MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220028

Cross-Language Terminology Mapping Between ICD-10-CN and SNOMED-CT

Na Hong[a,#], Yaoyun Zhang[b,#], Yuankai Ren[c], Li Hou[d], Changran Wang[f], Jing Li[e], Mui Van Zandt[e,*], Lei Liu[f,*]

[a] *Digital Health China Technologies Co. Ltd., Beijing, China*
[b] *Melax Technologies, Inc, Houston, TX, USA*
[c] *Jiangnan University, Suzhou, China*
[d] *Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China*
[e] *Real World Solutions, IQVIA, Durham, NC, USA*
[f] *Fudan University, Shanghai, China*
[#] *Co-first authors*
* *Corresponding authors*

Abstract

The objective of this study was to develop a hybrid method and perform an initial evaluation of mappings from the International Statistical Classification of Diseases, 10th revision, Chinese version (ICD-10-CN) to the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT). The methods used to perform mapping include reusing existing mappings, term similarity modeling for automatic mapping and manual review. We evaluated the results of automatic mapping and the coverage of the maps between two terminologies. Experimental results demonstrated that fine-tuning the pre-trained biomedical language model of PubmedBERT obtained the optimal performance, with a precision of 0.859, a recall of 0.773, and a F1 of 0.814. 100% 4-digit code ICD-10-CN terms were mapped to SNOMED-CT terms through exsit code mappings. Around 42.41% randomly selected 6-digit code ICD-10-CN terms had exact matches to corresponding SNOMED-CT terms, and we did not find appropriate SNOMED-CT terms for ICD grouping terms.

Keywords:

Terminology; Health Information Interoperability; Vocabulary, Controlled

Introduction

The OHDSI standard vocabularies are foundational terminology and ontology resources enable consistent coding and query of clinical data across disparate observational databases. There are more than 100 vocabularies, represent different domains of clinical data and multiple languages is supported, work together to harmonize clinical content behind OHDSI information model. The ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification) and SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms) are two mostly commonly used vocabularies for diagnoses and clinical finding terms standardization. ICD-10-CN is the Chinese version of ICD-10-CM[1], is adopted by OHDSI vocabulary system.

The ICD-10-CM is developed by the National Center for Health Statistics (NCHS), under authorization by the World Health Organization. ICD-10-CM is used by physicians and other healthcare providers to classify and code all diagnoses, symptoms and procedures recorded in conjunction with hospital care[2]. It provides a level of detail that is necessary for diagnostic specificity and morbidity classification. It is an administrative coding systems. In 2016, China released a Chinese version ICD-10-CM, as a national standard GB/T 14396-2016, it was purposed to standardize the disease related medical terms in China, ensure the consistency of data, and lay a foundation for promoting the development of future healthcare statistics and application. To keep use of consistent name, we adopted the brief name "ICD-10-CN" of this terminology in OHDSI vocabulary system.

SNOMED-CT is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED-CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world, developed by an international consortium[4; 9].

As ICD-10-CN is a recommended national standard of China, more and more Chinese diagnosis terms from EHR will be standardized using ICD-10-CN; However, ICD-10-CN and SNOMED-CT are not fully aligned in OHDSI vocabulary system. Therefore, implementing mappings from ICD-10-CN to SNOMED-CT is essential. Cross lingual terminology mapping is the process of finding correspondences between terminologies of different languages to allow them to interoperation, and the created mappings will facilitate future efficient standard implementation and research reuse.

However, cross lingual and cross terminology mapping work led to major semantic changes between coding systems due to the different design purpose, hierarchy, granularity and language expression differences[6].

A traditional method of cross-terminology linking is to use the semantic path within existing terminology hierarchies curated in knowledge bases and lexicon based semantic similarity between terms[10]. For example, Fun et al. proposed to use UMLS for inter terminology mapping between ICD9CM and SNOMED-CT. However, due to the term sparsity of such methods and the essential structural and semantic difference between ICD9CM and SNOMED-CT, a modest performance was obtained (recall: 43%, precision: 27%).

