CLASSIFICATIONS AND TERMINOLOGIES FOR MAPPING THE INDICATION AND ORPHAN CONDITION IN REGULATORY DOCUMENTS

Keerti Jadoenathmisier

Master Drug Innovation
Utrecht University
9393706

Writing Assignment (07/11/2022-09/12/2022)
Dutch Medicines Evaluation Board, Utrecht
Regulatory Science Department

Supervisors : Dr. ir. Marjon Pasmooij (MEB)
Dr. Lourens Bloem (UU)
Examiner : Prof. dr. Toine Egberts (UU)

Plain language summary

Informatie over de regulatoire aspecten van geneesmiddelen wordt op verschillende online platforms gepubliceerd. Zo heeft het geneesmiddelbewakingsorgaan van Europa, de European Medicines Agency (EMA) haar eigen website als voornaamste bron van informatie. Maar, andere instanties publiceren data omtrent geneesmiddelen op hun eigen website. Dit maakt het voor belanghebbenden vaak moeilijk om de voor hun relevante informatie te vinden. Om dit te vergemakkelijken heeft het nationale geneesmiddelbewakingsorgaan van Nederland, het College ter Beoordeling van Geneesmiddelen, in samenwerking met de Universiteit Utrecht een database ontworpen om de belangrijkste regulatoire gegevens van geneesmiddelen samen te brengen op een platform. Dit platform bevat informatie die betrekking heeft op alle door de EMA goedgekeurde geneesmiddelen van 1995 tot heden. Een belangrijke variabel voor de database, is de indicatie of aandoening waarvoor een geneesmiddel is goedgekeurd. Om dit aan de database toe te voegen moet deze informatie eerst gestructureerd worden. Om dit zo efficiënt mogelijk te doen, kan er gebruik gemaakt worden van ziekte classificaties en terminologieën. Er zijn echter verschillende classificaties en terminologieën beschikbaar om ziekten in kaart te brengen. Het doel van dit verslag is om een overzicht te geven van classificaties en terminologieën die het best voor regulatoire doeleinden gebruikt kunnen worden.

De classificaties en terminologieën die wij het meest geschikt achten in deze context zijn ICD-11, SNOMED-CT, MeSH, MedDRA, Orphanet en Disease Ontology. We hebben het ontstaan, doel en de structuur van deze ontologieën besproken en zijn in gegaan op hun voor- en nadelen. Vervolgens hebben we deze ontologieën in werking gezet en beoordeeld of ze geschikt zijn om de indicatie en/of aandoening waarvoor een geneesmiddel goedgekeurd is te structureren. Dit hebben wij gedaan voor de laatste 20 geneesmiddelen die t/m 17 november 2022 een wees aanwijzing kregen. Dat betekent dat deze geneesmiddelen te gebruiken zijn voor zeldzaam voorkomende aandoeningen. We hebben de weesaandoening van deze geneesmiddelen van de website van de Europese Commissie gehaald en de indicatie uit hun Summary of Product Characteristics (SmPC). Vervolgens hebben we in de verschillende ontologieën gezocht naar trefwoorden van de indicatie en weesaandoening en hebben we gekeken naar van hoeveel detail elke ontologie ons kan voorzien.

We hebben gevonden dat elke ontologie verschillende gradaties van detail biedt en dat dit direct gerelateerd is aan hun doel. Zo is MedDRA gericht op het voorzien van informatie omtrent de bijwerkingen van geneesmiddelen, terwijl Orphanet zich specifiek richt op het classificeren van weesaandoeningen. Dit maakt het kiezen van één ontologie om de aandoening of indicatie te structureren nogal moeilijk. Echter, uit alle geteste ontologieën gaf SNOMED-CT ons de meest omvattende beschrijving van de ziekten en voldoende hoeveelheid details. Op basis hiervan en de verschillende andere toepassingen die SNOMED-CT biedt, raden wij SNOMED-CT als meest geschikte ontologie aan voor regulatoire doeleinden.

Abstract

Background: To make data on medicines regulation and marketing approval openly available, the Dutch Medicines Evaluation Board and Utrecht University are developing the European Medicines Regulatory Database (EMRD). To efficiently map data on the orphan condition and indication of medicines approved by the European Commission, we deem to make use of a disease ontology. However, many different disease ontologies exist. The aim of this review was to identify, describe and assess the main ontologies eligible for regulatory purposes.

Methods: For an initial overview of disease ontologies, a search was conducted in FairSharing.org and ontologies that were eligible for regulatory purposes were selected. We supplemented these with ontologies identified in the scientific literature. Information on the different ontologies was compiled from their website, their user guides or scientific literature. Next, we extracted the orphan condition from the European Commission Union Register of medicinal products for human use and the indication from the initial Summary of Product Characteristics (SmPC) of the last 20 human medicinal products that received one or more orphan designation upon marketing authorisation until the 17th of November 2022. Keywords from the orphan conditions and indications were searched in the ontologies, through their own website and/or via the Unified Medical Language Systems (UMLS).

Results: We selected ICD-11, SNOMED-CT, MeSH, MedDRA, Orphanet and Disease Ontology as eligible ontologies for regulatory purposes. SNOMED-CT (91%) captured the most orphan conditions, while MedDRA (73%) covered the least conditions. A similar trend was observed for the indications, with SNOMED-CT (60%) capturing most indications and MedDRA (20%) the least.

Conclusions: Each ontology provides different degrees of detail, and this is directly related to the purpose for which they were created. SNOMED-CT was the most comprehensive and captured the most orphan conditions and indications of medicines in our review. Thus, we deem SNOMED-CT to be best suitable for mapping the orphan condition or indication of medicines in regulatory documents in general, and specifically for the EMRD. Importantly, given the frequent use of SNOMED-CT in healthcare, it may also facilitate interoperability between regulatory and healthcare data.

Table of contents

List of abbreviations

Abbreviation	Definition
AADC	Aromatic L-amino Acid Decarboxylase
CM	Content Model
CMV	Cytomegalovirus
CTV-3	Clinical Terms Version 3
DAG	Directed Acyclic Graph
DL	Description Logic
DO	Disease Ontology
EC	European Commission
EB	Epidermolysis Bullosa
EMA	European Medicines Agency
FSN	Fully Specified Name
GPA	Granulomatosis with polyangiitis
hATTR amyloidosis	Hereditary Transthyretin-Mediated Amyloidosis
HL7	Health Level 7
HLGT	High Level Group Term
ICD	International Classification of Diseases and Health Related Problems
ICD-11 MMS	ICD Mortality and Morbidity Statistics
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
LLT	Low Level Term
MedDRA	Medical Dictionary for Regulatory Activities
MeSH	Medical Subject Headings
MPA	Microscopic polyangiitis
NLM	National Library of Medicine
OBO	Open Biomedical Ontologies
ORDO	Orphanet Rare Disease Ontology
OWL	Web Ontology Language
PT	Preferred Term
SmPC	Summary of Product Characteristics
SMQ	Standardized MedDRA Query
SNOMED-CT	Systemized Nomenclature of Medicine-Clinical Terms
SOC	System Organ Class
UMLS	Unified Medical Language Systems
URI	Unique Resource Identifier
WHO	World Health Organisation

Introduction

Background

Drug regulatory agencies, including the European Medicines Agency (EMA) and the European Commission (EC), provide information on drug authorisation and specific product information on their websites, however they do so in a non-integrated matter. These data include information on attributes of a drug's development and life cycle, including registrational studies, the drug label and label changes, safety and adverse event history, regulatory review and agency interactions. Drug developers are increasingly using these data to gain insights into their expectations and applications of regulatory policy. However, the use of different sources of information leads to several challenges, including difficulties to link and analyse data, but also to validate the accuracy of data and can lead to possible duplication of work[1,2]. Other stakeholders besides drug developers, such as patients, health care professionals and academic researchers, would also benefit from a single platform which compiles data on the regulatory actions and activities of drugs[2].

To aid the establishment of such a platform of regulatory actions for European pharmaceuticals, the Dutch Medicines Evaluation Board collaborated with Utrecht University to develop a Regulatory Science Database to make data on medicines regulation and marketing approval openly available. The database, called European Database For Pharmaceutical Policy & Regulation, consists of key regulatory data variables covering all centrally authorised medicines by EMA from 1995-to date acquired from the European Commission Union Register, EMA website and EPARs. The variables include general information such as drug brand name and active substance, but also identifying information (e.g. application number), medicines designation (e.g. advanced therapy medicinal products or orphan designation), in addition to legal information and authorisation timing. To optimise the use of the database for different stakeholders, the condition and indication for which a medicinal product has been approved will also be added to the database. The condition is usually assigned before a drug is approved for the market, for example upon receiving the status of an orphan designation. The indication, on the other hand, is assigned upon marketing approval and is usually more specific than the condition. The condition and indication are supplied by the European Commission in the Union Register of medicinal products for human use and the Summary of Product Characteristics (SmPCs), respectively. To efficiently map the condition and indication of a drug, we deem to make use of disease classifications and terminologies. However, there are currently many different disease classifications and terminologies available which makes it difficult to specifically choose one for this purpose. The aim of this report is to give an overview of classifications and terminologies for diseases which can be used for regulatory purposes and discuss their advantages and disadvantages, as well as their interoperability. We further put these classifications and terminologies to action and assess how much detail they provide for classifying the therapeutic indication of recently authorized medicinal products. This will aid the decision of which system to use for extracting the indication and/or condition out of the SmPCs of drugs for the Regulatory Science Database.

Disease terminologies, ontologies and classifications

The rise of the digital age, led to an increase of individually developed standards for identification, citation and reporting of data, limiting the retrieval, reproducibility and reusability of research[3]. It is essential to have shared standards in health information management through coding so that users can access, combine and share comparable data with other healthcare organizations or settings. Terminologies, ontologies and classification systems are key in this function. A terminology is a collection of terms representing a concept. A terminology on its own does not necessarily have a structure and can range from being simple (e.g. one term for one concept) to more complex (e.g. multiple terms for one concept). Additional complexity is added if the same terms can represent different concepts. In this case the terminology adapts a certain structure. It can be structured as a taxonomy in which concepts are represented in a hierarchy (i.e. terms are represented using parent-child relations). Terminologies can also adapt a thesauri structure, which lists terms, their natural language meanings and relations of synonyms between terms[4]. Much effort has been invested in standardizing medical terminology to improve the representation of medical knowledge, storage in electronic medical records, retrieval, reuse and for effective transmission between users. Furthermore, standardization of medical terminology improves the efficiency for secondary uses such as research, public health and regulatory activities[5]. An ontology is a comprehensively structured terminology which provides additional complex relations between terms, as opposed to the simple parent-child relations in taxonomies. Ontologies are expressed in a formal language intended for computational use. They provide a robust foundation for describing knowledge classifications, including disease classifications. The majority of biomedical ontologies make use of a knowledge representation language such as the Open Biomedical Ontologies (OBO) format or Web Ontology Language (OWL), which are based on description logic (DL) and allow for logical relationships between terms[6]. Since the recognition of the value of the use of ontologies for representing disease classifications, there has been a rise in their adoption and use in the biomedical field. The huge diversity in disease ontologies is due to the variety of structures, strengths, limitations and uses of these classifications[7]. Box 1 gives a list of terms which will be encountered throughout this report. We included definitions of the terms which are relevant for our purposes.

Box 1

Classification A system that arranges or organizes related entities. Classifications can be based on ontologies.

Description logic A family of formal knowledge representation languages. It provides a logical formalism for ontologies.

Hierarchy Arrangement of terms that are represented as being ‘above’ or ‘below’ one another. In this context, the upper term is referred to as ‘parent’ of the term it precedes, the ‘child’.

Interoperability The ability of different systems operating effectively together and exchanging information with other one another.

Mono-hierarchy A hierarchy in which a child is only assigned to one parent.

Ontology A comprehensively structured terminology which provides additional complex relations between terms, as opposed to the simple parent-child relations in taxonomies.

Ontology mapping Process of finding similarities among concepts in different ontologies. Mapping is required for enabling interoperability of ontologies.

Poly-hierarchy A hierarchy in which a child can be assigned to multiple parents. Also referred to as *‘a poly-parental structure’.*

Post-coordination The ability to expand existing codes with additional information in order to provide more detail in an ontology.

Terminology A collection of terms representing a concept. Can differ in complexity and can be structured as a *taxonomy* in which concepts are represented in a hierarchy

Methods

For an overview of the available disease terminologies, a search was conducted in FairSharing.org. FairSharing.org is a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies and is maintained by a community consisting of stakeholders, representing academia, industry, funding agencies, standards organizations and more [3]. The following search query was used : Standard (Registry) and terminology artefact (Record type) and biomedical science (Subject) and homo sapiens (Species) and disease (Domain) and ready (Output status). From the resulting ontologies we selected the ontologies that can be used for general disease classification in the context of regulatory application. We therefore also added ontologies which were not retrieved by FairSharing.org, but were still deemed relevant by literature for our purpose. Information presented here on the different ontologies was either extracted from their website, the user guides or literature retrieved from Google Scholar or Web of Science.

