Toward a General-Purpose Geography-Focused OHDSI Infrastructure

Kyle Zollo-Venecek[1], Robert Miller, MS[2], William G. Adams, MD[3], Jay Greenfield, MA, PhD[4], Timothy B Norris, PhD[5], Polina Talapova, MD, PhD[1], Maksym Trofymenko[6], Andrew Williams, PhD[1]

[1] Tufts Clinical and Translation Science Institute (CTSI), Tufts Medicine, Boston, MA
[2] Minderoo Foundation: Collaborate Against Cancer Initiative
[3]Boston Medical Center/BU Chobanian & Avedisian School of Medicine, Boston, MA
[4]University of Miami Libraries, University of Miami Institute for Data Science and Computing
[5]CODATA: Committee on Data of the International Science Council
[6]SciForce

Background

The impact of our living environment on health often surpasses the influence of healthcare received, with location being a stronger predictor than genetics[1, 2]. However, OHDSI studies typically overlook regional factors like poverty, health policies, and environmental toxins, or rely on non-scalable solutions due to the lack of a geography-focused infrastructure compatible with OMOP CDM and OHDSI tools. Consequently, applying OHDSI's best practices to place-related data is challenging, introducing potential bias from unmeasured regional attributes. Moreover, without a suitable GIS infrastructure, investigating the health impacts of social and environmental drivers remains limited, despite the vast geographic reach and abundance of public geographic data within the community.

Despite the potential value of geospatial data analysis within the OMOP CDM field[3, 4], researchers face significant technological barriers due to the volume and heterogeneity of geospatial data sources[5]. Overcoming these challenges can be particularly daunting for researchers with limited resources or technical expertise.

The OHDSI GIS Workgroup has made notable progress in advancing geospatial data utilization within the OHDSI community[6]**.** This includes the development of a universal representation for geospatial data and software tools for geocoding addresses, data ingestion from a read/write inventory, and harmonization of diverse data structures. While the inventory does not store geospatial data directly, it enables users to access a wide range of publicly available datasets that record environmental exposures potentially influencing human health through metadata.

The workgroup is actively exploring integration pathways with the broader OHDSI ecosystem. This involves incorporating geospatial data into phenotyping and analytical software packages, empowering researchers to leverage geographic dimensions for person- and population-level health research. Through these efforts, the workgroup aims to foster collaboration, drive advancements, and generate impactful research and health outcomes.

Methods

Foundational Work

The foundation of our approach is a "universal representation" for geospatial data. This representation allows for the compilation of externally-published geospatial data sources into a consistent and extensible schema that is intended to persist outside of the OMOP CDM instance.

To facilitate the integration of geospatial data, we have enabled a standardized geocoding pipeline as well as created an extensible mechanism that automates the retrieval, translation, and ingestion of

diverse geospatial data, greatly reducing manual effort as well as increasing consistency of source data across disparate studies. With this we reduce duplicative work and increase reproducible analysis.

The aforementioned tooling leverages a functional metadata design that contains information about geospatial datasets that is needed to automate ingestion, such as source names, URLs, transformation logic, and the variables within each dataset. By standardizing and cataloging this metadata as a source repository, we enable efficient and predictable data management to serve as a foundation for OHDSI tooling integration.

OHDSI integration:

To drive the integration of our tooling with the OHDSI ecosystem:

We are leveraging existing use cases to identify intersections between the geospatial framework and the OHDSI ecosystem. One such use case is so-called “deep phenotyping” that investigates the role of environmental exposures as drivers of health behaviors and disease outcomes. In this and other use cases we are performing a landscape analysis in which we determine how environmental exposures might figure into and play out in OHDSI data analysis. This landscaping exercise will help determine the breadth and priority of areas of integration.

We are currently reviewing documentation and high-level code of the priority tooling to help determine the scope of work required for each specific integration effort and consequently help guide our roadmap.

To prioritize a community approach, we will draft proposed solutions and approach the responsible parties within that domain of the OHDSI ecosystem to collaboratively evaluate the feasibility and alignment. Once finalized, we will then conduct sprints towards development of these integration mechanisms including the development of testing suites. Further testing of maturity will be conducted through proof-of-concept network studies with OHDSI partners.

Results

Our initial use cases have led us to target intersections with the OHDSI tooling in the areas of computable phenotyping and reproducible analyses, specifically through ATLAS, and consequently Circe, as well as the HADES suite of R packages.

Our investigations have highlighted the inherent barriers of incorporating place-related data in the person-centric OHDSI tooling. So far from these investigations we have begun to examine the “cost” of one solution that has presented itself: the implementation of a new, pseudo-OMOP event table: **EXPOSURE_OCCURRENCE**. The function of this data structure is to temporarily populate place-related data as person-level exposures to be leveraged by the OHDSI tooling. This enables the process of OHDSI integration while maintaining manageable volumes of data, one of the core goals of the geospatial framework, as well as significantly assuaging the burdens of tooling integration.

Conclusion

Our work to date has demonstrated the value and feasibility of integrating a geospatial framework with the OMOP CDM. We have also identified several additional areas in which we can expand, including the integration of raster data into the model and supporting retrieval from API resources. Leveraging this success, we will continue to expand the framework's capabilities, address these constraints, and further enhance the integration and utilization of geospatial data within the OHDSI community.

