Quality issues on mappings between ICD10 and SNOMED CT

Emiliano Reynares[1,*,†], Andrea Splendiani[2,†]

[1]IQVIA, Provença 392, Barcelona, Spain
[2]IQVIA, Kirschgartenstrasse 14, Basel, Switzerland

Abstract

Research and clinical data are generated for a variety of needs, and as such make use of various medical vocabularies. Enabling complex research and analytics use cases requires semantic interoperability between data sources, which implies the definition and maintenance of vocabulary mappings. Both the vocabularies and the mappings are dynamic and involve a context, yet in many practical usages such a dynamic and context are not considered. Although evidence has been presented on possible quality issues on commonly used vocabulary mappings, it is unclear what the extent of such issues may be. As an initial assessment we compared the mappings relating an ICD10 term to a single SNOMED CT concept, as defined by OHDSI Standardized Vocabularies and SNOMED CT International. Our analysis found that 27.5% of the mappings do not match due to differences on the level of abstraction of the mappings (47% of the mismatches), slightly variations on the semantics of the terms (10% of the mismatches), evolution of the vocabularies (4% of the mismatches), and plain errors on the release of mappings (2% of the mismatches). Identification of the causes for the remaining mismatches (37%) will be tackled in future works. The lack of proper attention to mapping dynamics results in a lower quality of the resulting datasets, in ways that are very difficult to detect once a dataset has been generated. With this work, we made a step in quantifying the potential impact of such data quality issues, so that proper actions can be taken.

Keywords

vocabulary mappings, mappings quality, medical vocabularies, SNOMED, ICD10, OMOP

1. Introduction

Medical practice and biomedical research make use of a variety of vocabularies[1] to standardize terms and codes used to describe medical concepts, such as diseases, procedures, medications, and anatomical structures. These vocabularies ensure consistent and precise communication among healthcare professionals and systems, facilitating accurate documentation, data exchange, and analysis. Vocabularies are typically maintained by public and government agencies. For example, the World Health Organization (WHO) produces the International Classification of Diseases (ICD) [1] and SNOMED International develops the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) [2].

Different research and clinical data workflows generate data that is normalized using different vocabularies, for a set of historical and operational reasons, as emphasis may vary from clinical encounter documentation, to research, billing, safety and more. This variety prevents both interoperability of healthcare systems and the integration of data, such as delivering integrated patient histories for secondary use.

Enabling complex research and analytics use cases requires semantic interoperability between data sources, which implies the definition and maintenance of vocabulary mappings. While both the vocabularies and the corresponding mappings are dynamic by nature, and mappings also have a contextual aspect[2], such a dynamic and context are not usually considered in the practical usages [3, 4].

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
†These authors contributed equally.
emiliano.reynares@iqvia.com (E. Reynares); andrea.splendiani@iqvia.com (A. Splendiani)
0000-0002-5109-3716 (E. Reynares); 0000-0002-3201-9617 (A. Splendiani)

[1]While we use the word "vocabularies", we intended here the broader collection of vocabularies, terminologies, taxonomies, coding schemes, ontologies

[2]Mappings originate from a reason, and with assumptions, that results in different approximations.

Some evidence on possible quality issues on commonly used vocabulary mappings has been presented (cfr. related work section), but as of today there has not been an evaluation of the potential extent of these issues. [5, 6, 7, 8]. This work aims at providing a first quantitative evaluation of what the impact may be of neglecting mapping dynamics. While the impact ultimately depends on the use of data, we focus on the recurrence of such issues, e.g.: how prevalent is the problem respect to an overall mapping set.

2. Related work

The literature presents a variety of studies on the dynamic nature of vocabularies and mappings, and the challenges this poses for their development and management.

In [4], researchers compiled a collection of terms from the PubMed database to analyze the temporal dynamics of medical vocabularies. They observed temporal patterns in vocabulary development and usage that could be linked to major research breakthroughs, growing areas of interest, policy changes, or shifts in public health priorities.

