Demonstration of the OHDSI phenotype library

Gowtham A Rao

Introduction:

The Observational Health Data Sciences and Informatics (OHDSI) community has developed a publicly accessible, version-controlled Phenotype Library to guide real-world evidence towards the FAIR principles: Findability, Accessibility, Reproducibility, and Interoperability.[1] This library aims to foster the submission of high-quality cohort definitions, cataloguing of metadata, attribution and promotion of discovery and reuse in scientific research.

Within the OHDSI Phenotype Library (OHDSI PL), each entry represents a unique cohort definition identifiable by a stable, externally referenceable ID. Comprehensive metadata about each cohort definition is catalogued and made searchable for researchers.[2] Content in the library is subject to version control, with each version is assigned a specific DOI.

The OHDSI PL employs a community engagement and contribution process, crediting contributors via ORCID where available. Submitted cohort definitions are subject to a voluntary, open peer review process managed by the OHDSI Phenotype Development and Evaluation Workgroup. All cohort definitions are computable and portable and conform to the specifications of the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) promoting efficient implementation, standard terminology use, seamless conversions between computable and human-readable definitions, and consistent understanding of the logic.[3-6]

Metadata:

The library has the capacity to collect a wide range of metadata:

a) User/community/author-submitted metadata. This includes the short and long names of the cohort definition; the names or ORCID IDs of the contributor(s) and peer reviewer(s); a clinical description of the phenotype for which the cohort definition was designed; a concise explanation of the cohort definition's technical logic to help others understand the underlying code; the recommended study applications for these definitions; pertinent external links, such as OHDSI forum posts discussing the definition; any contributed summary output from OHDSI software such as CohortDiagnostics and/or PheValuator; community recommended tags; any relevant evaluation or peer reviewer comments; and notes on the community's experiences implementing the definition in research.

b) Librarian-assigned metadata, including a managed taxonomy of tags to promote systematic discovery and content navigation; the status of the definition (whether it's accepted, pending peer review, or deprecated), and its Digital Object Identifier (DOI).

c) Computer-generated metadata, which is currently only available for cohort definitions that adhere to the Circe-defined phenotype definition object model. This data encompasses a human-readable,

complete cohort definition logic; a list of domains used in the cohort definition; entry event code lists; comprehensive code lists; and resolved codes.

Maintenance:

Lifecycle:

Once peer-reviewed and accepted, Cohort Definitions become immutable. This differs from Cohort Definitions that have not undergone peer review; these definitions have the potential to evolve in future versions. However, within a referenced released version with an associated DOI, all cohort definitions maintain stability. Cohort Definitions accepted in a release can be deprecated or marked as an error in subsequent versions. Deprecated cohorts [D] remain valid but an alternate cohort might be suggested based on peer review feedback. These cohorts continue to be relevant for OHDSI studies and remain immutable and referenceable. On the other hand, an Error cohort [E] refers to an accepted cohort identified to have a previously unrecognized error. This is akin to a soft deletion, and such cohorts are not recommended for use in OHDSI studies. Despite the error, as accepted cohorts, they will persist and maintain their immutable status in the OHDSI library.

Technical infrastructure and version control:

The OHDSI Phenotype Library (PL) is hosted in a GitHub repository under the OHDSI organization (https://github.com/ohdsi/PhenotypeLibrary) and is encapsulated within the R package known as PhenotypeLibrary. This R package is an integral component of the OHDSI HADES ecosystem (https://ohdsi.github.io/Hades/), and in adherence with the HADES principles, it is designed to seamlessly integrate with other HADES packages. It's worth noting that this repository can be accessed directly using GitHub APIs without utilizing R, as illustrated by https://dash.ohdsi.org/phenotype-explorer.

The PhenotypeLibrary R package also includes a function, getPhenotypeLog(), which retrieves the cohort definitions and related metadata in a tabular format. The release process of this library is aligned with the HADES convention, employing a three-segment numbering system. The first segment signifies major breaking changes, like a full library overhaul, although no such changes are anticipated in the foreseeable future. The second segment is the most common change and is for cohort definitions. Importantly, once a cohort definition is accepted, it remains unchanged. The third segment is used when changes are limited to documentation but not to cohort definitions.

This library follows a regular release cycle, having launched approximately 15 releases since the establishment of major version 3 in 2022. Prior major versions, like version 2 (deprecated in 2022 - https://github.com/OHDSI/PhenotypeLibrary/tree/master-archive) and version 1 (deprecated in 2020 - https://github.com/OHDSI/PhenotypeLibrary/tree/legacy), are now archived and no longer in active use.

Quality checks:

The library is subject to regular quality checks to ensure that the cohort definitions are executable across the OHDSI network. This is done by executing a study package, named PhenotypeLibraryDiagnostics, that executed limited set of diagnostics within CohortDiagnostics. It is executed on volunteer data partners and the outcomes are uploaded to https://data.ohdsi.org/PhenotypeLibrary. This process guarantees that every cohort definition can be executed on the OMOP CDM v5.x platform.

Limitations:

Although the library attempts to guide real world evidence towards the FAIR principles, there are several limitations specifically on the conformance to a machine-readable semantic Resource Description Framework (RDF) standard[1]. In the absence of such conformance, it is hard to relate items in meta-data, or to find related cohort definitions with ease. As the library is evolving to such as desired future state, this limitation is perhaps best described as that of the current observational research standards. By promoting the standards set out in this paper, we aim to work towards a larger resource of information that will conform to the FAIR principles.

Conclusions

The OHDSI Phenotype Library is an open-science version-controlled cohort definition repository with robust community engagement and contribution, an embedded open peer review process using DOI and ORCID. Comprehensive metadata and peer review processes ensure cohort definitions are good quality and usable in observational research.

References:

1. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 2016. **3**(1): p. 160018.
2. Richesson, R.L., M.M. Smerek, and C. Blake Cameron, *A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications.* EGEMS (Wash DC), 2016. **4**(3): p. 1232.
3. Hripcsak, G., et al., *Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers.* Stud Health Technol Inform, 2015. **216**: p. 574-8.
4. Overhage, J.M., et al., *Validation of a common data model for active safety surveillance research.* J Am Med Inform Assoc, 2012. **19**(1): p. 54-60.
5. Newton, K.M., et al., *Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network.* J Am Med Inform Assoc, 2013. **20**(e1): p. e147-54.
6. Hripcsak, G., et al., *Facilitating phenotype transfer using a common data model.* J Biomed Inform, 2019. **96**: p. 103253.