In order to assess which ontologies can be used to map the indication or condition of a given drug out of the SmPC, we identified the last 20 human medicinal products that received one or more initial orphan designations until the 17th of November 2022. We extracted the orphan designation from the product page of the European Commission Union Register of medicinal products for human use. The therapeutic indication of the drugs was extracted from section 4.1 of their SmPC also published in the Union Register of medicinal products for human use of the European Commission. Keywords from the condition and therapeutic indication were searched in the ontologies to determine how much detail each ontology provides. The ontologies were either accessed through their own website or via the Unified Medical Systems Language (UMLS) or a combination of both (Fig. 1). For each ontology we concluded whether or not (yes/no) it represents the condition and indication based on a set of criteria. ‘Yes’ was assigned to an ontology if all aspects of the condition or indication were included (e.g. severity and mutations). However, if the ontology lacked the inclusion of at least one aspect, the ontology was assigned ‘no’ (e.g. severity is included, but not the mutations).

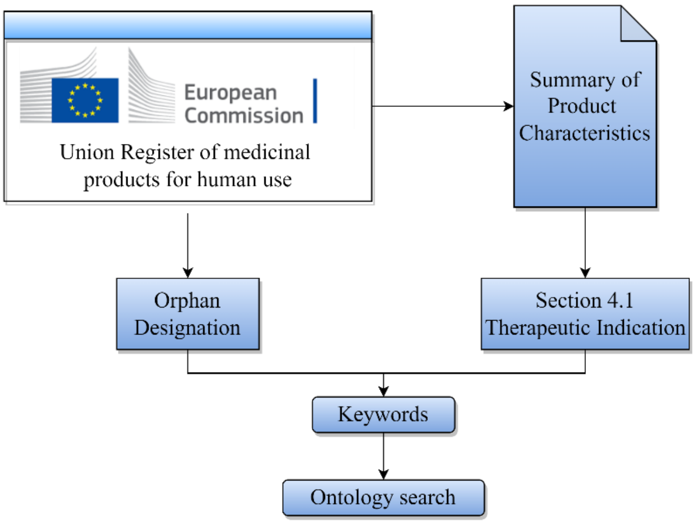
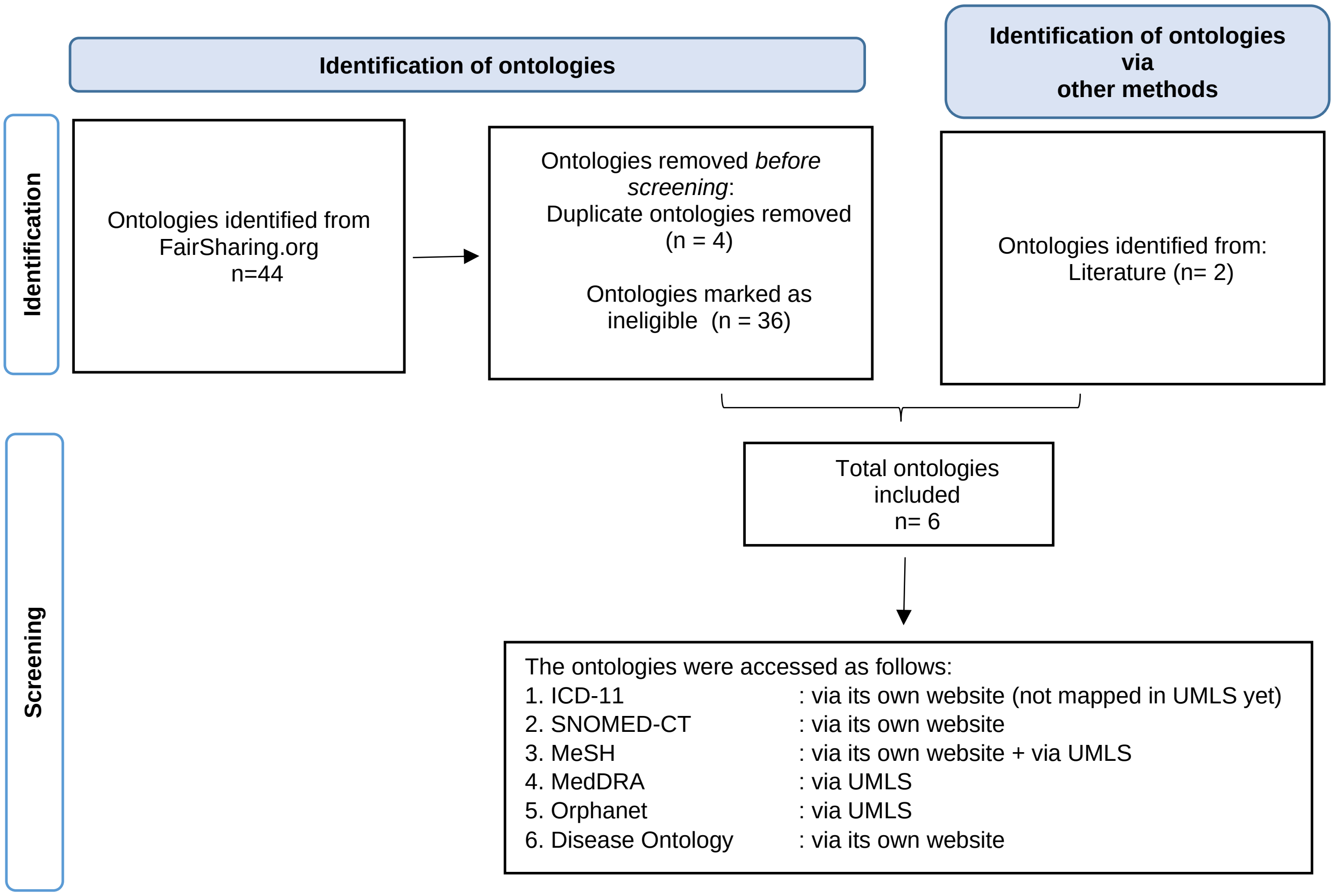

European Commission
Union Register of medicinal products for human use
Summary of Product Characteristics
Orphan Designation
Section 4.1 Therapeutic Indication
Keywords
Ontology search

Figure 1: *Overview of methodology used for retrieval of condition and therapeutic indication of the drugs, as well as the route of access to the different ontologies. Created with BioRender.com*

Results

Disease ontologies

The search query in FairSharing.org yielded 44 ontologies in total (Table S1). In this list, five versions of International Classification of Diseases (ICD) were present and thus four duplicates were removed. The majority of the ontologies identified by FairSharing.org were specific for certain use cases. For example, the results included ontologies which were specific for a certain disease domain (e.g. COPD ontology) or other purposes (e.g. clinical measurement ontology). As we aimed to identify ontologies that can be used for general disease classification for regulatory purposes, we eliminated the ontologies that did not meet this criterion. Furthermore, using 'disease' as domain in the search query in FairSharing.org omitted important ontologies such as Systemized Nomenclature of Medicine Clinical Terms (SNOMED-CT) and Medical Subject Headings (MeSH)[8]. Thus we expanded the selected ontologies of FairSharing.org with ontologies found in literature and ended up with the following ontologies: ICD, SNOMED-CT, MeSH, Medical Dictionary for Regulatory Activities (MedDRA), Orphanet and Disease Ontology (DO). Figure 2 depicts the workflow for the selection procedure of the ontologies and how each ontology was accessed. For each of these six ontologies the history, structure and advantages and disadvantages are described in the following sections.


Identification of ontologies
Identification of ontologies via other methods
Identification
Screening
Ontologies identified from FairSharing.org n=44
Ontologies removed *before screening*: Duplicate ontologies removed (n = 4)
Ontologies marked as ineligible (n = 36)
Ontologies identified from: Literature (n= 2)
Total ontologies included n= 6

The ontologies were accessed as follows:

1. ICD-11	: via its own website (not mapped in UMLS yet)
2. SNOMED-CT	: via its own website
3. MeSH	: via its own website + via UMLS
4. MedDRA	: via UMLS
5. Orphanet	: via UMLS
6. Disease Ontology	: via its own website

Figure 2: *Workflow of the identification and screening of ontologies*

International Classification of Diseases

History and purpose

The International Classification of Diseases and Health Related Problems (ICD) is a tool for recording, reporting and classifying health-related conditions. It originates from several early classification systems that stem back around 150 years. From its first revision (ICD-1), being a short list of causes of death, ICD has evolved into a comprehensive classification and terminology system. The initial purpose of ICD is to allow comparable statistics on mortality and morbidity data collected in different countries or regions and at different times, but over the decades its use cases have expanded over health statistics to decision support, resource allocation, reimbursement, guidelines and more. It is the most dominantly used classification system worldwide and contains categories for human diseases and disorders, health-related conditions external causes of illness or death, anatomy, sites, activities, medicines, vaccines and more[9,10]. ICD is updated every decade by the World Health Organization (WHO), who also mandates the implementation and use of the most recent revision of ICD in all member states of the WHO. ICD is one of the three reference classifications that forms the WHO-Family of International Classifications (WHO-FIC), along with the International Classification of Functioning, Disability and Health (ICF) and the health interventions in the International Classification of Health Interventions (ICHI). While the ICD classifies diseases and causes of death, the ICF describes functioning and disability in relation to a health condition and the ICHI describes and classifies health interventions[9].

Structure

The architecture of ICD-11 includes three integrated parts, namely a semantic network of biomedical concepts (Foundation), a traditional tabulation of hierarchical codes that derives from that network (Linearization) and a formal ontology that anchors the meaning of terms in the semantic network[10] (Fig. 3). The Foundation component is a multidimensional collection of all ICD entities, including diseases, disorders, injuries, external causes, signs and symptoms, but also functional descriptions and extension codes. Each entity of the Foundation has a unique Uniform Resource Identifier (URI) and is defined in a standard way. The Content Model (CM) provides the structured framework for this and includes required elements such as concept title, fully specified name, synonyms, parent and child relationships and a brief definition. The entities are organized into poly-hierarchies, meaning that a single term may have more than one conceptual parent. For example a disease such as oesophageal cancer, can be correctly classified to cancers (malignant neoplasm) or to conditions of the digestive system. In the same way, cerebral ischaemic conditions could be classified to the vascular system or to the nervous system. The entities may also have different types of relationships to other entities in in the Foundation Component and are thus not necessarily mutually-exclusive[10–12]. However, for statistical classification it is essential that content is mutually exclusive and exhaustive. This means that each concept must have one place in a hierarchy and thus only one parent assigned to them to avoid double counting. In order to account for this, a Linearization Component is derived from the Foundation. The Linearization Component contains the actual classifications or tabular lists that are generated from the Foundation Component. For example, the ICD Mortality and Morbidity Statistics (MMS) is one of the linearizations, but many other linearizations can be generated for particular purposes (e.g., Primary Care, Research, Dermatology, etc.). When linearizing from the Foundation, exclusiveness is achieved through mono-hierarchy where each concept has a single parent for inheritance. The exhaustive requirement for statistical classification is achieved through addition of residual categories, ‘other specified’ or ‘unspecified’. Apart from the URIs that are inherited from the Foundation Component, Linearizations have shorter hierarchical codes referred to as stem codes[10].

