ANNUAL REVIEWS

Annual Review of Biomedical Data Science

A Census of Disease Ontologies

Melissa A. Haendel,[1,2] Julie A. McMurry,[1] Rose Relevo,[1] Christopher J. Mungall,[3] Peter N. Robinson,[4] and Christopher G. Chute[5]

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon 97239, USA; email: haendel@ohsu.edu

[2]Linus Pauling Institute, Oregon State University, Corvallis, Oregon 97331, USA

[3]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

[4]The Jackson Laboratory, Farmington, Connecticut 06032, USA

[5]School of Medicine, School of Public Health, and School of Nursing, Johns Hopkins University, Baltimore, Maryland 21205, USA

Annu. Rev. Biomed. Data Sci. 2018. 1:305–31

First published as a Review in Advance on May 9, 2018

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-080917-013459

Keywords

disease, ontology, phenotype, precision medicine, disease taxonomy

Abstract

For centuries, humans have sought to classify diseases based on phenotypic presentation and available treatments. Today, a wide landscape of strategies, resources, and tools exist to classify patients and diseases. Ontologies can provide a robust foundation of logic for precise stratification and classification along diverse axes such as etiology, development, treatment, and genetics. Disease and phenotype ontologies are used in four primary ways: (*a*) search, retrieval, and annotation of knowledge; (*b*) data integration and analysis; (*c*) clinical decision support; and (*d*) knowledge discovery. Computational inference can connect existing knowledge and generate new insights and hypotheses about drug targets, prognosis prediction, or diagnosis. In this review, we examine the rise of disease and phenotype ontologies and the diverse ways they are represented and applied in biomedicine.

INTRODUCTION

Ontologies are knowledge classifications and have arisen from the need for tools to represent, query, and analyze data and knowledge. Ontology as a discipline goes back at least as far as Aristotle (384–322 BC), who developed conceptual taxonomies that are in some ways similar to modern bio-ontologies (1). More recently, computer scientists adopted the word "ontology" to denote a computational representation describing specific domains of knowledge (2). The distinction and evolution of taxonomies versus ontologies are summarized in Reference 3. Briefly, ontology complexity ranges from a simple controlled list of terms (**Figure 1*a***) to a hierarchical taxonomy (**Figure 1*b***), to asserted multiple-parentage hierarchies (**Figure 1*c***), and to logically defined multiproperty graphs (**Figure 1*d***). For most taxonomies, the relationship between levels of the hierarchy are rarely specified by logical formalisms but rather are typically broader-than

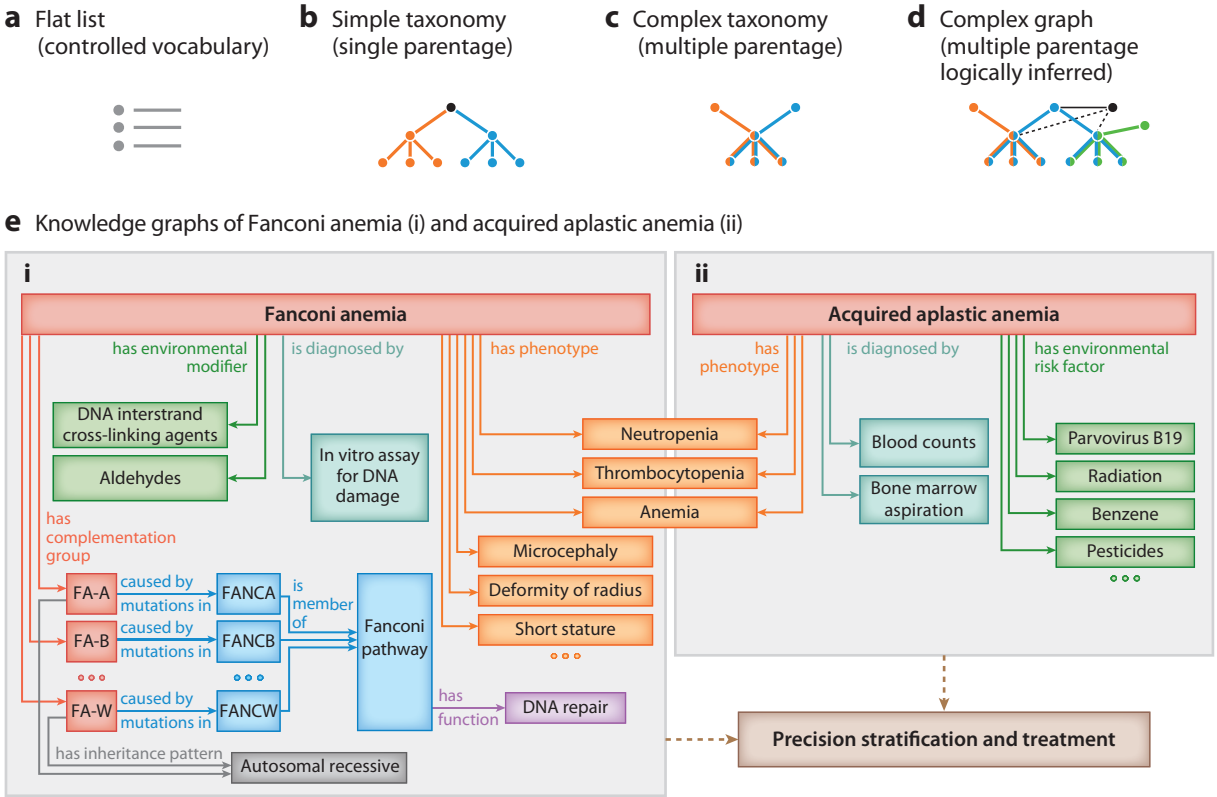

Figure 1

Ontology complexity ranges from (*a*) a simple controlled list of terms (e.g., OMIM) to (*b*) a hierarchical taxonomy (e.g., ICD-10), to (*c*) asserted multiple-parentage hierarchies (e.g., MeSH), and to (*d*) logically defined multiproperty graphs (e.g., NCIt). Panel *e* is an example of a domain-specific complex knowledge graph. In such a knowledge graph, each node is formally identified (identifiers are omitted from the figure for simplicity), and each relationship between nodes is also formally described in support of logical inference. Subpanel *i* represents Fanconi anemia (MONDO:0019391), an inherited disease with environmental modifiers. Subpanel *ii* represents rare acquired aplastic anemia (MONDO:0015610), a disease caused by environmental exposure. These two diseases share some but not all of the same phenotypes and similar, but not identical, environmental factors. A reasoner can utilize these disparate but linked attributes for precision classification of patients, for example, for assisting disease diagnosis. Abbreviations: FA, Fanconi anemia; ICD-10, International Classification of Diseases, Tenth Revision; MeSH, Medical Subject Headings; NCIt, National Cancer Institute Thesaurus; OMIM, Online Mendelian Inheritance in Man.

or narrower-than assertions, which can be expressed using the Simple Knowledge Organization System (SKOS; **http://www.w3.org/TR/skos-reference**). For our purposes, we use an inclusive definition of ontology—one that encompasses any taxonomy that utilizes description logic (DL), a computable subset of first-order predicate logic, to assert logical relationships between terms. This working definition would exclude simple controlled lists and simple taxonomies (**Figure 1*a*,*b***) but covers logically defined taxonomies and more complex graphs (**Figures 1*b*–*d***).

Early classification systems of disease and mental illnesses were developed in the eighteenth century by Linnaeus, de Sauvages, Vogel, and Cullen (4, 5). Carl Linnaeus developed the first modern medical classification that can be considered a true ontology of diseases (nosology). It divided diseases into 11 classes, 37 orders, and 325 species. Linnaeus' classification laid the foundation for what led to the first International Classification of Diseases (ICD; **http://www.who.int/classifications/icd**) in 1893 (6). Some current ontological definitions of "disease" include:

> "Any abnormal condition of the body or mind that causes discomfort, dysfunction, or distress to the person affected or those in contact with the person. The term is often used broadly to include injuries, disabilities, syndromes, symptoms, deviant behaviors, and atypical variations of structure and function." [National Cancer Institute Thesaurus (NCIt); 7]

> "A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown." [Medical Subject Headings (MeSH); 8]

As is clear from these definitions, a disease classification must utilize the relationship between a disease entity and individual phenotypic features (phenotypes, signs, symptoms, or endotypes) that commonly occur among those affected by the disease. An example of such discrete features, or so-called deep phenotyping (9), would be macular degeneration in the context of Gaucher disease (a disease entity). This example also highlights the fact that an individual phenotypic feature can be conceptualized as a disease (e.g., age-related macular degeneration) or as a feature of another disease, such as Gaucher disease. Note that clinical researchers often refer to a phenotype as a search/filter algorithm to define a study cohort or patient population, usually within the context of querying electronic health records (EHRs) (10–12).

A disease classification represented as an ontology must consistently utilize clinical presentations, specific findings, responses to treatment, biomarkers, and molecular characterizations as taxonomic features—either represented within the taxonomy itself (e.g., macular degeneration diseases) or as sets of annotations, as in the case of the Human Phenotype Ontology (HPO; **http://www.human-phenotype-ontology.org/**), which provides associations between diseases and phenotypes (10). In other words, one can either build into a disease classification all of the indicated disease attributes or separate them as annotations with metadata, evidence, and provenance. The use of DLs, such as the Web Ontology Language (OWL) (11), has enabled machines to reason over the ontology, which helps with detection of errors and automatic classification. DLs allow a curator to express more of the meaning in an ontology in a way that computers can understand. For example, without a DL, a machine may not know that when a curator creates two sibling classes (for example, `acquired metabolic disease` and `inherited metabolic disorder` as two subtypes of `metabolic disease`), the intention is for these to form mutually disjoint categories. A DL allows the curator to explicitly assert this, which may reveal some classes accidentally classified as both deeper in the ontology. Furthermore, a DL allows one to construct ontologies in a modular fashion, building up complex concepts from simpler ones. The concept of "metabolic disease" may be composed using DL constructs from the simpler concepts of "disease" and "metabolic process," with the latter coming from the Gene Ontology (GO). Similarly

for "disease of glucose metabolism." This, in turn, allows for the automatic classification of large parts of the disease ontology, leveraging other ontologies. It has been shown in other ontologies that this approach can avoid many errors of omission (13). Use of DLs allows numerous axes of disease classification, such as etiology, developmental origin, causal environmental factors, viruses or other insults, molecular mechanisms and genetics, primary anatomical site of the phenotypic features, response to treatment, biomarkers, heredity, rarity, etc.

ONTOLOGIES ON THE RISE

The use of ontologies for classifying disease has steadily increased, as evidenced by the increasing number of general and disease-specific ontologies (discussed below and listed in **Table 1**), as well as citations and mentions of specific disease-related ontology resources (**Figure 2**). Moreover, the rate of increase in articles about ontologies is twice the rate of all articles published (see **Supplemental Table 1**).

Supplemental Material >

One might ask, why are there so many different classifications of disease? The diversity of origins and primary uses for disease ontologies is reflected by the variety of structures, strengths, limitations, and uses of these classifications. Specific disease ontologies are often part of a larger ontology or semantic framework, and numerous ontologies exist for specific diseases or disease categories. A given disease entity may be a single entity in one classification but the primary subject of another. One reason for the diversity of classifications is the diversity of applications, and as such, many of these knowledge representations include much more than a single axis of disease classification. For example, sickle cell disease would be represented within SNOMED-CT (Systematized Nomenclature of Medicine–Clinical Terms; **http://www.snomed.org**), but the Sickle Cell Disease Ontology (14) is a separate ontology representing treatments used in the context of specific symptoms, as well as the genetic and environmental interactions that lead to phenotypic diversity in presentation. Ontologies such as the Sickle Cell Disease Ontology may include regional information as well, such as clinical assays and therapies that are relevant to resource-poor or rural care or local environmental variables such as diet.

