Vocabulary Versioning: Tracking Concepts over Time
Software Demonstration

Tom Seinen, Peter Rijnbeek
Department of Medical Informatics, Erasmus MC, The Netherlands

Background

The Observational Medical Outcomes Partnership (OMOP) standardized vocabulary provides consistency and transparency across the global network of disparate OMOP common data model (CDM) databases and is fundamental to conducting efficient and reproducible observational research. The standardized vocabulary consists of a library of more than 100 distinct vocabularies and includes the mappings of terms to a set of standard concepts from vocabularies including SNOMED CT[1], RxNorm (and extension) [2], LOINC [3], and the ICD10 Procedure Coding System [4]. The creation and maintenance of the standardized vocabulary is an arduous task. The vast number of concepts and the asynchronous intervals at which the individual vocabularies are released make combining the vocabularies and creating the mappings between them a challenge.

With the continuous improvement of the standardized vocabulary, the concepts are changing over time and directly influence all observational research performed on OMOP CDM databases. A concept change can be benign, such as the capitalization of a character in the concept description, but also destructive, such as a concept domain change. Concept sets and cohort definitions, and thus also individual studies, are in essence directly linked to one specific vocabulary version. This means that redefining the affected concept sets and cohort definitions is required when the concept changes are severe. Therefore, it is crucial that for any study, performed on multiple databases or repeated on the same database over time, the possible changes in individual concepts, caused by a discrepancy in vocabulary version, are taken into account.

Currently, only the latest version of the OMOP standardized vocabulary is available for download from *Athena*[1] and for each new version release only aggregated concept additions and changes are reported. Therefore, our objective was to enable the quick and efficient storage and retrieval of any vocabulary version and the ability to analyze and track individual concepts between multiple vocabulary versions.

Methods

Vocabulary versions – A total of 43 successive vocabulary versions were downloaded from *Athena* during the period from March 2020 to June 2022. Additional to the default vocabularies, we included MedDRA, ICD10, dm+d, and CTD.

Full release database – Inspired by the versioning system of SNOMED CT[5], a framework was designed that enabled the creation of a full release database from individual successive *Athena* vocabulary versions (snapshot releases), see Figure 1. The differences (additions, changes, and deletions) between each successive pair of snapshot releases were calculated and stored as a delta release. The full release database is created by stacking the delta releases, effectively only storing the vocabulary version differences. The table columns or fields in the full release database are identical to the OMOP standardized vocabulary tables, only two columns are added, that indicates the version for each row and whether it concerns an addition, a change, or a deletion. We created a lightweight R-package,

[1] https://athena.ohdsi.org/

VocabularyVersioning[2], for creating the full release database and performing database operations. Using the full release database all vocabulary versions can be efficiently stored, retrieved, and the difference between any two versions can be calculated. Furthermore, the full release table can be used to retrieve the history of each concept and enables further analysis of the standardized vocabulary.