Recently, with the fast development of deep learning based methods, state-of-the-art performances have been produced in various biomedical NLP (natural language processing) tasks such as named entity recognition, relation extraction and concept normalization[8]. In particular, significant performance improvements have been achieved by using the attention-based mechanisms[12] and BERT (Bidirectional Encoder Representations from Transformers)[7] based language models pre-trained on large scale biomedical datasets such as biomedical literature and clinical text. Such methods serve as an efficient complement to knowledge based methods. A study was conducted on linking terminologies between Hebrew[5] and English using attention-based deep learning methods and knowledge base information, following the methods proposed for Chinese entity linking[13]. However, both of them mainly focused on entities in online health communities. A pilot study linked Chinese medication terms to RxNorm based on lexicon matching, and manual review of semantically similar terms[14]. Different from other terminologies, ICD-10-CN has its unique characteristics, including both 4-digit coded terms similar to ICD-10-CM and an additional set of 6-digit codes continuously extending based on commonly used diagnoses terms in Chinese clinical settings. So far, few efforts have been made on linking terms in ICD-10-CN to SNOMED-CT.

In order to promote diagnosis terms interoperation, support OHDSI China research community attending cross-countries large-scale clinical studies, this study took the initiative effort to map concepts from and ICD-10-CN to SNOMED-CT. However, the transition from ICD-10-CN to SNOMED-CT posed a significant challenge due to the structure and semantic differences behind these two terminologies. Finding correspondence SNOMED-CT code is a challenging process because similar concepts are present at different nodes within a hierarchy or in different hierarchies, leading to possible inconsistent mappings. To address this, we developed a hybrid method that provides candidate SNOMED-CT term recommendations for each ICD-10-CN terms to facilitate mapping implementation effectively.

Methods

Data set

The two vocabularies are available from the OHDSI vocabulary system Athena[11]. There are in total of 34491 Chinese terms in ICD-10-CN and 1035027 English terms in SNOMED-CT.

In 34491 Chinese terms of ICD-10-CN, there are 22928 4-digit coded Chinese terms, 12511 6-digit extension coded Chinese terms that added according to commonly used diagnoses terms based on Chinese clinical usage. In addition, 1615 terms are used for grouping all the ICD terms. As displayed in Table 1.

Table 1– Distribution of different ICD term types

ICD-10-CN	Number of terms	Examples
4-digit code	22928	D73.2 慢性充血性脾大 Chronic congestive splenomegaly
6-digit code	12511	G00.903 耳源性脑膜炎 Otogenic meningitis
Grouping code	1615	J60-J70 外部物质引起的肺部疾病 Lung disease caused by external substances

Method Framework

In this study, with the purpose of implementing as complete mappings as possible from ICD-10-CN to SNOMED-CT, we used different method for different term types. The methods used to perform mapping implementation include reuse exist mapping, machine translation, term similarity modeling for automatically mapping and manual review.

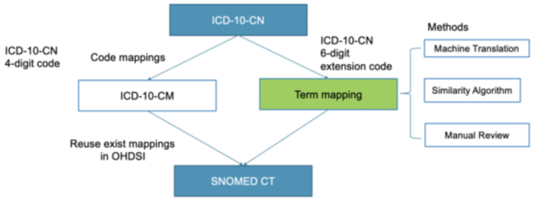

Figure 1– Overview of the ICD-10-CN to SNOMED-CT mapping framework

- Reuse existing mappings for 4-digit code

In terms of previous published mapping results, approximately 20,000 SNOMED-CT concepts have been mapped to ICD-10-CM[3]. For the Chinese ICD term/code with Standard WHO ICD-10-CM code(*4-digit code term*), we reused these mapping relations.

For example,慢性充血性脾大 (ICD-10-CN: D73.2) /Chronic congestive splenomegaly (ICD-10-CM: D73.2), the matched SNOMED term is Chronic congestive splenomegaly (SCTID: 191382009).