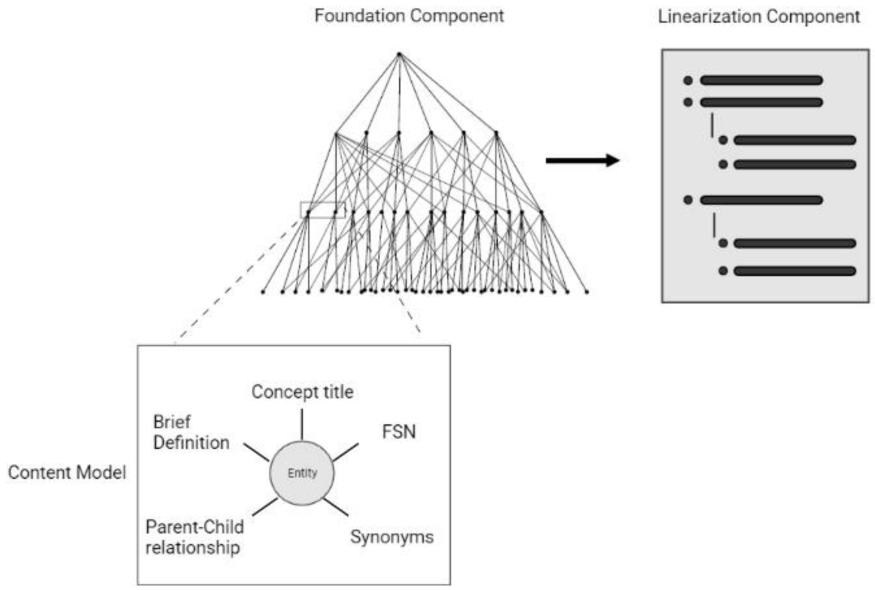


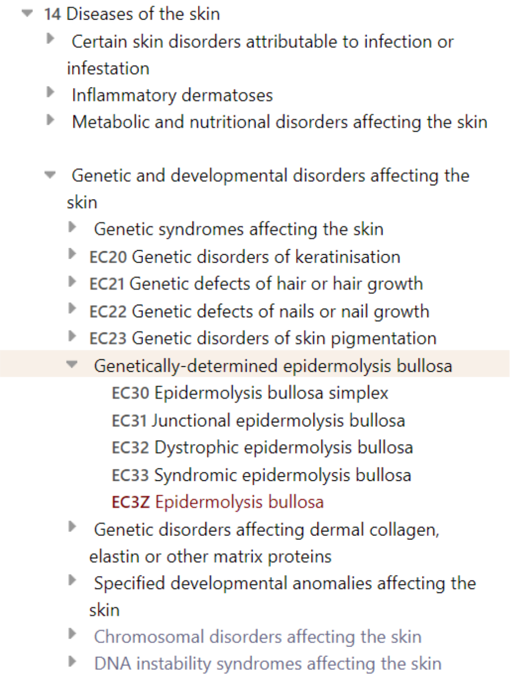
Figure 3: *A schematic depiction of the different elements of ICD-11. The Foundation component is the multidimensional collection of all ICD-11 entities. This demonstrates the great depth of ICD-11. Entities in the Foundation are defined using the structured framework provided by the content model. For statistical purposes, ICD-11 allows for the generation of a Linearized component with which mono-hierarchy is achieved. Created in BioRender.com*

ICD-11 further allows for post-coordination to enable adding more detail to an entity in a Linearization. Existing stem codes can be extended with other existing stem codes, which is referred to as clustering. For example 'proliferative diabetic retinopathy in Type 2 diabetes' is not precoordinated in ICD-11. To still code this using ICD-11, the user can combine the existing stem codes of 'Type 2 diabetes mellitus' and 'Proliferative diabetic retinopathy'. Thus additional detail can be added by combining base terms thereby increasing the expressive power of ICD. Other properties that can be used for post-coordination are called post-coordination axes and examples include disease severity, specific anatomy, histopathology, but also relationships such as 'has causing condition', 'has manifestation' or 'is associated with'. The post-coordination value sets (the allowed values for post-coordination axes), are usually hierarchies of entities from the Extension Codes branch in the Foundation or they are hierarchies from elsewhere in the Foundation[11,12]. An example is 'Pressure ulceration grade 4' which can be post-coordinated with manifestation site 'Sacral region' and other clinical details like 'Present on admission'.

We assessed the applicability of each ontology for mapping the indication and condition of 20 recently approved drugs. One of these drugs is Filsuvez and is indicated for treatment of partial thickness wounds associated with dystrophic and junctional epidermolysis bullosa (EB) in patients 6 months and older. Figure 4 gives an example of the information retrieved for EB in ICD-11 MMS. EB is assigned to one parent as is the rule for the linearization component and has five children which include the junctional and dystrophic form for which Filsuvez is indicated. We will use this example to demonstrate the structure and granularity of each ontology throughout this report.

Advantages and disadvantages

The 11th version of ICD is for the first time fully electronic and is expected to increase the accessibility of ICD, as well as its use with other classifications and terminologies, thereby increasing its global reach and standardization[11]. Another advantage of ICD-11 is the use a knowledge base, the Foundation. The ability of entities in the Foundation to be defined with attributes opens up the possibility for interoperability and comparability of data collected from different settings[13]. ICD-11 also allowed, for example, to easily incorporate the SARS-CoV-2 virus, the COVID-19 disease and various manifestations of the disease, which emerged after its release[10]. The option of post-coordination greatly increases the amount of detail that can be described by ICD-11. However, ICD-11 does not necessarily use DL for post-coordination and thus there is no computational way to identify equivalence of coding. This can lead to coding variability in which the same meaning can be expressed by different code combinations[10,13].



- ▼ 14 Diseases of the skin
  - ▶ Certain skin disorders attributable to infection or infestation
  - ▶ Inflammatory dermatoses
  - ▶ Metabolic and nutritional disorders affecting the skin
  - ▼ Genetic and developmental disorders affecting the skin
    - ▶ Genetic syndromes affecting the skin
    - ▶ EC20 Genetic disorders of keratinisation
    - ▶ EC21 Genetic defects of hair or hair growth
    - ▶ EC22 Genetic defects of nails or nail growth
    - ▶ EC23 Genetic disorders of skin pigmentation
    - ▼ Genetically-determined epidermolysis bullosa
      - EC30 Epidermolysis bullosa simplex
      - EC31 Junctional epidermolysis bullosa
      - EC32 Dystrophic epidermolysis bullosa
      - EC33 Syndromic epidermolysis bullosa
      - EC3Z Epidermolysis bullosa
    - ▶ Genetic disorders affecting dermal collagen, elastin or other matrix proteins
    - ▶ Specified developmental anomalies affecting the skin
    - ▶ Chromosomal disorders affecting the skin
    - ▶ DNA instability syndromes affecting the skin

Figure 4: *An example of the place of epidermolysis bullosa (EB) in the hierarchy of ICD-11-MMS. As this is the linearized component of ICD-11, EB is only assigned to one parent. It has five children which include the junctional and dystrophic form for which Filsuvez is indicated. Children located elsewhere in the Foundation can be seen in grey, while an unspecified residual category is depicted in red*

Systematized Nomenclature of Medicine Clinical Terms

History and purpose

The Systemized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is a clinical healthcare terminology that covers a wide range of clinical findings, symptoms, diagnoses, procedures, body structures and organisms, as well as aetiologies, substances, pharmaceuticals, devices and specimens. The origin of SNOMED-CT can be traced back to the Systematized Nomenclature of Pathology (SNOP) in 1965 which was created to describe the pathological observations in the categories aetiology, morphology, topography and function. Its success led to the expansion of SNOP to embrace all medical terms which gave rise to the Systematized Nomenclature of Medicine (SNOMED), followed two decades later, by a logic-based version SNOMED-RT. In parallel, the Clinical Terms Version 3 (CTV-3) was developed and the combination of SNOMED-RT and CTV-3 yielded SNOMED-CT[7]. SNOMED-CT is considered to be the most comprehensive medical terminology and is the intellectual property of the International Health Terminology Standards Development Organization[5]. The goal of SNOMED-CT is to make accurate recording and sharing of clinical and related health information, as well as semantic interoperability of health records, easier.

Structure

SNOMED-CT is organized as a multi-hierarchical ontology that enables concepts to be related to one another. The content of SNOMED-CT is represented using three types of components, the first being 'concepts' which represent clinical meanings that are organized into hierarchies. Concepts are arranged from the general to the more detailed within each hierarchy. Another component is 'descriptions' which link appropriate human readable terms to concepts. A concept may be associated with multiple descriptions, each of which serves as a synonym for the same clinical concept. Lastly, 'relationships' link concepts to other related concepts. These relationships provide formal definitions and other properties of the concept (Fig. 5). Every component has its unique numeric identifier[14].

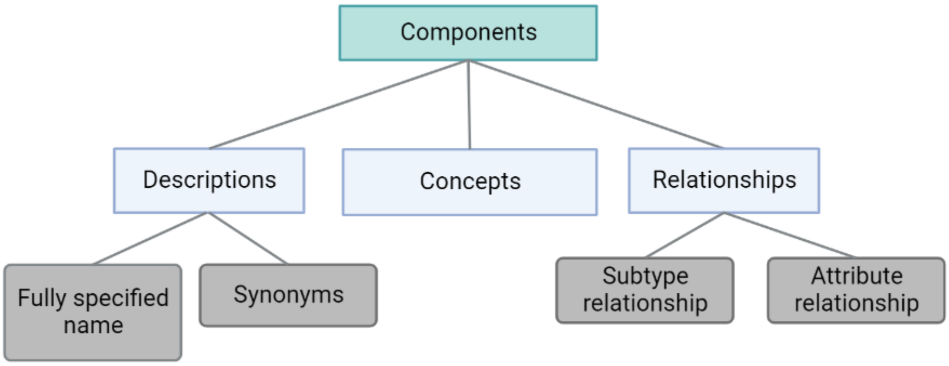
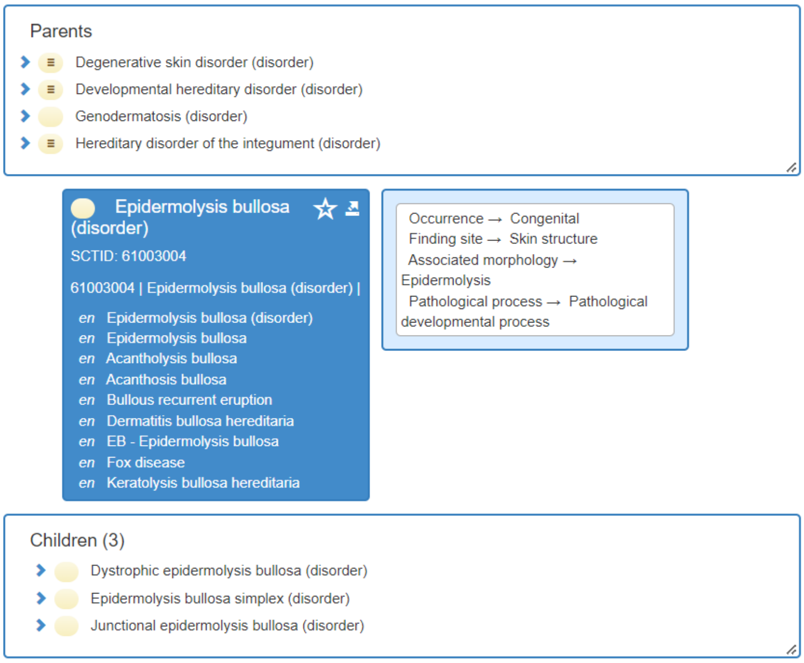


Figure 5: *Overview of the three components of SNOMED-CT: Concepts, Descriptions and Relationships and their characteristics. Created in BioRender.com*

SNOMED-CT uses DL for internal consistency and non-redundancy. The SNOMED-CT logical model defines the way in which each type of SNOMED-CT component is related and represented. As mentioned before, each component has a unique identifier, however, the concept identifier has a specific role as it is the code used to represent the meaning in clinical records, documents, messages and data. The human readable form of a concept is provided by the descriptions, of which there are two types. One description is the Fully Specified Name (FSN) and according to the logical model this is always unique, as each concept can have only one FSN, while the other type of description is synonyms, which are not unique. According to the logical model, a concept may have several synonyms. But out of all synonyms, one is marked as 'preferred' in a given language, dialect, or context of use, while the remaining synonyms are marked as 'acceptable'[14,15]. Associations between two concepts are represented by relationships. There are different types of relationships available within SNOMED-CT. One type is the subtype relationship which uses a 'is a' relationship. This relationship relates a concept to more general concepts and hereby define the hierarchy of SNOMED-CT concepts. According to the logical model, all active SNOMED-CT concepts have at least one 'is a' relationship, but can also be related to several other concepts. As a result, SNOMED-CT adopts a multi-hierarchical structure. Another type of relationship between SNOMED-CT concepts, is the attribute relationship. This relationship represent aspects of the meaning of a concept, such as 'causative agent', 'finding site' and 'associated morphology'. Unlike subtype relationships, attribute relationships are limited in their applicability to ensure consistent definitions that deliver reliable meanings.

The logic-based framework of SNOMED-CT further allows for post-coordination in which expressions contain two or more concept identifiers[5]. Post-coordination greatly increases the depth of detail that SNOMED-CT can represent by allowing additional clinical detail to be noted if required. This goes further then only combining different concept identifiers and includes addition of relationships between them[14].

Figure 6 depicts the data retrieved for EB in SNOMED-CT. The poly-parental structure of SNOMED-CT unlike ICD-11-MMS, allows EB to have multiple parents. It further includes the junctional and dystrophic form for which Filsuvez is indicated as children of EB. Additional detail on the disease is given using attributes, such as 'occurrence' and 'finding site'.