In this section, we describe and review a few of the major disease classifications currently in use, and refer the reader to **Table 1** for descriptions of a wide range of ontologies and disease classifications. We aim to describe history, structural design features, and precoordinated versus postcoordinated terms—for example, the ability to combine entities or the inclusion of a full spectrum of possibilities (for example, skin cancer in every skin location).

International Classification of Diseases

One of the most widely utilized disease classifications, ICD, is maintained by the World Health Organization (WHO), is used by over 100 countries, and has been translated into 43 languages. The Tenth Revision (ICD-10) is organized anatomically by organ system (an outcome of the mid-nineteenth-century debate over organization by anatomical versus constitutional or syndromic nature) (15) and was intended since its inception for the tabulation of statistics on mortality and, later, morbidity. In the late twentieth century, ICD was embraced for billing and reimbursement categories, which is popularly argued to have distorted its evolution and assignment in clinical practice. ICD can trace its roots to the *Natural and Political Observations Made Upon the Bills of Mortality* (16) in the early sixteenth century and largely retained its tabular, structural origins until the recent development of the Eleventh Revision (ICD-11).

Historical editions of the ICD are statistical classifications, not ontologies. Statistical classifications are mutually exclusive (to not double count things) and exhaustive (to have a place to put everything). The ICDs have achieved exclusiveness through a monohierarchy (single parentage)

Table 1 Disease ontology resources

Ontology[a]	Description[b]
MeSH (Medical Subject Headings; **http://www.nlm.nih.gov/mesh**)	A controlled vocabulary of over 25,000 concepts used to index records in MEDLINE provided by the US National Library of Medicine. MeSH is organized as a broader to narrower set of subject headings that can appear in more than one part of the hierarchy, most of which have a short description or usage notes. Qualifiers (subheadings) can be added to subject headings to refine the topic being annotated. MeSH contains an extensive C branch of diseases.
LOINC (Logical Observation Identifier Names and Codes) (117)	Identifies medical laboratory observations. The Regenstrief Institute first developed LOINC in 1994 in response to the demand for an electronic database for clinical care and management. LOINC is publicly available at no cost and is endorsed by the American Clinical Laboratory Association and the College of American Pathologists. Since its inception, LOINC has expanded to include not just medical laboratory code names but also nursing diagnoses, nursing interventions, outcome classifications, and patient care data sets.
MedDRA (Medical Dictionary for Regulatory Activities; **http://www.meddra.org**)	Provides a standardized international medical terminology to be used for regulatory communication and evaluation of data about medicinal products for human use. MedDRA was first developed in the 1990s by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use.
ORDO (Orphanet Rare Disease Ontology)	A structured vocabulary for rare diseases capturing relationships between diseases, genes, and other relevant features, jointly developed by Orphanet and the EBI. It contains information on nearly 10,000 cancers and related diseases, 8,000 single agents and combination therapies, and a wide range of other topics related to cancer and biomedical research. ORDO was first released in 2014.
ICF (International Classification of Functioning, Disability and Health; **http://www.who.int/classifications/icf/**) (118)	Represents diseases and provides a conceptual basis for the definition and measurement of health and disability as organized by patient-oriented outcomes of function and disability. ICF considers environmental factors as well as the relevance of associated health conditions in recognizing major models of disability. ICF's scope is across disciplines and countries. ICF was created in 2001, and a companion classification for children and youth (ICF-CY) was released in 2007.
MedGen (NCBI Medical Genetics) (119)	MedGen contains information about conditions and phenotypes related to medical genetics. Terms from the NIH's GTR, UMLS, HPO, Orphanet, ClinVar, and other sources are aggregated into concepts, each of which is assigned a unique identifier and a preferred name and symbol. MedGen provides links to such resources as the GTR, GeneReviews, ClinVar, OMIM, related genes, disorders with similar clinical features, medical and research literature, practice guidelines, consumer resources, and ontologies such as HPO and ORDO. MedGen began in 2012, and the latest release was in 2017.
ICD-O (International Classification of Diseases for Oncology) (120)	A classification of neoplasms used by cancer registries to record incidence of malignancy and survival rates. ICD-O has two coding axes to describe tumors: topography and morphology. It was first published in 1976 and is currently in its third revision, which was updated in 2011.
EFO (Experimental Factor Ontology) (121)	Provides a systematic description of many experimental variables available in EBI databases; supports the annotation, analysis, and visualization of data handled by many groups at the EBI; and is the core ontology for OpenTargets.org. EFO was first published in 2010.
DO (Human Disease Ontology) (122)	A general ontology used by model organism researchers and other communities for tagging disease entities for retrieval. DO comprises a single inheritance model based on etiology and does not include onset (e.g., early, late, metastasis, stages), severity (e.g., transient, acute, chronic), or compound disease terms, as might be seen in vocabularies such as ICD. DO was first released in 2003, with the latest update in 2017.

(*Continued*)

Table 1 (*Continued*)

Ontology[a]	Description[b]
RDO (Rat Disease Ontology) (123)	Provides the foundation for ten comprehensive disease area–related data sets at the Rat Genome Database Disease Portals. Two major disease areas are the focus of data acquisition and curation efforts each year, leading to the release of the portals. Implemented as an asserted polyhierarchy, RDO was first released in 2016.
DSM (Diagnostic and Statistical Manual of Mental Disorders) (124)	Defines and classifies mental disorders to improve diagnoses, treatment, and research. The fifth version was published in 2013 by the American Psychiatric Association.
IDO (Infectious Disease Ontology; **http://infectiousdiseaseontology.org**) (125, 126)	A suite of ontologies intended to be extended for specific infectious disease categories, such as malaria. Currently, there are over 500 classes, 20 individuals, and 39 properties. IDO was first released in 2010, with the latest update in 2016.
ND (Neurological Disease Ontology) (42)	A representational tool that addresses the need for unambiguous annotation, storage, and retrieval of data associated with the treatment and study of neurological diseases. ND was first released in 2013, with an update in 2015.
SCDO (Sickle Cell Disease Ontology) (14)	Currently under development, SCDO will establish (*a*) community-standardized sickle cell disease terms and descriptions, (*b*) canonical and hierarchical representation of knowledge on sickle cell disease, and (*c*) links to other ontologies and bodies of work. SCDO was first released in 2017.
PDON (Parkinson Disease Ontology) (127)	A comprehensive semantic framework with a subclass-based taxonomic hierarchy, covering the whole breadth of the Parkinson disease knowledge domain from major biomedical concepts to different views on disease features held by molecular biologists, clinicians, and drug developers. PDON was first released in 2015.
ADAR (Autism DSM-ADI-R Ontology) (128)	An ontology of autism spectrum disorder (ASD) and related neurodevelopmental disorders that extends an existing autism ontology to allow automatic inference of ASD phenotypes and Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria based on subjects' Autism Diagnostic Interview–Revised (ADI-R) assessment data.
ASDPTO (Autism Spectrum Disorder Phenotype Ontology) (129)	Encapsulates the ASD behavioral phenotype, informed by the standard ASD assessment instruments and the currently known characteristics of this disorder. ASDPTO was first published in 2014.
LDA (Ontology of Language Disorder in Autism; **http://purl.bioontology.org/ontology/LDA**)	An ontology assembled from a set of language terms mined from the autism literature. LDA was first released in 2008.
DERMO (Human Dermatological Disease Ontology) (130)	The most comprehensive dermatological disease ontology available, with over 3,500 classes available. There are 20 upper-level disease entities, with features such as anatomical location, heritability, and affected cell or tissue type. DERMO was first released in 2013 and was updated in 2016.
PDO (Pathogenic Disease Ontology) (131)	An ontology for describing both human infectious diseases caused by microbes. PDO was first released in 2013 and updated in 2016.
ABD (Anthology of Biosurveillance Diseases) (132)	Provides information on infectious diseases, disease synonyms, transmission pathways, disease agents, affected populations, and disease properties. Diseases are grouped into syndromic disease categories, organisms are structured hierarchically, and both disease transmission and relevant disease properties are searchable. ABD was first released in 2016.

(*Continued*)

Table 1 (*Continued*)

Ontology[a]	Description[b]
CVDO (Cardiovascular Disease Ontology) (133)	An ontology based on the OGMS (Ontology for General Medical Science) model of disease that is designed to describe entities related to cardiovascular diseases (including the diseases themselves, the underlying disorders, and the related pathological processes). CVDO is being developed at Sherbrooke University (Canada) and Inserm (France). CVDO was first released in 2016, with an update in 2017.
MFOMD (Mental Functioning Ontology–Mental Disease) (134)	Describes and classifies mental diseases annotated with DSM-IV and ICD codes. MFOMD was published in 2012.
CKDO (Chronic Kidney Disease Ontology; **http://purl.bioontology.org/ontology/CKDO**)	Assists routine data studies and case identification of chronic kidney disease in primary care. First released in BioPortal in 2017.
HORD (Holistic Ontology of Rare Diseases; **http://purl.bioontology.org/ontology/HORD**)	Describes the biopsychosocial state (i.e., disease, psychological, social, and environmental state) of persons with rare diseases in a holistic way. HORD was released in 2017.
OHD (The Oral Health and Disease Ontology) (135)	Represents the content of dental practice health records and is intended to be further developed for use in translational medicine. OHD is structured using BFO (Basic Formal Ontology) and uses terms from many ontologies, NCBITaxon, and a subset of terms from the CDT (Current Dental Terminology). OHD is in early development and was last updated in 2016.
RPO (Resource of Asian Primary Immunodeficiency Diseases) (136)	Represents observed phenotypic terms, sequence variations, and messenger RNA and protein expression levels of all genes involved in primary immunodeficiency diseases. RPO was published in 2008.
OCVDAE (Ontology of Cardiovascular Drug Adverse Events) (137)	A biomedical ontology of cardiovascular drug–associated adverse events. OCVDAE was first released in 2016.
DCO (Dispedia Core Ontology; **http://purl.bioontology.org/ontology/DCO**)	A schema for information brokering and knowledge management in the complex field of rare diseases. DCO describes patients affected by rare diseases and records expertise about diseases in machine-readable form. DCO was initially created with amyotrophic lateral sclerosis as a use case. DCO was first released in 2015.
COPDO (Chronic Obstructive Pulmonary Disease Ontology) (138)	Models concepts associated with chronic obstructive pulmonary disease in routine clinical databases. COPDO was published in 2015.
IDOBRU (Brucellosis Ontology) (139)	Describes the most common zoonotic disease, brucellosis, which is caused by *Brucella*, a type of facultative intracellular bacteria. IDOBRU is an extension of IDO and was published in 2011.
OGMD (Ontology of Glucose Metabolism Disorder) (140)	Represents glucose metabolism disorder and diabetes disease names, phenotypes, and their classifications. OGMD was created in 2009, with the latest update in 2017.
AI-RHEUM (Artificial Intelligence Rheumatology Consultant System Ontology) (141)	Contains findings, such as clinical signs, symptoms, laboratory test results, radiologic observations, tissue biopsy results, and intermediate diagnosis hypotheses, for the diagnosis of rheumatic diseases. AI-RHEUM is used by clinicians and informatics researchers and was released in 2015.
FILDO (Fibrotic Interstitial Lung Disease Ontology) (142)	An in-progress, four-tiered ontology proposed to standardize the diagnostic classification of patients with fibrotic interstitial lung disease. FILDO was published in 2017.
PEO (Pre-Eclampsia Ontology) (143)	Represents clinical features, treatments, genetic factors, environmental factors, and other aspects of the current knowledge in the domain of pre-eclampsia. PEO was published in 2016.
RPDO (Removable Partial Denture Ontology) (144)	Represents knowledge of a patient's oral conditions and denture component parts, originally developed to create a clinician decision support model. RPDO was published in 2016.

