JAMIA Open, 2025, **8(1)**, ooaf001
https://doi.org/10.1093/jamiaopen/ooaf001
Brief Communications

Brief Communications

Increased discoverability of rare disease datasets through knowledge graph integration

Ian Braun, PhD*[,1], Emily Hartley, BS[1], Daniel Olson, MPH[1], Nicolas Matentzoglu, PhD[2], Kevin Schaper, BS[3], Ramona Walls, PhD[1], Nicole Vasilevsky, PhD[1]

[1]Data Collaboration Center, Critical Path Institute, Tucson, AZ 85718, United States, [2]Independent Consultant, Semanticly, Athens 10563, Greece, [3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

*Corresponding author: Ian Braun, PhD, Data Collaboration Center, Critical Path Institute, 1840 E River Road, Suite 100, Tucson, AZ 85718, United States (ibraun@c-path.org)

Abstract

Objectives: Demonstrate a methodology for improving discoverability of rare disease datasets by enriching source data with biological associations.

Materials and Methods: We developed an extension of the Biolink semantic model to incorporate patient data and generated a knowledge graph (KG) comprising patient data and associations between biological entities in an existing KG, leveraging existing mappings and mapping standards.

Results: The enriched model of patient data can support a search application that is aware of biological associations and provides a semantic search interface to discover and summarize patient datasets within the broader biological context.

Discussion and Conclusion: Our methodology enriches datasets with a wealth of additional biological knowledge, improving discoverability. Using condition concepts, we illustrate techniques that could be applied to other entities within source data such as measurements and observations. This work provides a foundational framework for how source data can be modeled to improve accuracy of upstream language models for natural language querying.

Lay Summary

Healthcare datasets can be used for many different purposes in the pursuit of identifying treatments for rare diseases, but critically, the data must be found first. However, predicting user search patterns can be difficult. For example, they may be looking for gene mutation-specific or phenotype-specific or clinical trial-specific data. It is a challenge to ensure that all these potential connections are accounted for when representing a dataset within a data catalog. In this work, we demonstrate a method for connecting rare disease datasets with a public knowledge graph (KG) that includes a massive curated collection of relationships among biological concepts such as diseases, human anatomy, and genes. We describe a method for translating source data into a format that reuses a common data model and is compatible with the KG, and making datasets searchable via the expanded list of concepts that the KG helps account for.

Key words: knowledge graph; ontology; rare disease.

Background and significance

There are approximately 10 000 known rare diseases that affect about 10% of the population.[1] Rare diseases are underserved in terms of drug development and drug targets, with only about 10% having an FDA-approved drug indicated to treat the disorder. Rare disease data are often siloed, heterogeneous, and not interoperable, making it difficult to harmonize and compare data across different patient groups. Using standardized terminologies, such as ontologies, to structure and aggregate the data makes them more interoperable, compatible with external resources and knowledge, and supports complex querying to identify data for a given use case. This paper describes a collaboration between the Critical Path Institute (C-Path) and the Monarch Initiative[2] aimed at making rare disease data more FAIR[3] (Findable, Accessible, Interoperable, and Reproducible) as part of the Rare Disease Cures Accelerator-Data and Analytics Platform (RDCA-DAP).[4]

RDCA-DAP comprises a data catalog to promote discoverability and accessibility and analysis workspaces. There are currently 63 rare disease datasets available through the RDCA-DAP including patient registries, natural history studies, and clinical trials. Patient data on the RDCA-DAP is anonymized. C-Path utilizes a responsive curation approach where curation is prioritized based on data access requests and requests for specific curation benchmarks such as transformation into a common data model (CDM).

The Monarch Initiative provides the Monarch Knowledge Graph (KG), an aggregated graph representation of biological associations between entities such as diseases, phenotypes, and genes, from numerous data sources.[5] These associations are harmonized using the Biolink model,[6] a high-level schema

Received: November 5, 2024; **Revised:** December 23, 2024; **Editorial Decision:** January 2, 2025; **Accepted:** January 30, 2025

for life science that is intentionally extensible, and terms from Open Biological and Biomedical Ontologies (OBO)[7] as standard representations of concepts. The associations and metadata contained in the Monarch KG are available through a RESTful API[8] that supports interfaces such as entity and association browsers, phenotypic profile tooling, and natural language search interfaces[9] supported by large language models (LLMs).

