop and deploy their own branches which are adapted to their own needs by often including additional terms or other changes. For example, secondary healthcare providers in the US make use of ICD-10 Clinical Modifications (ICD-10-CM) for discharge summaries and reimbursement purposes which is maintained by the US Centres for Disease Control and Prevention (CDC) [35] (Table 4).

ICD-10 is organized in 21 top level chapters which represent disease systems and are denoted by roman numerals e.g. chapter IX contains terms related to diseases of the circulatory system. Terms within each chapter are often organized in one or more blocks which define a range of codes e.g. block I20-I25 encapsulates terms related to ischaemic heart disease. Individual ICD-10 terms can have up to seven characters. All ICD-10 codes always begin with a letter that is associated with the chapter which they belong to e.g. codes related to circulatory diseases begin with the character "I". This is followed by one or two numbers which further specify the category of the diagnosis. The remaining characters indicate the disease aetiology, anatomic site, severity or other relevant clinical detail. The first three characters are separated by the remaining characters by a decimal character. Within individual codes, the 5th or 6th character length codes represent terms with the highest level of specificity. In certain disease chapters such as obstetrics, a 7th character can be used to denote the type of encounter (e.g. initial vs. subsequent). Within three and four character codes,

Table 4 Comparison of ICD-10 and ICD-10-CM terms used to record heart failure

ICD-10-CM	ICD-10
I50.1 Left ventricular failure, unspecified I50.2 Systolic (congestive) heart failure I50.20 Unspecified systolic (congestive) heart failure I50.21 Acute systolic (congestive) heart failure I50.22 Chronic systolic (congestive) heart failure I50.23 Acute on chronic systolic (congestive) heart failure I50.3 Diastolic (congestive) heart failure I50.30 Unspecified diastolic (congestive) heart failure I50.31 Acute diastolic (congestive) heart failure I50.32 Chronic diastolic (congestive) heart failure I50.33 Acute on chronic diastolic (congestive) heart failure I50.4 Combined systolic (congestive) and diastolic (congestive) heart failure I50.40 Unspecified combined systolic (congestive) and diastolic (congestive) heart failure I50.41 Acute combined systolic (congestive) and diastolic (congestive) heart failure I50.42 Chronic combined systolic (congestive) and diastolic (congestive) heart failure I50.43 Acute on chronic combined systolic (congestive) and diastolic (congestive) heart failure I50.8 Other heart failure I50.81 Right heart failure I50.810 …… unspecified I50.811 Acute right heart failure I50.812 Chronic right heart failure I50.813 Acute on chronic right heart failure I50.814 …… due to left heart failure I50.82 Biventricular heart failure I50.83 High output heart failure I50.84 End stage heart failure I50.89 Other heart failure I50.9 Heart failure, unspecified	I50.0 Congestive heart failure I50.1 Left ventricular failure I50.9 Heart failure, unspecified

a "rubric" often denotes a number of other diagnostic terms that are associated with that code such as other related syndromes, synonyms for the disease or common terms. Finally, when a conclusive diagnosis was not possible, for example when the presenting symptoms did not meet the diagnostic criteria for one of the existing defined codes in the hierarchy, generic, broader "Not Otherwise Specified" codes can be used e.g. "I50.9 Heart failure, unspecified".

Working Across ICD Versions

A key challenge of working with longitudinal data that has been recorded using ICD is dealing with different versions of the same coding system e.g. ICD-9 and ICD-10 [36]. Major new versions of an ontology will, by definition, contain a substantial amount of new entities that can be used to record information (e.g. ICD-9-CM contains 13,000 codes while ICD-10-CM contains 68,000 codes) which will often be organized differently. As a result, there are often many additional codes (and often in higher fidelity than before) that can be used to define clinical concepts.