The achieved level maturity of geospatial data within OHDSI, along with the promising sentiment and results of this exercise, provide a bright outlook for the GIS Workgroup endeavor of integration with

OHDSI tooling and consequently lowering the barriers for incorporation of new kinds of evidence and promoting better health outcomes.

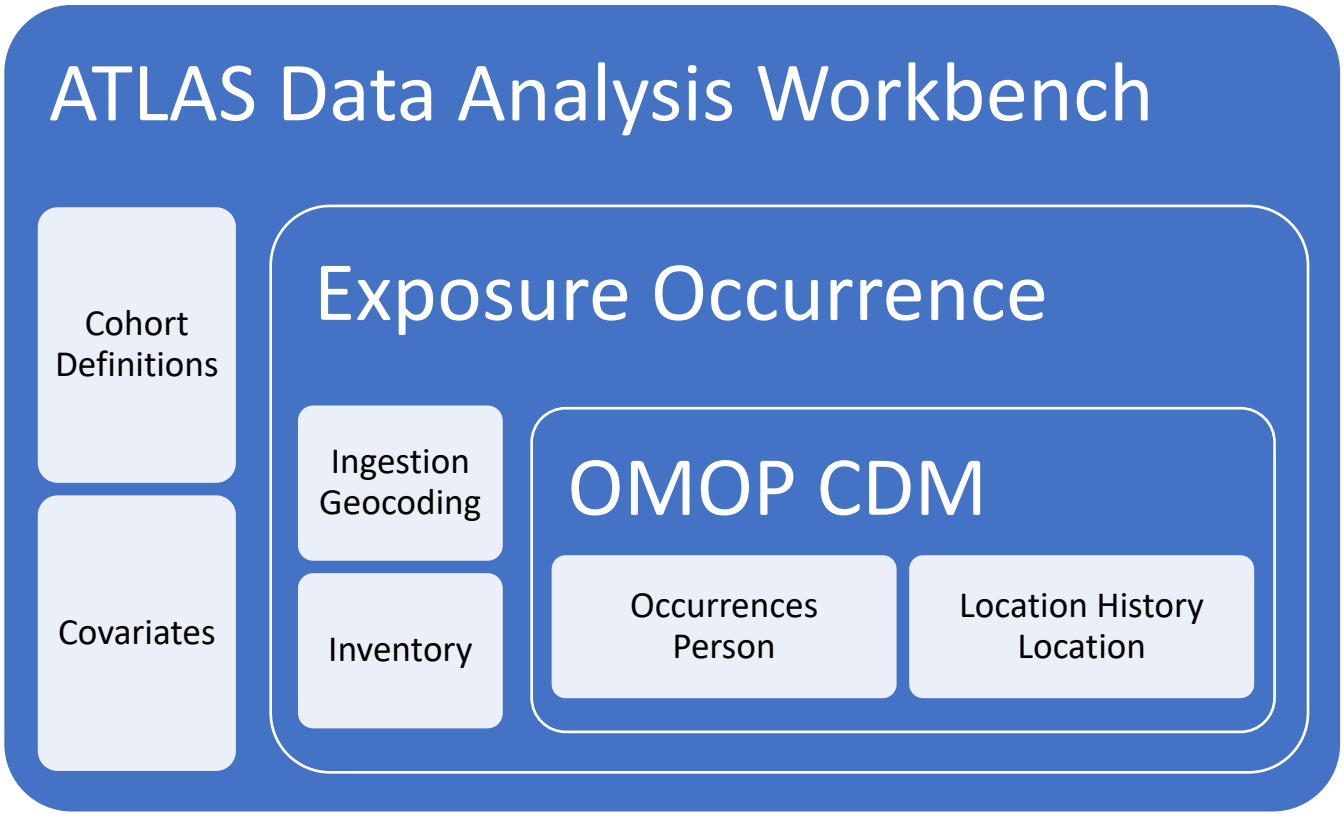

Figure 1: Intersectionalities between the geospatial framework and the OHDSI ecosystem under deep phenotyping

References

1. World Health Organization. Preventing Disease through Healthy Environments: a Global Assessment of the Burden of Disease from Environmental Risks [Internet]. 2016. Available from: https://www.who.int/quantifying_ehimpacts/publications/preventing-disease/en/
2. The National Institute for Occupational Safety and Health (NIOSH). Exposome and exposomics [Internet]. Available from: https://www.cdc.gov/niosh/topics/exposome/default.html.
3. Beaton M, Jiang X, Maura M, Xinzhuo W. Leveraging Location Data in OMOP to Incorporate ADI. OHDSI Symposium 2022; October 29, 2022; Virtual. Available from: https://www.ohdsi.org/wp-content/uploads/2022/10/29-beaton_jiang_maura_xinzhuo_Leveraging-Location-Data-in-OMOP-to-Incorporate-ADI_2022symposium-Maura-Beaton.pdf
4. Cho J, You SC, Lee S, Park D, Park B, Hripcsak G, et al. Application of Epidemiological Geographic Information System: An Open-Source Spatial Analysis Tool Based on the OMOP Common Data Model. International Journal of Environmental Research and Public Health [Internet]. 2020;17:7824. Available from: http://dx.doi.org/10.3390/ijerph17217824
5. Reynard D. Five classes of geospatial data and the barriers to using them. Geography Compass. 2018;12:e12364. Available from: https://doi.org/10.1111/gec3.12364
6. https://github.com/ohdsi/gis