(*Continued*)

Table 1 (*Continued*)

Ontology[a]	Description[b]
PCOSKB (PolyCystic Ovary Syndrome Knowledgebase; **http://pcoskb.bicnirrh.res.in**) (145)	Comprises genes, single nucleotide polymorphisms, diseases, gene ontology terms, and biochemical pathways associated with polycystic ovary syndrome, a major cause of female subfertility worldwide. PCOSKB was published in 2015.

Abbreviations: EBI, European Bioinformatics Institute; GTR, Genetic Testing Registry; HPO, Human Phenotype Ontology; MEDLINE, Medical Literature Analysis and Retrieval System Online; NCBI, National Center for Biotechnology Information; NCBITaxon, NCBI Taxonomy Database; NIH, National Institutes of Health; OMIM, Online Mendelian Inheritance in Man; UMLS, Unified Medical Language System.

[a]Ontology abbreviations are followed by the full name in parentheses, along with online sources and references.

[b]Descriptions are adapted from BioPortal (**https://bioportal.bioontology.org/**), Wikipedia, or the source's website.

so that each rubric or code has one and only one parent. This precludes multiple counting but creates arbitrary associations. For example, `gastric cancer` (ICD-10:C16) is a child of `malignant neoplasms` but is not, nor can be, a child of `digestive system diseases`. Modern ontologies conventionally embrace acyclic graphs with multiple parenting, as explained in the desiderata for terminologies in Reference 17. ICDs invoke residual categories, such as `not elsewhere classified` or `other specified` to address the exhaustive requirement for statistical classification;

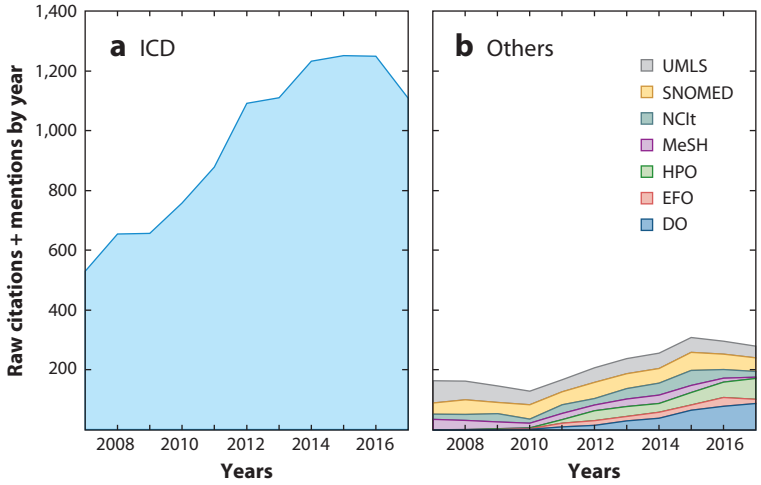

Figure 2

Ontology use over time. In aggregate, the mention and citations of disease ontologies have doubled over the last 10 years, with the overwhelming majority of mentions to ICD, and all others increasing markedly albeit at a much lower level. Ontologies greater than 10 years old are much more likely to be mentioned than cited. The reverse is true for ontologies under 10 years old; however, even in these circumstances, 20% of the papers that should formally cite an ontology do not. We used MEDLINE (PubMed) to identify the marker papers of each ontology. Once found, we used Scopus to find publications citing each marker paper. Additionally, we searched MEDLINE (PubMed) using MeSH subject headings and text strings to identify all mentions of these ontologies in MEDLINE (PubMed). We mapped the Scopus DOIs to PMIDs/PMCIDs using various services, deduplicated the list of citing/mentioning publications, and stratified by year. The search strategies, scripts, and raw data files are all available on GitHub at **https://github.com/monarch-initiative/ont-review**. Abbreviations: DO, Human Disease Ontology; EFO, Experimental Factor Ontology; HPO, Human Phenotype Ontology; ICD, International Classification of Diseases; MEDLINE, Medical Literature Analysis and Retrieval System Online; MeSH, Medical Subject Headings; NCIt, National Cancer Institute Thesaurus; PMCID, PubMed Central identification number; PMID, PubMed identification number; SNOMED, Systematized Nomenclature of Medicine; UMLS, Unified Medical Language System.

these too violate Cimino's desiderata (17). Some have argued that those desiderata were not intended to be applied to statistical classifications.

ICD-11 was formally introduced by WHO in November 2016 at the Revision Conference in Tokyo (18). This revision breaks away from the sixteenth-century tabular form to embrace a multitiered architecture, centered around an acyclic-graph semantic network called the foundation component. A monohierarchy is derived from that network to achieve the requirements of a statistical classification, which is called a linearization because it can be printed in a book from beginning to end. Residual categories are added only to a linearization to make it exhaustive. Presently, the Joint Linearization for Morbidity and Mortality Statistics is the primary derivative, and it looks and feels like traditional revisions of the ICD. However, the ICD-11 architecture can support an arbitrary number of linearizations, optimized for decision support, subspecialty care, quality improvement, or reimbursement.

The original architecture of ICD-11 envisioned an ontology component to semantically anchor the foundation component and, in turn, the linearizations. A memorandum of agreement was signed in 2010 with the SNOMED organization to partner in this process. Some researchers proposed to create a Common Ontology (19) to address the dissonance between a simple (SKOS-based) hierarchy (ICD) and a DL-based system (SNOMED); their idea was to invoke query logic [as in SQL (structured query language) set theory] rather than DL (20). A substantial prototype for the cardiovascular chapter was developed as a demonstration (21), but funding to complete this work for the entire foundation component was not available.

The goal of automated coding from EHRs to disease classifications such as ICD is partly realized in many commercial coding tools today. However, the new architecture implemented for ICD-11 supports a sophisticated information model around each rubric, comprising a profile of that disease or syndrome. Presently it is populated only with fully specified terms, synonyms, and limited human language definitions. Projects such as the Data Translator (OT3 OT-TR-16-001) from the National Center for Advancing Translational Sciences (NCATS) may catalyze the authoring of more complete disease information, including computable clinical criteria, that will enable API (application programming interface) access to disease concepts. While intended to support translational biomedical research and EHR data linkage, such tooling will also contribute to a computational ecosystem for fully automated classification of patients from underlying granular clinical data.

Various countries including the United States adapt the ICD for national use, namely the ICD-10 for Clinical Modification (ICD-10-CM) and corresponding adoptions for earlier ICD revisions. Implementing ICD-10-CM in the United States may offer marginal value for health care or biomedical data (22) and, in any event, was done nearly 25 years after the release of ICD-10 by WHO. ICD-11 will be formally released in July 2018, although its adoption in the United States may again be delayed in view of the expense and effort of ICD-10-CM.

Systematized Nomenclature of Medicine–Clinical Terms

SNOMED-CT is a compositional system of clinical findings, symptoms, diagnoses, procedures, body structures, and organisms, as well as other etiologies, substances, pharmaceuticals, devices, and specimens. It can trace its origins to the three compositional axes of the Standard Nomenclature of Diseases and Operations (SNDO) in 1932 (topology, etiology, and surgical procedures; 23), which was published through 1961 (24). From this grew the Systematized Nomenclature of Pathology (SNOP) (25), which was a four-axis compositional system replacing the surgical procedures of SNDO with morphology (histopathology) and function. SNOP was so successful that

the American College of Pathologists attempted to embrace all of medicine with the first version of SNOMED in 1976 (26). That expanded the SNOP axes to restore procedure and add disease and occupation. However, the first three versions of SNOMED were effectively tabular enumerations of terms with codes, arrayed in a crudely segmented hierarchy, although the third version, SNOMED International in 1993 (27), was relatively large, totaling 132,574 concepts compared to the few hundred in SNDO.

Given the size and complexity of SNOMED International as a compositional system, users quickly recognized the reality and likelihood of redundant coding; for example, appendectomy could be compositionally expressed in 17 different ways (28). This led researchers to consider whether the emerging practicality of DL could be leveraged to provide a computable framework for terminologies, enabling them to be internally consistent and nonredundant (29). From this grew the preliminary work to create SNOMED for Reference Terminology (SNOMED-RT), based on proprietary DLs from IBM called K-Rep (Knowledge Representation) (30). K-Rep became the basis for what is today called the EL++ DL, for which there now exists an OWL2 profile (**https://www.w3.org/TR/owl2-profiles**). EL++ is significantly more tractable for large-scale terminologies, as it is computable in polynomial time (31). However, EL++ is severely compromised in expressiveness relative to OWL2, of which it is a subset, lacking any abilities to represent class negation, cardinality restrictions, or disjoint properties, among others.

SNOMED-CT represents a union between SNOMED-RT and the UK National Health Service (NHS) Clinical Terms V3 (32); the latter is a large clinical terminology that is the successor to the NHS Read Codes (33). SNOMED-CT still uses the EL++ DL, although less than a quarter of its content is logically defined; the remainder is primitive. SNOMED-CT supports composition, although it now has a sophisticated clinical model of concepts that frame correct compositional syntax. Compositional expressions that are redundant with precoordinated terms can be computationally reduced to a canonical form (34), obviating historical concerns such as the 17 ways to express appendectomy.

More recently, SNOMED has embraced concept models for medical genetics (35) and harmonizing content for nursing models (36). It is exploring mechanisms to facilitate decision support in EHRs (37). Authors are also exploring evaluation techniques for larger terminologies (38) and the degree to which these are transferable to other ontologies such as Uberon (39). SNOMED has 30 member countries and is maintained in a variety of languages.

The Unified Medical Language System

The Unified Medical Language System (UMLS) is a suite of data and knowledge resources for biomedical concepts and terms maintained by the National Library of Medicine since 1990. Its core component since its inception is the Metathesaurus; the 2017AB release (40) contains 3.64 million concepts, and 13.9 million unique terms from 201 source vocabularies. These statistics have all increased by more than an order of magnitude from the original release. The structure and detail of the Metathesaurus have also evolved, at various times being published as bar-delimited ASCII (American Standard Code for Information Interchange) text files, SQL data sets, ASN.1 (Abstract Syntax Notation One) notation, and, since 2004, the canonical RRF (Rich Release Format) (41), which expands the ASCII sources to preserve source file transparency.