To enable this translation of data between ICD versions, the Centers for Medicare & Medicaid Services (CMS) curates and provides a set of General Equivalent Maps (GEMs, these are often referred to as *crosswalks*) [37]. GEMs can provide forward maps (e.g. ICD-9-CM to ICD-10-CM) and backward maps (e.g. ICD-10-CM to ICD-9-CM). The use of GEMs however is not straightforward as newer concepts that exist in ICD-10-CM might not always exist in ICD-9-CM and some ICD-9-CM concepts might map to a combination of more than one ICD-10-CM codes. For example, the ICD-9-CM

code "250.10 Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrolled" can potentially map to "E11.69 Type 2 diabetes mellitus with other specified complication" or "E13.10 Other specified diabetes mellitus with ketoacidosis without coma" ICD-10-CM codes. In their work, Fung et al. [38] show that the majority of ICD-10-CM codes are not represented in the forward map, and a significant portion of ICD-9-CM codes (25%) are not represented in the backward map e.g. the backward map provides 78,034 unique pairs of ICD-9-CM and ICD-10-CM codes (over three times more than the forward map), of which only 18,484 pairs (23.7%) are also found in the forward map.

Other Clinical Terminologies and Ontologies
A plethora of other clinical ontologies and terminologies exist that are used to record information related to health and healthcare. Information on drugs and medical devices is captured by RxNorm [39] in the US and the Dictionary of Medicines and Devices (DM+D) in the UK [40]. Similarly, surgical procedures and interventions in the US are recorded using the Current Procedural Terminology (CPT) [41] while in the UK using the Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures, 4th revision (OPCS-4) classification which is maintained by the NHS [42]. Molecular pathology testing data and metadata can be standardized by using the LOINC (Logical Observation Identifier Names and Codes) ontology [43]. Semi-structured data, such as reports from investigative radiology procedures, can also contain clinically significant information that can benefit from harmonization and a bespoke ontology, RadLex, has been created to enable the standardized recording of entities [44].

3 Defining Diseases in Electronic Health Records

EHR data offer a rich source of information for research as they capture a diverse set of information on diagnoses, laboratory measurements, procedures, symptoms, medication prescriptions alongside metadata related to healthcare delivery such as referrals. The process of transforming raw EHR data and extracting clinical information for research is referred as *phenotyping* and involves the creation of algorithms (referred to as *phenotyping algorithms)* that can either be deterministic (rule based) or probabilistic [2]. Rule-based algorithms often combine multiple pieces of information, alongside logic rules, to identify patients with a given disease [42].

The use of EHR however for research is associated with significant challenges as the data are often fragmented, recorded using different controlled clinical terminologies and have variable data quality and completeness [45]. Importantly, the purpose and processes in which data are generated and captured varies significantly. For example, primary care EHR are generated by the clinician for direct patient care but are influenced by local clinical guidelines while secondary care claims data are recorded by clinical codes which in turn operate based on a predefined coding protocol. This in turn might influence how data are recorded within each source and how data should be merged across sources [46]. For example, a study comparing the recording of non-fatal myocardial infarctions (AMI) in linked data from primary care, hospitalization records and a myocardial ischaemia national audit observed that only a third of AMI events were recorded in all three sources [47]. As a result of these challenges, researchers must both study the underlying processes that generate the data and perform robust validation across multiple layers of evidence.

3.1 The Need for Aggregated Code Representations

One of the many challenges of working with coded data is that related concepts (e.g. all manifestations of a particular disease) can be fragmented across the terminology used to record information. For example, tuberculosis related diagnoses in ICD-10 occur in four different ICD chapters (e.g. infections, skin diseases, diseases of the genitourinary system and diseases

of the musculoskeletal and connective tissue). Furthermore, when working with longitudinal data, researchers have to deal with changes within clinical terminologies and changes related to new major versions of ontologies such as the transition of ICD-9-CM to ICD-10-CM or SNOMED-CT concepts becoming inactive and replaced by newer alternative concepts. As a result, the creation of phenotyping algorithms to define diseases in complex EHR becomes significantly more challenging and requires a significant amount of resources.