[3] described the complexities of mapping medication information across different vocabularies due to the lack of up-to-date information, the difficulty in using rapidly evolving vocabularies, the differences in granularity between source and target vocabularies, and the need for continuous updates as vocabularies evolve.

Similarly, [5] highlighted the challenges of ensuring that the vocabularies used are updated and reflect the latest medical knowledge, their continuous evolution, the differences in their level of detail, the ambiguities and inconsistencies in mappings, and the lack of information about how, when, and why mappings were created.

[6] also examined the use of cross-references - i.e., used to link terms in one ontology to terms in another - in 30 ontologies from the Open Biological and Biomedical Ontologies (OBO) Foundry, finding that 10.7% of the one million cross-references analyzed are ontology mappings but the usage is semantically ambiguous and difficult to reuse.

[7] evaluated the mappings between the Medical Dictionary for Regulatory Activities (MedDRA) and ICD through the Unified Medical Language System (UMLS) and the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [9]. The study found that less than 30% of MedDRA preferred terms are mapped to ICD while less than 3% of the sampled ones could have exact matches in ICD, concluding that differences in granularity and focus between the vocabularies are the causes of this.

[8] also identified the challenges posed by semantic drift, i.e., changes in the meaning of terms across different versions of vocabularies, such as the disappearance of terms, lack of machine-readable traceability details, evolving meaning of terms that leads to semantic inconsistencies across versions, and the complexity of managing and integrating multiple versions of terminologies.

Despite this evidence on quality issues in commonly used vocabularies and their corresponding mappings, the extent of the issues for the vocabularies under consideration in this work remains unclear.

3. Approach for mappings analysis

We compared the mappings between the 10th revision of ICD and the SNOMED CT International Edition as defined by two reference sources.

ICD10 is a vocabulary for coding diseases, symptoms, abnormal findings, and external causes of injury or diseases. It is globally used for health management, epidemiology, and clinical purposes, helping in the systematic recording, analysis, interpretation, and comparison ofmortality and morbidity data across different regions and time periods.

SNOMED CT is a multilingual clinical healthcare terminology. It provides a standardized way to represent clinical phrases captured by healthcare providers, providing a consistent and accurate representation of clinical content and supporting better data analysis.

Mappings between ICD10 and SNOMED CT are arguably some of the most established ones because they serve complementary purposes in healthcare data management. ICD10 is primarily used for coding and reporting diseases and health conditions, while SNOMED CT is used for detailed clinical documentation. SNOMED is also used as target ontology to normalize different coding to, when multiple sources of data are aggregated for secondary use. These mappings facilitate the translation of detailed clinical information (captured in SNOMED CT) into standardized codes for reporting and billing (ICD10), or conversely the use of claims data in RWE studies. Overall, these mappings support various healthcare processes, including clinical documentation, billing, and statistical analysis.

In this evaluation, we retrieve IDC10 to SNOMED (and vice versa) mappings from two sources: the OHDSI Standardized Vocabularies and the SNOMED CT standard release package.

The OHDSI Standardized Vocabularies [10] centralizes the vocabularies, and their corresponding mappings, as used in the OMOP CDM and underpins a large amount of RWD generation. We retrieved the OHDSI Standardized Vocabularies version 2024-08-30 as used by the version 5.4 of CDM from the Athena OHDSI Vocabularies Repository. It maps the 2021 release of ICD10 with the 2024-02-01 release of SNOMED CT International Edition.

SNOMED CT provides SNOMED to ICD10 mappings as part of its standard releases. These mappings are validated by both the WHO and SNOMED International, and they are also used in the development of extensions (e.g.: SNOMED to ICD10 extensions), created and maintained by member countries [11]. We retrieved such mappings from the 2024-06-01 release of SNOMED CT International Edition (mappings relate to version 2016 of ICD10) from the official source.

