ICD10–SNOMED mapping pitfalls: Post-coordinated expressions and concept sets

Sigfried Gold, Tanner Zhang, Richard L. Zhu, Stephanie Hong, Harold P. Lehmann, Davera Gabriel, Tricia Francis, Lisa Eskenazi, Christopher G. Chute

Background

The National COVID Cohort Collaborative (N3C)[1], an open science community focused on analyzing patient-level data from many centers, organized in response to the pandemic in 2020, uses OMOP as its common data model (CDM), introducing it to hundreds of new researchers and analysts. The volume and diversity of research using N3C's data enclave has necessitated the production of hundreds of concept sets, collections of vocabulary codes used in cohort definitions and study protocols. One attempt to meet this need involved the automated import and conversion of concept sets from external sources such as the Healthcare Cost and Utilization Project and the Value Set Authority Center.[2] Though we have come to believe that automated conversion of concept sets cannot succeed without extensive clinical review, our attempt exposed issues in using OMOP vocabulary mappings that affect many researchers and concept set developers.

A core feature of the OMOP CDM and OHDSI approach to replicable research over an internationally distributed research network is the mapping of coded clinical concepts to standard vocabulary terms.[3 Ch. 5] Electronic health record data in the U.S. frequently represent clinical conditions using the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD10-CM). The OMOP `concept_relationship` table maps these to designated standard concepts from the Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT). When there is no single SNOMED term exactly matching a given ICD term, it may be mapped to a less granular SNOMED concept ("up-hill mappings"), or it may be mapped to more than one SNOMED term. These one-to-many ICD-SNOMED mappings do not imply the same relationship between the source and target concepts as one-to-one mappings do. As *The Book of OHDSI* explains[3 Sec. 5.3.1]

> "Some mappings connect a source concept to more than one Standard Concept. For example, ICD-9-CM 070.43 "Hepatitis E with hepatic coma" is mapped to both SNOMED 235867002 "Acute hepatitis E" as well as SNOMED 72836002 "Hepatic Coma." The reason for this is that the original source concept is a pre-coordinated combination of two conditions, hepatitis and coma. SNOMED does not have that combination, which results in two records written for the ICD9CM record, one with each mapped Standard Concept."

The text describes "Hepatitis E with hepatic coma" as the *pre-coordinated* combination of two conditions. These two concepts combined do not exist as a single coded concept in SNOMED CT but can be represented by a *post-coordinated expression* that contains both of them. So, a source record in `condition_occurrence` for this example would have `condition_source_concept_id` set to the `concept_id` for 070.43, but the `concept_id`s for the two SNOMED codes cannot both be put in the `condition_concept_id` column, so the ETL process will generate two records from the one source record. In order to identify occurrences of this condition, an analyst will either need to use the `condition_source_concept_id`, or use a query finding *two co-occurring records*, one with each of the target SNOMED concepts, which cannot be done using a single concept set.

Though fully automated conversion of concept sets from non-standard to standard vocabularies is not

advisable, we have frequently observed efforts to generate OMOP concept sets starting from sets of ICD concepts, for instance, when replicating published studies that report ICD codes used. Our goal in this report is to characterize the issues involving the use of one-ICD-to-many-SNOMED mappings and to explore the practical impact these have on OMOP studies.

Method

For our study, we analyzed the `concept_relationship` table, finding 'Maps to' relationships between ICD10-CM concepts and SNOMED CT concepts. Then we created a mapping table of each ICD10-CM concept and the list of all the SNOMED "standard" condition concepts it maps to.

The table was then joined with our local patient dataset to examine the validity of the mapping.

1. Join the mapping table described above with the local `condition_occurrence` table.
2. Among the multiple mapping records, pull the one-ICD-to-two-SNOMED mapped records.
3. Compare differences in cohort size when treating SNOMED concept pairs as synonyms (counting records with either code) as opposed to post-coordinated expressions (counting the co-occurrence of records with each of the codes.)

The findings were then discussed by a group of physicians and terminology experts.

Results

Among the existing 90,518 'Maps to' relationships in `concept_relationship` table (2022 Sep 10), 67,377 (74.4%) of them are one-to-one mappings and 23.0% of them are one-to-two mappings.

ICD10-CM concepts	map to	OMOP Standard SNOMED condition concepts
67,377		1
20,870		2
1,651		3
260		4

The range of differences in cohort sizes (method step 3 above) ranged widely, especially as we explored a variety of selection strategies regarding the co-occurence of source and target codes in the original records. For the three examples below, the distinct patient counts for records with either SNOMED code were in the neighborhood of four times greater than the count for co-occurring records of each code.

ICD10-CM concept and code	**SNOMED CT concepts and `concept_ids`**
Type 2 diabetes mellitus with ketoacidosis without coma (E11.10)	- Diabetic ketoacidosis without coma (201826) - Type 2 diabetes mellitus (4009303)
Adolescent idiopathic scoliosis, thoracolumbar region (M41.125)	- Adolescent idiopathic scoliosis (4067872) - Idiopathic scoliosis of thoracic and lumbar spine (37017436)
Candidiasis of skin and nail (B37.2)	- Candidiasis (433968) - Disorder of integument (4028387)

Conclusion

Problems with mapping single ICD codes to multiple SNOMED codes are likely to arise when attempting to convert concept sets from ICD to SNOMED or otherwise attempting to study conditions that must be represented by post-coordinated concept expressions. Further work is needed on three fronts:

1. Performing a more comprehensive analysis of the impact of the problem in actual practice.
2. Developing better educational materials to help avert mistakes made when researchers do not account for this issue.
3. Developing mechanisms in the OMOP vocabulary system and the OHDSI tool stack to allow for post-coordinated concept expressions.

References/Citations

1. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021 Mar 1;28(3):427–43.
2. Khatipov E, Madden M, Chiang P, Chuang P, Nguyen DM, D'Souza I, et al. Creating, Maintaining and Publishing Value Sets in the VSAC. In: AMIA. 2014.
3. OHDSI. The Book of OHDSI [Internet]. 2020th-04–16th ed. Observational Health Data Sciences and Informatics; 2020 [cited 2020 Jun 17]. 470 p. Available from: http://book.ohdsi.org