MedDRA-ICD Current Mapping Evaluation based on UMLS and OHDSI

Xinyuan Zhang[a], Yixue Feng[b], Fang Li[a], Cui Tao[a]

[a] School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA,
[b] Department of Computer Science, University of Virginia, Charlottesville, VA, USA

Abstract

With the increasing quantity and quality of biomedical data, there are numerious valuable information embeded in large amount of data sources. Combining the signals from various data warehouses can provide powerful foundation for secondary-data anlytics studies. The standardized database systems may have different data structures and rely on different terminologies. Hence, data integration of different databases becomes more and more important. In this paper, we provided a quantitative and qualitative analysis of the mapping situation based on the mapping tools from Unified Medical Language System (UMLS) and Observational Health Data Sciences and Informatics (OHDSI). We also further evaluated the non-mapping status to find the capacity of improvement for Meddra-ICD mapping.

Keywords:

International Classification of Diseases, Evaluation Studies, Terminology

Introduction

In this era of big data, vulable information is buried in numerious data warehouses. With the increasing quantity of data produced in the biomedical field, data integration and interoperation is becoming more and more important for clinical practice and research. However, most of the standardized database systems have different data structures and rely on different terminologies. In order to facilitate integrative analysis of data coming from a variety of sources, mapping the implemented terminologies is necessary, so that different data sources can communicate on the same level. For example, FDA's Adverse Event Reporting System (FAERS), a voluntary adverse events reporting system detects only a small portion of patients with limited data quality [1]. However, combining the symptom of patients from FAERS with other longitudinal observational databases can achieve completeness of information extraction. When using electronic health record (EHR) and electronic medical record (EMR) data for pharmacovigilance, combining the signal from EHR and EMR with FAERS is desirable.

International Classification of Diseases (ICD) records clinical information used by healthcare providers. Due to its use in medical billing, ICD is frequently used by healthcare providers and has been incorporated into many EHRs and EMRs as a way to capture diagnoses. The ICD coding systems are used for external reporting requirements, and can report diseases, injuries, and other health problems. However, ICDs are inadequate for studying adverse reactions due to their designed hierarchies and reporting rules. In this case, mapping ICD codes to a standardized terminology for patient safety research purposes becomes necessary.

Medical Dictionary for Regulatory Activities (MedDRA) is a terminology that records regulatory information pertaining to medicinal products. It is also a recommended terminology for adverse event reporting in several data sources, such as FAERS, Canada Vigilance database and EudraVigilance database. The benefits of using MedDRA as a reference terminology for secondary uses of EHR and EMR data are substantial.

U.S. National Library of Medicine (NLM) released and distributed official mapping through applications like Unified Medical Language System (UMLS) [2]. The Observational Health Data Sciences and Informatics (OHDSI) Standard Vocabulary [3] maps all the data to common data standards including EHR and MedDRA. Both OHDSI and UMLS Metathesaurus provide mapping structures for multiple terminologies. While the UMLS is more of a concept-based system, all the concepts are given a unified identification number. Both systems can help integrate vocabulories from different resources with different purposes. MedDRA is aimed at supporting adverse reaction analysis, while ICD codes are built for insurance reimbursement processes. They are collected for a different reason, resulting in different structures and physical formats. Using the mapping results from UMLS and OHDSI can accommodate both FAERS and EHR/EMR, allowing users to generate evidence from a wide variety of sources.

Julio et al. [4] gave a summary of several mapping results between MedDRA, ICD and NCI Thesaurus (NCIT), however, the direct mapping between MedDRA and ICD wasn't taken into account. Hence, in this paper, we provided a quantitative and qualitative analysis of the mapping situation based on both resources. During our research, we found that the mappings provided by UMLS and OHDSI are under development. Hence, in this paper, we also evalutated the non-mapping status.

Methods

The first step is to investigate the current statistics on MedDRA and ICD mapping in existing terminology services. - Unified Medical Language System (UMLS) and Observational Health Data Sciences and Informatics (OHDSI). The 2017AB UMLS Metathesaurus release is used in this analysis.

MedDRA

MedDRA has five-level hierarchical structure. The top level are System Organ Classes (SOCs), representing 26 broad classes grouped by etiology, manifestation site and purpose [5]. The middle two levels are High Level Group Terms (HLGTs) and

High Level Terms (HLTs), provide clinically relevant grouping of terms. Below them are Preferred Terms (PTs), which are distinct and unambiguous descriptors. Clinical pathologic or etiologic qualifiers are represented in this level, e.g., PT Rhinitis perennial, PT Rhinitis seasonal. The lowest level are the Lowest Level Terms (LLTs), with the maximum specificity. LLTs are the synonym, sub-element, or identical LLT to their parent PTs.