Parents

- Degenerative skin disorder (disorder)
- Developmental hereditary disorder (disorder)
- Genodermatosis (disorder)
- Hereditary disorder of the integument (disorder)

Epidermolysis bullosa (disorder)

SCTID: 61003004

61003004 | Epidermolysis bullosa (disorder) |

en Epidermolysis bullosa (disorder)
en Epidermolysis bullosa
en Acantholysis bullosa
en Acanthosis bullosa
en Bullous recurrent eruption
en Dermatitis bullosa hereditaria
en EB - Epidermolysis bullosa
en Fox disease
en Keratolysis bullosa hereditaria

Occurrence → Congenital
Finding site → Skin structure
Associated morphology → Epidermolysis
Pathological process → Pathological developmental process

Children (3)

- Dystrophic epidermolysis bullosa (disorder)
- Epidermolysis bullosa simplex (disorder)
- Junctional epidermolysis bullosa (disorder)

Figure 6: *Example of the representation of EB in SNOMED-CT. EB is assigned to multiple parents and includes the indicated junctional and dystrophic form of Filsuvez. Additional detail is given using attributes, such as 'occurrence' and 'finding site'.*

Advantages and disadvantages

SNOMED-CT successfully aids data storage and retrieval and is the most used terminology for data-interchange in electronic health records[16]. SNOMED-CT's concept based approach, rather than the use of terms for retrieval, overcomes the mismatch between terms found in documents and queries of different stakeholders[17]. The application of SNOMED-CT is much broader than only medical records. A recent study showed, for example, that SNOMED-CT can be used for categorizing clinical studies based on their indication and that this aids information retrieval from clinical study registries such as ClinicalTrials.gov[18]. SNOMED-CT further facilitates semantic interoperability through, for example, providing mappings to other classification systems such as ICD. It also collaborates with Health Level 7 (HL7), an international organization that develops standards for exchange, integration, sharing and retrieval of electronic health information[8]. The use of SNOMED-CT also raises challenges. It was, for example, reported that there is ambiguity of terms. Multiple very similar terms are used for different concepts, making it difficult to find the right concept. Furthermore, post-coordination in SNOMED-CT can be applied even when there is an existing pre-coordination available to represent the required means[15].

Medical Subject Headings

History and purpose

Medical Subject Headings (MeSH) is a controlled vocabulary used for indexing, cataloguing and searching for biomedical and health-related information and documents. MeSH provides a consistent way to find content with different terminology but the same concepts. It was produced by the National Library of Medicine (NLM) and came forth out of the need for a single subject list for books and periodical articles[19]. Its first official version was launched in 1954 and has been updated continually since. MeSH is used to categorise publications for libraries and other institutions around the world, but also to retrieve publications using the MeSH terms. The MeSH vocabulary is primarily designed for use by NLM to index citations for journal articles in the MEDLINE database and to search the MEDLINE data using PubMed.

Structure

MeSH consists of three major components or record types: Descriptors, Qualifiers and the Supplementary Concept Records (SCRs) (Fig. 7). Descriptors play a central role in MeSH vocabulary as they characterize the subject matter or content. They are organized in 16 categories and each category is further divided into subcategories, forming a hierarchical structure with up to 13 levels[8]. Each MeSH descriptor appears in at least one place in the hierarchy, but may also appear in additional places. Descriptors are divided in four classes. The first class are the main headings which indicate the subject of an indexed item, such as a journal article. So, they provide an indication of what the item is about. Main headings are used to index citations in the MEDLINE database and for cataloguing of publications. The second class of descriptors are publication characteristics, also known as publication types. In contrast to main headings, publication characteristics indicate what the indexed item is e.g. an historical article. They may include Publication Components, such as Charts, Publication Formats, such as Editorial and Study Characteristics, such as Clinical Trial. They function as metadata, rather than being about the content. The third class of descriptors are check tags which are used for tagging citations that contain certain categories of information. They do not appear in the MeSH tree. The fourth class of descriptors characterize the physical location such as continents, regions, countries, states and other geographic subdivisions. Together with descriptors, qualifiers, also known as subheadings, are used in MeSH. Qualifiers group together citations which cover a particular aspect of a subject. For example, Liver/drug effects indicates that it is not about the liver in general, but about the effect of drugs on the liver. SCRs are used to index chemicals, drugs and other concepts such as rare diseases for MEDLINE. Each SCR is linked to one or more descriptors by the Heading Mapped To (HM) field in the SCR and SCRs are thus not organised in a tree hierarchy. Class 1 of the SCR is dedicated to chemicals and primarily mapped to category D of descriptors. The second class is dedicated to Chemotherapy Protocols and are mapped to the MeSH heading "Antineoplastic Combined Chemotherapy Protocols" and to chemicals used in the protocols found in category D. The third class dedicated to diseases is mapped to category C and anatomical headings found in category A. The last class is dedicated to organisms (e.g., viruses) and is primarily mapped to the category B organism descriptors[20].

MeSH also has entry terms, which are synonyms, near-synonyms alternate forms and other closely related terms in a MeSH record. Entry terms are generally used interchangeably with the preferred term and are equivalent to the preferred term for purposes of cataloguing, indexing and retrieval. Another category of MeSH are concepts. Concepts group together terms in a MeSH record which are strictly synonymous. Each MeSH record consists of one or more concepts and each concept consists of one or more synonymous terms. Each concept has a preferred term and each record has a preferred concept[20].

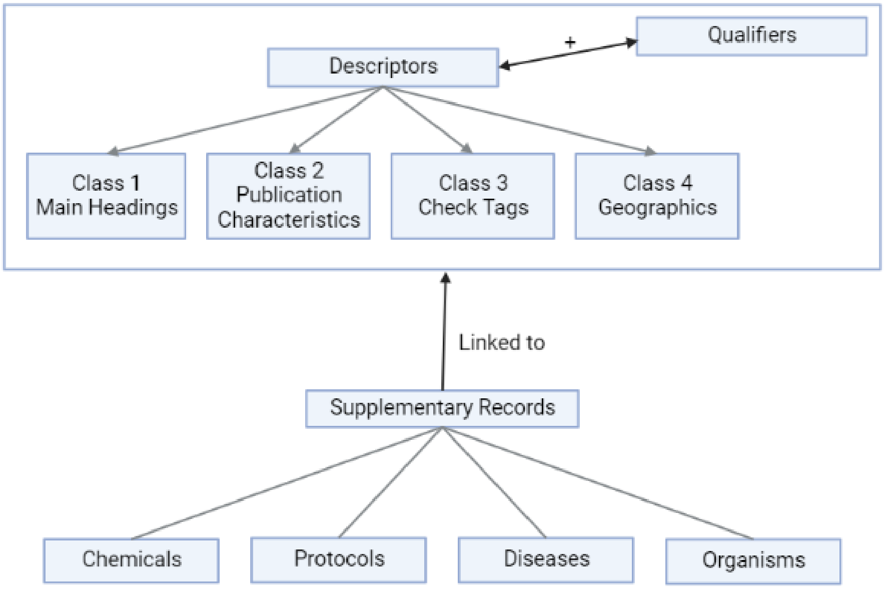
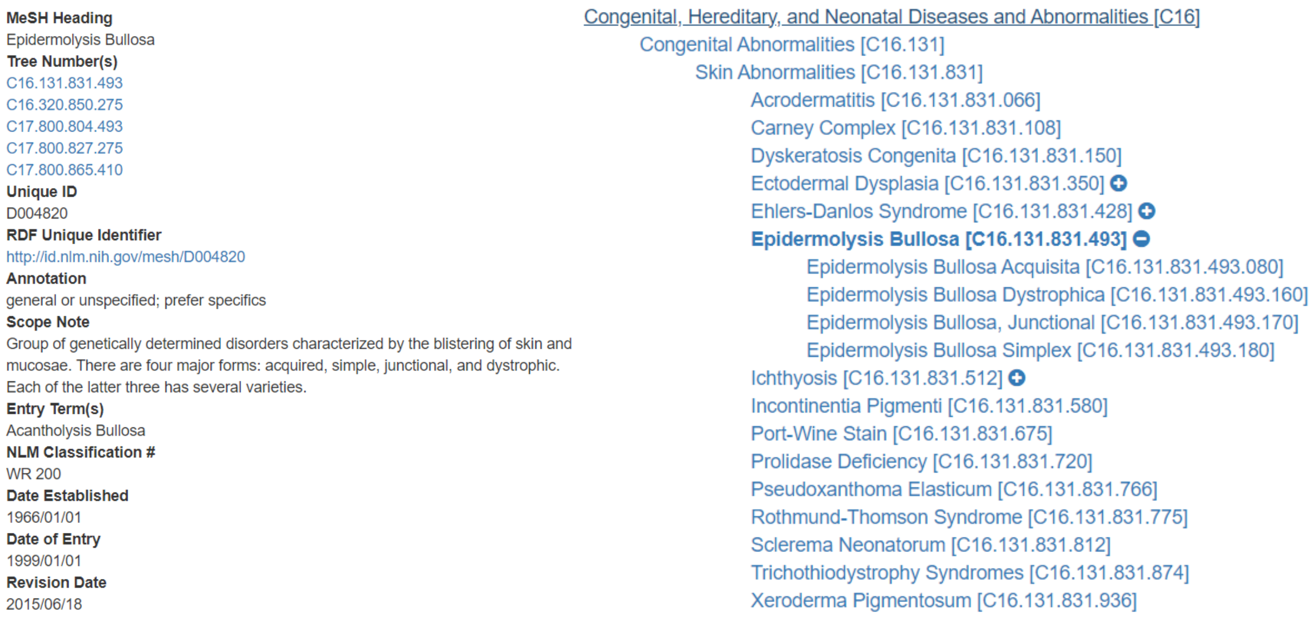


Figure 7: *Graphical overview of the three record types in MeSH: Descriptors, qualifiers and supplementary records. Created in BioRender.com*

An example of the retrieved data from a search on EB is given in Fig. 8. The left panel shows the details retrieved upon initial search. As can be seen from the multiple tree numbers, EB is part of different hierarchies, out of which one is depicted in the right panel. The entry level term would also lead the user to EB. In the right panel the children of EB are depicted with inclusion of the junctional and dystrophic form for which Filsuvez is indicated.



MeSH Heading
Epidermolysis Bullosa
Tree Number(s)
C16.131.831.493
C16.320.850.275
C17.800.804.493
C17.800.827.275
C17.800.865.410
Unique ID
D004820
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D004820
Annotation
general or unspecified; prefer specifics
Scope Note
Group of genetically determined disorders characterized by the blistering of skin and mucosae. There are four major forms: acquired, simple, junctional, and dystrophic. Each of the latter three has several varieties.
Entry Term(s)
Acantholysis Bullosa
NLM Classification #
WR 200
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2015/06/18

Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16]
- Congenital Abnormalities [C16.131]
  - Skin Abnormalities [C16.131.831]
    - Acrodermatitis [C16.131.831.066]
    - Carney Complex [C16.131.831.108]
    - Dyskeratosis Congenita [C16.131.831.150]
    - Ectodermal Dysplasia [C16.131.831.350] ⊕
    - Ehlers-Danlos Syndrome [C16.131.831.428] ⊕
    - **Epidermolysis Bullosa [C16.131.831.493]** ⊖
      - Epidermolysis Bullosa Acquisita [C16.131.831.493.080]
      - Epidermolysis Bullosa Dystrophica [C16.131.831.493.160]
      - Epidermolysis Bullosa, Junctional [C16.131.831.493.170]
      - Epidermolysis Bullosa Simplex [C16.131.831.493.180]
    - Ichthyosis [C16.131.831.512] ⊕
    - Incontinentia Pigmenti [C16.131.831.580]
    - Port-Wine Stain [C16.131.831.675]
    - Prolidase Deficiency [C16.131.831.720]
    - Pseudoxanthoma Elasticum [C16.131.831.766]
    - Rothmund-Thomson Syndrome [C16.131.831.775]
    - Sclerema Neonatorum [C16.131.831.812]
    - Trichothiodystrophy Syndromes [C16.131.831.874]
    - Xeroderma Pigmentosum [C16.131.831.936]

Figure 8: *Example of EB as descriptor in MeSH.* ***Left:*** *the representation of data upon searching for EB in MeSH. Different tree numbers indicate that EB is part of multiple hierarchies.* ***Right:*** *One of the hierarchies in which EB occurs with inclusion of the junctional and dystrophic form for which Filsuvez is indicated.*

Advantages and disadvantages

The main purpose of MeSH is to act as a tool to easily search for relevant publications in PubMed. Indeed the use of MeSH terms allows for a more focused search of literature[21]. Besides subject searching of data, MeSH is also used on the website of the EMA to enable browsing by therapeutic area for human medicines. Specific branches of the MeSH taxonomy tree are used for this purpose and some terms have been modified to facilitate easier searching by non-specialists[22]. This and the multi-parental hierarchy which increases the inclusivity of MeSH, makes it interesting for us to use as a tool to extract the therapeutic indication out of the SmPCs of drugs. However, for our purpose a possible pitfall is the fact that diseases in MeSH are present in the hierarchy of the descriptors, but also exist as supplementary concepts which are not part of the hierarchy but only linked to it. Previously, a difference was found between using supplementary concepts and descriptors in a MeSH query, with the latter showing a higher retrieval precision. Thus the methods used for drafting of queries by the user has a direct impact on the specificity and effectiveness of retrieved results[23].

