ICDO: Ontological representation of the International Classification of Diseases (ICD) and its application in English and Chinese healthy data standardization

Ling Wan[a,b], Edison Ong[b], Yongqun He[b]

[a] *OntoWise, Nanjing, Jiangsu, China,*
[b] University of Michigan Medical School, Ann Arbor, MI 48109, USA

Abstract

The ICD-9/10/11 are released by the WHO and used worldwide to support applications including health insurance classification. However, different countries develop their own modifications of the ICD system and these versions are often incompatible. In addition, the semantic relations among ICD disease terms are unclear, and how these terms are related to other entities such as anatomic entities are not defined. To address these issues, we developed an ICD ontology (ICDO) to logically represent ICD terms and their relations with anatomic entities, qualities, etc. Different from other disease ontologies, all ICD diseases are defined disease processes in ICDO. The current ICDO focuses on English and Chinese representation. As a use case, we used ICDO to integrate ICD related data from 33 regions in Jiangsu province in China. Our strategy was able to identify and standardize local ICD versions in these regions.

Keywords:

ICD; ontology; disease standardization

Introduction

The International Classification of Diseases (ICD), maintained by the World Health Organization (WHO), is the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes. ICD defines diseases, disorders, injuries and other related health conditions in the biomedical and clinical domains in a comprehensive and hierarchical fashion. The ICD has been continuously revised and published in a series of editions to reflect advances in health and medical science over time (1,2). ICD is the foundation for the identification of health trends and statistics in a global setting.

Many countries have adopted the ICD standard and developed their own modified versions. For example, there are the USA version of ICD-10-CM (3) and Germany version of ICD-10-GM (4). In China, there are different formats including National Standard V.1.1, GB/T14396-2016 and National Clinical Version 1.1 (Table 1). The availability of so many versions makes it difficult to standardize health records in China. This study focuses on the GB/T14396-2016, which is the ICD10 Chinese version authorized by the national administrative. Recently WHO released the ICD-11, which is the latest version of ICD and China reported to adapt the ICD11 version in 2019.

Table 1. Different ICD10 versions in China

No.	ICD10 versions in China
1	National StandardV.1.1
2	GB/T 14396-2016
3	National Clinical Version1.1
4	Beijing Clinical versionICD-10 V6.01
5	National Standard V.1.0
6	ICD10（2011modification）
7	National ICD10 V1.3
8	Shanghai ICD-10(2013 updated)
9	National RC020-1CD-10 Diagnostic code
10	Beijing version RC020-1CD-10 Diagnostic code
11	Beijing ICD10 V5.0
12	Guangdong ICD-10（2017）
13	National clinical ICD10 V.1.0
14	Ji'nan city,Shandong province ICD10 (For Health information data sharing and exchange)

Note: This source of this table comes from a survey by OAMAHA:http://www.sohu.com/a/302897591_324186, 2019-03-21, which is translated by author.

The ICD is used as the controlled terminology of diseases in the medical information platform in most healthcare administrations. There are many application systems that exist in hospitals, such as: HIS (health information systems) (5), LIS (laboratory information system) (6), PACS (A picture archiving and communication system), the EMR (Electronic medical records). These data can be integrated by the ICD framework.

On the other hand, both ICD codes and Diagnosis-related Groups (DRGs) are a major method for medical insurance control and the DRGs is dependent on the correctness of ICD (7). Due to its important role in many medical and clinical fields, a large amout of mapping effort is required to ensure interoperability among different ICD versions.

The semantic mapping among databases generated under two different coding systems (e.g., ICD10 and ICD11) is very difficult and generally requires manual intervention. The National Institutes of Health (NIH) refer such difficulty to the phenomenon of 'data wrangling' encompassing activities that make data more usable by changing their forms but not their meanings (8). Although great efforts have been made on this area, the obstacle still exists. The ICD terminology is composed of a code/value pair. Each ICD standard code corresponds to a unique disease name as a value. However, in reality, there are often multiple synonyms expressed for one disease in the natural language. For example, the ICD 11 code AA0Z has value of Infectious diseases of external ear, unspecified; the GB/T14396-2016 code H60.001 has value of 外耳疖 (external ear furuncle); the ICD 10 code H60.5 has value of acute otitis externa, noninfective. Due to the existence of polysemy in natural language (especially in Chinese), the code-value mapping often encounters ambiguity after using Extraction-Transfer-Load (ETL) tool for data integration, and results in improper matching. Particularly in China, these problems are mainly due to the different local ICD versions with private extensions to certain ICD terms. These modifications are made according to the internal clinical needs coming from different medical units. The large discrepancy among different versions might cause many problems, such as the appearance of the large amount of data with different values but the same code, or the same value with different codes. This also affects the accuracy of ICD-based DRG grouping, the accuracy of Medicare payments as well as the statistics accuracy of death causes.

In addition to the ICD, there are many disease discription models being developed and used. Hadzic et al. classify disease into four dimensions: (i) generic disease types; (ii) phenotypes that are mainly based on observations to describe the various symptoms of the disease; (iii) etiology that is a strictly scientific basis of pathogenic factors, mainly including two categories - genetic factors and environmental factors; (iv) treatment that is a possible effective measures against a particular disease (9). These four dimensions together can describe the overall knowledge of a disease field. Fang et al. learned from the classification of SNOMED (10,11) and ICD to improve and make a new disease description model. On the basis of the axis, the general disease discription model of Hadzic was improved, and two basic characteristics of complications and detection methods were added, and the symptoms, signs, staging, sex, age, acute and chronic and onset time were classified as clinical manifestations (12).

Ontology is likely the best approach to solve the issue of semantic mapping among different databases and terminology systems. A formal biomedical ontology is a set of computer and human-interpretable terms that represent entities and relations among the entities in a biomedical domain. Ontologies have emerged to be critical to biomedical and clinical data standardization, management, integration, and analysis. Two different databases or terminologies may be formed based on different organizational principles and are unlikely or difficult to form an agreement about what each piece of information refers to and how they can be aligned. The inability to achieving interoperability can severely compromise the goals of information integration and aggregation. Such issue is difficult to solve internally or among the two databases (8). However, the usage of community-based and consensus-based ontologies provides a feasible way to solve the term mapping and information integratoin issues.