ICD

ICD-10-CM is the United States' clinical modification of the ICD-10. It has been expanded to include health-related conditions and to provide greater specificity at the sixth- and seventh-character level [6]. The terms contains rich detailed information, including new-added laterality. As its the main component, the Tabular List divides codes into 21 chapters based on body system or condition. Fig 1 shows the composition of 7- character code of ICD-10-CM.

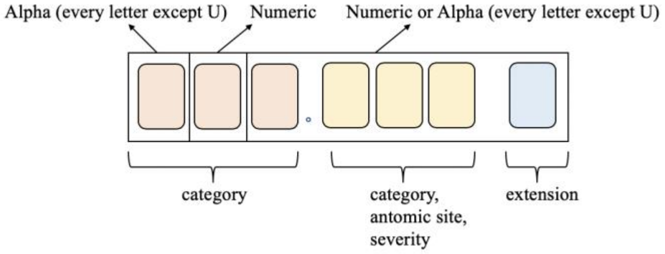

Alpha (every letter except U)
Numeric
Numeric or Alpha (every letter except U)
category
category,
antomic site,
severity
extension

Figure 1 The composition of ICD-10-CM code

The rules of ICD-10-CM coding are: the first character is always alpha, the second character is always numeric, the remaining 5 digits may be any combination of alpha/numeric, and all codes require a decimal after the third character. The codes and categories within **S52** could serve as specific examples:

S52 Fracture of forearm
S52.5 Fracture of lower end of radius
S52.52 Torus fracture of lower end of radius
S52.521 Torus fracture of lower end of *right* radius
S52.521A Torus fracture of lower end of *right* radius, *initial encounter*, closed fracture

In the above example, S52 is the category. The fourth and fifth characters of "5" and "2" provide additional clinical detail and anatomic site. The sixth character "1" indicates laterality, i.e., right radius. The seventh character, "A", is an extension which, in this example, means "initial encounter".

UMLS

Each concept in the the UMLS Metathesaurus is associated with a unique identifier, with different specificity. Concept Unique Identifier (CUI) is of the highest level - it groups terms together by semantic meaning. So terms with the same meaning share the same CUI.

The terms are first filtered by vocabulary source to retrieve MedDRA, ICD9CM and ICD10CM terms. Then corresponding MedDRA and ICD terms with common CUI are extracted as mapped terms.

OHDSI

OHDSI Common Data Model (CDM) classifies medical vocabularies from different sources into one common format as well as a common representation (terminologies, vocabularies, coding schemes).

The Standard Vocabularies from OHDSI store all terminologies in the CONCEPT table. Semantic relationships between terms are defined in the CONCEPT_RELATIONSHIP table. In this mapping extraction step, both direct mapping and indirect mapping were tried. In the indirect mapping method, we used SNOMED as a intermediary data source. MedDRA is first mapped to SNOMED, an intermediary data source, then to ICD using "MedDRA - SNOMED eq" and "Maps to" concept relationships. The direct mapping was used for defining "MedDRA – ICD" relationships.

Mapping

Since both ICD 9 and ICD 10 are included in EMR systems, PT and LLT are both mentioned in the mapping relationship. We created multiple mapping pairs: MedDRA PT - ICD9CM, MedDRA PT - ICD10CM, MedDRA LLT - ICD9CM and MedDRA LLT - ICD10CM. Since the PTs level from MedDRA is the most frequently used terms to describe adverse reactions for drug safety, the unique number count after converting LLT terms to PT level was also calculated.

Evaluation

The relationship types of semantic scope between two non-synonymous concepts (e.g. A and B) could be classified into four different forms which are shown in Figure 2: A is partially overlapping with B (Situation I), A is broader than B (Situation II), A is narrower than B (Situation III), and A is irrelevant with B (Situation IV). To further explore the non-mapped MedDRA terms, 100 PT terms from each SOC category were randomly chosen. The "similar terms" function from the UMLS Metathesaurus was used as a standard to compare the non-mapping terms from MedDRA and similar terms from ICD codes. One experienced clinician annotated the reasons why these terms were not mapped into seven categories accordingly: "Exact Match", "PT term broader than ICD term", "PT term narrower than ICD term", "Partial Overlap", "Totally Irrelevant", "No Response" or "Other reasons", where "No Response" means there was no returned results from similar term search.