Medical Dictionary for Regulatory Activities

History and intended use

The Medical Dictionary for Regulatory Activities (MedDRA) is an internationally accepted medical terminology for biopharmaceutical regulatory purposes. MedDRA was developed to address the need for a single international standard for reporting adverse events[24]. As late as the early 1990s most organizations that processed regulatory data used an international adverse drug reaction terminology in combination with morbidity terminology. For Europe this meant the use of the World Health Organization's Adverse Reaction Terminology (WHO-ART©) in combination with ICD-9. This differed to what, for example, the United States and Japan were using for these purposes. The use of multiple terminologies made it difficult to integrate data on adverse effects across the pre- and post-marketing stages[25]. In the mid-1990s the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) expanded a pre-existing terminology used by the British regulatory agency and refined its capabilities in representing adverse events, to produce MedDRA[25]. MedDRA terminology can be used in all phases of development of medical products for human use, excluding animal toxicology. MedDRA's scope includes medical, health-related and regulatory concepts related to such products.

Structure

The MedDRA hierarchy consists of five levels which provide increasing specificity as it descends (Fig. 9). At the highest level of the hierarchy is located the System Organ Class (SOC). This is the broadest level and consists of groupings based on aetiology, manifestation site and purpose. The 27 SOCs represent parallel axes that are not mutually exclusive. This characteristic allows a term to be represented in more than one SOC and to be grouped by different classifications (e.g., by aetiology or manifestation site). A SOC is directly related to at least one High Level Group Term (HLGT) with no restriction on the number of links to HLGTs. The HLGTs act as descriptor of one or more High Level Terms (HLTs) related by anatomy, pathology, physiology, aetiology or function. The single medical concept level of MedDRA is the Preferred Term (PT). It represents distinct descriptors for a symptom, sign, disease, diagnosis, therapeutic indication, investigation, surgical or medical procedure and medical, social, or family history characteristic[24,25]. The PT level groups synonymous terms or equivalent terms to provide a horizontal equivalence relationship. The PT level is preferred for data analysis and retrieval[26]. The lowest level of the hierarchy includes the Lowest Level Terms (LLTs), which reflect how an observation might be reported in practice. LLTs are related to their parent PT as synonyms (different terms for the same concept inherent in the PT), lexical variants (different word forms for the same expression), Quasi-synonyms (terms that are not precisely the same meaning as another term, but are treated as synonymous in a given terminology) or sub-concept (provide more detailed information such as anatomic specificity). Each LLT is linked to only one PT. Each term in MedDRA has a unique non-expressive code[24].

To aid a standardized approach for data retrieval from MedDRA, Standardized MedDRA Queries (SMQs) were developed. SMQs are groupings of MedDRA terms, at the PT level that relate to a defined medical condition or area of interest. They act as a starting point for analysis by helping to identify cases of interest for further medical analysis. SMQs begin with a definition of the condition or area of interest, from which a candidate list of MedDRA terms is built including terms related to signs, symptoms, diagnoses, syndromes, physical findings, laboratory and other physiologic test data, etc.

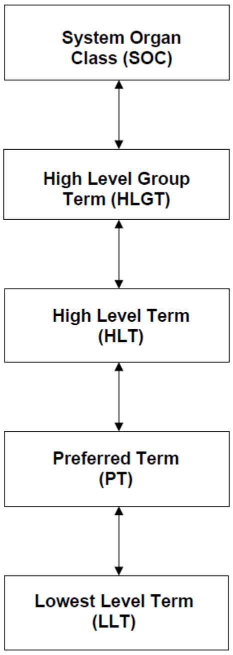
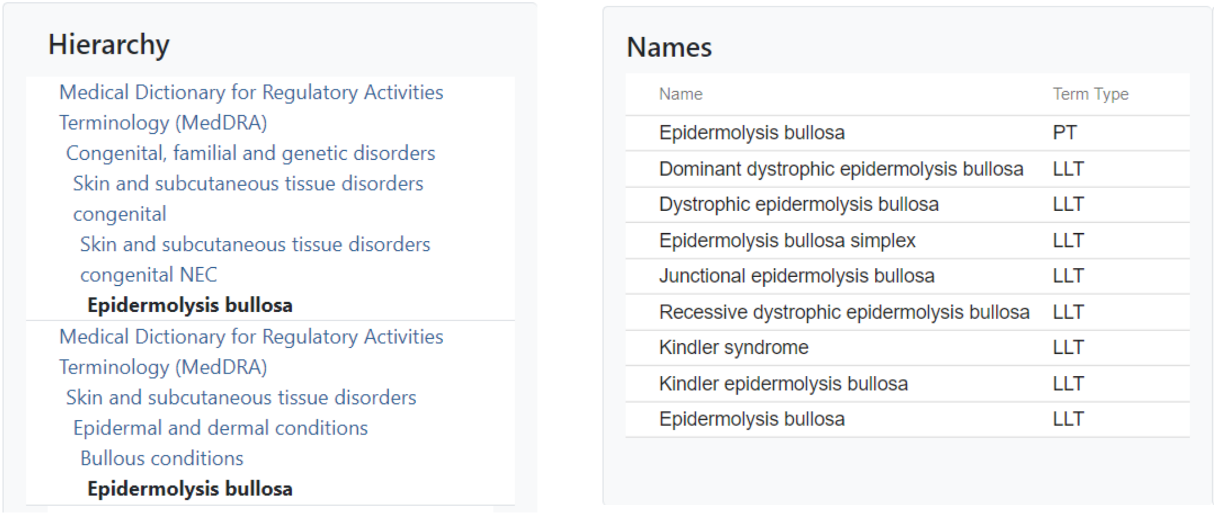


Figure 9: *Overview of the structural hierarchy of MedDRA. The PT level is preferred for data analysis and retrieval. Adapted from MedDRA. Introductory Guide MedDRA Version 25.1. (2022).*

Figure 10 shows the information provided by MedDRA on EB. It has to be noted that MedDRA was accessed via UMLS, as its individual ontology is not openly available without a license. UMLS will be described in more detail below. Nevertheless, it can be seen that EB is part of different hierarchies in MedDRA (Fig. 10, left panel). However, no children are given and junctional and dystrophic EB are merely present as LLTs (Fig. 10, right panel).



Hierarchy

- Medical Dictionary for Regulatory Activities Terminology (MedDRA)
  - Congenital, familial and genetic disorders
    - Skin and subcutaneous tissue disorders congenital
      - Skin and subcutaneous tissue disorders congenital NEC
        - **Epidermolysis bullosa**
- Medical Dictionary for Regulatory Activities Terminology (MedDRA)
  - Skin and subcutaneous tissue disorders
    - Epidermal and dermal conditions
      - Bullous conditions
        - **Epidermolysis bullosa**

Names

Name	Term Type
Epidermolysis bullosa	PT
Dominant dystrophic epidermolysis bullosa	LLT
Dystrophic epidermolysis bullosa	LLT
Epidermolysis bullosa simplex	LLT
Junctional epidermolysis bullosa	LLT
Recessive dystrophic epidermolysis bullosa	LLT
Kindler syndrome	LLT
Kindler epidermolysis bullosa	LLT
Epidermolysis bullosa	LLT

Figure 10: Left: *The placement of EB in the hierarchy of MedDRA.* ***Right:*** *Different LLTs of EB. Junctional and dystrophic EB are not part of MedDRA's hierarchy, but only exist as LLTs.*

Advantages and disadvantages

MedDRA has steadily moved into regulatory frameworks, including that of the EMA. Since 2003 all reporting of adverse drug reactions has to be documented using the MedDRA. This includes Suspected Unexpected Serious Adverse Reactions in clinical trials and post-authorization Individual Case Safety Reports, as well as in the SmPCs. In the latter, MedDRA terms are used in several sections (e.g. “Undesirable Effects”) and it can also be used as coding to estimate the frequency of events[25]. However, a disadvantage of MedDRA is that data retrieval has shown to be difficult. This is mainly due to the multi-axial nature of MedDRA in which different PTs may be present in different groupings within the same SOC or they may be located in more than one SOC. This poses challenges for identifying clinically related terms[26]. Although MedDRA includes several diseases in its hierarchy, it is more oriented towards clinical safety and pharmacovigilance. This makes it likely less applicable for our purpose as it is not optimized for addressing other topics of regulatory interest, such as the mapping of diseases[25].

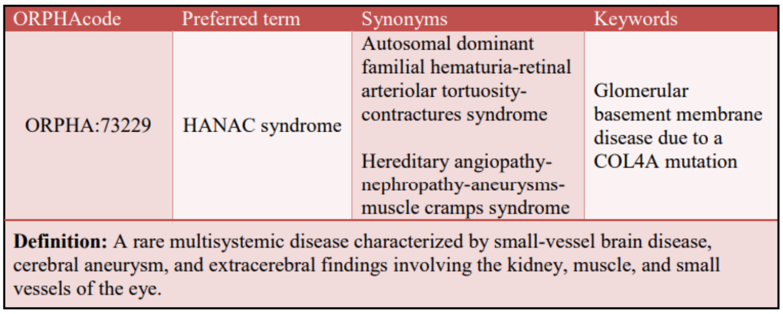
Orphanet

History and purpose

Historically, there has been a lack of medical and scientific knowledge about rare diseases. There are around 6000 rare disorders and most are genetic in origin[27]. Although many classification systems such as ICD and SNOMED-CT now see the need for incorporating rare diseases in their terminology, up till 1997 there was no reference portal for information on rare diseases. Orphanet addressed this issue and dedicated a database to rare diseases populated from literature and validated by international experts[28]. They developed and maintain the Orphanet nomenclature of rare diseases. This standardised system aims at providing a specific terminology for rare diseases. Orphanet also maintains the Orphanet classification of rare diseases. This is a multi-hierarchical and poly-parental structure built on the Orphanet nomenclature and is organised by medical specialty according to diagnostic and therapeutic relevance.

Structure

Within the Orphanet nomenclature, each clinical entity is assigned a set of elements. First, a unique and time-stable ORPHAcode is assigned to each clinical entity. This is accompanied by a preferred term, which is the most generally accepted name according to the literature and as adopted by the medical community. Other elements include synonyms, keywords (other significant clinical terms for a disorder or a group of disorders) and a definition (Fig. 11).

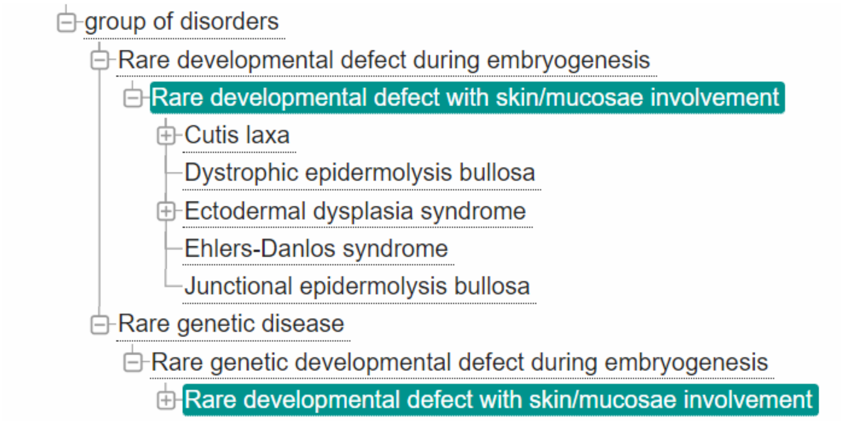


ORPHAcode	Preferred term	Synonyms	Keywords
ORPHA:73229	HANAC syndrome	Autosomal dominant familial hematuria-retinal arteriolar tortuosity-contractures syndrome Hereditary angiopathy-nephropathy-aneurysms-muscle cramps syndrome	Glomerular basement membrane disease due to a COL4A mutation
Definition: A rare multisystemic disease characterized by small-vessel brain disease, cerebral aneurysm, and extracerebral findings involving the kidney, muscle, and small vessels of the eye.			

Figure 11: *Representation of entities in the Orphanet nomenclature of rare diseases.*

The Orphanet classification is based on this nomenclature and is organised to three hierarchical levels: Group of disorders, Disorder and Subtype of a disorder. The group of disorders level encompasses clinical entities which share a set of common features. They are assigned to a category based on one general feature and to clinical groups based on other features such as similar aetiology, course and outcome. Clinical entities can be included in several classification groups, but they are assigned to one group as a preferential parent. The disorder levels provide more detail and characterise clinical entities by a set of homogeneous phenotypic abnormalities and evolution, allowing a definitive clinical diagnosis. Within this level, clinical entities can be labelled as a disease, clinical syndrome, malformation syndrome, morphological anomaly, biological anomaly or particular clinical situation in a disease or syndrome. The last level of the hierarchy, subtype of a disorder, provides a subdivision of a disorder according to clinical, etiological or histopathological characteristics.