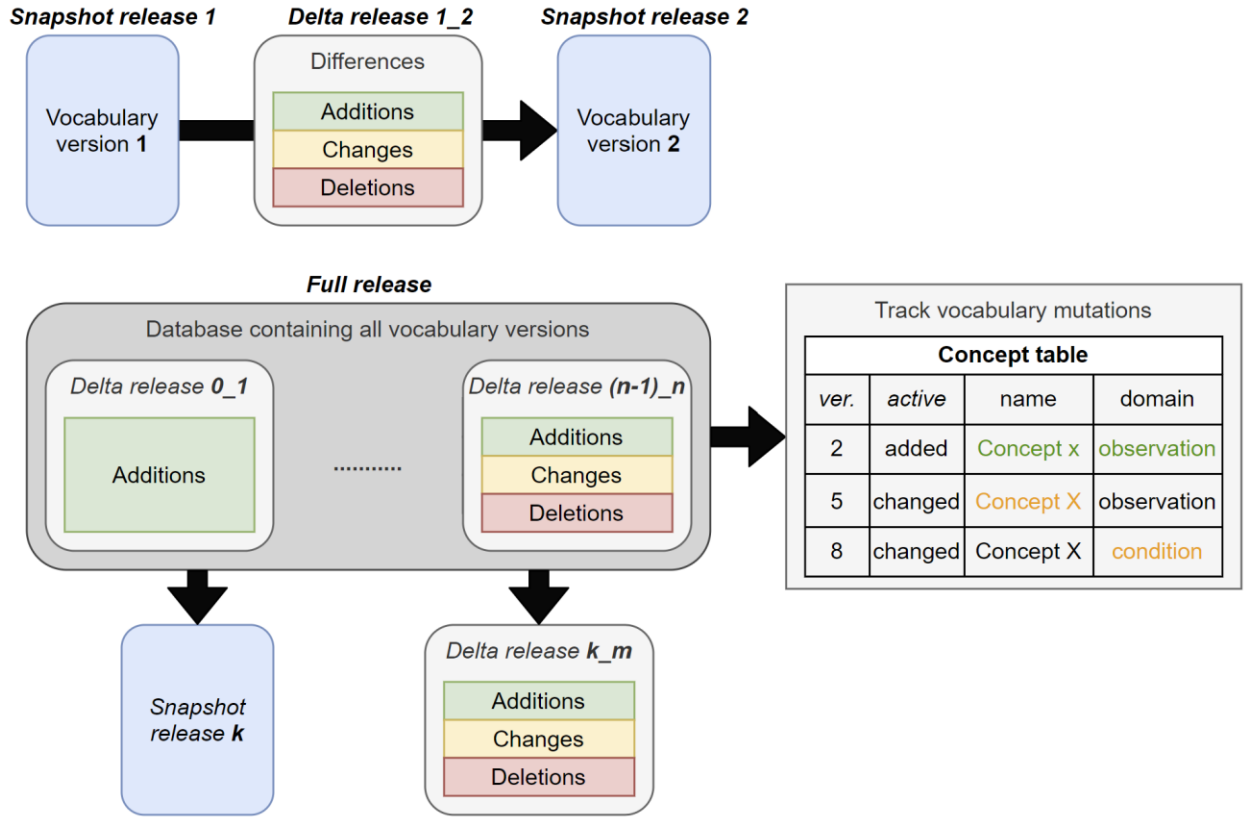

Figure 1. Visualization of the vocabulary versioning framework, including the creation of the full release database, retrieval of any snapshot release and delta release, and the ability to track individual concepts over time.

Data visualization and interactive tracker – All analysis, such as aggregating the number of mutations over time or looking up the history of a concept, can be performed directly on the database, e.g. in SQL or R. However, repeated analysis, data visualizations, or more complex queries are better predefined to improve the user experience and efficiency. For this reason, we created an R-shiny application that performs as an interactive layer on top of the database and enables the visualization of the vocabulary table changes over time, the history lookup of specific concepts, and the tracking of entire concept sets and cohort definitions from a connected *OHDSI WebAPI*[3] instance. The overview page of the R-shiny application, with the total number of rows in the vocabulary and the number of mutations by table and mutation type, can be found in Figure 2.

[2] https://github.com/mi-erasmusmc/VocabularyVersioning
[3] https://github.com/OHDSI/WebAPI

