Evaluating the impact of different vocabulary versions on cohort definitions and CDM

Dmitry Dymshyts[1], Frank DeFalco[1], Anna Ostropolets[1], Gowtham Rao[1], Azza Shoaibi, Clair Blacketer [1,2]

[1]Janssen Research & Development, Raritan, NJ; [2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, NL

Background

Standard phenotypes are developed by defining cohort definitions that identify populations of interest. Cohort definitions rely on concept sets as part of inclusion and exclusion criteria definitions. Concept sets are lists of concepts in addition to logic settings for each concept. Logic settings provide a way to specify that descendants of a concept should be included, mapped concepts should be included, or rather that a concept and its descendant or mapped concepts should be excluded. Using these logic settings, a concept set is resolved to a simple list of concepts when it is evaluated against a version of the OHDSI standardized vocabularies[1]. As the OHDSI standardized vocabularies are continuously updated[2], it impacts concept sets and cohort definitions. OHDSI phenotype library[3] contains cohort definitions should reflect the same clinical idea regardless of the vocabulary version used. Thus, we assessed changes in cohorts from OHDSI Phenotype library by running the PhenotypeChangesInVocabUpdate R package[4].

Methods

The concept sets were extracted from the OHDSI Phenotype library in the form of JSON expressions of concept sets used in cohorts. Note, the cohorts instantiated in Atlas can be extracted as well by PhenotypeChangesInVocabUpdate R package.

The cohort definitions were grouped by the vocabulary version they were most likely based on. We assume that vocabulary version becomes active in the next month after the release. This way, for example, if cohort was created on Jan-2023, it uses the Sep-2022 vocabulary version, while the cohort from Feb-2023 is using Jan-2023 vocabulary. Currently the vocabulary version the cohort is based on is not stored, so we need to make assumption based on dates, which is not accurate.

Note, we didn't run the cohorts, and we only know the counts of affected concepts in data sources licensed to Janssen Research & Development and standardized to the The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM). These concepts might be a part of initial event criteria, or part of inclusion or restriction criteria, this way we don't know how much it affects the overall cohort generation, but the general assumption is that the higher number of occurrences a concept has, the more it impacts a cohort generation.

The cohort definitions were assessed based on the following metrics:

1. <u>Presence of non-standard concepts in concept set expression.</u> Here we don't compare vocabulary versions and look only at the current vocabulary. Concepts could become non-standard after the vocabulary update, or source codes were used intentionally. Both scenarios need to be addressed, in the first we need to update concept sets, in second to update the OHDSI standardized vocabulary as it's not granular enough to reflect all the concepts needed.
   The concept set definition JSON isn't updated with the vocabulary update, so user will not see changes in Atlas. So, it's necessary to run this tool to see concepts changed to non-standard.

2. Changes in Included Source Codes. This reflects alterations in the chain from a clinical event to cohort event:
   Clinical event as a source concept (ICD10CM, for example) –[*mapping*]-> **standard_concept** – *[inclusion rules with concept hierarchy*]-> **cohort event.**
   Thus, different source concepts will be captured if either mapping or the standard hierarchy was changed.
   The following source concepts were used: ICD10, ICD10CM, CPT4, HCPCS, NDC, ICD9CM, ICD9Proc, ICD10PCS, ICDO3, JMDC, LOINC, since these are prevalent in our database network.
3. Concepts included in concept sets changed their domains. Only concepts that have related source codes from our datasets are shown, since we are interested only in concepts associated with real events and hence play a practical role.

Results

343 out of 599 cohorts present in OHDSI Phenotype library have at least one change described above detected.

We identified 630 concepts used in concept sets across the cohorts which had become non-standard in subsequent vocabulary releases. These changes happen both due to external decisions (SNOMED editorial policy changes) and internal decisions (made by OHDSI vocabularies). Note, the same concept can affect multiple cohorts.

2665 unique related source concepts were added, these changes are due to improvement of concept mapping.
854 unique source concepts were removed from concept sets. This happens mostly due to change of concept mapping.

114 included concepts changed their domain. Mostly these are Conditions changed to Observation in the recent vocabulary release when Symptoms and Victim status were changed to the Observation domain.