- Bridging concept hierarchies for narrow-to-broad mapping of 6-digit code

Since ICD-10-CN is a classification system which structured in tree hierarchy, each 6-digit code ICD term belong to an upper 4-digit code ICD term. For example, 耳源性脑膜炎(ICD-10-CN: G00.903) could not be directly mapped to SNOMED-CT, however, it belongs to the class of 细菌性脑膜炎(ICD-10-CN: G00.9) with corresponding English version: Bacterial meningitis, unspecified(ICD-10-CM:G00.9), which has exist matched SNOMED term Bacterial meningitis (SCTID: 95883001). Therefore, The term 耳源性脑膜炎(ICD-10-CN: G00.903) could be narrow-to-broad mapped to SNOMED term Bacterial meningitis (SCTID: 95883001).

- Term similarity model for exact mapping of 6-digit code

To build an automatic model of concept mapping between ICD-10-CN and SNOMED-CT, a gold-standard corpus was first manually curated. Then following a typical workflow of automatic concept mapping, a small set of candidate terms (in SNOMED-CT) semantically most similar to the query term (in ICD-10-CN) was first generated (top 10). After that, a ranking based classification model was generated to determine the final mapping terms among the candidates. Details of each step are as follows:

(1) Manual annotation: For the 6-digit extension code in ICD-10-CN, 1172 concepts were randomly selected from the entire 12511 concepts and were manually checked. Since few Chinese terms exist in UMLS, for the convenience of manual annotation and to facilitate later automatic model, the 1172 concept terms were also translated into English using the online portal of

google translation. The annotation process went through several rounds to come up with consistent mapping standards. Two annotators worked independently and the final cohen-kappa score of inter-annotator agreement was 0.84. Discrepancies were resolved together with the third annotator. In total, 497 (42.41%) of them had semantic-equivalent mapping concepts in SNOMED-CT. Examples of the 6-digit concepts and their mapping results are listed in Table 2.

Table 2– Examples of 6-digit concepts in ICD-10-CN and mappings in SNOMED-CT

ICD-10-CN	English translation	SNOMED-CT
滤泡性膀胱炎	cystitis follicular	follicular cystitis
颈部皮肤良性肿瘤	Benign tumor of neck skin	benign neoplasm of skin of neck
腘窝恶性肿瘤	popliteal fossa tumor	#
输尿管支架移位	ureteral stent migration	#

(2) Candidate generation: To generate candidate concepts from SNOMED-CT, a search engine with indexed SNOMED-CT terms of the problem semantic group collected from UMLS was built. By querying the search engine, the semantic similarity would be calculated between the ICD-10-CN terms translated into English and each SNOMED-CT term using the BM25 information retrieval model. The top 10 most semantically similar SNOMED-CT terms were be returned.

(3) Candidate re-ranking: The query term q in ICD-10-CN and retrieved candidate terms T={ t_i, i in [1,10] } form pairs (q, t_i) , for which a ranking based binary-classifier will be built to determine if they are semantically equivalent.

Usually, when mapping entities in clinical text or biomedical literature to standard terminologies, only a small percentage of entities (e.g., around 3%-5%) can not be mapped and they can be handled by pre-collected keyword list and semantic type filtering. However, more than half of concepts in 6-digit ICD-10-CN code can not match to SNOMED. Therefore, two aims need to be targeted in candidate re-ranking:

Let S_{qti} stands for the score output for (q, t_i) from the classifier, a threshold S_{th} is needed to first determine if q has any mapping with T: If $S_{qti} =< S_{th}$ for all t_i in T, then q has no mapping; Else for those t_i in T with $S_{qti} > S_{th}$, concept of t_i with the max(S_{qti}) will be labeled as the mapping for q.

Several pre-trained language models based on BERT (Ref. *Evaluation method*) were fined tuned using our manual annotated gold standard pairs to build the classifier. q and t_i were considered as two short sentences, their semantic representation vectors (V_q,, V_{ti}) were connected by a sentence separator [seg] as input to the model - V_q [seg] V_{ti}. Cosine similarity score calculated for the transformed input vectors was used as the objective function for term ranking.

Evaluation method

We evaluated the performance of term similarity-based automatic mapping and the average mapping coverage between two terminologies.

As mentioned in the Dataset section, 1172 Chinese concepts of 6-digit extension codes in ICD10-CN were randomly selected, automatically translated in to English and manually examined whether they could exactly match any concept in SNOMED-CT.