Many disease-related ontologies exist, including Human Disease Ontology (DOID) (13,14), Monarch Disease Ontology (MONDO) (15), and the Ontology of General Medical Science (OGMS) (16). In DOID and MONDO, diseases are treated as disposition, which is a realizable entity that bears in some material entity and can be realized in a life process (8). However, in the setting of ICD usage, diseases have already occurred and are not disposition per se. OGMS includes two high level terms: disease and 'disease course', where disease is asserted as a disposition and 'disease course' as a process.

To find a semantic mapping method between different ICD versions, here we report the development of an ICD ontology (ICDO) to address the issues of database interoperability and data integration as listed above. Given that ICD is mainly applicable to statistical analysis and disease grouping for healthcare insurance, we present in this paper our disease design pattern that combines the advantages of the above disease description models. Our disease design pattern in ICDO is based on the understanding that the disease in ICD is a human pathological process that realizes disease disposition. Such process is composed of a group of entities, which has reversible decomposition. These entity are 'anatomical structure', 'pathological anatomical entity', 'etiology', 'quality' and 'syndrome'. Therefore, all the ICD terms are defined as subclasses of the ICDO "disease process" class, which is then defined as a subclass of the imported OGMS term 'pathological bodily process' (16). In this manuscript, we detail our ICDO developmental strategy and provide a comprehensive use case to illustrate the usage of the ICDO.

Methods

General ICDO development strategy

Our ICDO development closely followed the WHO ICD 10/11 classification and principles. The ICDO development used the eXtensive Ontology Development (XOD) strategy (17), which emphasizes the reuse and alignment of ontology terms and semantic relations, ontology design patterns, and community effort. Specifically, we aligned the ICDO terms with Basic Formal Ontology (BFO) and BFO-compatible ontologies (8). Ontofox (18) was used to extact terms from existing ontologies that were then imported and reused in ICDO.

We focused our first stage ICDO development on the specific area of external ear diseases as a proof-of-concept. This early stage ICDO prototype includes all diseases related to external ear part in: (i) ICD11 under the class, "Disease of the ear and mastoid process", coded from AA00 to AA6Z, (ii) ICD10 under the "external ear diseases", and (iii) GB/T 14396-2016.

The Protégé OWL editor (http://protege.stanford.edu) was used to visualize ICDO, add new ICDO terms, edit imported terms and merge imported ontologies. ICDO-specific terms were generated using new ICDO identifiers with the prefix "ICDO_" followed by 7-digit auto incremented numbers. The Hermit reasoner was used for consistency checking and reasoning (http://hermit-reasoner.com/).

ICDO format, source code, and deposition

ICDO is expressed using the W3C standard Web On-tology Language (OWL2) (http://www.w3.org/TR/owl-guide/). The current ICDO source code is openly available at GitHub: http://github.com/icdo/ICDO.

The ICDO ontology is deposited in the NCBO BioPortal website: https://bioportal.bioontology.org/ontologies/ICDO, as well as the ontology repository website Ontobee (19): http://www.ontobee.org/ontology/ICDO.

Application of ICDO for mapping and standardizing different versions of disease classifications

The health systems in several regions of the Jiangsu Province in China used different modified versions of ICD10. To standardize the coding systems from these regional health information platforms, we identified three regional ICD10 modification coding systems, and used ICDO to model and standardize these coding systems.

ICDO query and analysis

Description Logic (DL) query was used to query the knowledge built in ICDO. The DL query function in the Protégé-OWL editor was used for the implementation.

Results

General disease definition of disease development strategy

First we performed a survey of how the term "disease" is defined in different ontologies and dictionaries (Table 1). It is clear that the nature of disease is defined differntly. In four ontologies including DOID, OGMS (16), MONDO, and EFO (Experimental Factor Ontology) (20,21), disease are all defined as a disposition. In the Semanticscience Integrated Ontology (SIO) (22), disease is defined as an outward manifestatoin of one or more disorders. Disease has also been defined as a disorder by itself or pattern of abnormality (Table 1).

Table 2. Survey of disease definitions

Source	Definition
DOID, OGMS, and MONDO	A disease is a disposition (I) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.
EFO	A disease is a disposition that describes states of disease associated with a particular sample and/or organism.
SIO	disease is the outward manifestation of one or more disorders.
Exposure Ontology (23)	A disease is a pattern of abnormal functioning, or abnormal localization of normal functioning, and/or abnormal localization of constituents when compared to other members of that species.
Dictionary (https://www.dictionary.com)	A disorder of structure or function in a human, animal, or plant, especially one that produces specific signs or symptoms or that affects a specific location and is not simply a direct result of physical injury.

In OGMS, there are two disease-related terms 'disease course', and 'pathological bodily process'. The term 'disease course is defined as "The totality of all processes through which a given disease instance is realized". However, it is unclear what the "all processes" in the definition means. It is possibly that some of the processes are not directly related to disease. The OGMS term 'pathological bodily process' is defined as "A bodily process that is clinically abnormal". The diseases listed in ICDO have already happened, and are not an upcoming event. Given that the ICD is used primarily for post-disease recording and insurance filing purposes, we think that the disease in ICD is primarily meant to be a type of pathological bodily process; therefore, the disease in ICD can be better regarded as a "disease process" under OGMS 'pathological bodily process'.

In ICDO, based on the nature of ICD and its applications, we focus on the representation of disease processes instead. Therefore, the term 'disease process' becomes our major term, which is defined in ICDO as follows:

> *Disease process =def. a pathological bodily process that occurs in a specific anatomic location, realizes a disease disposition, has abnormal bodily phenotype, and results in a pathological anotomic entity.*

Therefore, all the specific diseases in ICDO are all defined as disease processes, which are different from other disease description frameworks. As a result, ICDO represents all disease names from ICD11, ICD10, GB/T14396 as disease processes, often abbreviaetd with the suffix "DP" in ICDO term labels.