The OHDSI Standardized Vocabularies define mappings as relationships allowing a crosswalk between equivalent concepts. Equivalent concepts mean they carry the same meaning, and their descendants cover the same semantic space. The OHDSI Standardized Vocabularies maps concepts to a broader target when the equivalent one is not available. For example, ICD10CM W61.51 'Bitten by goose' has no equivalent in the SNOMED vocabulary, so it is mapped to SNOMED 217716004 'Peck by bird' losing the context of the bird being a goose [12].

On the other hand, SNOMED CT to ICD10 map is a directed set of associations from SNOMED CT source concepts to ICD10 target classification terms. The SNOMED CT source domains for the map are limited to subtypes of 'clinical finding', 'event' and 'situation with explicit context'. SNOMED CT also defines a set of 8 mapping categories: (i) 'Map of source concept is context dependent', (ii) 'Map source concept cannot be classified with available data', (iii) 'Map source concept is outside of the scope of target classification', (iv) 'Map source concept is properly classified', (v) 'Mapping guidance from World Health Organization is ambiguous', (vi) 'Source concept has been retired from map scope', (vii) 'Source SNOMED concept is ambiguous', (viii) 'Source SNOMED concept is incompletely modeled' [11, 13].

Since semantics of mappings varies across sources, we restricted our analysis to mappings relating an ICD10 term to a single SNOMED CT concept. Only active mappings from OHDSI were considered for the analysis. From SNOMED CT, only active mappings where the map source concept was properly classified were considered for the analysis, i.e.: 'map category' equals to the 447637006 SNOMED CT concept. This approach is conservative, in that considering 1:1 mappings arguably over-represents simple cases (at the same time we would expect discrepancies between two versions be less relevant than discrepancies across versions, for vocabularies that have been in service for decades). It also provides a common ground in terms of intended operational semantics in both contexts, as both mappings could be easily used to translate data coded in ICD10 to SNOMED.

4. Results

The OHDSI Standardized Vocabularies define 18,579 unique ICD10 to SNOMED CT mappings, while SNOMED CT defines 123,106 unique mappings to ICD10. But as mentioned previously, our approach focuses on analyzing the common subset of ICD10 codes that are mapped to a single SNOMED CT code by both systems. Consequently, we identified that 9,568 ICD10 codes are mapped to a single SNOMED CT code by the OHDSI Standardized Vocabularies, and 1,941 are mapped from a single SNOMED CT

