

Metadata standards in OMOP to increase reproducibility in exposomics studies

Anne E Thessen¹, Nico Matentzoglou², Corey Cox¹, Madan Krishnamurthy¹, Bryan Laraway¹, Sruthi Magesh¹, Julie McMurray¹, and Melissa Haendel¹

¹University of North Carolina, Chapel Hill NC USA, ²Independent Consultant at Semantics.ai

Background

Advances in genomics and phenomics have transformed disease diagnosis and treatment. Despite estimates that environmental factors may be a larger predictor of disease mortality than genetics alone¹, exposure data remain difficult to integrate into medical research and care. This is partially due to the lack of standards available for linking health data and environmental data when compared to genomics and phenomics. In OMOP, exposure information is linked to a participant record using the proposed external exposure, location, and location history tables. The environmental data sets used to calculate exposures can be discovered using the Gaia Catalog and processed to patient-level values using repeatable workflows in GaiaDB. While several efforts have successfully documented metadata for geospatial data sets (such as the OHDSI GIS WG), many details in how environmental data are made available and prepared for analysis still go undocumented, depending on downstream requirements (e.g. comparability of results, reproducibility). Furthermore, the community has not yet agreed on a standard to share such metadata beyond specific projects such as OMOP. This poster will present a prototype standard for representing geolocated environmental metadata that will support consistency and reproducibility in environmental health studies. Lessons learned from this activity will provide feedback to the OHDSI GIS WG.

Methods

Prototype development was driven by use cases gathered from participation in the OHDSI GIS Working Group, the ESIP Geodata 4 Health Cluster, and the GA4GH Human Exposomics Data Standards Study Group, a review of the environmental epidemiology literature, a review of project documentation from C-HER², Amadeus³, CLAD⁴, HEW-DA⁵, and DEGAUSS⁶, and interviews with environmental epidemiologists. We selected a toy example of five U.S. academic health centers collaborating to study the impact of heat stress to guide development and defining the problem. In this example, these sites co-author a single OHDSI network analysis using daily maximum air temperature during the summer from 2018-2022. This example compared two different methods to calculate environmental variables for each patient: 1) Daily Tmax via the DeGAUSS Daymet container (Daymet V4 R1), and 2) Daily Tmmx via the amadeus R package (GridMET / METDATA). Each site executed the same geocoding and environmental data analysis locally and assigned temperatures to patients based on their residential address - one numeric value per person-day. The toy example is available in GitHub⁷.

Results

The two different pipelines examined in the toy example resulted in different temperature values for the same person at the same site (43.91 C Daymet DEGAUSS vs 42.65 GridMET Amadeus). This discrepancy (1.26 C) was enough to shift a 95th percentile heat-day classification resulting in different definitions of “extreme heat”. While this is not unexpected, the inconsistent metadata payloads from each of the pipelines makes it difficult to reveal the reason for this difference. Closer examination of metadata between the sites showed differences in day-boundary convention, extraction rules, and reference period for the 95th-percentile threshold that were not captured in the current pipeline metadata. This is likely only a subset of the metadata needed to explain these differences.

Conclusion

To address this problem, we propose to develop a metadata standard using LinkML⁸ (a highly descriptive, program-agnostic modeling language) to describe essential metadata that is emitted as a by-product of exposure data pipelines - a metadata “sidecar”. Tools that ignore the sidecar would be unaffected; tools that consume it would be able to validate, compare, and audit the methodology behind each value. One of the barriers to understanding the impact of environmental exposures on health is the difficulty in reproducibility and comparing results across studies and data sets, as demonstrated by this work. This proposed sidecar approach makes methodological differences explicit, enabling a coordinator (human or machine) to detect each divergence prior to pooling and decide whether to re-run, exclude, or pool with method-aware weighting. Exposure values can then become structured, citable, comparable scientific objects, not numerical values reported in free-text methods sections.

References

1. Sarigiannis D, Karakitsios S, Anesti O, Stem A, Valvi D, Sumner SCJ, et al. Advancing translational exposomics: bridging genome, exposome and personalized medicine. *Hum Genomics* [Internet]. 2025 Apr 30;19(1):48. Available from: <http://dx.doi.org/10.1186/s40246-025-00761-6>
2. Hanson HA, Ramsay J, Grant J, Davis M, Agbaje JO, Maguire D, et al. Centralized Health and Exposomic Resource (C-HER): Analytic and AI-ready data for external exposomic research [Internet]. *arXiv [stat.AP]*. 2025. Available from: <http://arxiv.org/abs/2511.03750>
3. amadeus [Internet]. [cited 2026 Jun 4]. Available from: <https://niehs.github.io/amadeus/>
4. All of Us Research Program [Internet]. [cited 2026 Jun 4]. Available from: <https://allofus.nih.gov/article/center-for-linkage-and-aquisition-of-data>
5. Scientists A. National Institute of Environmental Health Sciences. [cited 2026 Jun 4]. NIH Health and Extreme Weather (HEW) Data Accelerator. Available from: <https://www.niehs.nih.gov/research/programs/chords/hew-data>
6. Brokamp C. DeGAUSS: Decentralized Geomarker Assessment for Multi-Site Studies. *J Open Source Softw* [Internet]. 2018 Oct 3;3(30):812. Available from: <http://dx.doi.org/10.21105/joss.00812>
7. examples/heat at main · monarch-initiative/envar [Internet]. Github; [cited 2026 Jun 4]. Available from: <https://github.com/monarch-initiative/envar/tree/main/examples/heat>
8. Mungall C. LinkML - Linked data Modeling Language - LinkML - Linked data Modeling Language [Internet]. [cited 2023 Jun 5]. Available from: <https://linkml.io/>