In this study, ICDO is mainly used to support data standardization among different ICD versions. ICDO aims to standardize clinical data from international multi-center and also data generated under different ICD local and modified versions in China. To support the general interoperability goal, we have included ICD10 and ICD11 terms in both English and Chinese languages in the ICDO.

ICDO top level design and structure

ICD 10/11 has different classification and principles in top level design and therefore we closely have followed OBO to develop ICDO top level hierarchy. Extending from the formal definition and classificaiton of this ICDO *"disease process"* term, we generated an upper level ICDO hierarchical structure (Fig. 1). ICDO reused many terms from existing ontologies such as the BFO (8), OGMS (16), UBERON (24), PATO (Phenotype And Trait Ontology, an ontology of phenotypic qualities (properties, attributes or characteristics), https://github.com/pato-ontology/pato/). The top-level terms were aligned with BFO.

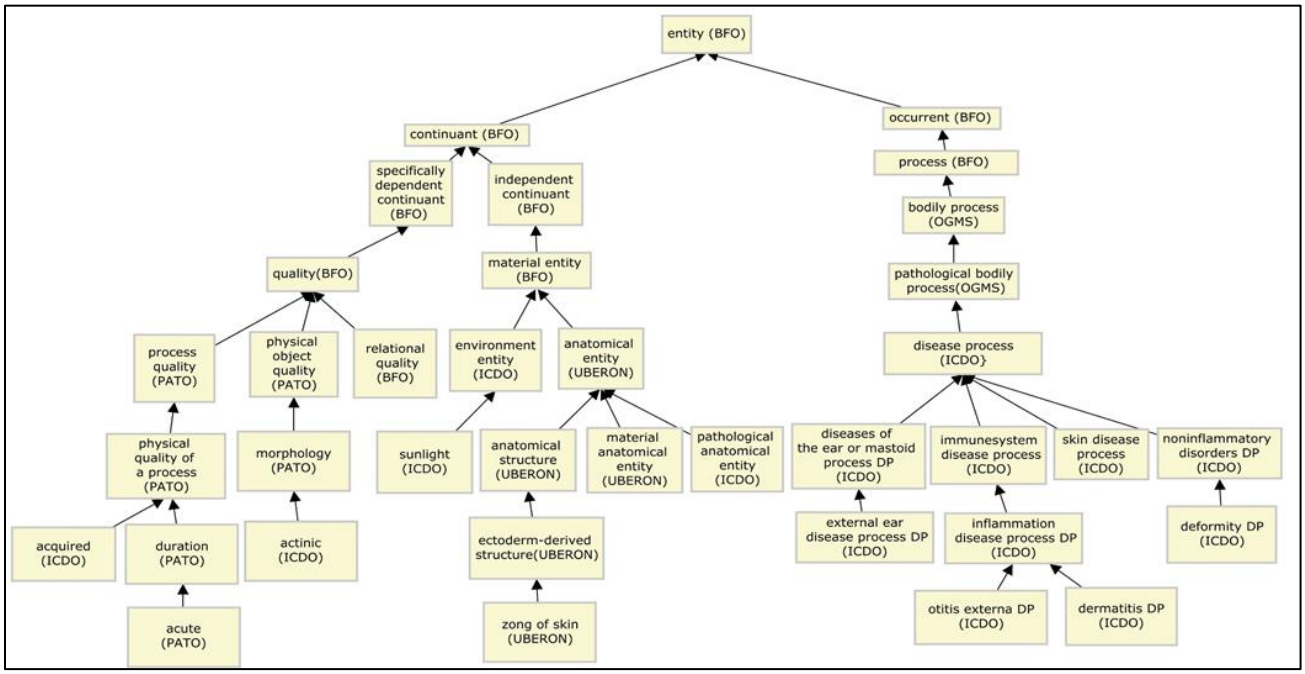

Fig 1. ICDO top level hierarchical structure.

ICDO general design pattern for diseases

In ICDO, a disease process was composed of four major elements: etiology entity, quality, anatomical stucture and pathological anatomical entity. The disease pattern of ICDO was shown in Fig. 2:

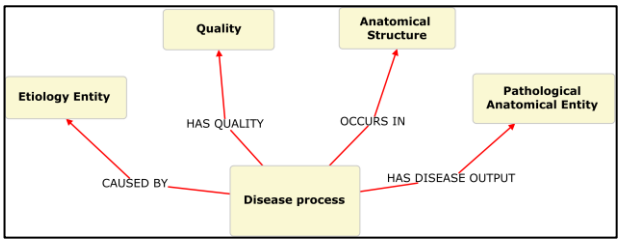

Fig 2. ICDO disease process pattern

We represented different diease processes following the disease process pattern (Fig. 2). For example, the ICD term 'granuloma of external ear canal' is defined in ICDO as *"granuloma of external ear canal DP"*, which is a granuloma disease process that *"occurs in" some "external ear canal"* and *"has disease output" some "granuloma"*. The *"necrotic external ear otitis"* is an otitis disease process *"occurs in" some "external ear"* and *"has quality" some "necrotic"*. Note that the 'has quality' is indeed a shortcut relation where the quality is not the quality of the process per se. Instead, the quality is the quality of the anatomical entity of the patient.

ICDO also has different design strategy for diseases compared to DOID and MONDO. In general, DOID and MONDO do not discomposed a disease term into different entity components such as etiology entity, quality and anatomical structure. We paid a lot of attention to each of these issues and developed our specific strategeis. In addition, due to the nature of ICD usage in clinical disease classificaiton and insurance filing, we have designed many special design patterns for the ICDO generation. Some of the special design patterns, together with the approaches proposed in ICDO, are described in next session.