Ten-fold cross validation was used to evaluate the performance of automatic mapping. Precision, Recall and F1 were reported as defined in Equation (1-3). Specifically, precision represents the percentage of correct mappings in the total number of mappings automatically predicted by the system; recall represents the percentage of correct mappings in the total number of gold standard annotations; F1 is the harmonic mean of precision and recall.

$$precision = \frac{\#correct\ mapping}{\#\ predicted\ mapping} \quad (1)$$

$$recall = \frac{\#correct\ mapping}{\#gold\ standard\ mapping} \quad (2)$$

$$F1 = \frac{2*precision*\ recall}{precision+recall} \quad (3)$$

The following methods have been implemented and compared for automatic mapping:

UMLS_Synonym: Since each concept in UMLS has a list of synonyms, a baseline method was implemented: if q and t_i were an exact match or both were synonyms of the same concept, the concept of the term t_i was labeled as the mapping for q.

Fine-tuning variations of BERT based language models: including BERT based on open text, BioBERT fined tuned on biomedical literature, BioClinicBERT fined tuned on both biomedical literature and clinical text, and PubmedBERT built from scratch on biomedical literature using the BERT framework.

In addition, the performances of top 10 candidate terms returned in the candidate generation step were also compared as a reference baseline.

Results

A total of 22928 4-digit coded ICD-10-CN concepts have been mapped to SNOMED-CT by reusing existing released mappings between ICD-10-CN and ICD-10-CM and between SNOMED-CT and ICD-10-CM [3]. These maps have been updated in the OHDSI terminology portal of Athena.

Performances of multiple models for automatic mapping between terms of ICD-10-CN 6-digit code and terms in SNOMED-CT were shown in Table 2. Top 10 candidate terms returned by BM25 obtained the highest recall and a precision (P@10) of 0.424. On the other hand, exact match to UMLS synonyms obtained the highest precision and a much lower recall of 0.614. Among the four pre-trained language models, fine-tuning three of them obtained lower F1 than the UMLS_synonym baseline. However, fine-tuning PubmedBERT achieved the highest F1 score of 0.814 among all the implemented methods. Therefore, a pipeline was formed by using the UMLS_synonym method to produce accurate mappings first, then applying the PubmedBERT model to map the rest terms, followed by manual review.

Table 2– Performance of automatic mapping

Method	P	R	F1
Top10_BM25	0.424	***1.000***	0.596
UMLS_synonym	***1.000***	0.614	0.761
BERT	0.803	0.672	0.732
BioBERT	0.835	0.694	0.758
BioClinicBERT	0.845	0.656	0.738
PubmedBERT	0.859	0.773	***0.814***

Overall, 12511 concepts of 6-digit coded Chinese extension have been mapped to corresponding SNOMED-CT terms by bridging through their upper 4-digit code concepts within the tree hierarchy of ICD-10-CN (Narrow-Broad mapped, 100%). Exact mapping percentage based on manual review is 41.42%.

The other 1615 codes that represent names of groups of terms in a tree structure, do not have appropriate SNOMED-CT concepts to map. As shown in Table 3.

Table 3–Mappings results with different term types

ICD-10-CN	Number of concepts	Mapping type/Percentage
4-digit code	22928	Exact mapping (100%)
6-digit code	12511	Narrow-Broad mapping (100%) Exact mapping (~41.42%)
Grouping code	1615	No mapping (0%)

Discussion

We proposed a hybrid method of cross lingual and cross terminology mapping, and presented the initial mapping results of ICD-10-CN and SNOMED-CT. We will also analyze their similarities, differences and challenges in mapping implementation, and outline future directions to improve interoperability between them.

In order to understand why the performance of several pre-trained language models is lower than the UMLS_synonym baseline, and why PubmedBert outperforms other methods, we calculated the percentage of words in the gold standard mapping annotations covered by the vocabulary of each language model. The coverage is shown in Figure 2. For words not covered in the vocabularies, their representations were initialized uniformly with the same vector without any differentiation. Both BioBERT and BioClinicBERT were fine-tuned based on BERT using the same vocabulary. Despite that they have been fine-tuned, their vocabulary coverage was lower and only modest improvements have been obtained compared to using BERT. The performance of BioClinicBERT was worse than that of BioBERT, which may be due to the use of ICD-10-CN standard terminology for mapping, instead of concept mentions in clinical text. PubMedBERT was trained from scratch using biomedical literature, its vocabulary coverage was much higher, and had the best performance. Further increase in vocabulary coverage may lead to performance improvements in the future.