To enable the scalable definition of diseases in EHR, using all available ICD diagnosis codes, a layer above source ICD codes has been developed by Bastarache et al. [48] that provides phenotype codes (*phecodes*) grouppings. Phecodes were originally developed in ICD-9-CM and derived partially from the Agency for Healthcare Research and Quality Clinical Classification Software for ICD-9-CM (CCS) [49]. Phecodes are manually curated, hierarchically organized groupings of ICD codes aiming to capture common adult diagnoses to facilitate phenome-wide genetic association studies (PheWAS) [50]. Phecodes version 1.2 condenses roughly 15,500 ICD-9-CM codes and 90,000 ICD-10-CM codes into 1867 phecodes. Subsequent research mapped phecodes to ICD-10 and ICD-10-CM codes [51] and phecodes have been shown to produce robust genotype–phenotype associations compared with other relevant approaches [52].

3.2 Bridging Molecules to Phenotypes

Phenotypes typically require aggregation of structured data fields in clinical records as described in the preceding section. Phenotypic inferences can be made based on an interpretation of lab test results, medications prescribed, diagnoses, and clinical notes. To make such inferences using a programmatic approach requires connecting phenotypes to structured representations of those clinical record elements. The OBO Foundry includes relevant ontologies for bridging molecules to phenotypes. The Chemical Entities of Biological Interest (ChEBI) ontology covers molecules and their roles while the Drug Ontology (DrON) captures the relationships between the molecules defined in ChEBI and the drugs where the molecules are active ingredients and also links to RxNorm terms (from the National Library of Medicine [39]). The human disease ontology (DO) has database-cross references to ICD-9 and ICD-10 codes as well as to SNOMED. The Monarch Disease Ontology (MonDO [53]) connects DO with additional disease resources (e.g., Orphanet [54], OMIM [55]). Genotyping results can be interpreted through the Gene Ontology (GO [56]) to identify the processes affected by mutations. The Ontology for Biomedical Investigations (OBI) [19] can be used to link lab test results with specimens and assays. Anatomy-based data can be interpreted through Uberon [57], a species neutral anatomy ontology, or the Foundational Model of Anatomy (FMA [58]) which is focused on human anatomy. The Human Phenotype Ontology [59] provides representation of phenotypes and connects to many of these listed OBO Foundry ontologies as well as clinical terminologies.

4 Application of Standards to Aid Machine Learning

Representing words as numerical vectors based on the contexts in which they appear has become the de facto method of natural language processing approaches. A survey of word embeddings for clinical text provides some good pointers on other approaches [60].

Learnt representations of controlled clinical terminologies can be used as the basis for features in machine learning. In order to utilize the information located in free text, it has to be converted to structured representation. This transformation however needs to take into consideration the structure of the clinical terminology itself as it provides essential contextual information. Artificial Intelligence approaches are increasingly being used to learn and predict

phenotypes. An example of deep learning applied to EHR records is BEHRT [61], a deep neural sequence transduction model capable of simultaneously predicting the likelihood of 301 phenotypes (originally developed in the CALIBER resource [62]) in a patient's future visits. When trained and evaluated on the data from nearly 1.6 million individuals, BEHRT was able to show a striking improvement in terms of average precision scores for different tasks over the existing state-of-the-art deep EHR models. In addition to its scalability and improved accuracy, BEHRT enables personalized interpretation of its predictions. Its flexible architecture enables it to incorporate multiple heterogeneous concepts (e.g., diagnosis, medication, measurements, and more) to further improve the accuracy of its predictions; its (pre-)training results in disease and patient representations can be useful for future studies (i.e., transfer learning).

Tensor factorization methods such as Limestone and Granite have also provided phenotype predictions [63, 64]. EHR data do not always directly and reliably map to medical concepts that clinical researchers need or use. Some recent studies have focused on EHR-derived phenotyping, which aims at mapping the EHR data to specific medical concepts; however, most of these approaches require labor intensive supervision from experienced clinical professionals. Furthermore, existing approaches are often disease-centric and specialized to the idiosyncrasies of the information technology and/or business practices of a single healthcare organization. Limestone [64], a nonnegative tensor factorization method to derive phenotype candidates with virtually no human supervision. Limestone represents the data source interactions naturally using tensors (a generalization of matrices) and investigates the interaction of diagnoses and medications. The resulting tensor factors are reported as phenotype candidates that automatically reveal patient clusters on specific diagnoses and medications. Using the proposed method, multiple phenotypes can be identified simultaneously from data.