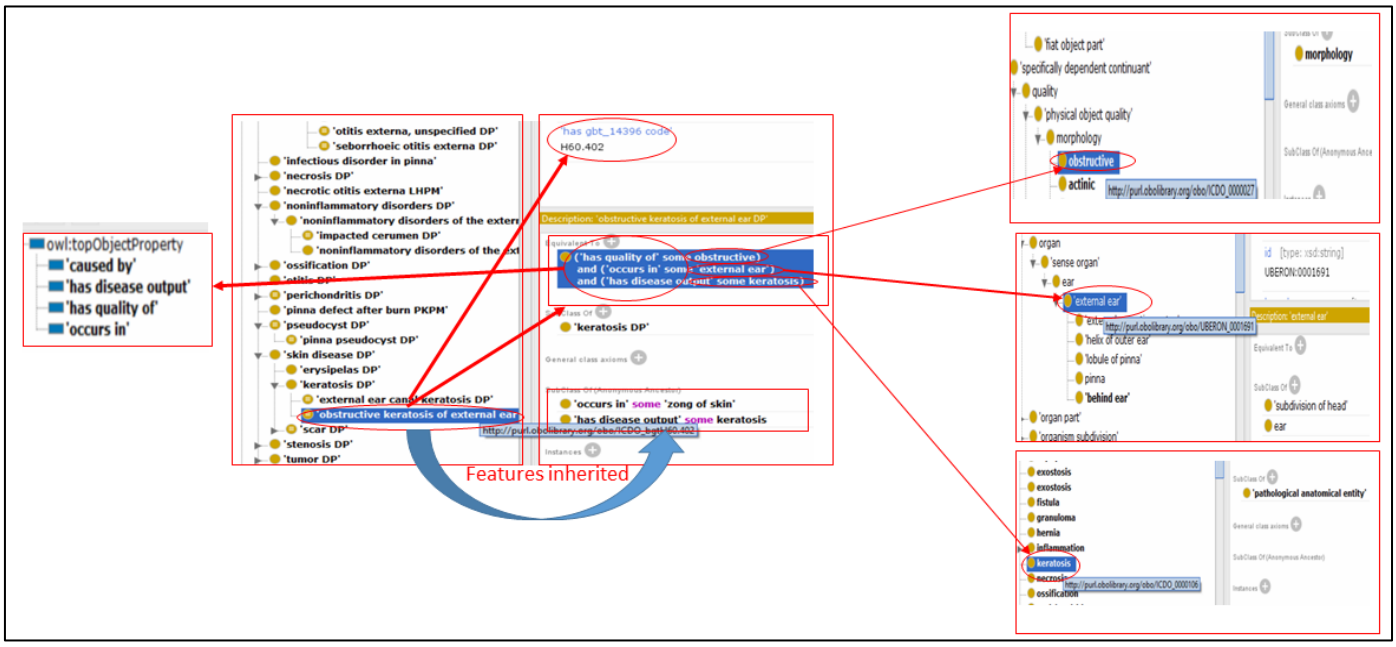

Fig. 2. Disease process modeling in ICDO. In this example, the term 'obstructive keratosis of external ear DP' is represented using the design pattern, including disease quality, occurs in location, and disease material output.

ICDO strategy to represent special ICD10 disease classes

Besides general disease classifications, ICD includes many special terms such as "classified elsewhere", "other specified" and "unspecified". ICDO has implemented special strategies to handle the mentioned special terms.

The first special terms, "classified elsewhere", were treated as obsoleted terms in ICDO. We believe that the disease classification must be clear and consistent among various disease categories. The definition of "classified elsewhere" is confusing because there is no obvious and proper disease category for "elsewhere". To ensure the classification integrity, a disease term can be classified under multiple disease categories based on varies definitions and applications, but it should not be classified as an undefined category, "elsewhere". To balance the mapping process among various ICD versions and proper handling of the undefined category, we added all the 'disease classified elsewhere' terms in ICDO but made them as obsoleted terms in the ontology.

There are also many ICD terms labeled as "other specified". Logically speaking, all ICD terms should be classified into specific classes and there should not exist any 'other' class. We consider this type of "other specified" terms class as logical error and put all the terms under this class into their parent class. In other words, we generated an ICDO term "other specified" and put it under the obsolete to support mapping among to existing ICD versions. To ensure the continuity of the various versions of the ICD in the conversion adaptation process, in the data adaptation process, this obsolete class term may still participate in the operation to ensure the accuracy of data mapping.

Many ICD terms are labeled as "unspecified". For their corresponding classes, we have been able to determine the parent classes. However, due to the limitations of current definitions and the lack of knowledge, these "unspecified" terms have not given any specific description as of now, and can be mapped to their parent terms in ICDO.

The ICD10/11 has the Extension Codes used to support clinical treatment, such as the organ laterality in different ICD versions and the "special anatomy" in ICD11. The laterality commonly found in various ICD versions includes "left", "right", "bilateral", "unilateral, unspecified", "unspecified laterality". Additionally, ICD11 introduced a term "special anatomy", which includes anatomical synonyms and possible anatomical structure of disease. We built the laterality as the quality of the disease in ICDO, and adopted the synonym of the specific anatomy as a synonym label in annotation in anatomical structure if UBERON have not included the synonym. Other extension codes exist such as distribution and regional. For example, abscess of right external ear DP can be defined to have axiom assertion of:

> *"abscess of external ear DP" and ("occurs in anatomical side" some "right side of anatomical entity")*

Such design properly handles the issue of anatomic literality.

With the focus of disease process, ICDO also has a natural advantage of defining different disease stages, or the beginning, middle, and end of a disease process. Such a process aspect supports real life disease representation.

ICDO mapping process and use case application

The development of ICDO starts from designing the disease pattern first, then decomposing the terms into different components of the semantic equivalence terms. Then we used the Ontobee annotator (http://www.ontobee.org/annotate) to discompose these disease terms into different components and tag the components according to our disease design pattern (Fig. 2). Finally, we established relationships between the components of the disease terms by creating objective properties and annotations following the disease desgin pattern. This entire process involves the decomposition of a disease name and then establish the logical relations among the components. (Fig. 3).