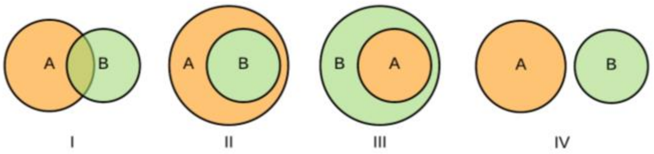


Figure 2 Relationships of semantic scope between two non-synonymous concepts

Results

Mapping Statistics

Mappings created were MedDRA PT - ICD9CM, MedDRA PT - ICD10CM, MedDRA LLT - ICD9CM and MedDRA LLT - ICD10CM. After removing duplicates and combining results, there were 4496 MedDRA PT terms and 18327 MedDRA LLT terms that had a mapping to at least one ICD term in UMLS, and 4240 unique MedDRA PT terms and 187 LLT terms in OHDSI.

Combining all UMLS and OHDSI mappings, there were a total of 19574 unique terms. Cross checking with the March 2018 MedDRA release, these included 5679 PT and 19572 LLT terms (In MedDRA, all PT terms are also present at the LLT level. There were 2 terms from UMLS that were not present in the latest MedDRA release). Using the official MedDRA LLT-PT mapping files, all LLT terms were converted into PT terms. Analysis showed that a total of 6305 unique PT terms were mapped in either UMLS or OHDSI, covering 27.31% of all MedDRA PT terms (See Table 1 for details).

We also calculated and graphed the trend of MedDRA - ICD mappings in UMLS over the years (Figure 3)., With the increase of terms being added each year into the UMLS Metathesaurus, the percentage of terms being mapped is slowly decreasing.

System Organ Class (SOC) Distribution

Each MedDRA PT term can be mapped to one of the 26 System Organ Class (SOC) categories. We also placed the mapped and unmapped PT terms under 26 SOC levels (Figure 4). “Pregnancy, puerperium and perinatal conditions”, “Ear and labyrinth disorders” and “Congenital, familial and genetic disorders” had mapping percentages above 50%, highest among all 26 SOCs. “General disorders and administrative site conditions”, “Investigations and Product issues” had the lowest mapping coverage, below 10%.

The results of non-mapping reasons evaluation are summerized in Table 2, 8 SOC was included in the evaluation: “Blood and lymphatic system disorders”, “Cardiac disorders”, “Congenital, familial and genetic disorders”, “Ear and labyrinth disorders”, “Endocrine disorders”, “Eye disorders”, “Gastrointestinal disorders”, and “General disorders and administration site conditions”.

Table 1– Mapping Summary from UMLS and OHDSI.

	MedDRA PT		MedDRA LLT	
	UMLS	OHDSI	UMLS	OHDSI
ICD9CM	2777	3478	13366	151
ICD10CM	3722	3611	13568	160
Combined	4496	4249	18327	187
Total Combined Unique Terms	19574			
	5679		19572	
Convert LLT to PT	6305			

Table 2– Evaluation Summary of Non-Mapping terms.

Non-Mapping Reason	**Number Count**
PT term narrower than ICD term	236
PT term broader than ICD term	48
Partial Overlap	161
Totally Irrelevant	185
Other reasons	37
Exact match	24
No response	185
Total	876

Discussion

The number of terms being added each year into the UMLS Metathesaurus has been increasing since 2010. Before the 2016 release, there were no LLT mappings in UMLS. With this increase however, the percentage of terms being mapped has been slowly decreasing since 2010, which shows the need for creating more mapping relations between terms.

For the MedDRA PT term summarized under 26 SOC categories, we can see that “Investigations” is especially notable, for it has the most number of PT terms of all SOCs, and only 4.82% were mapped in either UMLS and OHDSI. Investigations describes concepts related to medical conditions, procedures and qualitative results, (e.g. Insulin C-peptide decreased, Total lung capacity normal). Given that some Investigations terms have similar linguistic structure, future mapping steps can be specifically tailored to identify the procedure and corresponding results in each term.