Standards in the form of biomedical ontologies can be used directly for analysis of annotated data. The most visible form of this approach is in the enrichment analysis of gene expression data using annotations of proteins and genes with the Gene Ontology. Those analyses while very successful do not take advantage of relationships encoded in the ontologies. Recent work has been done however using ontology-based network analysis and visualization for COVID-19 analysis [65]. In a similar vein, in the AI-driven cell ontology brain data standards project, ontologies are being used to capture results of analysis and learn more through reasoning [66].

Knowledge graphs provide the ability to connect clinical terminologies and encodings in EHR with biomedical ontologies and standards. For example, a knowledge graph framework has been developed for COVID-19 focused around molecular and chemical information, enabling users to conduct complex queries over relevant biological entities as well as machine learning analyses to generate graph embeddings for making predictions. This framework can also be applied to other problems in which siloed biomedical data must be quickly integrated for different research applications, including future pandemics [67].

5 Future directions

The proper use of standards is an active area of research. In a recent call for proposals, the issue of relating real-world data (RWD) (e.g., EHR, claims, and digital health technologies) between different sources was raised as not just an issue of mapping but also transforming the data and the underlying definition of its meaning as these can be similar but not identical. Even if standards are used, proper use of data from multiple sources will rely heavily on human interpretation and efforts are still needed for fully reliable computer-driven approaches. In this chapter, the emphasis has been on data for phenotyping. The same concerns and considerations about the choice and application of standards need to be applied for genotyping and genomics. Linkages of this type of data to clinical terminologies are

either non-existent or in their infancy. There are standards for file formats and some relevant OBO Foundry ontologies exist (e.g., OBI, Sequence Ontology[68]) which should aid the ultimate goal of combining phenotyping and genotyping/genomics.

A fundamental difference between clinical terminologies/coding systems such as SNOMED-CT and ICD with OBO Foundry ontologies such as the Basic Formal Ontology (BFO) or the Disease Ontology (DO) is the modeling approach. SNOMED and ICD are representing information collected by a health care worker whereas BFO and DO are representing what happened or exists in the world. The former fits well with data models while the latter provides a common grounding in reality. It remains a challenge to leverage the benefits of both clinical standards like SNOMED-CT and OBO Foundry ontologies. SNOMED has greater adoption in the clinical area but lacks the semantic rigor and breadth (for example in genomic technologies) than OBO Foundry ontologies. The use of database cross-references in OBOF ontologies to SNOMED-CT does provide a bridge.

Resources for further reading:
We provide below several resources for further reading on topics covered in this chapter:

- Bodenreider and colleagues [69] provide an excellent overview and discussion of recent developments in SNOMED-CT, LOINC and RxNorm.
- Aspden and colleagues discuss the topic of healthcare data standards in depth and provide examples of their application in healthcare [7].
- Standards are by their nature about classes of concepts. However, when working with RWD, attention needs to be placed on their application to instances to establish when the diagnosis or even the patient being referred to is the same or different. This topic is covered in detail by Ceuster [70].
- Practical applications and theoretical background for applied ontology especially in the biomedical area can be found in Smith, Arp, and Spears Building Ontologies with Basic Formal Ontology [71].
- Hemingway and colleagues provide a detailed overview with examples on how electronic health records are utilized for early and late translational cardiovascular research [72].