In order to illustrate the process, we extracted terms from three local ICD10 versions genereated by two districts and one city (Pukou district, Liuhe district and Jiangyin city) in Jiangsu province, China, and performed the mapping. The China administrative departments from different regions often publish and adopt their own ICD versions. Therefore, the harmonization of the ICD in China is a complex and difficult issue due to varies local versions and custom modifications. As shown in Table 1, there are 14 different ICD local and modified versions in China. Among them, GB/T-14396-2016 is a Chinese version of ICD10 modification required by the Chinese government since February 2017. Some local versions listed in Table 1 are used for clinical purposes and the others used for adminstrative statistics purposes. For example, in Jiangsu province in China, we identified more than 30 regional datasets but they used different locally modifed ICD versions. Even though the Jiangsu province has the most advanced health informatics system among all the provinces in China, many incorrect code-value pairs existed in these local coding systems and need to be mapped to the GB/T-14396-2016 coding system.

(i) Same code but different values: For example, in the local Liuhe district coding system, the two Chinese disease term "坏死性外耳炎" and "恶性外耳炎" (English translation: "necrotic external ear otitis" and "malignant otitis externa" respectively) have the same code H60.200. However, the code H60.200 corresponds to the "malignant otitis externa by ICD10 or GB/T14396-2016, and the term "necrotic external ear otitis" should be coded as H60.900 ("otitis, externa").

(ii) Same values but different codes: For exmaple , in the local Jiangyin city coding system, the Chinese disease term "后天性外耳畸形" (English translation: "acquired deformity of ear externa") has two codes: H61.303 and H61.101. However, according to ICD10 or GB/T14396-2016, the correct code of this disease name should be H61.101. The code H61.303 even does not exist in ICD10 or GB/T14396-2016.

These two types of errors shown above widely exist in the local Chinese ICD modified versions. For example, even for the same external ear branch, there are 36 errors in Pukou district, 4 errors in Liuhe disctrict and 14 errors in Jiangyin city local coding systems. Note that there are only 22 code-value pairs in ICD10 and 41 code-value pairs in GB/T14396-2016 for the external ear-related diseases. Considering the total number of over 20,000 terms in ICD10 and 14 different local ICD10 versions in China (*Table 2*), it is a huge effort to manually correct these code-value pair errors. the pair errors have become a major issue to support data integration and systematic statistical data analysis in China.

In this study, we developed an ICDO-based semantic disease name mapping algorithm (ISDNA), as shown in Fig. 4, with the aim to solve the disease name mapping issue as illustrated in the use cases of Jiangsu Province, China. First, the ISDNA algorithm first accepted some disease names in a specific langugae (e.g., English, Chinese) as input. The input names were then decomposed via natual language process (NLP) to different components using Ontobee Annotator. For example, a disease name could be broken down into the anatomic entity term as the location of the disease process, quality of the patient or the anatomic bodily entity of the patient, and abnormal pathological entity as the output of the disease. These identified components were then mapped to their corresponding ontology terms and IDs. Based on the axioms defined in ICDO, we can use ontology reasoners to automatically infer the ICDO codes that the input disease name belongs to (Fig. 5). Our ISDNA algorithm is able to identify exact matches that perfectly mapped or semantically inferred the parent terms of the matched disease names from the Pukou district, Liuhe district and Jiangyin city in Jiangsu province, China to the corresponding ICDO terms. Next we will provide two examples to illustrate the features and performance of the ISDNA algorithm.

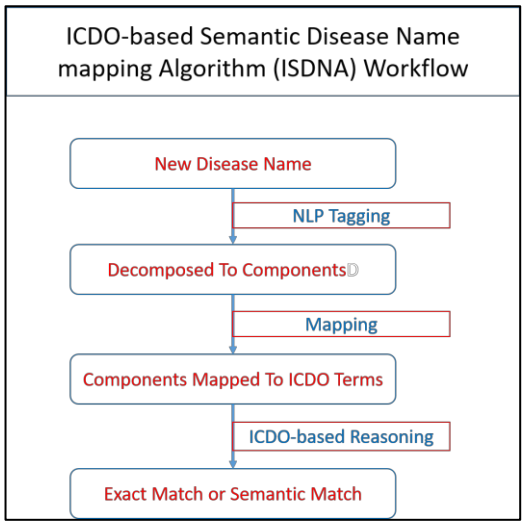

Fig. 4. ISDNA algorithm workflow

Fig. 5. provides the first example of how we can infer a specific name to a perfectly matched ICDO term. Specifically, in this use case, "cellulitis of external ear" (Chinese name: "外耳蜂窝织炎") is an input disease term. It was first split into two components: *"cellulitis anatomic entity"* and *"external ear"*. Given the nature of these two terms, the following two axioms could be assigned:

'occurs in' some 'external ear'

'has disease output' some 'cellulitis anatomic entity'

Based on these two axioms, the ICDO reasoner was able to infer "cellulitis of external ear" as an exact match to the ICDO *"cellulitis of external ear DP"* (Chinese name "外耳蜂窝织炎") term with the GB/T14396 code H60.100, ICD10 code H60.1, and ICD11code AA01. Then we can select one of the code from them according to our needs.

Fig. 6. provides another example that has the input disease name "Pinna defect after burn" (Chinese name: "烧伤后耳廓缺损"), a term from the Pukou district local coding system in our use case. Similarly, our algorithm started with spliting the long disease name to three components: *"pinna"*, *"defect"*, and *"after burn"*, which could then be mapped to their corresponding ontology terms in ICDO. Note that the term "after burn" is an defined as an synonym of the "acquired after burn", which is a subclass of the quality "acquired". The term "Pinna defect after burn" is not included in either ICD or ICDO. But for demonstrative purpose, the term "Pinna defect after burn" was introducted (Fig. 6) to simulate the situation that there is no perfect matching. After direct mapping, there was no exact match for this disease name. However, after running the Hermit reasoner available in the Protégé OWL editor, we were able to infer this disease name to be subclass of the ICDO *"acquired deformity of pinna DP"* term (Chinese name: *"获得性耳廓畸形"*) with ICD11 code AA41. Given that there was no exact match for "Pinna defect after burn", the ICDO *"acquired deformity of pinna DP"* term was defined as the preferred semantically matched ICDO term for the input disease name.