True to its name, the Metathesaurus is a joining of terms from source vocabularies, such as MeSH and the ICDs, to generate a semantically harmonized union of terms and concepts. The UMLS distinguishes concepts, which have a unique meaning, from terms, which have similar words and order after lexical simplification. A concept may have many associated terms that are effectively synonyms (42). Many synonyms are human language translations, such as Spanish,

where those human languages are explicitly tagged. One consequence of this semantic reduction is that the Metathesaurus can function as a mapping from source terminologies with shared concepts even if the exact terms significantly differ. Additionally, concepts are structured as a hierarchy, supporting near-meaning mappings from a highly specific term to a term parent in another source terminology. Despite frequent attempts by investigators and NLP practitioners to use the Metathesaurus as an ontology, this often proves impractical because the hierarchy can be cyclic. If the hierarchy ordering of a source terminology declares concept B to be a child of A, while another source declares A to be a child of B, both relationships will be deliberately instantiated in the UMLS to preserve source transparency; the RRF format disambiguates which source asserts which relations.

A second component of the UMLS, the Specialist Lexicon (43), is designed to support the lexical normalization of text for natural language processing (NLP). It comprises a suite of software tools, a knowledge resource of regular lexical forms (case, tense, number, etc.) for the English language, and a database of more than 30,000 terms with regular and irregular lexical forms. The Specialist Lexicon is used to generate lexically similar versions of text to define terms for the Metathesaurus.

The third major component of the UMLS is the Semantic Network (44). This is an upper ontology of biomedicine (45) that presently comprises 133 biomedical semantic types (such as disease or syndrome) in a hierarchy of 54 relationships. These numbers have changed little over the 27-year history of the UMLS, consistent with the stability of an upper ontology. Each concept in the Metathesaurus is assigned a semantic type, which can serve to disambiguate homonyms such as "cold" disease (`disease or syndrome`) and "cold" temperature (`natural phenomenon or process`).

The UMLS has been used widely in scholarly analyses of terminologies and ontologies. It has also been used to enhance or develop applications, such as EHRs, classification tools, dictionaries, language translators, and text mining (46, 47).

National Cancer Institute Thesaurus

NCIt (**https://ncit.nci.nih.gov**; 48) is produced by the National Cancer Institute (NCI) and was derived originally from the NCI Metathesaurus in 2001. Within the NCI, a variety of independent coding systems had arisen, and NCIt was designed to assist uniform data and activity annotation of cancer research across NCI's programs. The NCIt today is a widely utilized standard for coding, knowledge reference, and public reporting. For example, NCIt is a key feature within the international Clinical Data Interchange Standards Consortium terminology, is used by the US Food and Drug Administration (FDA) for drug approval applications, and is included within the Federal Medication Terminologies and the Japanese National Council for Prescription Drug Programs for medication coding. A large part of this ontology incorporates standardized drug information, which is used to facilitate clinical trial research and registration such as to code ClinicalTrials.gov. NCIt is also used to code studies within the Cancer Genomic Data Commons (49), other cancer-based public databases, and new initiatives such as KidsFirst (**https://commonfund.nih.gov/kidsfirst**) and the Variant Interpretation for Cancer Consortium of the Global Alliance for Genomics and Health (**http://cancervariants.org/**).

The NCIt covers a wide variety of concepts in clinical care, translational and basic research, and public information and administrative activity, with over 100,000 textual definitions and over 400,000 relationships between terms. NCIt represents over 10,000 cancers and related diseases and 8,000 single agents and combination therapies related to the diseases. NCIt is a comprehensive taxonomy of neoplasm types, including a range of very broad terms such as `carcinoma` (NCIT:C2916) and highly specific subtypes such as `Thyroid Gland Mucosa-Associated`

Lymphoid Tissue Lymphoma (NCIT:C7601). Subsets of the NCIt facilitate tasks in a variety of contexts, such as the neoplasm subset that forms a core reference set of approximately 1,400 cancer classification terms (50). The NCIt Browser (51) enables user navigation of concepts and displays the text definitions, logical relationships, synonyms, coding subsets, and mappings (see below).

The NCIt is a rich DL-based ontology and was released in OWL2 in 2017. The NCIt has a fairly large number of subontologies, such as Anatomic Structure, System, or Substance; Biochemical Pathway; Abnormal Cell; Disease, Disorder or Finding (where the neoplasm branch mentioned above is included); Diagnostic or Prognostic Factor; Chemotherapy Regimen or Agent Combination; Molecular Abnormality; Organism (for nonhuman reference); and others. The ontology contains a rich set of transitive role properties between concept pairs, as well as nontransitive association properties. For example, the logical definitions of Thyroid Gland Mucosa-Associated Lymphoid Tissue Lymphoma (NCIT:C7601) include related findings, anatomic sites, genetic abnormalities, associated diseases, and cellular origins, among many others:

Disease_Has_Finding: Primary Lesion

Disease_Has_Associated_Anatomic_Site: Thyroid Gland

Disease_Has_Normal_Cell_Origin: Marginal Zone B-Lymphocyte

Disease_May_Have_Associated_Disease: Hashimoto Thyroiditis

Disease_May_Have_Cytogenetic_Abnormality: t(3;14)(p14.1;q32)

The rich knowledge structure contained within the NCIt makes it amenable to modern quality assurance and development processes that leverage OWL semantics (52, 53). The NCIt has also been used to evaluate the quality of the annotations on the Common Data Element value sets in cancer research (54). The NCIt has been used in an increasing number of applications that leverage the rich semantics to predict activities of cancer patients (55), facilitate precision medicine (49), support diagnosis (56), and crowdsource cancer variants (57).

NCIt was recently released as an Open Biomedical Ontologies (OBO) ontology (58) to support translational and interoperable use across basic and clinical research (59). This release is a direct transformation of the NCIt using OBO-style term Internationalized Resource Identifiers and annotation properties, with direct references to terms from other domain-specific OBO Library ontologies [e.g., Cell Type Ontology (60) and the Uberon Anatomy Ontology (61)]. The release also contains NCIt, ICD for Oncology, and Oncotree mappings and subsets that also include OBO Uniform Resource Identifiers in their logical axioms.

Human Phenotype Ontology

HPO (10) provides terms that allow deep phenotype descriptions of patient and disease characteristics. The set of HPO terms, or phenotypic profile, is the basis of comparison for identifying candidate diseases and variant prioritizations. Numerous clinical labs and diagnostic tools use HPO and the HPO-provided disease–phenotype associations for this purpose. The HPO was initially published in 2008 with the goal of integrating phenotypic data for translational research and diagnostics. It has since become a standard used by projects such as the NIH (National Institutes of Health) Undiagnosed Diseases Program and Network, the 100,000 Genomes Project, DECIPHER (Database of Genomic Variation and Phenotype in Humans using Ensembl Resources), Deciphering Developmental Disorders, and the European Reference Networks, and it has been translated into such languages as Japanese, Chinese, French, and Spanish (10). Following

Table 2 The most important components of the HPO (Human Phenotype Ontology) term `atrial septal defect`

Item	Example	Explanation
ID	HP:0001631	Accession number for this term
Name	Atrial septal defect	Preferred label
Synonyms	ASD, Atrial septum defect	Other commonly used names for the same concept
Text definition	Atrial septal defect (ASD) is a congenital abnormality of the interatrial septum that enables blood flow between the left and right atria via the interatrial septum.	Human-readable definition of the concept
Xref	ICD-10:Q21.1	Cross-reference to synonymous terms in other databases
`is_a`	`abnormality of cardiac atrium` (HP:0005120); `abnormality of the atrial septum` (HP:0011994)	One or more so-called parent terms, i.e., more general terms located directly above this term in the ontology
Logical definition	`'has part' some ('closure incomplete' and ('inheres in' some 'interatrial septum') and ('has modifier' some abnormal))`	Computable definition of the concept

experiences in the 100,000 Genomes Project, the NHS is now using HPO for patients with rare diseases (62).

The HPO has four subontologies, `Phenotypic abnormality`, `Clinical modifier`, `Mode of inheritance`, and `Mortality/Aging`, with 12,299 terms and 15,976 subclass relations between the terms. Each term has a text definition and, in many cases, synonyms and comments (see **Table 2**). Additionally, 5,717 of the terms have DL-based class definitions that provide computable definitions of HPO terms based on concepts from ontologies for anatomy, biochemistry, pathology, cell types, proteins, and biological functions. Each term in the `Phenotypic abnormality` subontology describes a specific phenotypic abnormality (sign, symptom, laboratory or imaging finding, behavioral abnormality, etc.). Thus, the terms of the HPO describe the individual components of diseases rather than the diseases themselves. The links to diseases are provided in the form of annotations, i.e., computational assertions that a certain disease is associated with a set of HPO terms. Each annotation is supplemented with metadata reflecting the provenance of the assertion and, in many cases, with attributes such as the typical age of onset or the overall frequency of a given feature in all patients with some disease [e.g., about 20–50% of patients with Noonan syndrome have `Pulmonary valve stenosis` (HP:0001642)]. The HPO project currently provides a total of 131,827 annotations to 6,996 diseases with identifiers from OMIM (Online Mendelian Inheritance in Man) (63), 2,664 from Orphanet (64), and 47 chromosomal disorders from DECIPHER (65). Although the initial focus of the HPO was Mendelian disease, it is being extended to common disease. A pilot using concept recognition in PubMed abstracts was used to derive 132,620 HPO annotations for 3,145 common diseases (66).

DISEASE AND PHENOTYPE ONTOLOGIES IN ACTION

Using Disease Ontologies for Search and Retrieval

Perhaps the simplest kind of computer-based search is a search for a text string in a document, where a string is simply a sequence of characters such as "Fanconi anemia." This type of search does not always work well for medical purposes because almost all concepts in medicine have

multiple synonyms. A search in MEDLINE (Medical Literature Analysis and Retrieval System Online) using only text strings retrieves different numbers of results for "Fanconi anemia" and "Fanconi's anemia" or for the gene symbol "RAD51C" and an alternate symbol "BROVCA3" for the same gene. Most ontologies define synonyms for the concepts in the ontology. An ontology-aware search like the PubMed interface to MEDLINE will return identical results for any of the synonyms of a given concept. In general, ontologies present the concepts within the ontology as a hierarchy of related terms, whereby more specific terms are presented as children of general terms; for instance, "Fanconi anemia, complementation group B" is a specific form of "Fanconi anemia" and would be represented as a child of the latter term in a disease ontology. This relationship allows ontology-driven searches to return not only matches to a given concept (say, all abstracts in PubMed that are annotated with "Fanconi anemia") but also to all of the more specific, descendant concepts (e.g., "Fanconi anemia, complementation group B" and the other forms of Fanconi anemia). Historically, this has been the most important use for vocabularies designed for identifying the scientific literature (i.e., MeSH). Information retrieval within and across a large number of resources such as research registries and EHRs can also benefit from integration with common ontologies to annotate and index each resource (67–69).

Cohort Identification

Now ubiquitous in health care settings, EHRs capture various aspects of the clinical care encounter. EHR data are increasingly used for translational research and to develop learning health care systems that aim to improve patient care based on existing system-wide data (70). However, substantial effort is required to extract, encode, and preprocess phenotypic information from EHRs prior to statistical analysis and integration with other data sources (genomic, metabolomic, etc.). Further, routine clinical care data do not necessarily constitute a perfect rendering of the biological observables of the patient, as clinicians seek to best describe the patient for the purposes of obtaining labs or insurance to provide the best care for their patients. As in the case of the eMERGE consortium, phenotyping algorithms are sets of rules and filters that exploit information in billing and diagnosis codes, clinical notes, and other unstructured data, laboratory measurements, and procedure and medication data (71) and can help better identify cases and controls for specific medical conditions (72). The Accrual to Clinical Trials ontology aims to streamline cohort definitions by harmonizing labs across clinical sites within the context of i2b2 (Informatics for Integrating Biology and the Bedside) (73, 74). Disease and phenotype ontologies are increasingly used as a critical component of text mining and phenotyping algorithms, which use recognition of synonyms and subsumption of more specific terms under general ancestor terms to improve the recognition of medical concepts in text. This is covered extensively in this volume by Shah and coauthors (75).