Figure 12 gives an example of the classification in Orphanet. Here, EB is not present as a parent of dystrophic and junctional EB. These forms of EB for which Filsuvez is indicated, are present as a subtype of the disorder 'rare developmental defect with skin/mucosae involvement'.



- group of disorders
  - Rare developmental defect during embryogenesis
    - Rare developmental defect with skin/mucosae involvement
      - Cutis laxa
      - Dystrophic epidermolysis bullosa
      - Ectodermal dysplasia syndrome
      - Ehlers-Danlos syndrome
      - Junctional epidermolysis bullosa
  - Rare genetic disease
    - Rare genetic developmental defect during embryogenesis
      - Rare developmental defect with skin/mucosae involvement

Figure 12: *Representation of classification levels in Orphanet junctional and dystrophic EB. EB itself is not present as a parent of the two forms.*

Orphanet Rare Disease Ontology (ORDO) is a representation of the data in the Orphanet information system, formalised as an OWL ontology. It forms a resource for the computational analysis of rare diseases and integration of the Orphanet nomenclature into health and research information systems[29]. Concepts from the Orphanet database form a distinct class in ORDO and are associated with other classes using a set of defined object properties. In ORDO there are 10 super classes: clinical entity, which is central, group of disorders, disorder, subtype of disorder, age of onset, epidemiology, genetic material, geography, inheritance and inactive clinical entity. Group of disorders, disorder and subtype of disorder are subclasses of the clinical entity super class and have poly-parental relations. The remaining super classes have annotations. For example the super class age of onset can have the annotation adolescent, adult, all ages etc. The inactive clinical entity super class encompasses clinical entities that have been excluded from Orphanet, because they are for example deprecated or are no longer considered rare. Examples of relationships between entries in ORDO include 'part-of', 'has_age_of_onset' and 'disease-causing germline mutation(s) in'.

Advantages and disadvantages

Since its establishment, Orphanet has grown considerably and is implemented in over 40 countries. The European Commission Expert Group on Rare Diseases has accepted Orphanet as the most appropriate nomenclature for clinical coding of rare diseases in Europe. Orphanet is also recognized by the International Rare Disease Research Consortium as a recognised resource. Orpha nomenclature is further used in European legislative texts concerning rare diseases[30]. Orphanet also facilitates interoperability by including cross-references to other international terminologies such as ICD, SNOMED-CT and MeSH[31]. However, Orphanet is focussed on providing information on rare diseases for medical professionals, researchers and decision makers. Its scope is thus limited to monitoring and reporting rare diseases and does not extend to general disease terminology. This means that non-rare diseases are not part of the Orphanet terminology, which could be a disadvantage for our purpose. When the to be extracted condition or indication of a drug is not rare it will probably not be covered by Orphanet.

Disease ontology

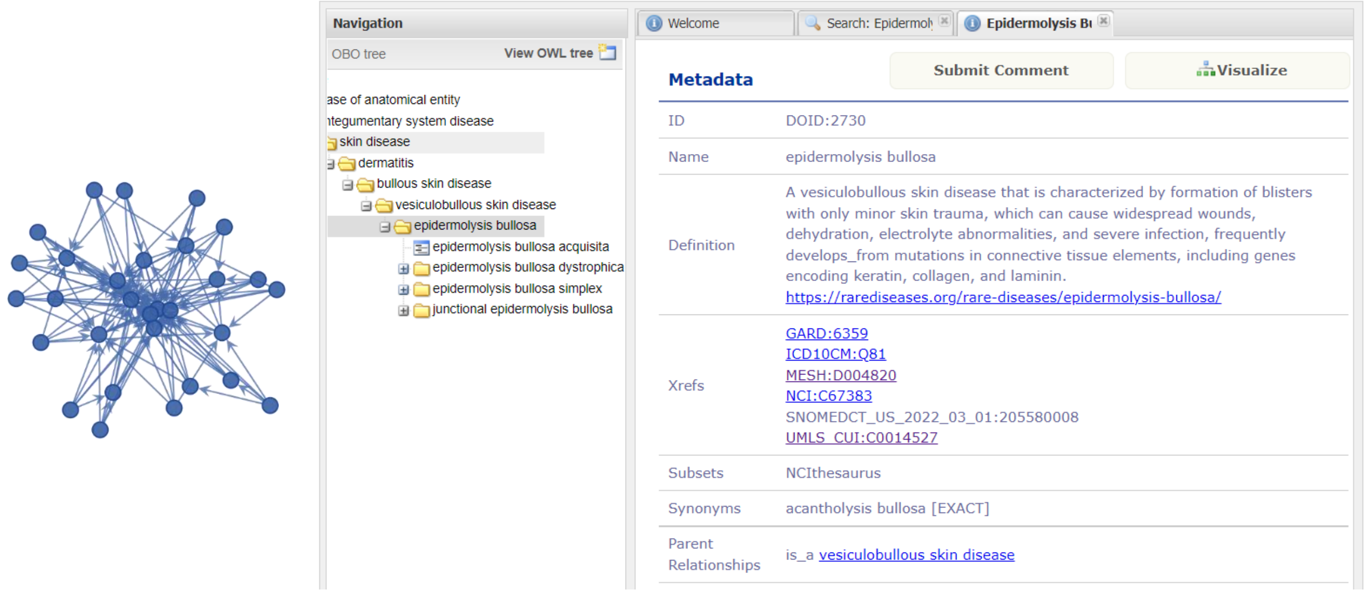
History and purpose

The Human Disease Ontology (DO) is an ontology that is focussed on representing common and rare disease concepts. It was established in 2003 and is a project of the University of Maryland School of Medicine, Institute for Genome Sciences. DO integrates biomedical data that is associated with human diseases from multiple taxonomies and terminologies. DO therefore only includes concepts of diseases and does not include the progression and manifestation of the disease as part of its definition[32,33]. The initial version of DO was a combination of disease terms extracted from ICD-MMS, SNOMED-CT and MeSH. Expansion led to the inclusion of disease terms from multiple other biomedical resources, including Orphanet[34].

Structure

DO is organized in eight nodes representing cellular proliferation, mental health, anatomical entity, infectious agent, metabolism and genetic diseases and physical disorders and syndromes. Individual disease terms in DO have traceable, stable identifiers (DOIDs). DO's terms are represented in a directed acyclic graph (DAG), allowing each term to have multiple parents (Fig. 13, left panel). Within the DAG, terms are linked by relationships in a hierarchy with interrelated subtypes. The logical definitions in DO are referred to as axioms and are supported with description logic. DO contains 'equivalent to' and 'subclass of' axioms to describe relevant disease drivers in the different nodes. Within the axioms, the relations are defined by a specific relation ontology. Examples include 'derives from', 'disease has basis in' and 'transmitted by'. This makes it possible to investigate indirect links between diseases and provides better understanding of complex diseases. In line with its purpose, DO provides mappings to synonymous disease concepts in other ontologies which are included in DO. These are referred to as cross-references (xrefs)[35].

An example of the retrieved data from a search on EB is given in Fig. 13, right panel. On the left the place of EB with its children dystrophic and junctional EB are depicted. On the right a definition, the xrefs and synonyms are given.



ase of anatomical entity
ntegumentary system disease
skin disease
dermatitis
bullous skin disease
vesiculobullous skin disease
epidermolysis bullosa
epidermolysis bullosa acquisita
epidermolysis bullosa dystrophica
epidermolysis bullosa simplex
junctional epidermolysis bullosa

Metadata

ID	DOID:2730
Name	epidermolysis bullosa
Definition	A vesiculobullous skin disease that is characterized by formation of blisters with only minor skin trauma, which can cause widespread wounds, dehydration, electrolyte abnormalities, and severe infection, frequently develops_from mutations in connective tissue elements, including genes encoding keratin, collagen, and laminin. https://rarediseases.org/rare-diseases/epidermolysis-bullosa/
Xrefs	GARD:6359 ICD10CM:Q81 MESH:D004820 NCI:C67383 SNOMEDCT_US_2022_03_01:205580008 UMLS_CUI:C0014527
Subsets	NCIthesaurus
Synonyms	acantholysis bullosa [EXACT]
Parent Relationships	is_a vesiculobullous skin disease

Figure 13: Left: *An example of the Directed Acyclic Graph structure of DO. DO is organized in eight nodes.* ***Right:*** *Example of the representation of pneumonia in DO. On the left the hierarchical relationship of pneumonia is depicted. The mappings to other ontologies are found under Xrefs.*

Advantages and disadvantages

The strength of DO lies in its focus on the aetiology of diseases. It particularly addresses the need for incorporating genetic and environmental causes of diseases in its ontology. For example, the integration of the online mendelian inheritance in man ontology (OMIM), which is a standard reference for human genes and genetic diseases, is maintained with high priority[36]. In addition, environmental factors that underlie disease aetiology, captured in ontologies such as the exposure ontology (ExO), are also incorporated in DO[35]. This could be an advantage for capturing certain details of the condition or indication for which a drug is advised and could thus be interesting for our purpose. However, a caveat of DO is that it only considers pre-coordination, meaning that existing concepts cannot be combined to yield new concepts. Although DO aims to connect diseases on different levels, the absence of the ability for post-coordination may limit the necessary detail needed for certain use cases and would mean that in our case we are dependent on the updates of DO to provide us with the latest details of diseases[37].

Ontology repositories

The majority of ontologies described above can be found in ontology repositories, which act as a facility where ontologies can be stored, retrieved and managed. Examples of such ontology repositories are BioPortal and UMLS[38,39]. An advantage of ontology repositories is, for example, that it provides users with a single platform to access a broad range of, in this case, biomedical ontologies. This saves the user time and effort for evaluating and comparing individual ontologies. UMLS was the first biomedical ontology repository which was organised by concept and its broad coverage makes it useful for linking between ontologies[40]. Here we describe UMLS in detail as it is the largest and most heavily used repository of biomedical ontologies[41].

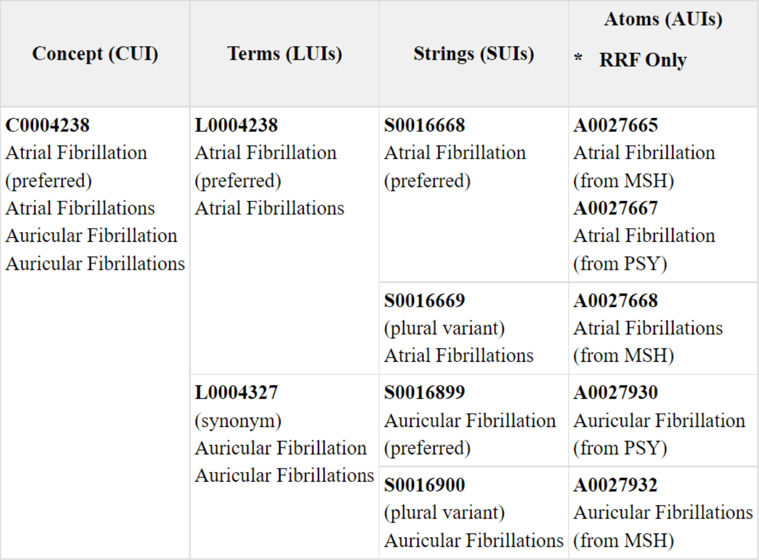
Unified Medical Language Systems

History and purpose

Ontologies relating to biomedical data have proliferated significantly. Six ontologies are described in detail above and it already becomes clear that different systems use different terms and codes to serve different purposes. The heterogeneity of diseases in different databases leads to hinderance in the utilization of biomedical data. UMLS provides a solution for this by integrating many health and biomedical controlled vocabularies to enable interoperability between computer systems[31]. UMLS is an initiative of the NLM and has been maintained by them since 1990. The main purpose of UMLS is to support mapping between various terminologies and is thus not designed as a ontology system[7].

Structure

UMLS consists of three knowledge sources. The biggest component of UMLS is the Metathesaurus. This large biomedical thesaurus joins names, meaning and useful relationships of biomedical concepts from numerous source vocabulary systems. It functions as a mapping tool between synonymous names of the same concept from different vocabulary systems[7,31]. The Metathesaurus preserves the meanings, concept names and relationships from its source vocabularies. This means, for example, that if two different source vocabularies use the same name for different concepts, the Metathesaurus retrieves both of the meanings and states which meaning is present in which source vocabulary. Thus, the Methathesaurus preserves the different views on the same concept as it might be useful for different tasks. The Metathesaurus is organized by concepts, which have a meaning. These meanings can have different terms assigned to them in different vocabularies. Thus the purpose is to connect different terms for the same concept from many different vocabularies. Each concept in the Metathesaurus has a unique and permanent concept identifier (CUI). In addition, a lexical unique identifier (LUI) is assigned to each lexical variant or different word for the same concept. Each concept is linked to at least one lexical variant, but can also be linked to many of each of these[39]. The Metathesaurus also takes into account different spellings of terms or lexical variant and also if they are noted in upper or lower case. These are called string sets and have string unique identifiers (SUIs). On top of that every term from every source vocabulary is given an atom unique identifier (AUI). These form the basic building blocks of the thesaurus[42] (Fig. 12). In addition to the synonymous relationships between concepts, the Metathesaurus also describes additional relationships between concepts. The relationships can be between concepts from the same source vocabulary and between concepts in different vocabularies.



Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 (plural variant) Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

Figure 12: *An example of atrial fibrillation with its connected Concept, Term, String and Atom Identifiers in UMLS. The AUIs represent how the terms appear in different ontologies. In this case only AUIs from MeSH and Psychological Index Terms are depicted, but in reality the user will be provided with a list of and access to all ontologies in which atrial fibrillation is represented. MSH=MeSH, PSY= Psychological Index Terms.*

The second component of the UMLS is the Semantic Network which acts as the upper ontology. It consists of broad categories, or semantic types, that provide a consistent categorisation of all concepts represented in the UMLS Metathesaurus, and a set of useful and important relationships, or semantic relations, that exist between semantic types. Each Metathesaurus concept is assigned at least one semantic type. The semantic network adapts a hierarchical structure, in which children are linked to their parents by a ‘is a’ link. The third component is the SPECIALIST Lexicon and Lexical Tools. This component is designed to provide lexical normalization of text for natural language processing (NLP). Coverage includes both commonly occurring English words and biomedical vocabulary[39].

Disease ontologies in action

Each of the described ontologies have their own specific purpose and therefore also provide different levels of detail. For the Regulatory Science Database it is of interest to assess which of the above described ontologies can be used to map the indication or condition of a given medicine. Keywords from the condition and indication were searched in the ontologies to determine how much detail each ontology provides. The results can be found in supplementary material 2 *(provided as Excel file).*

For 30% of the drugs we analysed, there was a negligible or no difference between their condition and indication. This is for example the case for Pyrukynd which has as condition, treatment of pyruvate kinase deficiency, while its indication only adds information on the population. In this case, adult patients. For Nulibry, the condition and indication are both determined as treatment of molybdenum cofactor deficiency type A.

For each condition and indication of a medicine we labelled whether or not it was captured by a certain ontology using 'yes' or 'no', respectively (Fig. 13). In total there were 22 orphan conditions and 20 indications. All ontologies were able to capture most of the conditions of the medicines. SNOMED-CT, ICD-11, MeSH and Orphanet captured the most conditions, with the former capturing 91% of the conditions we analysed. MedDRA, on the other hand, covered the least amount of conditions, alongside DO. Regarding the indications, SNOMED-CT included more than half of the indications analysed, while MedDRA, Orphanet and ICD-11 were the least inclusive.

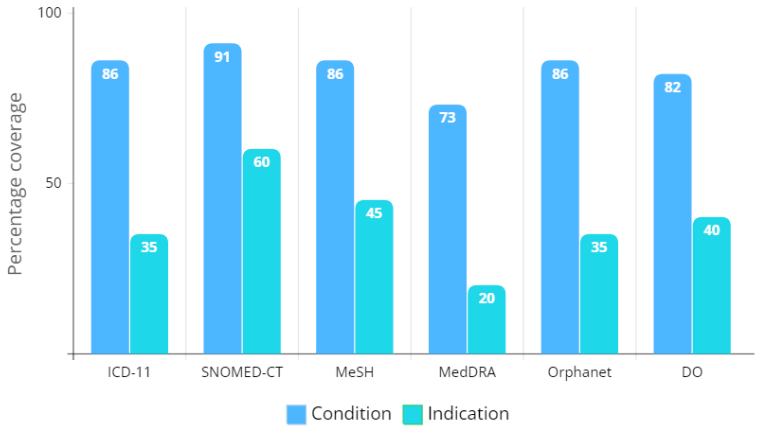


Figure 13: *Graph depicting the percentage of the analysed conditions and indications covered by ICD-11, SNOMED-CT, MeSH, MedDRA, Orphanet and DO*

ICD-11 was able to capture the main concepts of the conditions. For example, cytomegalovirus (CMV) infection, myasthenia gravis and immunoglobulin A (IgA) nephropathy. It also accurately captured rare conditions such as pyruvate kinase deficiency. However, it did not provide adequate details to capture certain indications. This was the case for hereditary transthyretin-mediated (hATTR) amyloidosis for which it did not indicate the type of polyneuropathy associated with it. For aromatic L-amino acid decarboxylase (AADC) deficiency and microscopic polyangiitis (MPA) as well as granulomatosis with polyangiitis (GPA), ICD-11 did not indicate the disease severity. ICD-11 further also did not provide information about LMNA and ZMPSTE24 mutations associated with Hutchinson-Gilford progeria syndrome. It has to be noted that in the linearization component of ICD-11, which we used, each child is assigned to only one parent.

With SNOMED-CT we were able to capture enough details to address the indication in most cases. For example, where ICD-11 lacked the type of polyneuropathy associated with hATTR amyloidosis, SNOMED-CT did include this detail. Furthermore, SNOMED-CT along with Orphanet were the only ontologies to include methotrexate toxicity. Only in some cases SNOMED-CT lacked certain details. For example, just like ICD-11, it did not account for the severity of AADC deficiency, MPA and GPA and mutations associated with Hutchinson-Gilford progeria syndrome. SNOMED-CT also did not provide details on clinical manifestations defined in indications, such as the UPCR in IgA nephropathy.

For MeSH, the details we could capture varied. For instance, some conditions like CMV infection, were part of a hierarchy in MeSH, while others, such as molybdenum cofactor deficiency, complementation group A, only existed as a supplementary concept or even only as an entry term such as pyruvate kinase deficiency. Intriguingly, MeSH adds detail to some indications with its attributes 'may be treated by' and 'may be prevented by'. This linked CMV infection to the drugs presented in its indication.

The way diseases are presented in MedDRA differs from that of the other ontologies. For example, MedDRA links CMV infection as being caused by drug reaction with eosinophilia and systemic symptoms (DRESS). It provides no additional detail on the disease and gives no relationships. In cases were the condition was represented, MeSH lacked additional details to capture the indication, as was the case for haemophilia A and AADC deficiency.

Orphanet could capture rare conditions with lots of detail. For example, the severity of haemophilia A was included in its hierarchy. In the case of CMV infection, which is not a rare condition, Orphanet still provided 'cytomegalovirus disease in patients with impaired cell mediated immunity deemed at risk' in its hierarchy. Other non-rare conditions such as IgA nephropathy could, however, not be found in Orphanet.

DO was able to accurately capture most of the conditions and provided enough detail for certain indications. For example, DO provided additional detail of dystrophic epidermolysis bullosa and junctional epidermolysis bullosa in the form of children. However, in other cases DO lacked the detail needed to capture the indications, as is the case for AADC deficiency which was not assigned a degree of severity in DO. Interestingly, in the hierarchy of DO there are big leaps present. In the same example, AADC is a direct child of 'inherited metabolic disorder', whereas in ICD-11 the direct parent of AADC is 'disorders of catecholamine synthesis'. A similar leap can be seen for CMV infection which is a direct child of 'viral infectious disease' in DO, but a child of 'disease caused by betaherpesvirinae' in SNOMED-CT.

Discussion

This report provides an overview of ontologies that can be used to classify diseases. The ontologies described here are ICD-11, SNOMED-CT, MeSH, MedDRA, Orphanet and Disease Ontology. The background, structure and advantages and disadvantages of each ontology are provided. We further gave an overview of UMLS as ontology repository. Lastly, we assessed the ability of each ontology to capture the condition and indication of 20 drugs that were assigned an orphan designation. Based on our assessment, we deem SNOMED-CT as most suitable for this purpose, while ICD-11, MeSH, MedDRA, Orphanet and Disease Ontology seem less applicable in this context.

All ontologies were able to capture the majority of conditions, regardless of whether the condition is rare or not. This was, however, not the case for Orphanet which did not capture the condition unless it was rare. This is in line with its purpose. However, if chosen for regulatory purposes, Orphanet alone will not be able to cover all diseases and would need to be used in addition to another, broader, ontology. One option for efficient application of Orphanet would be to make a distinction between orphan conditions and non-orphan conditions and decide that Orphanet is used for mapping the former, while another ontology is used for non-orphan drugs. Here, UMLS can be used as a tool to standardize an entity by mapping it to a standard concept from the UMLS methathesaurus and provide cross-references to Orphanet and the chosen ontology. However, before adapting this method in practice its feasibility would first need to be assessed.

The inclusion of indications in the ontologies differed. SNOMED-CT gave the necessary depth of information for most of the indications. This is due to a number of reasons. First, SNOMED-CT adapts a multi-hierarchical structure, making it possible to link one concept to multiple parents. These additional relationships with other parents provide a broader overview of a given concept. The use of attributes such as ‘causative agent’ further aid the representation of the totality of a concept. Furthermore, where other ontologies gave additional detail as synonymous terms, SNOMED-CT included this as part of its hierarchy. This is for example the case with type I and type II familial amyloid polyneuropathy which exist as entry terms in MeSH, but lead to a hierarchy that does not capture these details. Although, MeSH is used on the website of the EMA, it was in our analysis the next best in covering the indications of the medicines. Overall, it provides lots of detail and its strength lies in its attributes, for example 'may_be_prevented by' and 'may_be_treated_by'. These directly link a disease to one or multiple drugs. DO, which specifically focusses on the description of human diseases, shows potential to be used for regulatory purposes. Ideally, DO would include more details of diseases and be updated more frequently to be optimal for our purpose. MedDRA was the least inclusive ontology regarding the indication of drugs. This clearly demonstrates that classification of diseases is not the primary purpose of MedDRA. Even though it is designed for regulatory activities, its use cases are limited to reporting of adverse effects and are thus not applicable for our purpose

The details of the indication of the drugs that were encountered and which were most often missing in the ontologies, were the severity of the disease, but also mutations associated with a disease. One can imagine that in some instances these details are decisive for the mapping of an indication. The inclusion of these aspects in the ontologies would therefore be desirable. We further saw that if drugs are indicated for use in a certain population, e.g. paediatric or adult, this could not be distinguished by the ontologies. In certain cases the manifestation of a disease is the same in paediatrics and adults and therefore no distinction is made between the populations. However, having a distinction in population would be advantageous for regulatory purposes. The same holds true for indications associated with specific clinical features, which can probably better be described by ontologies such as the Human Phenotype Ontology, but is missing in the ontologies we assessed.

The differences in the structure of the ontologies could have an impact on the coverage of diseases. For example, we browsed through the linearization component of ICD-11 in which each concept is assigned to only one parent and thus part of only one hierarchy limiting the overall description of a disease. This is in contrast to for example SNOMED-CT where, as mentioned before, a concept can be part of different hierarchies. Another difficulty we encountered when browsing ICD-11, was its lack of capturing synonymous words. For example, a search for 'chronic myeloid leukaemia', did not retrieve the accurate hierarchy of this disease, but instead led to multiple extension codes. When searching 'chronic myelogenous leukaemia' instead, the hierarchy is presented. The way diseases are represented in MeSH is quite different than in the other ontologies. In MeSH a separate tree is dedicated to diseases, but diseases can also be present as a supplementary concept which is not part of the hierarchy, but are mapped to a term in a hierarchy. This is also reflected in the results for the indications, where some diseases are presented as part of a hierarchy, while others are supplementary concepts which are linked to a hierarchy.

For choosing either one of the ontologies as tool to map the condition or indication of drugs, the above mentioned aspects were taken into consideration. Ontologies provide different degrees of detail, which is directly related to their purpose. MedDRA, with a focus on drug adverse events, is the least applicable for our purpose. The lack of certain details in DO limits its use for regulatory purposes, while the lack of hierarchical relationships in MeSH is also not desirable for our purpose. This leaves us with ICD-11 and SNOMED-CT. Although ICD-11 benefits from various post-coordination options similar to SNOMED-CT to expand its coverage, in SNOMED-CT we achieved high coverage of diseases with the existing pre-coordinated concepts. Based on our analysis we thus deem SNOMED-CT to be best suitable for mapping the condition or indication of drugs. This conclusion is further substantiated by a recent study which, similar to our approach but on a larger scale, attempted to categorize clinical studies using their condition terms with SNOMED-CT. They found that using SNOMED-CT accurately categorised the condition in clinical studies and achieved higher coverage than MeSH terms[18]. Furthermore, Nictiz, the National Release centre which is responsible for distributing and managing SNOMED-CT in the Netherlands, implements its use to retrieve data from electronic health records in observational studies. The orphan conditions we analysed here were all included in SOMED-CT, however for expanding the coverage of the indication, SNOMED-CT can be used in combination with Orphanet, as discussed earlier.