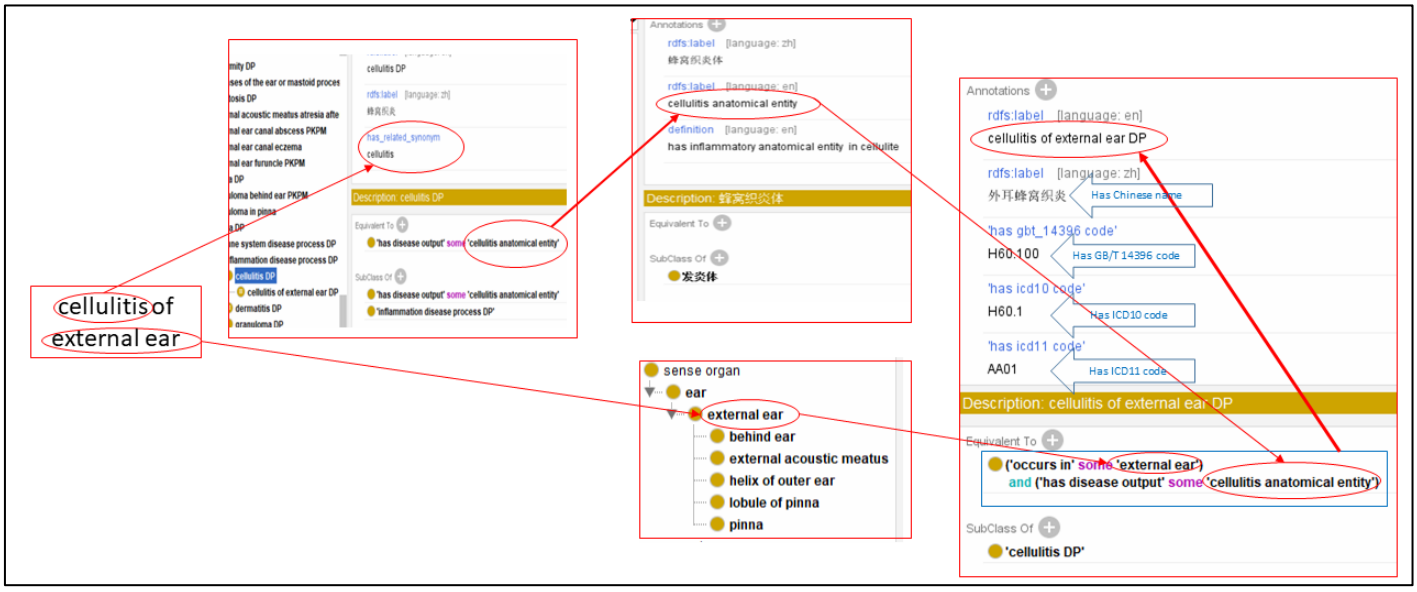

Fig.5. ISDNA inferred terms exact mapping include different codes and languages come from different codeing systems.The candidate term "cellulitis of external ear"will be discomposed into components as "cellulitis DP" and "external ear" after NLP first.Then they are mapped to respect terms in ICDO according to the dimensions disgned by disease pattern as "cellulitis DP" and "external ear".Finally inferred to "cellulitis of external ear DP" by axioms in ICDO .User can selects different ICD code by application requirement.

Note that in the above two examples, the NLP process was preformed manually. In the future, we plan to develop an automatic NLP process to achieve the same NLP results, which is not within the scope of current study.

ICDO query and analysis

In this case, we demonstrate how to use the Description Logic (DL) query in the Protege-OWL editor to identify from ICDO specific diseases that occurs in the external ear canal, or called external acoustic meatus. Basically, this DL query identified those diseases that meet this axiom requirement:

"occurs in" some "external acoustic meatus"

The *"external acoustic meatus"* is the formal anantomical structure term in UBERON and has synonam *"external ear canal"* in ICDO.

As shown in Fig. 7, a total of 14 diseases were identified, including 4 testing terms in Chinese. This example shows that ICDO is able to serve as a platform and knowledge system for computational programs like DL query to perform semantic query and analysis.