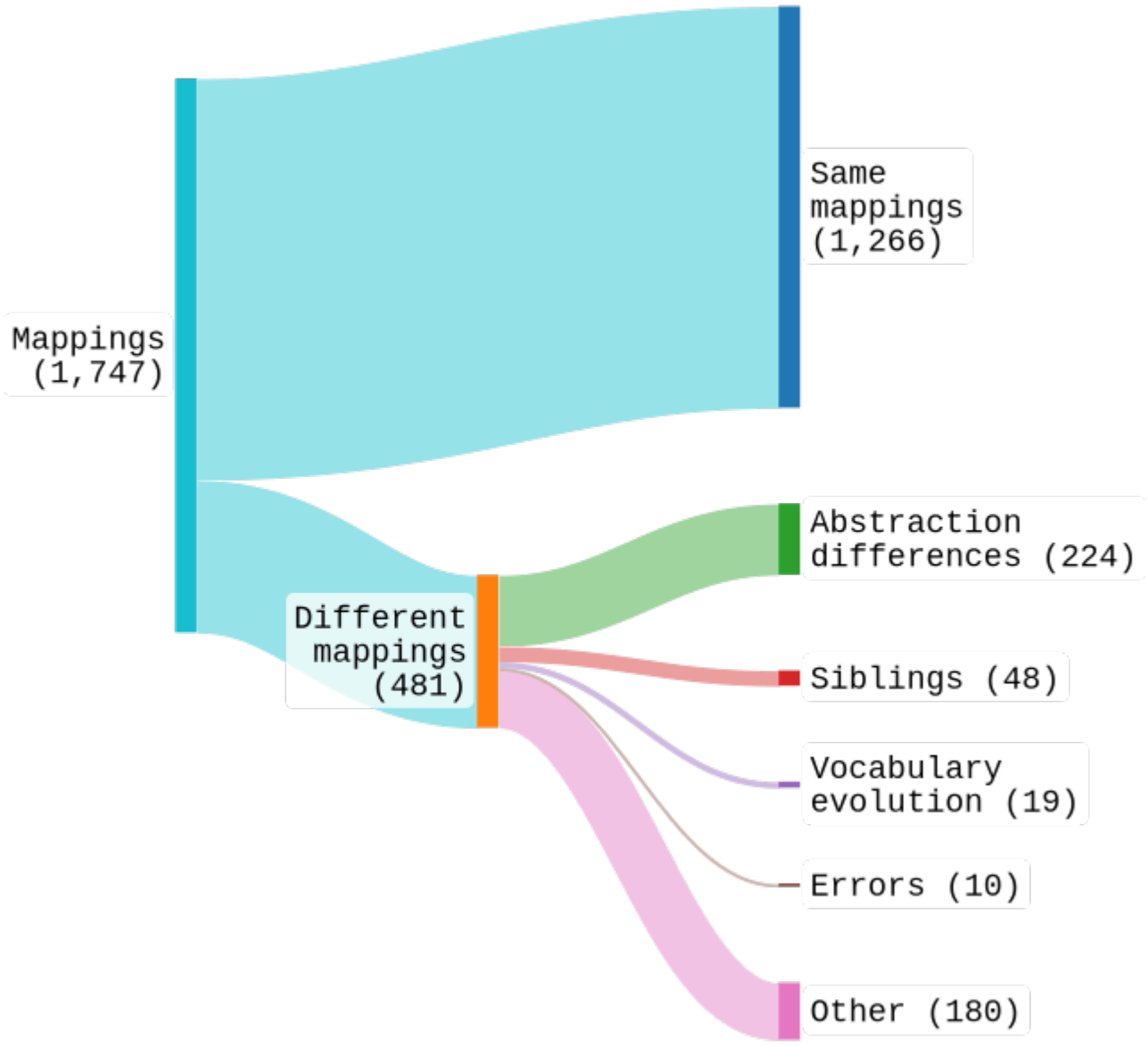

Figure 1: Mappings by category, showing number of instances on each category.

code by SNOMED CT. There is an overlap of 1,747 ICD10 codes that are mapped to a single SNOMED CT code by both systems.

Of them, 1,266 ICD10 codes are mapped to the same SNOMED CT code by both systems, i.e.: the mappings match, while the remaining 481 ICD10 are mapped to different SNOMED CT codes: i.e.: 27.5% of mappings mismatch. Our analysis found that 27.5% of the mappings do not match due to differences on the level of abstraction of the mappings (47% of the mapping mismatches), slightly variations on the semantics of the terms, i.e.: mapping targets are siblings (10% of the mapping mismatches), evolution of the vocabularies (4% of the mapping mismatches), and plain errors on the release of mappings (2% of the mapping mismatches). Identification of the causes for the remaining mapping mismatches (37%) will be tackled in future works.

Figure 1 describes the mappings mismatches. The rest of the section details and provides examples on each of the identified mismatch categories.

147 of the mapping mismatches (147/481*100=30.6%) are due to OHDSI mapping targets which are direct ancestors of the SNOMED counterpart. For example, A02.8 ('Other specified salmonella infections') is mapped to 302231008 ('Salmonella infection (disorder)') and 763772002 ('Invasive non-typhoidal salmonellosis (disorder)') by OHDSI and SNOMED, respectively.