For the evaluation of unmapped terms, there were 24 “exact match” out of 876 PT terms. For instance, our evaluater annotated “Hyperthyroidism due to ectopic thyroid nodule” to be the exact match of “Ectopic hyperthyroidism” from PT term. It suggested that there are still blank area for the official mapping technique. The different granularities of two coding systems could explain the non-mapping causes for “PT term broader than ICD term”, and “PT term narrower than ICD term”. Given an example of “Thyroid Diseases” from ICD, and “Haemorrhagic thyroid cyst” from PT terms. However, since we only have one evaluater, the accuracy remains to be tested. We plan to include at least two annotators to calcualte their agreement score in our future study.

Conclusion

In our study, we first evaluated the mapping situation between MedDRA and ICD using official mapping tools provided by UMLS and OHDSI. The overall percenrage of PT terms mapped in either UMLS or OHDSI is 27.31% of all MedDRA PT terms. We further evaluated the non-mapping terms and found out there are still terms that could be exact match pairs, suggesting the expansion capacity for MedDRA to ICD mapping.

Figure 3: UMLS PT term mapping over the year 2009 – 2018.

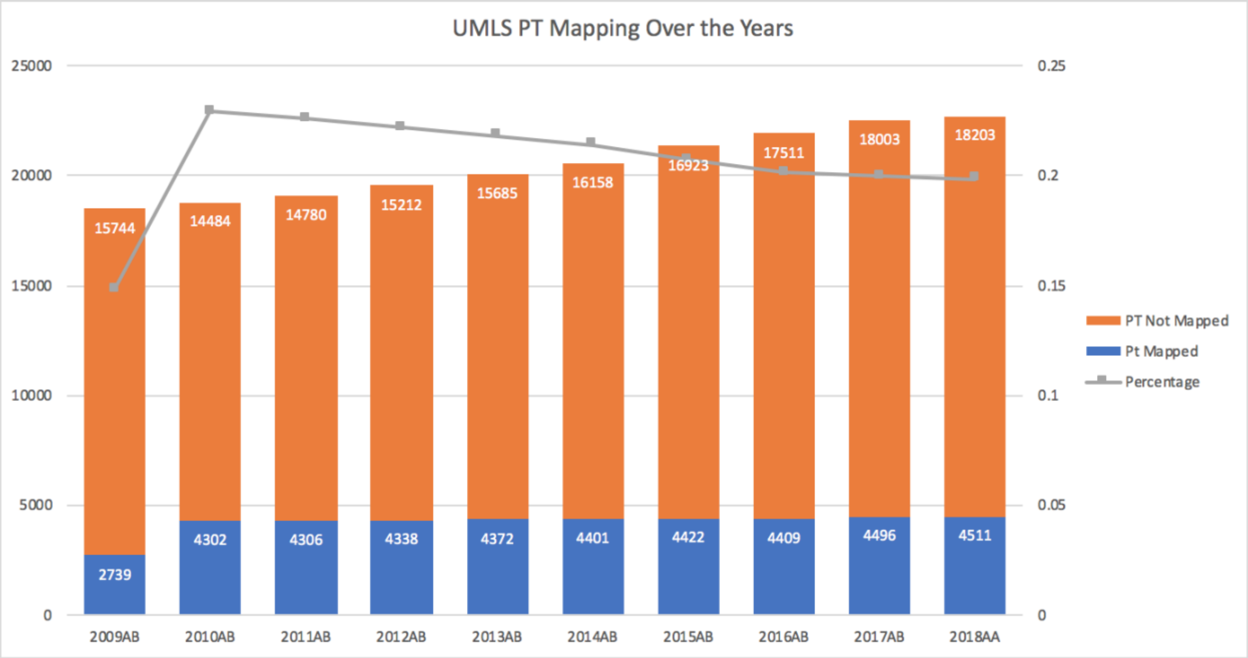
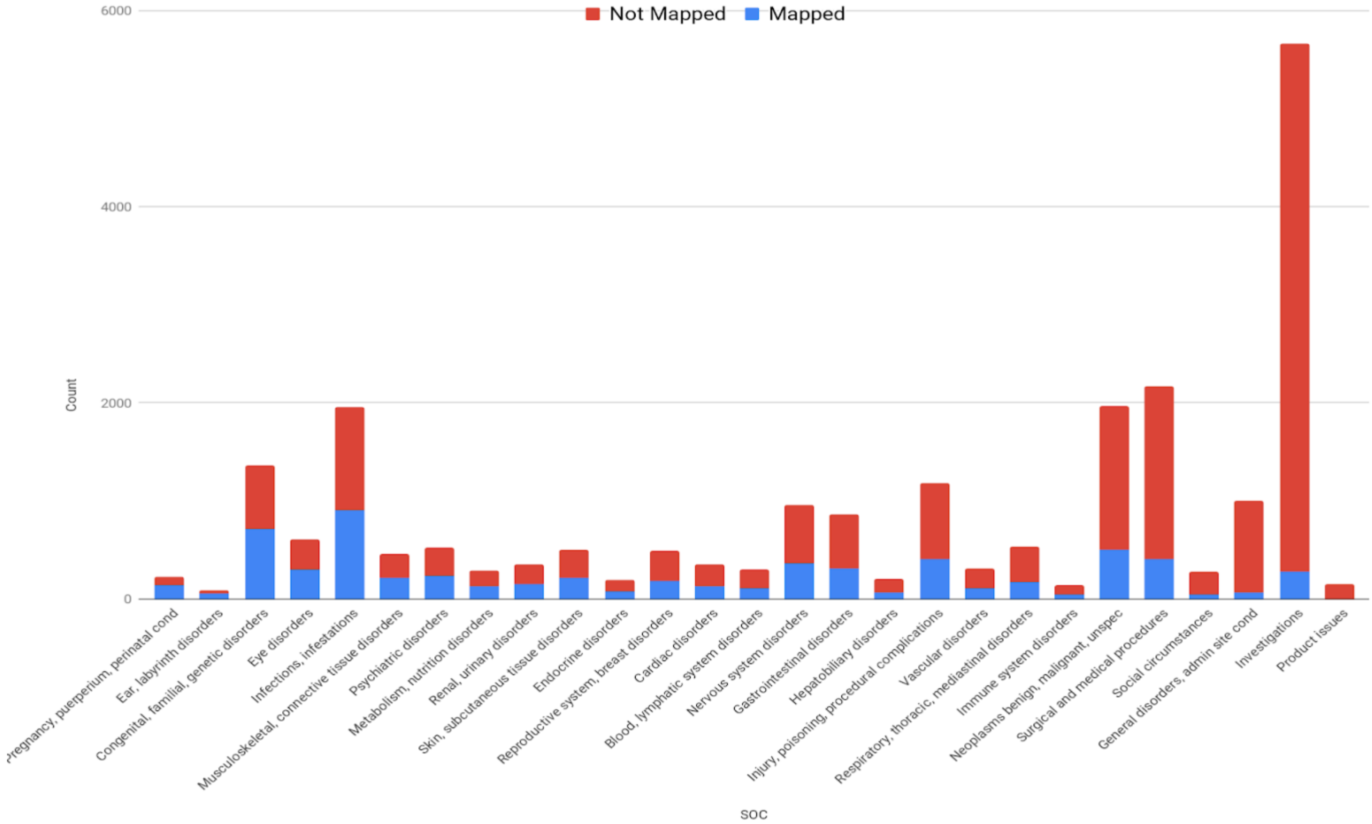