Using Ontologies for Patient Diagnostics, Care, and Decision Support

An extremely diverse spectrum of ontology-based efforts aims to support improved patient diagnosis, care management, and decision support. For example, Zhang et al. (76) implemented an ontology-based monitoring system that integrates patient data, medical knowledge, and patient assessment criteria for continuous management of chronic disease. Another example is the integration of alert information with EHR data, piloted with atrial fibrillation alerts (77), to classify alerts so that the most important ones are better highlighted. Ontology-based classification can also help triage patients in the emergency room (78), diagnose and monitor progress of Parkinson patients (79), and enable caregivers and clinicians to provide adaptive care to dementia patients

by monitoring their cognitive and behavioral status (**http://www.demcare.eu/**). Ontologies are increasingly used within clinical decision support systems, especially in the context of telemedicine or to support improved evidence-based decision making. Some interesting examples are consultation systems to support medical care onboard seagoing vessels (80) and ultrasound diagnosis in obstetrics and gynecology (81). Ongoing work on the ontologies aims to integrate multiple clinical practice guidelines to provide improved clinical decision support (82) and electronic clinical quality measures using the UMLS and the SNOMED-CT core problem list (83).

Some biomedical ontologies can have greatly expanded utility when applied in innovative ways. For example, the logical structure of the HPO provides a powerful foundation for computational phenotype analysis (84–87) and has been especially useful for the development of rare disease diagnostic tools, such as Exomiser (88–90), Phevor (91), Face2Gene (**http://face2gene.com/**), and numerous others. In these contexts, sets of HPO terms used to describe a patient provide the basis of comparison against known diseases (90, 92) and other patients. The Global Alliance for Genomics and Health Matchmaker Exchange (**http://www.matchmakerexchange.org**) utilizes HPO for *N*-of-1 (single-patient) global patient discovery (93), and new applications are being developed to support patient-led matchmaking using the HPO (**http://www.phenotypr.com**).

Use of Disease Ontologies in Information Exchange Standards

While ontologies provide a standard for describing data, they are not themselves a standard for computational exchange of data. A standard for exchanging data encoded with ontology terms is also needed; for instance, should we send a list of ontology terms in JSON (JavaScript Object Notation), XML (Extensible Markup Language), or some other format? What other metadata (such as the provenance of the data) or associations (such as with a genotype) are required, and what data structures can be used for data transmission? Exchange standards for phenotypic and genomic data for translational research and genomic medicine are still emergent, and no single standard has been widely accepted at the time of publication. However, several standards have obtained wide community participation and are under active development. Clinical data exchange between organizations is now routine, although proprietary systems (e.g., Epic) or the network of Health Information Exchanges leveraging the Consolidated Clinical Document Architecture (94), which in turn references standard ontologies and terminologies, should be bound to exchange documents in XML. The Fast Healthcare Interoperability Resources (**https://hl7.org/fhir/**) is a next-generation RESTful (Representational State Transfer) standard for exchanging health care information electronically that provides an abstraction layer on top of EHR information systems and is currently being extended for genomic data (95).

Observational Health Data Sciences and Informatics is an international collaboration that has developed a common data model allowing EHR data from multiple centers to be integrated and analyzed with standardized tools and ontologies (96). The Global Alliance for Genomics and Health is striving to develop standards, policies, and ethical frameworks for responsible genomic data sharing (97). For example, Phenopackets (**http://www.phenopackets.org**) is an extensible exchange format for exchange of data relating to patients, diseases, and phenotypic abnormalities together with metainformation such as age of onset. Increasingly, the community is converging on a suite of interoperable ontologies that define or cross-reference their terms by means of a compatible set of reference ontologies. For example, this can enable RDF (Resource Description Framework)-linked data approaches to integrate data from source ontologies for querying and analytics, such as in Reference 98 and the University of Pennsylvania's new project Transforming and Unifying Research with Biomedical Ontologies (**http://upibi.org/turbo/**).

DISEASE ONTOLOGY CHALLENGES AND OPPORTUNITIES

Precision of Disease Concepts

One of the biggest issues with many disease-related ontology resources is that using them effectively for data annotation requires clearly defined terms. However, many ontologies leave disease entities and groupings undefined; this is the case for just under half of terms in the Human Disease Ontology (DO), for just over half of terms in Orphanet's Rare Disease Ontology (ORDO), and for three-quarters of neoplasm terms in NCIt. Moreover, when definitions are provided, they often lack the precision necessary for curation or do not align with the taxonomic classification. In many cases the label/disease name acts as proxy for the definition, but this is unreliable as the label may be ambiguous alone. Using a formal language such as OWL allows for specification of definitions in a logical language, providing benefits for humans and machines, but it is not always possible to provide these. This is often easier with diseases such as cancer (10,000 of 15,000 neoplasm terms in NCIt have OWL definitions), whereas with syndromic conditions, the definition is frequently more nuanced or statistical in nature. For example, compare definitions for pancreatic cancer from different sources:

> "A carcinoma arising from the exocrine pancreas. The overwhelming majority of pancreatic carcinomas are adenocarcinomas." (NCIt)

> "Pancreatic cancer shows among the highest mortality rates of any cancer, with a 5-year relative survival rate of less than 5%. By the time of initial diagnosis, metastatic disease is commonly present. Established risk factors include a family history of pancreatic cancer, a medical history of diabetes type 2, and cigarette smoking." (OMIM)

Similarly, compare definitions for Ehlers–Danlos syndrome (EDS):

> "An inherited connective tissue disorder characterized by loose and fragile skin and joint hypermobility." (NCIt)

> "Group of inherited disorders of the connective tissue; major manifestations include hyperextensible skin and joints, easy bruisability, friability of tissues with bleeding and poor wound healing, calcified subcutaneous spheroids, and pseudotumors." (UMLS)

For pancreatic cancer, the NCIt definition is based on etiology, while the OMIM definition is about disease attributes and prognostic indicators. For EDS, both definitions describe phenotypic features of the diseases, but to varying degrees.

Defining ontological concepts is a socio-technical challenge that requires community coordination and adoption. Defining disease entities is nontrivial and different resources have different criteria for doing so. For example, OMIM requires that there be two published cases and one of the following: "first mutation to be discovered, high population frequency, distinctive phenotype, historic significance, unusual mechanism of mutation, unusual pathogenetic mechanism or distinctive inheritance" (63, p. D794). ClinGen (Clinical Genome Resource; **http://www.clinicalgenome.org**) (99) has a lumping and splitting working group, which has defined rigorous criteria for the creation of disease entities (different molecular mechanisms, phenotypic variability, segregation, and clinical management), essentially using genotypic, phenotypic, and care attributes to define independent disease entities. However, even with robust processes for disease definitions, there is a fundamental gradient between disease and health. Further, disease has been ontologically described as a process, event, state, disposition (e.g., a propensity), and a pathological anatomical

structure. Schulz et al. (100) approached this by enabling these three conceptual views of disease to be conflated (as demonstrated in the context of SNOMED-CT, where there is significant confusion between pathological structures, dispositions, and processes) and to be distinguished only when useful, such as in certain diseases. All of these efforts speak to the fundamental problem that disease definitions must sit at the intersection of robust descriptive patient attributes and ease of use, both by clinicians and by ontologists and other curation communities.

Ontology Mapping

While there is now a plethora of ontologies to choose from, each having specific advantages or domain coverage, this also leaves the community with the specific problem of not knowing which ontology to use for a given use case. Most disease classifications are not designed to be combined for cross-disease or cross-domain computational use. However, the cross-domain scope is important for both diagnosis and care: Across history, all kinds of disease subtypes have been reassigned to another domain when the underlying cause was uncovered. Many disease ontologies or databases contain overlapping content, with the same or similar disease entities and disease groupings present in each resource, with distinct identifiers. To combine these resources, researchers use mappings (also known as cross-references) to link identifiers from one resource to another. For example, `Ehlers-Danlos syndrome` in the Orphanet ORDO ontology is cross-referenced to `Ehlers-Danlos syndrome` in ICD-10 (Q79.6).

A new disease ontology, Monarch Disease Ontology (MonDO), recently was assembled to address the need for a single cohesive ontology that combines the respective strengths of existing partially overlapping ontologies. The original version of MonDO was created using algorithmic means (101) but has recently transitioned to a fully curated ontology with ongoing algorithmic support. MonDO covers multiple disease types, including rare, common, cancerous, and infectious disease, and is openly available as an axiomatized OWL ontology. It is updated regularly to ensure inclusion of new disease entities coming from numerous sources. It is utilized by resources such as ClinGen and the Monarch Initiative (102), which have a need to create computable definitions of disease for diagnostic purposes.

Table 3 highlights the diversity and precision of synonyms and mappings available from each ontology source for pancreatic cancer and EDS. For each of these example diseases, we systematically compared 11 ontologies: DO, ICD-10, ICD-11, MedDRA (Medical Dictionary for Regulatory Activities), MeSH, MonDO, NCIt, OMIM, Orphanet, SNOMED, and UMLS. Despite the similarities between records of the same disease concept, the differences pose challenges to automatic integration: Labels, definitions, synonyms, classification, and cross-references all partially overlap (for the same concept).

For example, the EDS records in the respective 11 ontologies together comprise 4 distinct labels, 7 distinct textual definitions, 61 distinct synonyms, and distinct external records over 14 other ontologies. EDS record differences range from minor (labels, hyphens, and casing) to major (conceptual differences in definition and hierarchy); MonDO aims to address these differences in a unified record (MONDO:0020066). Across the ontologies, we found significant conceptual differences in EDS classification, both in terms of subtypes of EDS and in terms of which parent terms EDS is classified under.

We repeated this comparison for pancreatic cancer records in the 11 ontologies. Taken together, the pancreatic cancer records comprise 4 distinct labels, 5 distinct definitions, and 59 distinct synonyms, and reference distinct external records distributed over 16 other ontologies. Differences range from minor (labels and casing) to major (conceptual differences in definition and hierarchy); MonDO aims to address these in a unified record (MONDO:0009831). **Table 3** does not

Table 3 Assessment of synonyms, hierarchies, and mappings across ontology sources for Ehlers–Danlos syndrome and pancreatic cancer[a]

	General information			Ehlers–Danlos syndrome				Pancreatic cancer			
Source ontology	Has definition[b]	Hierarchy type	Language translation	Xrefs	Synonyms	Parents	Descendants	Xrefs	Synonyms	Parents	Descendants
DO	Yes	Single	No	6/11	0/2	1/1	6/6	4/11	0/9	2/2	37/37
ICD-10	~Yes[c]	Single	No	–	–	1/1	–	–	–	1/1	–
ICD-11	Yes	Multi	Planned	–	1/2	2/2	26/26	–	2/7	3/3	23/23
MedDRA	No	Multi	Yes	–	–	2/2	–	–	–	1/2	–
MeSH	~Yes[d]	Single	Yes	–	22/23[e]	3/4	16/21	–	12/25[e]	3/3	14/14
MonDO	Yes	Multi	Planned	–	0/2	7/7	36/36	–	8/8	2/2	44/44
NCIt	Yes	Multi	No	–	–	2/2	7/7	–	3/3	2/2	68/68
OMIM[f]	Yes	Flat	No	–	7/7	–	–	2/2	–	–	–
Orphanet	No	Multi	No	2/5	–	6/6	24/24	5/5	0/2	7/7	10/10
SNOMED	No	Multi	Yes	–	2/11	8/8	21/21	–	5/5	3/3	47/47
UMLS	No	Multi	Yes	–	27/46	2/2	7/8	–	50/62	–	6/6

Abbreviations: DO, Human Disease Ontology; ICD-10/-11, International Classification of Diseases, Tenth Revision/Eleventh Revision; MedDRA, Medical Dictionary for Regulatory Activities; MeSH, Medical Subject Headings; MonDO, Monarch Disease Ontology; NCIt, National Cancer Institute Thesaurus; OMIM, Online Mendelian Inheritance in Man; SNOMED, Systematized Nomenclature of Medicine; UMLS, Unified Medical Language System; Xref, external reference (cross-reference).