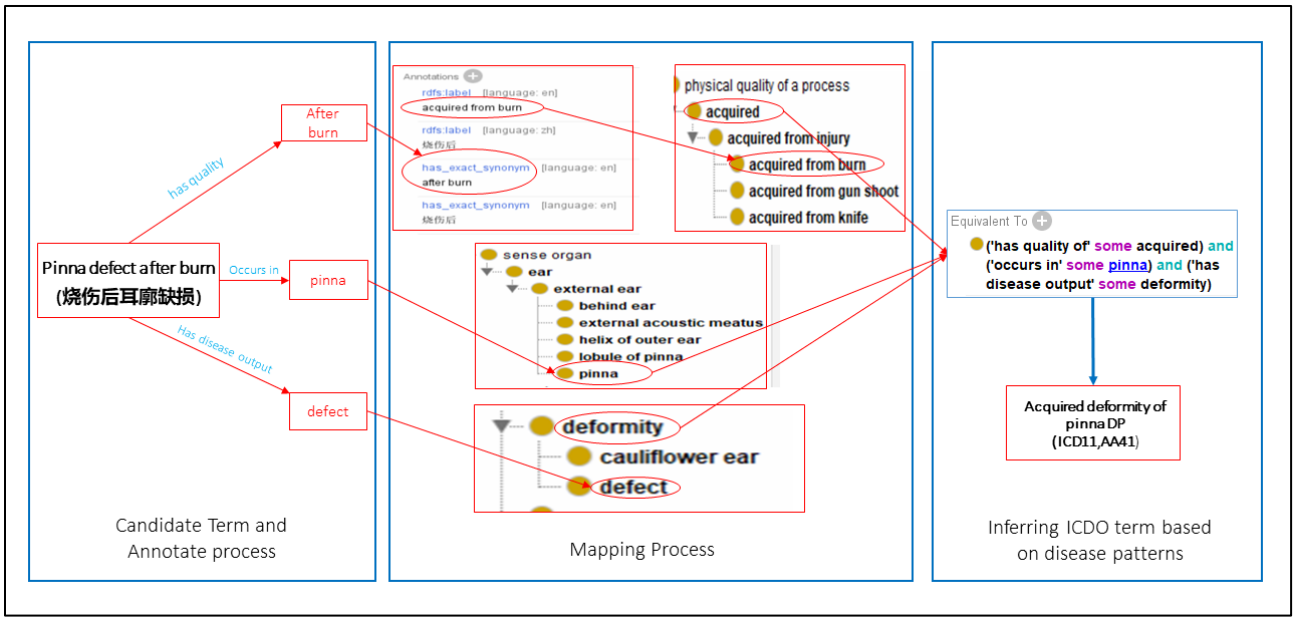

Fig. 6. ISDNA mapping of a candidate term to ICD standard term and code. In this example, the term 'Pinna defect after burn' (Chinese name: "烧伤后耳廓缺损") was used as the input. The term was decomposed into three components "after burn", "pinna", and "defect", where were then mapped to 'acquired from burn', pinna, and defect in ICDO, respectively. These three ICDO terms provide the quality, location of the disease process, and the output of the disease process term. Based on the axiom definition, our ontology reasoner was able to match this name to 'acquired deformity of pinna DP' (AA41 in ICD11). Note that this term is not an exact match.

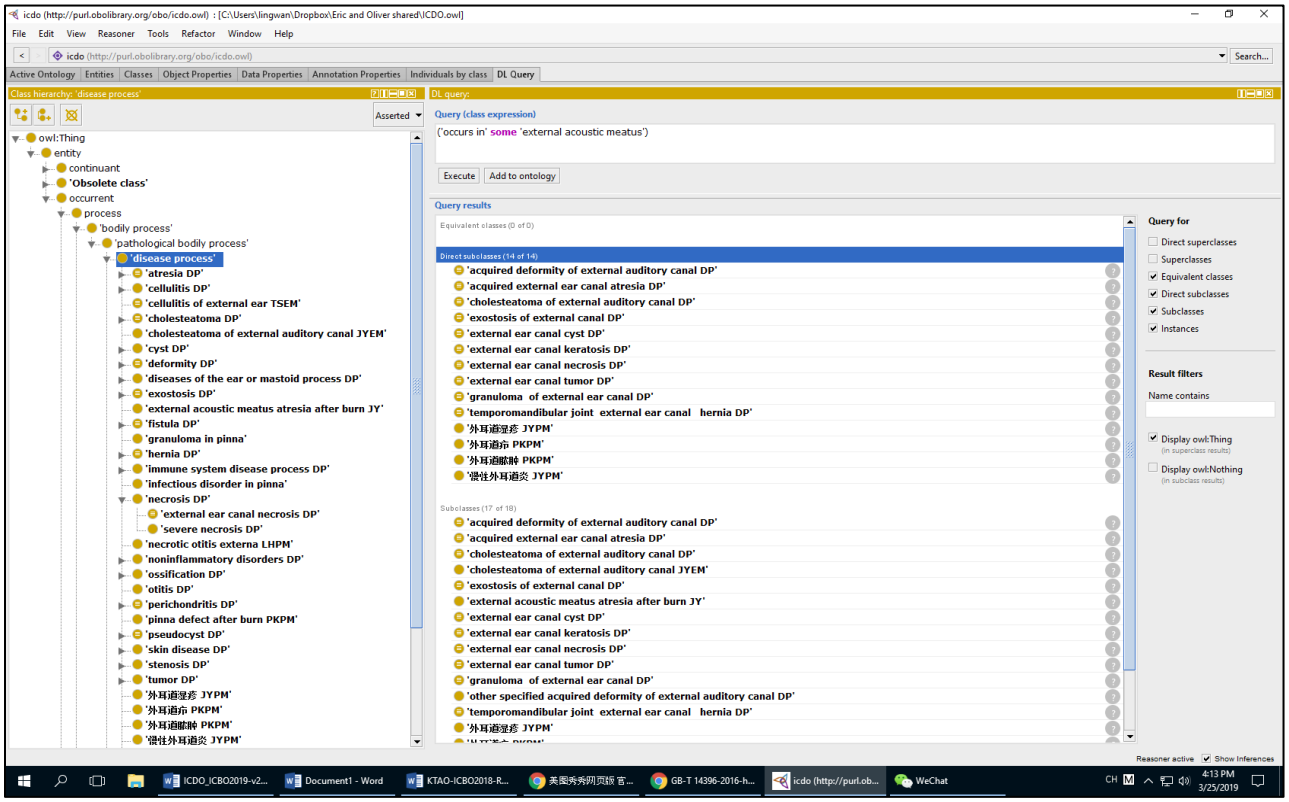

Fig. 7. DL query of ICDO looking for all disease "occurs in external acoustic meatus". This query was performed using DL query in Protege-OWL editor 5.2 (http://protege.stanford.edu/).

Discussion

In this manuscript, we presented our development of the ICDO ontology with the aim to standardize ICD disease records and support health record integration and analysis. We also proposed and tested a semantic analysis based on ICDO using the function of reasoner. It realized the interpretation of terms at the semantic level by reasoner between entities by axioms. ICDO improves the mapping accuracy, supports exact and semantically preferred mapping. and provides a useful application in terms of the standardization of heterogeneous data between different ICD versions.