Figure 4: Mapping Distribution among SOC level.



Acknowledgements

This research was partially supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011829, 1R01LM012607, R01AI116794, 7R01LM009012.

References

[1] Center for Drug Evaluation and Research. "Questions and Answers on FDA's Adverse Event Reporting System (FAERS)." U S Food and Drug Administration Home Page. June 04, 2018. Accessed July 30, 2018. https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/.

[2] Unified Medical Language System (UMLS) n.d. https://www.nlm.nih.gov/research/umls (accessed 4 Sep 2018).

[3] OHDSI – Observational Health Data Sciences and Informatics n.d. https://www.ohdsi.org/ (accessed 3 Sep 2018).

[4] Reis JC Dos, Pruski C, Da Silveira M, et al. The DyKOSmap approach for analyzing and supporting the mapping maintenance problem in biomedical knowledge organization systems. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2015;7540:163–75. doi:10.1007/978-3-662-46641-4_12

[5] MedDRA Maintenance and Support Services Organization. Introductory Guide to MedDRA Version 14.0. March, 2011.http://www.who.int/medical_devices/innovation/MedDRAintroguide_version14_0_March2011.pdf (accessed 13 Nov 2018).

[6] The National Center for Health Statistics (NCHS). ICD-10-CM Official Guidelines for Coding and Reporting FY 2019. https://www.cdc.gov/nchs/icd/data/10cmguidelines-FY2019-final.pdf (accessed 14 Nov 2018).

Address for correspondence

Corresponding author: Cui Tao.
Email address: cui.tao@uth.tmc.edu