[a]The Xref and synonym columns summarize the extent and precision of Xrefs and synonyms, respectively, provided by the respective records in the source ontologies. The denominators in these columns refer to the total number of Xrefs or synonyms listed by the source ontology; the numerators refer to the number of these that we manually assessed as equivalent to the concept (not narrower, broader, related, or obsolete). In some cases, something akin to a qualifier (narrower, broader, related, or obsolete) is already provided by the source; in such cases, if we found the qualifier to be accurate by manual review, it was counted toward the numerator. The notation "–" means that the denominator was zero (no such term was found). Source data and scripts for these assessments are available at **https://github.com/monarch-initiative/ont-review**.

[b]"Has definition" refers to whether the record has a textual definition in addition to a label; some of the ontologies also have logical definitions (not shown).

[c]Clinical information is provided, but it is not as concise as a conventional definition.

[d]The scope note is an approximation of a definition.

[e]The MeSH field `entry terms` is the closest to synonyms; numbers are especially high due to casing and punctuation variants.

[f]OMIM does not have a record for the group of diseases which comprise Ehlers–Danlos syndrome; these counts reflect those for a specific subtype (`Ehlers–Danlos Syndrome, Kyphoscoliotic Type, 1`; OMIM:225400).

capture the nuanced semantic drift between concepts. For example, NCIt and MonDO have distinct unique concepts for a neoplasm (which is not necessarily malignant/cancer) and for familial and somatic forms, whereas in other resources they may be conflated.

MonDO captures the valid union of all external references (Xrefs) but also provides a curated assessment for where the two diseases are truly logically equivalent and guaranteed to be one-to-one. Moreover, unlike other disease terminologies, MonDO (*a*) describes the provenance of these relationships and (*b*) makes explicit the nature of the relationships to other ontology terms. Although flat lists of undifferentiated Xrefs and synonyms are indeed helpful for manual browsing, machines need to know which of these are broader, narrower, exact, related to, derived from, etc. While a record may have many Xrefs listed, the designation `equivalent to` in MonDO means that, both algorithmically and by manual quality control, the relationship is logically consistent. While extensive synonyms can appear useful, in many cases these may mislead, as the synonyms are proxies for either descendant or parent terms; in other cases, hierarchies are implied but are difficult to use deterministically. The specificity and rigor provided by MonDO allows the user to make informed decisions about what sources of data (annotated to these various ontologies) might be appropriate to combine and which combinations would result in conclusions that are spurious at best or incorrect diagnoses at worst.

Mappings are created using a variety of different approaches, ranging from completely manual to semiautomated and completely automated. In general, expert-provided mappings are expected to contain fewer errors, but they are hard to keep current and may commonly have omissions, especially as referenced vocabularies evolve. Automated approaches typically leverage lexical similarity or graph-based matching. An example of a mapping tool is UMLS metamap (**https://metamap.nlm.nih.gov/**) (the same techniques often underpin named-entity recognition approaches; see below). Sometimes mappings are created by the same creators of the ontology and distributed alongside it (for example, in the commonly used OBO format, these appear as Xrefs). In other cases, mappings may be created by third parties. Examples of an automated approach by a third party are the mappings that can be explored in the BioPortal mappings system (**http://bioportal.bioontology.org/mappings**), but these still require review by experts and computational quality assurance (103). Another example of a system that can be used to explore mappings is the OxO web-based tool, a companion to the Ontology Lookup Service (**https://www.ebi.ac.uk/spot/oxo/**). The precise semantics of mappings vary according to who produces them—even within an ontology. Sometimes the intent is to provide one-to-one mappings between identical disease concepts, but in practice, mappings rarely have this desirable property, and imprecision and semantic drift mean that mappings cannot reliably be treated as equivalence relations. In some cases, mappings are annotated with a semantic relationship: For example, ORDO annotates mappings to OMIM as being either exact, broad-to-narrow, or narrow-to-broad.

Disease Stratification and Discovery

Ever increasing amounts of genomic, imaging, EHR, and other data are becoming available for disease research and health care, but existing computational infrastructure is not well prepared to fully exploit the potential of this deluge. Recent advances in genomic research have led to discoveries of disease subtypes with therapeutic implications, but this knowledge is currently difficult to integrate into hospital information technology and even academic research databases. An example of a disease that could benefit from tighter integration is systemic lupus erythematosus (SLE), an autoimmune condition with severe and varied manifestations. The FDA has approved only one drug for SLE treatment in the past 50 years, and several recent phase III trials of novel therapeutics proved unsuccessful, perhaps because of an underlying molecular heterogeneity of this

disease. Precision medicine aims to identify disease subgroups that have well-defined pathogenetic abnormalities and thus are likely to have a relatively homogeneous response to targeted treatments. In principle, this will allow clinical trials to be conducted on specific subgroups, thereby maximizing the response and minimizing adverse effects. A recent study on blood transcriptome profiling of 158 pediatric lupus patients identified seven molecular subgroups and suggested stratified treatment approaches for individual subgroups (104). One of the challenges to making the maximum use of these findings is the fact that there is currently no terminology or ontology that can be used to identify patient samples and results in studies conducted at different centers, and there is no way of identifying patients assigned to these subgroups in clinical terminologies such as SNOMED-CT. SLE is a great example of where it would be useful to have computational definitions of subgroups linked to disease and phenotype ontologies in a way that would support translational bioinformatics and research.

What researchers need are machine-readable disease definitions that would support and accelerate the entire endeavor of precision medicine. The most prominent proposal was made in 2011 by the Committee on a Framework for Developing a New Taxonomy of Disease (105). The committee suggested an information commons in which data on large numbers of patients would be derived from normal clinical care (e.g., EHRs or patient-derived data) and a knowledge network that would integrate the clinical data with other data and knowledge about fundamental biological processes. Precision medicine is about finding well-defined stratifications (subgroups of patients and the treatments that work best for them); the most useful classifications of disease are those that make these stratifications possible. Current taxonomies serve many other needs such as hospital administration and billing and are not necessarily the optimum foundation for the disease taxonomy of the future. Instead, the committee suggested that an ideal taxonomy for precision medicine would be based on clinical, environmental, genetic, genomic, and epigenomic data and knowledge. While this vision has not yet been fully realized, programs such as the NCATS Data Translator initiative are exploring strategies for developing such a taxonomy.

While current disease ontologies are far from perfect, they already enable sophisticated analysis of medical data in many ways. One of the classic uses of GO was to determine the characteristic GO terms (i.e., the terms that show statistically significant overrepresentation) in a set of overexpressed genes in a high-throughput experiment using microarrays or RNA sequencing (for recent examples, see References 106–108). These methods have been extended to integrate GO, HPO, or DO annotations to identify terms enriched with disease-relevant data (109, 110). Further, HPO disease–phenotype annotations can be used to integrate phenotypic analysis into larger genomic investigations of disease. For instance, phenotypic analysis was used to classify the pathomechanism of chromosomal deletions as either enhancer adoption or gene dosage based on an analysis of tissue-specific enhancers and genes adjacent to the deletions as compared to the phenotypic features of the patients (111). GREAT (Genomic Regions Enrichment of Annotations Tool) performs functional interpretation of regulatory regions by enrichment analysis with various ontology terms (112). GIANT (Genome-Scale Integrated Analysis of Gene Networks in Tissues) investigates genome-wide functional interaction networks for human tissues and cell types and allows visualization of interacting genes weighted according to functional similarity (113). Several other recent works show how gene–disease associations can be used for predicting therapeutic targets (114) or can be analyzed or predicted using ontological analysis (115, 116).

CONCLUSIONS

There are numerous disease ontologies with varying degrees of logical complexity and focus of disease areas. Some disease ontologies contain phenotypic and other disease attributes, whereas

others create separate annotations to support their use. Many disease and phenotype ontologies are in wide use across a variety of biomedical applications, and their use is increasing as better methods of patient stratification, care management, diagnostics, and data integration are required with new medical advances. However, defining disease for robust computational use has been nontrivial, and new ontological and statistical methods are needed. Reconciling existing resources' definitions of disease in the context of data integration or multiplatform analyses has been difficult, but methods are emerging that can reduce redundancy and best leverage the advantages of each ontology. Determining when to retire a disease ontology standard—or even when or how to evolve one—is nontrivial and dependent on social, technical, and financial factors. Disease and phenotype ontologies are key resources in biomedical data science for data classification and inference of new knowledge.

DISCLOSURE STATEMENT

C.G.C. led the Revision Steering Group for ICD-11 at the WHO from 2007 to 2017.

ACKNOWLEDGMENTS

Special thanks to Lilly Winfree for assisting with editorial management, to Nicole Vasilevsky for curation, and to Tim Putman for publication identifier mapping scripts.