Our use case focused on the different ICD10 local versions used in some local adminstrative healthcare information plateform in Jiangsu province China. Not every disease has a clear physical product "disease output" when entering clinical observation. For example, inflammation is an immune with multiple symptoms and are sometimes difficult to fully express in natural language. However, the physical entity of this inflammation is clearly present as specific anatomy structure in ontology. In ICDO we defined the output of inflammation process with *"inflammatory anatomical entity"* and asserted axiom in the form of *"physical pathological object" "occurs in" some "anatomical structure" "caused by" "inflammation process"*.

In clinical practice, our disease pattern can cover most disease types. Particularly there is a class of diseases which does not have pathological abnormalities in specific anatomical structures but have systemic symptoms. We designed a special dimension "syndrome" in ICDO disease pattern for this class of diseases. In the current stage of ICDO with the focus of external ear disease, there is no syndrome included. However, this situation will be carefully examined and appropriately handled in the future when we extend the ICDO to fullly cover all disease in ICD10 or ICD 11.

While many ICD terms can be clearly defined as disease processes, there may be many concerns in terms of using disease process for other scenarios. For example, the disease process may not be able to represent a disease output such as the size or mass of a tumor like external ear canal tumor. However, in this case, we can semantically define a disease process like 'external ear canal tumor disease process' that 'has output' of some tumor that 'has size' or 'has mass' of some specific values. Using this strategy, we can semantically link the disease process to the physical tumor (or other anatomic entity) and its qualities like size or mass. Another concern is that a disease process may be diagnosed before the result of the process becomes manifested. Note that ICD is typically used to represent the health outputs rather than unidentified or undiagnosed health issues. The fact that a disease is not diagnosed does not mean that the disease process does not occur. If the disease is not diagnosed, we may not be able to use the disease process term; however, it may not be necessary to use according to the ICD guidelines.

Conclusions

Ontology is clearly a very good tool for solving the problem of semantic mapping between different ICD versions. ICDO will improve the usability and interoperability among various ICD systems. ICDO can also be used for data standardization and analysis of international multi-center clinical trials between different languages in different counties, data normalization processing before DGRs grouping, data normalization and in hospital internal information systems, and data standardization for regional health information platform. The disease design pattern in ICDO can provide effective contributions to the medical data mining and retrospective researches.

Acknowledgements

We appreciate the discussion and editorial revision by Ms. Meng Liu.

Address for correspondence

LW and YH are co-corresponding authors. Their emails addresses are wanlingeric@qq.com and yongqunh@med.umich.edu.

References

1. Percy, C., Holten, V.v., Muir, C.S. and Organization, W.H. (1990) International classification of diseases for oncology.
2. Trott, P. (1977) International classification of diseases for oncology. Journal of clinical pathology, 30, 782.
3. Cao, L. and Morley, J.E. (2016) Sarcopenia is recognized as an independent condition by an international classification of disease, tenth revision, clinical modification (ICD-10-CM) code. Journal of the American Medical Directors Association, 17, 675-677.
4. Dilling, H. and Freyberger, H.J. (2012) Taschenführer zur ICD-10-Klassifikation psychischer Störungen. Bern (Huber).
5. Haux, R. (2006) Health information systems–past, present, future. International journal of medical informatics, 75, 268-281.
6. Vermeer, H.J., Thomassen, E. and de Jonge, N. (2005) Automated processing of serum indices used for interference detection by the laboratory information system. Clinical chemistry, 51, 244-247.
7. Aiello, F.A. and Roddy, S.P. (2017) Inpatient coding and the diagnosis-related group. Journal of vascular surgery, 66, 1621-1623.
8. Arp, R., Smith, B. and Spear, A.D. (2015) Building ontologies with basic formal ontology. Mit Press.
9. Hadzic, M. and Chang, E. (2005), Proceedings of the 38th Annual Hawaii International Conference on System Sciences. IEEE, pp. 143a-143a.
10. Spackman, K.A., Campbell, K.E. and Côté, R.A. (1997), Proceedings of the AMIA annual fall symposium. American Medical Informatics Association, pp. 640.
11. Stearns, M.Q., Price, C., Spackman, K.A. and Wang, A.Y. (2001), Proceedings of the AMIA Symposium. American Medical Informatics Association, pp. 662.
12. lin, F.A.H.W.J.W.F. (2009) Method Research of Constructing Clinical Disease Domain Ontology (In Chinese). Journal of intellgence, Vol.28.
13. Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J. and Vasant, D. (2014) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research, 43, D1071-D1078.
14. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G. and Kibbe, W.A. (2011) Disease

Ontology: a backbone for disease semantic integration. Nucleic acids research, 40, D940-D946.
15. Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N. and Engelstad, M. et al.(2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic acids research, 45, D712-D722.
16. Ceusters, W. and Smith, B. (2015), MIE, Vol. 210, pp. 155-159.
17. He, Y., Xiang, Z., Zheng, J., Lin, Y., Overton, J.A. and Ong, E. (2018) The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. Journal of biomedical semantics, 9, 3.
18. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A. and He, Y. (2010) OntoFox: web-based support for ontology reuse. BMC research notes, 3, 175.
19. Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A. and He, Y. (2017) Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic acids research, 45, D347-d352.
20. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A. and Parkinson, H. (2010) Modeling sample variables with an Experimental Factor Ontology. Bioinformatics, 26, 1112-1118.
21. Malone, J., Rayner, T.F., Zheng Bradley, X. and Parkinson, H. (2008), Proceedings of the Eleventh Annual Bioontologies Meeting. Toronto, Canada.
22. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N. et al. (2014) The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of Biomedical Semantics, 5, 14.
23. Mattingly, C.J., McKone, T.E., Callahan, M.A., Blake, J.A. and Hubal, E.A.C. (2012). ACS Publications.
24. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E. and Haendel, M.A. (2012) Uberon, an integrative multi-species anatomy ontology. Genome biology, 13, R5.