OMOP GIS Vocabulary Package for Observational Studies in Health Care and Public Health

Maksym Trofymenko, MD[1], Polina Talapova, MD, PhD,[1,2,3], Andrew Williams, PhD[3],
[1]SciForce, Kharkiv, Ukraine
[2]Kharkiv National Medical University, Kharkiv, Ukraine
[3]Tufts Clinical and Translational Science Institute, Boston, MA, United States

Background

Understanding the factors influencing health outcomes extends far beyond traditional clinical data. Spatial factors, environmental exposures, and social determinants of health (SDOH) are increasingly recognized as playing a critical role in shaping individual and community health[1,2]. However, integrating these diverse data types with traditional health records continues to pose significant challenges. Existing healthcare data often lacks the standardized terminologies and structures necessary to capture the complexities of these multifaceted influences.

This gap hinders researchers from exploring the interplay between these determinants and health outcomes. For instance, analyzing the association between air pollution (environmental exposure) and respiratory illness (health outcome) requires linking data on patient diagnoses with information about air quality in their geographic location. Similarly, investigating the connection between socioeconomic status (SDOH factor) and access to preventative care necessitates integrating data on income levels with healthcare utilization records.

Objective

To overcome the limitations of current healthcare data and enable a more comprehensive approach to health research, we introduce the OMOP GIS Vocabulary Package[3]. This innovative solution offers a unified framework for integrating three critical data types into existing OMOP-compliant databases, enhancing data interoperability and supporting more robust research efforts, especially for spatial data, environmental data and Social determinants of health.

Methods

Within the package content, the OMOP GIS Vocabulary is a compilation of terminologies related to geography, boundaries, and spatial elements. The PostGIS extension entities are used as basic geographical structures (points, lines, and polygons with their complex representations). All real locations were added manually using relationships provided by PostGIS.

The Exposome vocabulary, previously known as Toxin Vocabulary[4], is a proceeded and standardized the Toxin and Toxin Target Database (T3DB), which contains over 3,000 toxins, described by 41,000+ synonyms, including pollutants, pesticides, drugs, and food toxins, each linked to specific toxin target records. Each ToxCard provides extensive information such as chemical properties, toxicity values, molecular and cellular interactions, and medical details, spanning more than 90 metadata fields. For concept identification, CAS codes were used due to their compatibility with GIS data and the extensive CAS Registry. Toxins lacking CAS codes were assigned unique T3DB identifiers, while manually added Classification concepts received auto-generated codes.

The SDOH vocabulary seamlessly integrates key components from recognized standards such as Social Determinants of Health Ontology (SDOHO)[5], the Agency for Healthcare Research and Quality (AHRQ) indicators[6], and Environmental Justice Index (EJI) Indicators[7], including the Social Vulnerability Index (SVI) variables[8]. This integration ensures a rich, multi-dimensional perspective, capturing a wide spectrum of determinants from socioeconomic status to healthcare access, and from educational opportunities to neighborhood and built environment. The concepts in the SDOH Vocabulary are part of the newly introduced 'Phenotypic Feature' domain to expand the scope of OHDSI Vocabularies. This domain refers to a semantic category that captures observable characteristics, encompassing a broad range of environmental, social, and personal health determinants not adequately represented by traditional medical domains.

All three vocabularies within the OMOP GIS Vocabulary Package are curated by the OHDSI GIS working group, ensuring consistent terminology and adherence to OMOP standards.

Results

The OMOP GIS vocabulary includes 159 concepts categorized within the Observation OMOP domain via different concept classes, such as 'Geometry Item' (e.g., 'LineString', 'Polygon', '2D Geometry'), 'Location' (e.g., 'Administrative Boundary', 'County'), and 'Geom Relationship' (e.g., 'Within', 'Adjacent to'). Concepts are interconnected via hierarchical relationships ('Polygon' - 'Is a' - '2D Geometry') and supplemental GIS-specific relationships ('LineString' - 'Is geometry of' - 'International Border'). Targeted vocabularies for mapping include SDOH, OSM and SNOMED.

The OMOP Exposome Vocabulary features concepts classified under the 'Observation' domain and the 'Substance' concept class within OMOP. The vocabulary's structure includes over 79,000 internal relationships linking toxins to target cellular and tissue structures, proteins, associated medical conditions, biological processes, and toxin categories. Moreover, there are 1,800 "Maps to" relationships that facilitate the integration of the Toxin Vocabulary with OHDSI Standardized Vocabularies such as SNOMED CT, RxNorm, and RxNorm Extension. Exposome concepts that lack direct counterparts are retained as standard concepts.