71 of the mapping mismatches (71/481*100=14.8%) are due to OHDSI mapping targets which are transitive ancestors of the SNOMED counterpart. For example, B97.2 ('Coronavirus as the cause of diseases classified to other chapters') is mapped to 27619001 ('Disease caused by Coronaviridae (disorder)') and 713084008 ('Pneumonia caused by Human coronavirus (disorder)') by OHDSI and SNOMED, respectively.

5 of the mapping mismatches (5/481*100=1%) are due to SNOMED mapping targets which are direct ancestors of the OHDSI counterpart. For example, M19.19 ('Post-traumatic arthrosis of other joints, site unspecified') is mapped to 699262001 ('Post traumatic osteoarthritis (disorder)') and 840396000 ('Posttraumatic arthropathy (disorder)') by OHDSI and SNOMED, respectively.

1 of the mapping mismatches (1/481*100=0.2%) are due to SNOMED mapping targets which are transitive ancestors of the OHDSI counterpart. M40.39 ('Flatback syndrome, Site unspecified') is mapped to 203665007 ('Flatback syndrome (disorder)') and 249711007 ('Flattened lordosis (finding)') by OHDSI and SNOMED, respectively.

48 of the mapping mismatches (48/481*100=10%) are due to OHDSI and SNOMED mapping targets with a direct common ancestor, i.e.: the mapping targets are siblings. For example, B02.0 ('Zoster encephalitis') is mapped to 230176008 ('Herpes zoster encephalitis (disorder)') and 98541000119101 ('Herpes zoster myelitis (disorder)') by OHDSI and SNOMED, respectively.

13 of the mapping mismatches (13/481*100=2,7%) are due to mapping target codes that do not exist on the SNOMED CT release used by OHDSI. For example, the M54.02 code ('Panniculitis affecting regions of neck and back, cervical region') is mapped by OHDSI to the 202770004 code ('Panniculitis of neck (disorder)'), while it is mapped to the 317151000119105 ('Inflammation of subcutaneous fatty tissue of neck and thorax (disorder)') code by SNOMED.

13 of the mapping mismatches (13/481*100=2,7%) are due to mapping target codes that do not exist on the SNOMED CT release used by OHDSI. For example, the M54.02 code ('Panniculitis affecting regions of neck and back, cervical region') is mapped by OHDSI to the 202770004 code ('Panniculitis of neck (disorder)'), while it is mapped to the 317151000119105 ('Inflammation of subcutaneous fatty tissue of neck and thorax (disorder)') code by SNOMED.

10 of the mapping mismatches (10/481*100=2,1%) are due to what could be considered as plain mapping errors by OHDSI, since the proposed SNOMED target codes are invalid on the SNOMED CT releases used by the two mappings sources we used for mappings comparison. Table 1 shows these mappings mismatches.

5. Conclusions and future works

This work provides a quantitative estimate of the impact of neglecting mapping dynamics and provenance. We focused on mappings between two ontologies that are widely used in healthcare, derived from the two of the most relevant resources at the same time of availability. Arguably lack of attention to provenance and versions implies that both mappings could be used arbitrarily "at some point" in a complex data pipeline (perhaps including multiple actors) and therefore the mismatch between the two is a direct proxy of data quality issues in terms of hidden lack of semantic coherence in the resulting dataset.

We took a rather conservative approach in considering simple 1:1 mappings.

We found that the number of mismatches (30%) is significant, which complements anecdotal evidence of some cases found in previous works. Of these 30%, about a third can be reconducted to different levels of abstraction that are reasonable, in that the two sources of mappings address different level of abstraction (clinical vs population studies). A surprising 4% of mismatches is due to terms obsoletion, that is a surprising number given the two sources of mappings are synchronous in terms of release,

Table 1
Invalid mapping targets by OHDSI.