References

1. The health information technology for economic and clinical health act (HITECH act). PsycEXTRA Dataset. American Psychological Association (APA); 2009. https://doi.org/10.1037/e500522017-001.
2. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21:221–30.
3. ISO 8601-1:2019. In: ISO [Internet]. 2019 [cited 31 Jan 2022]. Available: https://www.iso.org/standard/70907.html.
4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
5. National Institutes of Health (NIH). In: National Institutes of Health (NIH) [Internet]. [cited 31 Jan 2022]. Available: https://www.nih.gov/.
6. Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database . 2021;2021. https://doi.org/10.1093/database/baab069.
7. Institute of Medicine (US) Committee on Data Standards for Patient Safety, Aspden P, Corrigan JM, Wolcott J, Erickson SM. Health Care Data Standards. National Academies Press (US); 2004.
8. McGlynn EA, Lieu TA, Durham ML, Bauck A, Laws R, Go AS, et al. Developing a data infrastructure for a learning health system: the portal network. J Am Med Inform Assoc. 2014;21:596–601.
9. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. ieeexplore.ieee.org; 2013. p. 326–31.
10. OMOP Common Data Model. [cited 31 Jan 2022]. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/.
11. Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. J Biomed Inform. 2019;100S: 100002.
12. McGuinness DL, Van Harmelen F, Others. OWL web ontology language overview. W3C recommendation. 2004;10: 2004.

13. Miles A, Bechhofer S. SKOS simple knowledge organization system reference. W3C Recommendation. 2009 [cited 22 Feb 2022]. Available: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:66505.
14. Semantic web - W3C. [cited 22 Feb 2022]. Available: https://www.w3.org/standards/semanticweb/.
15. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011;39:W541–5.
16. Wikipedia contributors. Abox. In: Wikipedia, The Free Encyclopedia [Internet]. 19 Nov 2021. Available: https://en.wikipedia.org/w/index.php?title=Abox&oldid=1056049124.
17. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43:D1071–8.
18. Hogan WR, Hanna J, Joseph E, Brochhausen M. Towards a consistent and scientifically accurate drug ontology. CEUR Workshop Proc. 2013;1060:68–73.
19. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The ontology for biomedical investigations. PLoS ONE. 2016;11:e0154556.
20. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016;44:D1214–9.
21. Giannangelo K. Healthcare code sets, clinical terminologies, and classification systems, 3rd ed. American Health Information Management Association; 2014.
22. Classification Systems : U.S. Bureau of Labor Statistics. 30 Sep 2015 [cited 14 Jan 2022]. Available: https://www.bls.gov/opub/hom/topic/classification-systems.htm.
23. ICD-10 Version:2019. [cited 14 Jan 2022]. Available: https://icd.who.int/browse10/2019/en.
24. Nouraei SAR, Hudovsky A, Virk JS, Chatrath P, Sandhu GS. An audit of the nature and impact of clinical coding subjectivity variability and error in otolaryngology. Clin Otolaryngol. 2013;38:512–24.
25. Benson T. Principles of health interoperability HL7 and SNOMED. Springer London; 2010.
26. Read Codes—NHS Digital. [cited 5 Mar 2021]. Available: https://digital.nhs.uk/services/terminology-and-classifications/read-codes.
27. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform. 2013;46:87–96.
28. Kalet IJ. Chapter 4—Biomedical Information Access. In: Kalet IJ, editor. Principles of Biomedical Informatics. 2nd ed. San Diego: Academic Press; 2014. p. 397–478.
29. Hansen DP, Kemp ML, Mills SR, Mercer MA, Frosdick PA, Lawley MJ. Developing a national emergency department data reference set based on SNOMED CT. Med J Aust. 2011;194:S8-10.
30. NHS Digital. The NHS Digital SNOMED CT Browser. [cited 14 Jan 2022]. Available: https://termbrowser.nhs.uk/?.
31. Compositional Grammar—Specification and Guide—Compositional Grammar - SNOMED Confluence. [cited 14 Jan 2022]. Available: https://confluence.ihtsdotools.org/display/DOCSCG/Compositional+Grammar+-+Specification+and+Guide.
32. Karlsson D, Nyström M, Cornet R. Does SNOMED CT post-coordination scale? Stud Health Technol Inform. 2014;205:1048–52.
33. Wikipedia contributors. SNOMED CT. In: Wikipedia, The Free Encyclopedia [Internet]. 23 Dec 2021. Available: https://en.wikipedia.org/w/index.php?title=SNOMED_CT&oldid=1061690432.
34. ICD-11. [cited 22 Feb 2022]. Available: https://icd.who.int/en.
35. ICD-ICD-10-CM - International classification of diseases, tenth revision, clinical modification. 11 Feb 2022 [cited 22 Feb 2022]. Available: https://www.cdc.gov/nchs/icd/icd10cm.htm.
36. Cartwright DJ. ICD-9-CM to ICD-10-CM codes: what? why? how? Adv Wound Care. 2013;2:588–92.
37. ICD-10-CM and ICD-10 PCS and GEMs Archive. [cited 22 Feb 2022]. Available: https://www.cms.gov/Medicare/Coding/ICD10/Archive-ICD-10-CM-ICD-10-PCS-GEMs.
38. Fung KW, Richesson R, Smerek M, Pereira KC, Green BB, Patkar A, et al. Preparing for the ICD-10-CM transition: automated methods for translating ICD codes in clinical phenotype definitions. EGEMS (Wash DC). 2016;4:1211.
39. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof. 2005;7:17–23.
40. Spiers I, Goulding J, Arrowsmith I. Clinical terminologies in the NHS: SNOMED CT and dm+ d. British J Pharmacy. 2017;2:80–7.
41. Association AM. Current procedural terminology: CPT. Am Med Ass. 2007.
42. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS ONE. 2014;9: e110900.
43. Huff SM, Rocha RA, McDonald CJ, De Moor GJ, Fiers T, Bidgood WD Jr, et al. Development of the logical observation identifier names and codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5:276–92.
44. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics. 2006;26:1595–7.

45. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20:117–21.
46. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019;26:1545–59.
47. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346: f2350.
48. Bastarache L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. Annu Rev Biomed Data Sci. 2021;4:1–19.
49. HCUP-US Tools & Software Page. [cited 22 Feb 2022]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp.
50. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31:1102–11.
51. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med Inform. 2019;7: e14325.
52. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS ONE. 2017;12: e0175508.
53. Vasilevsky N, Essaid S, Matentzoglu N, Harris NL, Haendel M, Robinson P, et al. Mondo disease ontology: harmonizing disease concepts across the world. CEUR Workshop Proceedings. CEUR-WS; 2020. Available: http://ceur-ws.org/Vol-2807/abstractY.pdf.
54. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28:165–73.
55. McKusick VA. Mendelian inheritance in man and its online version. OMIM Am J Hum Genet. 2007;80:588–604.
56. Gene Ontology Consortium. The gene ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49:D325–34.
57. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012;13:R5.
58. Cook DL, Mejino JLV, Rosse C. The foundational model of anatomy: a template for the symbolic representation of multi-scale physiological functions. Conf Proc IEEE Eng Med Biol Soc. 2004;2004:5415–8.
59. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83:610–5.
60. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Inform. 2019;100S: 100057.
61. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep. 2020;10:7155.
62. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Health. 2019;1:e63–77.
63. Henderson J, Ho JC, Kho AN, Denny JC, Malin BA, Sun J, et al. Granite: diversified, sparse tensor factorization for electronic health record-based phenotyping. In: 2017 IEEE international conference on healthcare informatics (ICHI); 2017. p. 214–23.
64. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. J Biomed Inform. 2014;52:199–211.
65. Wang Z, He Y. Precision omics data integration and analysis with interoperable ontologies and their application for COVID-19 research. Brief Funct Genom. 2021;20:235–48.
66. Aevermann BD, Novotny M, Bakken T, Miller JA, Diehl AD, Osumi-Sutherland D, et al. Cell type discovery using single-cell transcriptomics: implications for ontological representation. Hum Mol Genet. 2018;27:R40–7.
67. Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, et al. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns (N Y). 2021;2: 100155.
68. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The sequence ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6:R44.
69. Bodenreider O, Cornet R, Vreeman DJ. Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018;27:129–39.
70. Ceusters, W. The place of Referent Tracking in Biomedical Informatics. 2020. https://doi.org/10.31219/osf.io/q8hts.
71. Arp R, Smith B, Spear AD. Building ontologies with basic formal ontology. MIT Press; 2015.
72. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. Eur Heart J. 2018;39:1481–95.