The SDOH Vocabulary integrates various variables from AHRQ, SDOHO, EJI and SVI, along with newly developed hierarchical nodes, into a cohesive ontology. The SDOH-specific concept classes are meticulously designed to help researchers pinpoint the origin and intended use of each term within the SDOH framework. Each concept class, identified by a unique concept_class_id, serves a distinct purpose. For instance, constructs and determinants derived from AHRQ provide foundational elements for evaluating health outcomes. Similarly, constructs created by the GIS Working Group (WG) offer specialized tools for understanding geographic and social contexts. These constructs represent abstract frameworks used to organize and measure determinants, while determinants are concrete factors directly influencing health. Items within these determinants further break down the categories into specific, measurable elements, allowing for detailed analysis. Additionally, item values provide precise metrics for these elements, such as employment status classifications. This integration results in nodes with a general structure of Construct -> Determinant -> Item/Geo Item -> Item Value, intersecting and enhancing each other across standards. The SDOH vocabulary incorporates over 8,000 concept associations, establishing a complex hierarchy that integrates seamlessly with external OMOP vocabularies, enhancing the representation and analysis of SDOH. This intricate structure includes primary components relevant to demographics, education, geographic location, health, physical environment, population, and social and community context,

and interfaces with major vocabularies like SNOMED, ATC, CPT4, HCPCS, LOINC, Nebraska Lexicon, OMOP Extension, PPI, Race, UK Biobank and OMOP GIS.

Conclusion

The OMOP GIS Vocabulary Package enables the downloading, browsing, and integration of spatial, environmental, and social determinants data with the OMOP CDM, allowing measurements at individual and population levels. The Package significantly advances health research by promoting a holistic approach to data exploration, enabling researchers to investigate the combined influence of social circumstances, environmental exposures, and geographical locations on health outcomes. Its standardized terminologies enhance communication and collaboration across studies. The ongoing development, including planned testing with diverse use cases and continuous refinement, ensures its relevance and adaptability in the evolving healthcare research landscape. Future plans include an end-to-end approach for using the GIS Vocabulary Package in studies.

References

1. Hadley MB, Nalini M, Adhikari S, Szymonifka J, Etemadi A, Kamangar F, Khoshnia M, McChane T, Pourshams A, Poustchi H, Sepanlou SG, Abnet C, Freedman ND, Boffetta P, Malekzadeh R, Vedanthan R. Spatial environmental factors predict cardiovascular and all-cause mortality: Results of the SPACE study. PLoS One. 2022 Jun 24;17(6):e0269650. doi: 10.1371/journal.pone.0269650.
2. Chelak K, Chakole S. The Role of Social Determinants of Health in Promoting Health Equality: A Narrative Review. Cureus. 2023 Jan 5;15(1):e33425. doi: 10.7759/cureus.33425.
3. OHDSI. GIS vocabularies. Available from: https://github.com/OHDSI/GIS/tree/main/vocabularies. Accessed 21 June 2024.
4. Talapova P, Trofymenko M, et al. A Toxin Vocabulary for the OMOP CDM. OHDSI Symposium 2023. Available from: https://www.ohdsi.org/wp-content/uploads/2023/10/Talapova-Polina_A_Toxin_Vocabulary_for_the_OMOP_CDM_2023symposium-Polina-Talapova.pdf. Accessed 21 June 2024.
5. Dang Y, Li F, Hu X, Keloth VK, Zhang M, Fu S, Amith MF, Fan JW, Du J, Yu E, Liu H, Jiang X, Xu H, Tao C. Systematic design and data-driven evaluation of social determinants of health ontology (SDoHO). J Am Med Inform Assoc. 2023 Aug 18;30(9):1465-1473. doi: 10.1093/jamia/ocad096.
6. Kronick R. AHRQ's Role in Improving Quality, Safety, and Health System Performance. Public Health Rep. 2016 Mar-Apr;131(2):229-32. doi: 10.1177/003335491613100205.
7. Agency for Toxic Substances and Disease Registry (ATSDR). Environmental Justice Index: Indicators. Available from: https://www.atsdr.cdc.gov/placeandhealth/eji/indicators.html. Accessed 21 June 2024.
8. Mah JC, Penwarden JL, Pott H, et al. Social vulnerability indices: a scoping review. BMC Public Health. 2023;23:1253. doi:10.1186/s12889-023-16097-6.