ICD10 code	Map target by OHDSI
H54.1-Severe visual impairment, binocular	813871000000108
H54.5- Severe visual impairment, monocular	813881000000105
M12.80-Other specific arthropathies, not elsewhere classified, multiple sites	36941000119103
M19.00-Primary arthrosis of other joints, multiple sites	36941000119103
P13.1-Drainage of female perineum	944051000000105
P94.8-Other disorders of muscle tone of newborn	935461000000101
U12.9-COVID-19 vaccines causing adverse effects in therapeutic use, unspecified	1324661000000105
U84.7-Resistance to multiple antimicrobial drugs	1034561000000105
Y98-Lifestyle-related condition	987891000000105
Z62.0-Inadequate parental supervision and control	288431000119102

while they present a discrepancy in terms of ontology versions of only a few months. The number of errors (10) is also surprising, given the very conservative focus on simple, widely used mappings.

These findings suggest that the lack of provenance and versioning of mappings can have a significant impact (even in the order of 20%), on the quality of downstream data in integration processes. It is hence important that more discipline around mappings be in place. In future works, we intend to inspect possible root causes of such mismatches, as well as extend the analysis to other mappings.

Acknowledgments

The study is funded by IQVIA. We extend our gratitude to all team members who have contributed to the review of this article.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] WHO, international statistical classification of diseases and related health problems (ICD)., n.d. URL: https://www.who.int/standards/classifications/classification-of-diseases.

[2] SNOMED international, systematized nomenclature of medicine - clinical terms (SNOMED CT), n.d. URL: https://www.snomed.org/what-is-snomed-ct.

[3] H. Saitwal, D. Qing, S. Jones, E. V. Bernstam, C. G. Chute, T. R. Johnson, Cross-terminology mapping challenges: a demonstration using medication terminological systems, J. Biomed. Inform. 45 (2012) 613–625.

[4] A. Khakimova, X. Yang, O. Zolotarev, M. Berberova, M. Charnine, Tracking knowledge evolution based on the terminology dynamics in 4p-medicine, Int. J. Environ. Res. Public Health 17 (2020) 7444.

[5] H. Sajjad, S. Hong, S. Anil, E. G. B. Laleci, M. Charles, Gray Alasdair J.G., McGuinness Deborah L., P. Eric, D. Christel, F. Kerstin, A framework for evaluating and utilizing medical terminology mappings, in: Studies in Health Technology and Informatics, Studies in Health Technology and Informatics, IOS Press, 2014.

[6] A. Laadhar, E. Abrahão, C. Jonquet, Investigating one million XRefs in thirthy ontologies from the OBO world, in: Proceedings of 11th International Conference on Biomedical Ontologies, ICBO 2020, 2020, pp. 1–12.

[7] X. Zhang, Y. Feng, F. Li, J. Ding, D. Tahseen, E. Hinojosa, Y. Chen, C. Tao, Evaluating MedDRA-to-ICD terminology mappings, BMC Med. Inform. Decis. Mak. 23 (2024) 299.
[8] D. Unni, V. Touré, P. Krauss, K. Crameri, S. Österle, SPHN strategy to unravel the semantic drift between versions of standard terminologies, in: 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS), 2024, pp. 11–22. URL: http://ceur-ws.org/Vol-3890/#paper-2.
[9] OHDSI, observational medical outcomes partnership common data model (OMOP CDM), n.d. URL: https://ohdsi.github.io/CommonDataModel/.
[10] OHDSI, standardized vocabularies., n.d. URL: https://github.com/OHDSI/Vocabulary-v5.0/wiki.
[11] SNOMED international, SNOMED CT to ICD-10 map, n.d. URL: https://www.snomed.org/maps.
[12] OHDSI, the book of OHDSI, n.d. URL: https://ohdsi.github.io/TheBookOfOhdsi/.
[13] SNOMED international, complex and extended map from SNOMED CT reference sets, n.d. URL: https://confluence.ihtsdotools.org/display/DOCRELFMT/5.2.3.3+Complex+and+Extended+Map+from+SNOMED+CT+Reference+Sets.