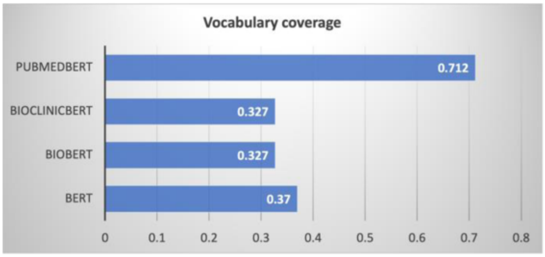

Figure 2 – Vocabulary coverage of each language model

Another challenge faced by the semantic similarity score-based model is that it is difficult to use a single threshold to distinguish between partially matched terms and semantically equivalent terms. Domain knowledge based heuristics need to be summarized and integrated with deep learning methods as a guide for result optimization.

Limitations and future work. Due to the differences in granularity, language expression and organization principles between ICD-10-CN and SNOMED-CT, it is impossible to always have a one-to-one correspondence between ICD-10-CN codes and SNOMED-CT codes. To address this challenge, our next direction of work is to increase the post-coordination support for ICD-10 Chinese extension codes that usually contain a combination of expressions and modifiers.

This research is an ongoing work. With the improvement of model capabilities and manual curation, the mapping coverage will continue to increase. After the next step of strong verification, the mappings will be published through the OHDSI vocabulary system.

Conclusions

This study introduces the preliminary mapping results from ICD-10-CN to SNOMED-CT, in order to bridge the gap between Chinese and English diagnostic concepts, and further support Chinese researchers to conduct efficient and reproducible observational research globally. The results of the study will be made public through the OHDSI vocabulary system, enabling researchers and multiple healthcare partners to play a role in the future.

Acknowledgements

Thanks to the OHDSI community and OHDSI China Working Group to support this study.

Address for correspondence

Corresponding authors:
Lei Liu, Fudan University, Liulei@fudan.edu.cn;
Mui Van Zandt, Real World Solutions, IQVIA, mvanzandt@us.imshealth.com.

References

[1] The release of GB/T 14396-2016 ICD-10, in, 2016.

[2] International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), in, CDC, 2021.

[3] SNOMED-CT to ICD-10-CM Map, (2021).

[4] T. Benson, *Principles of health interoperability HL7 and SNOMED*, Springer London, 2010.

[5] Y. Bitton, R. Cohen, T. Schifter, E. Bachmat, M. Elhadad, and N. Elhadad, Cross-lingual Unified Medical Language System entity linking in online health communities, *Journal of the American Medical Informatics Association* **27** (2020), 1585-1592.

[6] S.E. Bowman, Coordination of SNOMED-CT and ICD-10: getting the most out of electronic health record systems, *Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems/AHIMA, American Health Information Management Association* (2005).

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[8] U. Hahn and M. Oleynik, Medical Information Extraction in the Age of Deep Learning, *Yearb Med Inform* **29** (2020), 208-220.

[9] D. Lee, N. de Keizer, F. Lau, and R. Cornet, Literature review of SNOMED-CT use, *Journal of the American Medical Informatics Association* **21** (2014), e11-e19.

[10] B.T. McInnes, T. Pedersen, and S.V. Pakhomov, UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity, *AMIA Annu Symp Proc* **2009** (2009), 431-435.

[11] OHDSI, OHDSI Athena, (2021).

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, *arXiv preprint arXiv:1706.03762* (2017).

[13] J. Xu, L. Gan, M. Cheng, and Q. Wu, Unsupervised medical entity recognition and linking in Chinese online medical text, *Journal of healthcare engineering* **2018** (2018).

[14] Y. Zhang, J. Li, and M. Van Zandt, NCCD-RxNorm: Linking Chinese Clinical Drugs to International Drug Vocabulary, in: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2020, pp. 1752-1756.