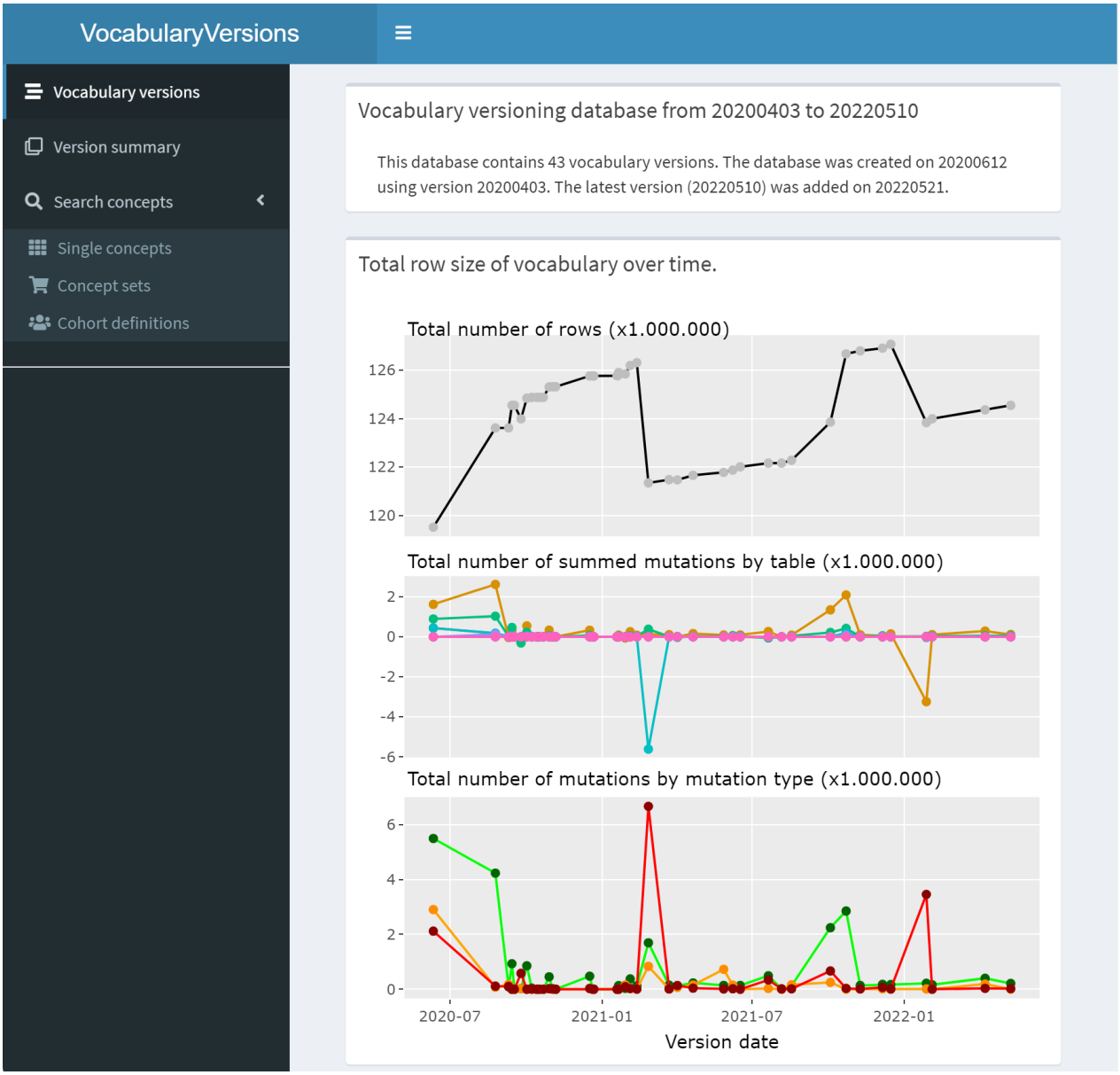

Figure 2. R-shiny application is used as an interactive layer on top of the full release database.

Results

As an example, we illustrate the concept changes over time for the concept ids 37311060 (suspected COVID-19) and 37311061 (COVID-19) in Figure 3. These two SNOMED CT concepts were subject to change over the course of the first months of the covid-19 pandemic. In both concepts, the concept names changed twice, which should not be a severe change if the meaning is similar. However, the concept with id 37311060 also changes from condition to observation in vocabulary version 20200611 and becomes standard. This can be considered a breaking change as cohort definitions are defined using concept ids within a specified domain.

VersionDate	Status	CONCEPT_ID	CONCEPT_NAME	DOMAIN_ID	VOCABULARY_ID	CONCEPT_CLASS_ID	STANDARD_CONCEPT	CONCEPT_CODE
All		All	All	Al	All	All	All	All
20200403	Reference	**37311060**	Suspected disease caused by severe acute respiratory coronavirus 2	Condition	SNOMED	Context-dependent		840544004
20200611	Changed	**37311060**	Suspected disease caused by 2019-nCoV	Observation	SNOMED	Context-dependent	S	840544004
20210226	Changed	**37311060**	Suspected COVID-19	Observation	SNOMED	Context-dependent	S	840544004
20200403	Reference	**37311061**	Disease caused by severe acute respiratory syndrome coronavirus 2	Condition	SNOMED	Clinical Finding	S	840539006
20200611	Changed	**37311061**	Disease caused by 2019-nCoV	Condition	SNOMED	Clinical Finding	S	840539006
20210226	Changed	**37311061**	COVID-19	Condition	SNOMED	Clinical Finding	S	840539006

Figure 3. Example illustrating the ability to track individual concepts between vocabulary versions. The names of the two concepts were changed between vocabulary versions and the first concept listed also has a domain and standard change.

Conclusion

We build the vocabulary versioning framework to store and retrieve any OMOP standardized vocabulary version using a full release database, inspired by the SNOMED CT versioning system. The creation of the full release database is simple yet powerful as it also enables endless possibilities of analyses of individual and aggregated concepts at one point or over time. Detecting changes between vocabulary versions, that can potentially break study definitions, is essential for performing reliable observational research on OMOP CDM databases. Our tool can both be used to inform researchers about affected studies and to inform database management on concept changes that might impact mappings and the conversion from the source database to the OMOP CDM.

References/Citations

1. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 2006;121:279.
2. Liu S, Ma W, Moore R, et al. RxNorm: prescription for electronic drug information exchange. *IT professional* 2005;7(5):17-23.
3. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry* 2003;49(4):624-33.
4. Averill RF, Mullin RL, Steinbeck BA, et al. Development of the ICD-10 procedure coding system (ICD-10-PCS). *Topics in health information management* 2001;21(3):54-88.
5. Benson T, Grieve G. Snomed ct. Principles of Health Interoperability: Springer, 2021:293-324.