LITERATURE CITED

1. Robinson PN, Bauer S. 2011. *Introduction to Bio-Ontologies*. Boca Raton, FL: CRC
2. Gruber TR. 1993. A translation approach to portable ontology specifications. *Knowl. Acquis.* 5(2):199–220
3. Cornet R, Chute CG. 2016. Health concept and knowledge management: twenty-five years of evolution. *Yearb. Med. Inform.* 2016:S32–41
4. Munsche H, Whitaker HA. 2012. Eighteenth century classification of mental illness: Linnaeus, de Sauvages, Vogel, and Cullen. *Cogn. Behav. Neurol.* 25(4):224–39
5. Starkstein SE, Berrios GE. 2015. The "preliminary discourse" to *Methodical Nosology*, by François Boissier de Sauvages (1772). *Hist. Psychiatry* 26(4):477–91
6. Knibbs GH. 1929. The International Classification of Disease and Causes of Death and its revision. *Med. J. Aust.* 1:2–12
7. NCI (Natl. Cancer Inst.). *Class: disease or disorder*. Term Defin., National Cancer Institute Thesaurus (NCIt) OBO Edition. **http://purl.obolibrary.org/obo/NCIT_C2991**
8. NLM (Natl. Libr. Med.). *Disease*. Term Defin., Medical Subject Headings (MeSH). **https://identifiers.org/MESH:D004194**
9. Robinson PN. 2012. Deep phenotyping for precision medicine. *Hum. Mutat.* 33(5):777–80
10. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, et al. 2017. The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* 45(D1):D865–76
11. Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S. 2012. *OWL 2 web ontology language primer*. Tech. Rep., World Wide Web Consort. (W3C), Dec. 11
12. Robinson JR, Wei WQ, Roden DM, Denny JC. 2018. Defining phenotypes from clinical data to drive genomic research. *Annu. Rev. Biomed. Data Sci.* 1:69–92
13. Mungall CJ, Gkoutos G, Smith C, Haendel M, Lewis S, Ashburner M. 2010. Integrating phenotype ontologies across multiple species. *Genome Biol.* 11(1):R2
14. Mulder N, Nembaware V, Adekile A, Anie KA, Inusa B, et al. 2016. Proceedings of a sickle cell disease ontology workshop—towards the first comprehensive ontology for sickle cell disease. *Appl. Transl. Genom.* 9:23–29
15. WHO (World Health Organ.). *History of the development of the ICD*. Geneva: WHO. **http://www.who.int/classifications/icd/en/HistoryOfICD.pdf**

16. Graunt J. 1939. *Natural and Political Observations Made Upon the Bills of Mortality*, ed. WF Willcox. Baltimore, MD: Johns Hopkins Univ. Press
17. Cimino JJ. 1998. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf. Med.* 37(4–5):394–403
18. WHO (World Health Organ.). 2016. *ICD-11 revision conference report*. Conf. Rep., 12–14 Oct., Tokyo, Japan. Geneva: WHO. **http://who.int/classifications/network/meeting2016/ICD-11RevisionConferenceReportTokyo.pdf**
19. Rodrigues J-M, Schulz S, Rector A, Spackman K, Üstün B, et al. 2013. Sharing ontology between ICD 11 and SNOMED CT will enable seamless re-use and semantic interoperability. *Stud. Health Technol. Inform.* 192:343–46
20. Rodrigues J-M, Robinson D, Della Mea V, Campbell J, Rector A, et al. 2015. Semantic alignment between ICD-11 and SNOMED CT. *Stud. Health Technol. Inform.* 216:790–94
21. Rodrigues J-M, Schulz S, Rector A, Spackman K, Millar J, et al. 2014. ICD-11 and SNOMED CT Common Ontology: circulatory system. *Stud. Health Technol. Inform.* 205:1043–47
22. Chute CG, Huff SM, Ferguson JA, Walker JM, Halamka JD. 2012. There are important reasons for delaying implementation of the new ICD-10 coding system. *Health Aff.* 31(4):836–42
23. Jordan EP, ed. 1932. *Standard Nomenclature of Diseases and Operations*. New York: McGraw-Hill. 1st ed.
24. Thompson ET, Hayden AC. 1961. *Standard Nomenclature of Diseases and Operations*. New York: McGraw-Hill. 5th ed.
25. Comm. Nomencl. Classif. Dis. 1965. *Systematized Nomenclature of Pathology*. Skokie, IL: Coll. Am. Pathol.
26. Côté RA, Sharpe WD, eds. 1976. *Systematized Nomenclature of Medicine: SNOMED*. Skokie, IL: Coll. Am. Pathol.
27. Côté RA, ed. 1993. *The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International*. Northfield, IL: Coll. Am. Pathol.
28. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. 1994. Toward a medical-concept representation language. The Canon Group. *J. Am. Med. Inform. Assoc.* 1(3):207–17
29. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH. 1996. Gálapagos: computer-based support for evolution of a convergent medical terminology. *AMIA Annu. Symp. Proc.* 1996:269–73
30. Campbell KE, Cohn SP, Chute CG, Shortliffe EH, Rennels G. 1998. Scalable methodologies for distributed development of logic-based convergent medical terminology. *Methods Inf. Med.* 37(4–5):426–39
31. Baader F, Brandt S, Lutz C. 2005. Pushing the EL envelope. *Proc. Int. Jt. Conf. Artif. Intell., 19th, 30 July–5 Aug., Edinb., Scotl.*, pp. 364–69. San Francisco: Morgan Kaufmann
32. Kudla KM, Blakemore M. 2001. SNOMED takes the next step. *J. AHIMA* 72(7):62, 64–68
33. Benson T. 2011. The history of the Read codes: the inaugural James Read Memorial Lecture 2011. *Inform. Prim. Care.* 19(3):173–82
34. Spackman KA. 2001. Normal forms for description logic expressions of clinical concepts in SNOMED RT. *AMIA Annu. Symp. Proc.* 2001:627–31
35. Campbell JR, Talmon G, Cushman-Vokoun A, Karlsson D, Scott Campbell W. 2016. An extended SNOMED CT concept model for observations in molecular genetics. *AMIA Annu. Symp. Proc.* 2016:352–60
36. Hardiker N. 2016. Harmonising ICNP and SNOMED CT: a model for effective collaboration. *Stud. Health Technol. Inform.* 225:744–45
37. Martínez-Salvador B, Marcos M, Mañas A, Maldonado JA, Robles M. 2016. Using SNOMED CT expression constraints to bridge the gap between clinical decision-support systems and electronic health records. *Stud. Health Technol. Inform.* 228:504–8
38. Ochs C, Geller J, Perl Y, Chen Y, Xu J, et al. 2015. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. *J. Am. Med. Inform. Assoc.* 22(3):507–18
39. Elhanan G, Ochs C, Mejino JLV Jr., Liu H, Mungall CJ, Perl Y. 2017. From SNOMED CT to Uberon: transferability of evaluation methodology between similarly structured ontologies. *Artif. Intell. Med.* 79:9–14
40. NLM (Natl. Libr. Med.). 2014. *2017AB UMLS® release notes and bugs*. Release Doc., updated Nov. 6, 2017. **https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html**

41. NLM (Natl. Libr. Med.). 2009. *Metathesaurus*. Fact Sheet, updated Apr. 12, 2016. **https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/**
42. Jensen M, Cox AP, Chaudhry N, Ng M, Sule D, et al. 2013. The neurological disease ontology. *J. Biomed. Semant.* 4(1):42
43. NLM (Natl. Libr. Med.). 2006. *SPECIALIST Lexicon*. Fact Sheet, updated May 24, 2012. **https://www.nlm.nih.gov/pubs/factsheets/umlslex.html**
44. NLM (Natl. Libr. Med.). *The UMLS Semantic Network*. Fact Sheet, updated Oct. 5, 2015. **https://semanticnetwork.nlm.nih.gov/**
45. McCray AT. 2003. An upper-level ontology for the biomedical domain. *Comp. Funct. Genom.* 4(1):80–84
46. Pecina P, Dušek O, Goeuriot L, Hajič J, Hlaváčová J, et al. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artif. Intell. Med.* 61(3):165–85
47. Merabti T, Abdoune H, Letord C, Sakji S, Joubert M, Darmoni SJ. 2011. Mapping the ATC classification to the UMLS metathesaurus: some pragmatic applications. *Stud. Health Technol. Inform.* 166:206–13
48. Ceusters W, Smith B, Goldberg L. 2005. A terminological and ontological analysis of the NCI Thesaurus. *Methods Inf. Med.* 44(4):498–507
49. Jensen MA, Ferretti V, Grossman RL, Staudt LM. 2017. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* 130(4):453–59
50. NCI (Natl. Cancer Inst.). *NCIt Neoplasm Core terminology files*. File Release, accessed 6 Mar. 2018, NCI, Rockville, MD. **https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/Neoplasm/About_Core.html**
51. NCI (Natl. Cancer Inst.). *NCI Term Browser*. Terminol. Resour. **https://ncit.nci.nih.gov/ncitbrowser/pages/multiple_search.jsf?nav_type = terminologies**
52. Min H, Zheng L, Perl Y, Halper M, De Coronado S, Ochs C. 2017. Relating complexity and error rates of ontology concepts: more complex NCIt concepts have more errors. *Methods Inf. Med.* 56(3):200–8
53. Mougin F, Bodenreider O. 2008. Auditing the NCI thesaurus with semantic web technologies. *AMIA Annu. Symp. Proc.* 2008:500–4
54. Jiang G, Solbrig HR, Chute CG. 2012. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. *J. Am. Med. Inform. Assoc.* 19:e129–36
55. Min H, Mobahi H, Irvin K, Avramovic S, Wojtusiak J. 2017. Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. *J. Biomed. Semant.* 8(1):39
56. Bertaud Gounot V, Donfack V, Lasbleiz J, Bourde A, Duvauferrier R. 2011. Creating an ontology driven rules base for an expert system for medical diagnosis. *Stud. Health Technol. Inform.* 169:714–18
57. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, et al. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* 49(2):170–74
58. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11):1251–55
59. NCI (Natl. Cancer Inst.). *NCIt OBO edition*. File Release. **https://github.com/NCI-Thesaurus/thesaurus-obo-edition**
60. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, et al. 2016. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* 7(1):44
61. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 13(1):R5
62. Davies SC. 2017. *Annual report of the Chief Medical Officer 2016: generation genome*. Annu. Rep., U.K. Dep. Health, London. **https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/631043/CMO_annual_report_generation_genome.pdf**
63. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43:D789–98
64. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. 2017. Clinical practice guidelines for rare diseases: the Orphanet database. *PLOS ONE* 12(1):e0170365
65. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, et al. 2014. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* 42:D993–1000

66. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, et al. 2015. The Human Phenotype Ontology: semantic unification of common and rare disease. *Am. J. Hum. Genet.* 97(1):111–24
67. Rodriguez JC, González GA, Fresno C, Llera AS, Fernández EA. 2016. Improving information retrieval in functional analysis. *Comput. Biol. Med.* 79:10–20
68. Liaw S-T, Taggart J, Yu H. 2014. EHR-based disease registries to support integrated care in a health neighbourhood: an ontology-based methodology. *Stud. Health Technol. Inform.* 205:171–75
69. Mao Y, Lu Z. 2017. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *J. Biomed. Semant.* 8(1):15
70. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, et al. 2013. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J. Am. Med. Inform. Assoc.* 20:e147–54
71. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, et al. 2011. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. *AMIA Annu. Symp. Proc.* 2011:274–83
72. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* 12(7):e0175508
73. Chen W, Kowatch R, Lin S, Splaingard M, Huang Y. 2015. Interactive cohort identification of sleep disorder patients using natural language processing and i2b2. *Appl. Clin. Inform.* 6(2):345–63
74. Cui L, Bozorgi A, Lhatoo SD, Zhang G-Q, Sahoo SS. 2012. EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. *AMIA Annu. Symp. Proc.* 2012:1191–200
75. Banda JM, Seneviratune M, Hernandez-Boussard T, Shah NH. 2018. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu. Rev. Biomed. Data Sci.* 1:53–68
76. Zhang Y-F, Gou L, Zhou T-S, Lin D-N, Zheng J, et al. 2017. An ontology-based approach to patient follow-up assessment for continuous and personalized chronic disease management. *J. Biomed. Inform.* 72:45–59
77. Rosier A, Mabo P, Temal L, Van Hille P, Dameron O, et al. 2016. Remote monitoring of cardiac implantable devices: ontology driven classification of the alerts. *Stud. Health Technol. Inform.* 221:59–63
78. Wunsch G, da Costa CA, Righi RR. 2017. A semantic-based model for triage patients in emergency departments. *J. Med. Syst.* 41(4):65
79. Alexiou A, Psiha M, Vlamos P. 2015. Towards an expert system for accurate diagnosis and progress monitoring of Parkinson's disease. *Adv. Exp. Med. Biol.* 822:151–64
80. Carletti G, Giuliodori P, Di Pietri V, Peretti A, Amenta F. 2016. An ontology-based consultation system to support medical care on board seagoing vessels. *Int. Marit. Health* 67(1):14–20
81. Maurice P, Dhombres F, Blondiaux E, Friszer S, Guilbaud L, et al. 2017. Towards ontology-based decision support systems for complex ultrasound diagnosis in obstetrics and gynecology. *J. Gynecol. Obstet. Hum. Reprod.* 46(5):423–29
82. Abidi S. 2017. A knowledge-modeling approach to integrate multiple clinical practice guidelines to provide evidence-based clinical decision support for managing comorbid conditions. *J. Med. Syst.* 41(12):193
83. Lin Y, Staes CJ, Shields DE, Kandula V, Welch BM, Kawamoto K. 2015. Design, development, and initial evaluation of a terminology for clinical decision support and electronic clinical quality measurement. *AMIA Annu. Symp. Proc.* 2015:843–51
84. Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, et al. 2013. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research* 2:30
85. Smedley D, Oellrich A, Köhler S, Ruef B, Sanger Mouse Genet. Proj., et al. 2013. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database* 2013:bat025
86. Bauer S, Köhler S, Schulz MH, Robinson PN. 2012. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics* 28(19):2502–8
87. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, et al. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85(4):457–64

88. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, et al. 2014. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* 6(252):252ra123
89. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, et al. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* 10(12):2004–15
90. Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, et al. 2016. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.* 18(6):608–17
91. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, et al. 2014. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* 94(4):599–610
92. Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, et al. 2017. Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.* 376(1):21–31
93. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, et al. 2015. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.* 36(10):915–21
94. Health Level Seven Int. *HL7 implementation guide for CDA® release 2: consolidated CDA templates for clinical notes*. Prod. Brief, updated Feb. 6, 2018, Ann Arbor, MI. **http://www.hl7.org/implement/standards/product_brief.cfm?product_id=379**
95. Boyce RD, Ryan PB, Norén GN, Schuemie MJ, Reich C, et al. 2014. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. *Drug Saf.* 37(8):557–67
96. Chakrabarti S, Sen A, Huser V, Hruby GW, Rusanov A, et al. 2017. An interoperable similarity-based cohort identification method using the OMOP common data model version 5.0. *Int. J. Healthc. Inf. Syst. Inform.* 1(1):1–18
97. Clin. Cancer Genome Task Team Glob. Alliance Genom. Health. 2017. Sharing clinical and genomic data on cancer—the need for global solutions. *N. Engl. J. Med.* 376(21):2006–9
98. Kock-Schoppenhauer A-K, Kamann C, Ulrich H, Duhm-Harbeck P, Ingenerf J. 2017. Linked data applications through ontology based data access in clinical research. *Stud. Health Technol. Inform.* 235:131–35
99. Nussbaum RL, Rehm HL. 2015. ClinGen and genetic testing: Drs. Nussbaum and Rehm reply. *N. Engl. J. Med.* 373(14):1379
100. Schulz S, Spackman K, James A, Cocos C, Boeker M. 2011. Scalable representations of diseases in biomedical ontologies. *J. Biomed. Semant.* 2:S6
101. Mungall CJ, Koehler S, Robinson P, Holmes I, Haendel M. 2016. k-BOOM: A Bayesian approach to ontology structure inference, with applications in disease ontology construction. bioRxiv 048843. **https://doi.org/10.1101/048843**
102. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, et al. 2017. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 45(D1):D712–22
103. Faria D, Jiménez-Ruiz E, Pesquita C, Santos E, Couto FM. 2014. Towards annotating potential incoherences in BioPortal mappings. In *The Semantic Web: ISWC 2014*, ed. P Mika, T Tudorache, A Bernstein, C Welty, C Knoblock, et al., pp. 17–32. Lect. Notes Comput. Sci. 8797. Cham, Switz.: Springer Int.
104. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, et al. 2016. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* 165(3):551–65
105. Natl. Res. Counc. 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: Natl. Acad. Press
106. Clare AJ, Wicky HE, Empson RM, Hughes SM. 2017. RNA-sequencing analysis reveals a regulatory role for transcription factor *Fezf2* in the mature motor cortex. *Front. Mol. Neurosci.* 10:283
107. Eidsaa M, Stubbs L, Almaas E. 2017. Comparative analysis of weighted gene co-expression networks in human and mouse. *PLOS ONE* 12(11):e0187611
108. Menche J, Guney E, Sharma A, Branigan PJ, Loza MJ, et al. 2017. Integrating personalized gene expression profiles into predictive disease-associated gene pools. *NPJ Syst. Biol. Appl.* 3:10
109. Yu G, Wang L-G, Yan G-R, He Q-Y. 2015. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31(4):608–9

110. LePendu P, Musen MA, Shah NH. 2011. Enabling enrichment analysis with the Human Disease Ontology. *J. Biomed. Inform.* 44:S31–38
111. Ibn-Salem J, Köhler S, Love MI, Chung H-R, Huang N, et al. 2014. Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.* 15(9):423
112. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28(5):495–501
113. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, et al. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47(6):569–76
114. Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, et al. 2016. Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. *J. Biomed. Semant.* 7:8
115. Opap K, Mulder N. 2017. Recent advances in predicting gene–disease associations. *F1000Research* 6:578
116. Hu Y, Zhou W, Ren J, Dong L, Wang Y, et al. 2016. Annotating the function of the human genome with Gene Ontology and Disease Ontology. *Biomed Res. Int.* 2016:4130861
117. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, et al. 1996. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin. Chem.* 42(1):81–90
118. Maritz R, Aronsky D, Prodinger B. 2017. The International Classification of Functioning, Disability and Health (ICF) in electronic health records: a systematic literature review. *Appl. Clin. Inform.* 8(3):964–80
119. Mitchell JA, Loughman WD, Epstein CJ. 1980. GENFILES: a computerized medical genetics information network. II. MEDGEN: the clinical genetics system. *Am. J. Med. Genet.* 7(3):251–66
120. Jouhet V, Mougin F, Bréchat B, Thiessard F. 2017. Building a model for disease classification integration in oncology, an approach based on the national cancer institute thesaurus. *J. Biomed. Semant.* 8(1):6
121. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, et al. 2010. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26(8):1112–18
122. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, et al. 2015. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* 43:D1071–78
123. Hayman GT, Laulederkind SJF, Smith JR, Wang S-J, Petri V, et al. 2016. The Disease Portals, disease–gene annotation and the RGD disease ontology at the Rat Genome Database. *Database* 2016:baw034
124. APA (Am. Psychiatr. Assoc.). 2013. *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, VA: APA. 5th ed.
125. Cowell LG, Smith B. 2010. Infectious Disease Ontology. In *Infectious Disease Informatics*, ed. V Sintchenko, pp. 373–95. New York: Springer-Verlag
126. Gordon CL, Pouch S, Cowell LG, Boland MR, Platt HL, et al. 2013. Design and evaluation of a bacterial clinical infectious diseases ontology. *AMIA Annu. Symp. Proc.* 2013:502–11
127. Younesi E, Malhotra A, Gündel M, Scordis P, Kodamullil AT, et al. 2015. PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. *Theor. Biol. Med. Model.* 12(1):20
128. Mugzach O, Peleg M, Bagley SC, Guter SJ, Cook EH, Altman RB. 2015. An ontology for Autism Spectrum Disorder (ASD) to infer ASD phenotypes from Autism Diagnostic Interview-Revised data. *J. Biomed. Inform.* 56:333–47
129. McCray AT, Trevvett P, Frost HR. 2014. Modeling the autism spectrum disorder phenotype. *Neuroinformatics* 12(2):291–305
130. Fisher HM, Hoehndorf R, Bazelato BS, Dadras SS, King LE Jr., et al. 2016. DermO; an ontology for the description of dermatologic disease. *J. Biomed. Semant.* 7:38
131. Takatsuki T, Saito M, Kumagai S, Takayama E, Ohshima K, et al. *A RDF-based portal of biological phenotype data created in Japan*. Presented at Int. Semant. Web Conf., 15th, Kobe, Japan
132. Daughton AR, Priedhorsky R, Fairchild G, Generous N, Hengartner A, et al. 2017. An extensible framework and database of infectious disease for biosurveillance. *BMC Infect. Dis.* 17(1):549
133. Barton A, Barton A, Rosier A, Rosier A, Burgun A, Ethier J-F. 2014. The cardiovascular disease ontology. In *Formal Ontology in Information Systems*, ed. P Garbacz, O Kutz, pp. 409–14. Front. Artif. Intell. Appl. 267. Amsterdam: IOS

134. Ceusters W, Smith B. 2010. Foundations for a realist ontology of mental disease. *J. Biomed. Semant.* 1(1):10
135. Schleyer TK, Ruttenberg A, Duncan W, Haendel M, Torniai C, et al. 2013. An ontology-based method for secondary use of electronic dental record data. *AMIA Jt. Summits Transl. Sci. Proc.* 2013:234–38
136. Keerthikumar S, Raju R, Kandasamy K, Hijikata A, Ramabadran S, et al. 2009. RAPID: Resource of Asian Primary Immunodeficiency Diseases. *Nucleic Acids Res.* 37:D863–67
137. Wang L, Li M, Xie J, Cao Y, Liu H, He Y. 2017. Ontology-based systematical representation and drug class effect analysis of package insert-reported adverse events associated with cardiovascular drugs used in China. *Sci. Rep.* 7(1):13819
138. Fu X, Batista-Navarro R, Rak R, Ananiadou S. 2015. Supporting the annotation of chronic obstructive pulmonary disease (COPD) phenotypes with text mining workflows. *J. Biomed. Semant.* 6:8
139. Lin Y, Xiang Z, He Y. 2011. Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology. *J. Biomed. Semant.* 2(1):9
140. Lin Y, Sakamoto N. 2009. Ontology driven modeling for the knowledge of genetic susceptibility to disease. *Kobe J. Med. Sci.* 54(6):E290–303
141. Porter JF, Kingsland LC 3rd, Lindberg DA, Shah I, Benge JM, et al. 1988. The AI/RHEUM knowledge-based computer consultant system in rheumatology. Performance in the diagnosis of 59 connective tissue disease patients from Japan. *Arthritis Rheum.* 31(2):219–26
142. Ryerson CJ, Corte TJ, Lee JS, Richeldi L, Walsh SLF, et al. 2017. A standardized diagnostic ontology for fibrotic interstitial lung disease: an international working group perspective. *Am. J. Respir. Crit. Care Med.* 196:1249–54
143. Mizuno S, Ogishima S, Nishigori H, Jamieson DG, Verspoor K, et al. 2016. The Pre-Eclampsia Ontology: a disease ontology representing the domain knowledge specific to pre-eclampsia. *PLOS ONE* 11(10):e0162828
144. Chen Q, Wu J, Li S, Lyu P, Wang Y, Li M. 2016. An ontology-driven, case-based clinical decision support model for removable partial denture design. *Sci. Rep.* 6:27855
145. Joseph S, Barai RS, Bhujbalrao R, Idicula-Thomas S. 2016. PCOSKB: A KnowledgeBase on genes, diseases, ontology terms and biochemical pathways associated with PolyCystic Ovary Syndrome. *Nucleic Acids Res.* 44(D1):D1032–35