These changes were aggregated into excel tables, see their description below:

1) **Non-Standard Nodes.**

This table lists non-standard concepts used in the concept set definition.

Table 1. Non-standard nodes used in concept set definition and their replacement mapping

cohortdefinitionid	865
cohortname	Major Non Cardiac Surgery, adults, inpt stay, no ED
conceptsetname	Knee arthroplasty
isexcluded	1
includedescendants	1
node_concept_id	4304358
node_concept_name	Diagnostic procedure
drc	2828512517
mapsToConceptId	4176642
mapsToConceptName	Procedure by intent

Node_concept_id - the concept used in concept set expression.
"drc" is a total number of descendant concepts of node concept_id

In this example diagnostic procedures were initially excluded from the concept set, so we need to find how to represent this logic in the new vocabulary version.

2) **Mapping difference:**

Table 2. Included Source concepts removed or added.

COHORTID	300
COHORTNAME	Heavy menstrual bleeding (menorrhagia) events
CONCEPTSETNAME	Heavy menstrual bleeding (menorrhagia)
CONCEPTSETID	0
SOURCE_CONCEPT_ID	44827910
RECORD_COUNT	29746615
ACTION	Added
SOURCE_CONCEPT_NAME	Excessive or frequent menstruation
SOURCE_VOCABULARY_ID	ICD9CM
SOURCE_CONCEPT_CODE	626.2
OLD_MAPPED_CONCEPT_ID	4078455
OLD_MAPPED_CONCEPT_NAME	Finding of menstrual bleeding
OLD_MAPPED_VOCABULARY_ID	SNOMED
OLD_MAPPED_CONCEPT_CODE	276319003
NEW_MAPPED_CONCEPT_ID	197607
NEW_MAPPED_CONCEPT_NAME	Excessive and frequent menstruation
NEW_MAPPED_VOCABULARY_ID	SNOMED
NEW_MAPPED_CONCEPT_CODE	266601003

In this example ICD9CM concept was originally mapped to a very broad target concept (Finding of menstrual bleeding). With the vocabulary refresh it gained meaningful target concept (Excessive and frequent menstruation) that was picked up by a concept set definition. This is the example of a positive impact of the OHDSI vocabulary evolution.

3) Domain Change

This table shows included concepts changed their domain, so the event table with the new domain should be used.

Table 3. Standard concepts changed their domain.

cohortid	691
cohortname	Transverse myelitis or symptoms indexed on symptoms or diagnosis
conceptsetname	Symptoms for Transverse Myelitis
conceptsetid	5
conceptId	437113
conceptName	Asthenia
vocabularyId	SNOMED
sourceConceptCode	R53.1
sourceConceptName	Weakness
sourceVocabularyId	ICD10CM
oldDomainId	Condition
newDomainId	Observation
sourceConceptRecordCount	166531707

Here “437113|Asthenia” concept changed its domain from Condition to Observation, so the concept set “Symptoms for Transverse Myelitis” needs to be used with Observation table as well.

Conclusion

Now because of the vocabulary update 57% of OHDSI Phenotype library cohorts have different concept set resolution comparing to the one they had once created. With different vocabulary version concept sets might capture different clinical events (source codes) due to hierarchy changes, concepts become non-standard, mapping changes; concept sets might resolve in different domains. We recommend evaluating the impact of vocabulary evolution on the OHDSI Phenotype library (and other phenotype repositories that may exist across the community) at each new vocabulary release to identify and potentially revise phenotype algorithms to account for concept and source code migration.

References

1. OHDSI 2023, *OHDSI Standardized Vocabularies GitHub Wiki,* accessed 27 May 2024, <https://github.com/OHDSI/Vocabulary-v5.0/wiki>
2. OHDSI 2024, *OHDSI Standardized Vocabularies release notes,* accessed 27 May 2024, <https://github.com/OHDSI/Vocabulary-v5.0/releases>
3. OHDSI 2024, *The OHDSI phenotype library GitHub.io package description,* accessed 27 May 2024, <https://ohdsi.github.io/PhenotypeLibrary/>
4. OHDSI 2024, *PhenotypeChangesInVocabUpdate package GitHub page,* accessed 27 May <https://github.com/OHDSI/PhenotypeChangesInVocabUpdate>