When implementing SNOMED-CT, some aspects have to be considered. First, SNOMED-CT is available in different languages, but for our purpose the International Edition in English is most suitable. It is also of interest to note that SNOMED-CT is updated twice yearly. The types of changes made include new concepts, new descriptions, new relationships between concepts and new reference sets, as well as updates and retirement of any of these components. The updates are always accompanied by release files which include the valuable additional data that was not supported by the earlier format.

Once applied for the Regulatory Science Database, it is of importance that SNOMED-CT can be adapted for different purposes in the future. For example, the European Health Data Space aims to provide a data sharing framework with common standards and practices for the use of electronic health data by patients, researchers and other stakeholders. It is thus of additional value that SNOMED-CT can map this type of data as well, in order to facilitate interoperability between the Regulatory Science Database and healthcare data.

Acknowledgements

I would like to thank my supervisors, Marjon Pasmooij from the MEB and Lourens Bloem from the UU for their guidance, valuable suggestions and discussions during the whole project. I would also like to thank the team of Lourens for compiling the list of the 20 drugs analysed in report. Lastly, I gratefully acknowledge the feedback on the writing process received from Stefan Verweij and the valuable insights from Jarno Hoekman.

References

1. Roberts, K. *et al.* A vision for integrated publicly available information on regulated medical products. *Clin. Transl. Sci.* **15**, 1321–1327 (2022).
2. Robertson, A. S., Reisin Miller, A. & Dolz, F. Supporting a data-driven approach to regulatory intelligence. *Nat. Rev. Drug Discov.* **20**, 161–162 (2021).
3. Sansone, S. A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **37**, 358–367 (2019).
4. Ghomari, L. Z. & Ghomari, A. R. Ontology versus terminology, from the perspective of ontologists. *Int. J. Web Sci.* **1**, 315 (2012).
5. Awaysheh, A. *et al.* A review of medical terminology standards and structured reporting. *J. Vet. Diagnostic Investig.* **30**, 17–25 (2018).
6. Bonner, S. *et al.* A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Brief. Bioinform.* 1–34 (2022) doi:10.1093/bib/bbac404.
7. Haendel, M. A. *et al.* A Census of Disease Ontologies. *Annu. Rev. Biomed. Data Sci.* **1**, 305–331 (2018).
8. Dalianis, H. Medical Classifications and Terminologies. in *Clinical Text Mining* vol. 7 497–508 (2017).
9. WHO. International Classification of Diseases Eleventh Revision (ICD-11). *Geneva: World Health Organization* https://icdcdn.who.int/icd11referenceguide/en/html/index.html#copyright-page (2022).
10. Chute, C. G. & Çelik, C. Overview of ICD-11 architecture and structure. *BMC Med. Inform. Decis. Mak.* **21**, 1–7 (2021).
11. Harrison, J. E., Weber, S., Jakob, R. & Chute, C. G. ICD-11: an international classification of diseases for the twenty-first century. *BMC Med. Inform. Decis. Mak.* **21**, 1–10 (2021).
12. WHO. *WHO-FIC Content Model Reference Guide*. https://icd.who.int/browse11 (2021).
13. Fung, K. W., Xu, J. & Bodenreider, O. The new International Classification of Diseases 11th edition: A comparative analysis with ICD-10 and ICD-10-CM. *J. Am. Med. Informatics Assoc.* **27**, 738–746 (2020).
14. International Health Terminology Standards Development Organization. SNOMED CT Starter Guide. *Snomed* 1–56 (2022).
15. Rossander, A., Lindsköld, L., Ranerup, A. & Karlsson, D. A State-of-the Art Review of SNOMED CT Terminology Binding and Recommendations for Practice and Research. *Methods Inf. Med.* 76–88 (2021) doi:10.1055/s-0041-1735167.
16. Khorrami, F., Ahmadi, M. & Sheikhtaheri, A. Evaluation of SNOMED CT content coverage: A systematic literature review. *Stud. Health Technol. Inform.* **248**, 212–219 (2018).
17. Koopman, B., Bruza, P., Sitbon, L. & Lawley, M. Towards semantic search and inference in electronic medical records: An approach using concept-based information retrieval. *Australas. Med. J.* **5**, 482–488 (2012).
18. Liu, H. *et al.* Ontology-based categorization of clinical studies by their conditions. *J. Biomed. Inform.* **135**, (2022).
19. Nelson, S. J., Johnston, W. D. & Humphreys, B. L. Relationships in Medical Subject Headings (MeSH). 171–184 (2001) doi:10.1007/978-94-015-9696-1_11.
20. National Library of Medicine. Introduction to MeSH. https://www.nlm.nih.gov/mesh/introduction.html (2021).
21. Chang, A. A., Heskett, K. M. & Davidson, T. M. Searching the literature using medical subject headings versus text word with PubMed. *Laryngoscope* **116**, 336–340 (2006).
22. European Medicines Agency. European public assessment reports: background and context. https://www.ema.europa.eu/en/medicines/what-we-publish-when/european-public-assessment-reports-background-context.
23. Darmoni, S. J. *et al.* Improving information retrieval using medical subject headings concepts: A test case on rare and chronic diseases. *J. Med. Libr. Assoc.* **100**, 176–183 (2012).
24. MedDRA. Introductory Guide MedDRA Version 25.1. (2022).

25. Harrison, J. & Mozzicato, P. MedDRA®: The tale of a terminology: Side Effects of Drugs Essay. *Side Eff. Drugs Annu.* **31**, xii (2009).

26. Bousquet, C., Souvignet, J., Sadou, É., Jaulent, M. C. & Declerck, G. Ontological and non-ontological resources for associating medical dictionary for regulatory activities terms to SNOMED clinical terms with semantic properties. *Front. Pharmacol.* **10**, 1–21 (2019).

27. Vasant, D. *et al.* ORDO: An Ontology Connecting Rare Disease, Epidemiology and Genetic Data. *Phenotype data ISMB2014* (2014).

28. Rath, A. *et al.* Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).

29. Orphanet. Procedural document on the Orphanet nomenclature and classification of rare diseases. *Orphanet J. Rare Dis.* **02**, 1–7 (2020).

30. Orphanet. The portal for rare diseases and orphan drugs. https://www.orpha.net/consor/cgi-bin/Education_AboutOrphanet.php?lng=EN.

31. Xiang, J., Zhang, J., Zhao, Y., Wu, F. X. & Li, M. *Biomedical data, computational methods and tools for evaluating disease-disease associations*. *Briefings in Bioinformatics* vol. 23 (2022).

32. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).

33. Kibbe, W. A. *et al.* Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**, D1071–D1078 (2015).

34. Schriml, L. M. & Mitraka, E. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. *Mamm. Genome* **26**, 584–589 (2015).

35. Schriml, L. M. *et al.* The Human Disease Ontology 2022 update. *Nucleic Acids Res.* **50**, D1255–D1261 (2022).

36. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).

37. Raje, S. & Bodenreider, O. Interoperability of disease concepts in clinical and research ontologies: Contrasting coverage and structure in the disease ontology and SNOMED CT. *Stud. Health Technol. Inform.* **245**, 925–929 (2017).

38. Whetzel, P. L. *et al.* BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**, 541–545 (2011).

39. National Library of Medicine. UMLS® Reference Manual. https://www.ncbi.nlm.nih.gov/books/NBK9681/ (2021).

40. Amos, L., Anderson, D., Brody, S., Ripple, A. & Humphreys, B. L. UMLS users and uses: A current overview. *J. Am. Med. Informatics Assoc.* **27**, 1606–1611 (2021).

41. Bodenreider, O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inform.* 67–79 (2008) doi:10.1055/s-0038-1638585.

42. Konopka, B. M. Biomedical ontologies - A review. *Biocybern. Biomed. Eng.* **35**, 75–86 (2015).

Supplementary Material 1

Table S1: *Summary of disease ontologies with a description compiled via FairSharing.org*

Ontology	Description
Autism DSM-ADI-R ontology	Represents DSM IV diagnostic criteria for autistic disorder and ASD criteria for Autism Spectrum Disorder
Breast tissue cell lines ontology	Contains a comprehensive list of cell lines derived from breast tissue, both normal and pathological, with cross relation to classes- genetic variation, pathological condition, genes, chemicals and drugs
Cancer chemoprevention ontology	A vocabulary that is able to describe and semantically interconnect the different paradigms of the cancer chemoprevention domain
CareLex controlled vocabulary	Contains controlled vocabulary terms from National Cancer Institute used to classify clinical trial electronic content (documents, images, etc)
Clinical measurement ontology	Standardizes morphological and physiological measurement records generated from clinical and model organism research and health programs
Clinical trials ontology	Describes clinical trials in the field of neurodegeneration
Common terminology criteria for adverse events	A coding system for reporting adverse events that occur in the course of cancer therapy
COPD ontology	Ontology used for modelling concepts associated to chronic obstructive pulmonary disease in routine clinical databases
Disease ontology	Standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics, underlying mechanisms and related medical vocabulary disease concepts.
Drug database for inborn errors of metabolism ontology	A database of therapeutic strategies and treatments for inborn errors of metabolism. These strategies are classified by mechanism and outcome
Epilepsy ontology	Ontology about the epilepsy domain and epileptic seizures
Epilepsy and seizure ontology	An application ontology developed to support epilepsy focused informatics tools for patient care and clinical research
Exposure ontology	Facilitates the centralization and integration of exposure data to inform understanding of environmental health
Genetic glycol-diseases ontology	Focuses on the molecular aetiology, pathogenesis, and clinical manifestations of genetic diseases and disorders of glycan metabolism and developed as a knowledge-base for this scientific field
Human phenotype ontology	Provides a structured and controlled vocabulary for the phenotypic features encountered in human hereditary and other disease

International Classification of Disease (version 9-11)	Allows the recording, reporting and grouping of conditions and factors that influence health
Infection disease ontology malaria	An application ontology for malaria extending the infectious disease ontology (IDO)
International Harmonization of Nomenclature and Diagnostic criteria	Standard reference for nomenclature and diagnostic criteria in toxicologic pathology
Logical observation identifier names and codes	A common language (set of identifiers, names, and codes) for clinical and laboratory observations
MedlinePlus health topics	Provides information on the symptoms, causes, treatment and prevention for a wide range of diseases, illnesses, health conditions and wellness issues
MedRA	Used to analyse individual medical events (e.g., "Influenza") or issues involving a system, organ or etiology (e.g., infections) using its hierarchical structure
NCI thesaurus	Covers vocabulary for clinical care, translational and basic research, and public information and administrative activities
Online medelian inheritance in man ontology	A comprehensive, authoritative compendium of human genes and genetic phenotypes as well as the relationship between them
Ontology for general medical science	An ontology of entities involved in a clinical encounter. OGMS includes very general terms that are used across medical disciplines
Ontology for genetic disease investigations	Is used to model scientific investigation, especially Genome-Wide Association Studies (GWAS), to discover genetic susceptibility factors to disease. It models the genetic variants, polymorphisms, statistical measurement, populations and other elements that are essential to determine a genetic susceptibility factor in GWAS study
Ontology for genetic susceptibility factor	Is an application ontology to model/represent the notion of genetic susceptibility to a specific disease or an adverse event or a pathological biological process. OGSF is built from a combination of three ontologies: the Ontology of Geographical Region (OGR), the Ontology of Glucose Metabolism (OGMD), and the OGDI
Ontology of cardiovascular drug adverse events	Is an ontology of adverse events associated with cardiovascular disease drugs. It extends the Ontology of Adverse Events (OAE)
Ontology of glucose metabolism disorder	Includes disease names, phenotypes and their classifications
Orphanet rare disease ontology	Provides a structured vocabulary for rare diseases capturing relationships between diseases, genes and other relevant features which will form a useful resource for the computational analysis of rare diseases
Parkinson's disease ontology	Created to represent and model the Parkinson's Disease knowledge domain. This ontology covers

	major biomedical concepts from molecular to clinical features of the disease as well as the different views on disease features held by molecular biologists, clinicians and drug developers
Pathway ontology	Covers all types of biological pathways, including altered and disease pathways, and to capture the relationships between them
Paediatric terminology	Terms associated with paediatrics, representing information related to child health and development from pre-birth through 21 years of age
Pharmacovigilance ontology	Connects known facts on drugs, disease, adverse drug events and their molecular mechanisms
PhenX phenotypic terms	Standard measures related to complex diseases, phenotypic traits and environmental exposures
Physician data query	Is part of NCI's comprehensive cancer information database, which contains expert summaries on a wide range of cancer topics,
Radiation oncology ontology	Covers the radiation oncology domain with a strong focus on re-using existing ontologies
Radiology gamuts ontology	Is a knowledge resource for radiology diagnosis and contains differential-diagnosis listings for imaging findings in all body systems
Sickle cell disease ontology	Supports the building of databasing and clinical informatics in SCD
Sleep domain ontology	An application ontology for the domain of Sleep Medicine
Symptom ontology	Designed for understanding the close relationship of Signs and Symptoms, where Signs are the objective observation of an illness, understanding that at times, the same term may be both a Sign and a Symptom