The Observational Medical Outcomes Partnership (OMOP) CDM[10–12] uses concepts from a defined set of vocabularies, such as the Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT).[13] The OMOP2OBO[14] project produced a set of mappings between concepts in OMOP CDM vocabularies and terms from OBO ontologies, with the majority of the mappings using the Human Phenotype Ontology (HPO)[15] and the Mondo Disease Ontology (Mondo)[16] as the target ontologies. These mappings were created through a variety of methods such as manual curation and automated string matching, and metadata related to the matching approach is included. These mappings include both strict exact semantic matches (eg, a disease term in SNOMED to the equivalent disease term in Mondo) as well as less direct relationships, such as between a laboratory test (measurement concept in the OMOP CDM) and the phenotype it is primarily assessing (eg, blood pressure measurement and hypotension).

We describe a pipeline for representing a subset of the available information in RDCA-DAP datasets mapped to the OMOP CDM using an extension of the Biolink model to support interoperability with the Monarch KG and Monarch API. The purpose of this interoperability is to support tooling for data discovery across CDMs on RDCA-DAP. We demonstrate how this pipeline is used to provide indexed search terms for comprehensive dataset discoverability, and support an interface built on top of the Monarch API and demonstrate semantic search over the datasets requestable on the platform, both at the dataset-metadata level (eg, the diseases that a given dataset is focused on), and at the patient data-level (eg, the phenotypes that a particular patient exhibited and diagnoses that they received).

Materials and methods

Biolink extension

The LinkML modeling language[17] was used to develop an extension of the Biolink schema to incorporate additional associations from patients to phenotypes and diseases, as well as additional dataset-level metadata relevant for the rare disease datasets contributed to and published by C-Path (eg, study type, CDM, disease area, etc.).

OMOP2OBO SSSOM mappings

A Simple Standard for Sharing Ontology Mappings (SSSOM)[18]-compliant mapping file was generated from the OMOP2OBO Condition Occurrence Mappings.[19] This translation included only unambiguous mappings where the mapping metadata could be represented using the SSSOM model and the Semantic Mapping Vocabulary[20] to describe the mapping evidence. Distinct combinations of mapping category, mapping evidence, and ontology logic in the OMOP2OBO mapping set were converted to SSSOM metadata elements (Table S1), to translate the semantics of the original mapping into the SSSOM metadata model where applicable. Simple Knowledge Organization System (SKOS)[21] predicates used were selected to be minimally assertive (eg, exact string matches were assumed to be "close matches," and all others were specified as "related." The Mapping Commons[22] infrastructure was used to create a repository to maintain these mappings and the translation to SSSOM logic from the source spreadsheets.

Data ingestion and transformation

A pipeline was developed to ingest 5 datasets adhering to the OMOP CDM and transform a subset of the data (the PERSON and CONDITION_OCCURRENCE tables) into the Biolink extension model. Conditions using SNOMED concepts were mapped directly to phenotype terms from HPO and/or disease terms from Mondo using the SSSOM OMOP2OBO mapping set (Figure 1). The outputs of this ingestion pipeline are associations between the patients and diseases or phenotypes that are defined by the Biolink extension and compatible with the Monarch KG to support patient-level queries that additionally utilize the Monarch KG or Monarch API and infrastructure (Figure 2).

Patient and dataset profiles

Using the combination of the Monarch KG and the Biolink schema extension described here, patient-level profiles were derived as the set of ontology terms associated with a patient. These profiles include (1) the phenotypes directly associated with a person, (2) the diseases directly associated with a person, (3) any phenotypes associated with those diseases or any more specific subclasses of those diseases, (4) the anatomical entities or systems related to those phenotypes or diseases, and (5) the causal or associated genes related to those phenotypes or diseases. Dataset-level profiles were also generated that include (1) the disease area(s) of the dataset, (2) any phenotypes associated with those diseases or any more specific subclasses of those diseases, (3) the anatomical entities or systems related to those diseases, and (4) the causal or associated genes related to those diseases (Figure 2).

Results

Combined KG

A joint KG using the RDF format was constructed to combine rare disease dataset metadata, the patient-level associations derived from OMOP CDM datasets and SSSOM mappings into the OBO ontologies, and associations from the existing Monarch KG (Figure 1). The SSSOM mapping sets were included in the KG, as metadata on the patient-level associations that can be used to subset these associations, eg, to identify all associations from patients to phenotypes where the HPO term was mapped from a SNOMED concept with predicate "close match" or "exact match" only. The combined KG enables queries spanning information present in both the patient data and the biological associations present in the Monarch KG, eg, finding patients with phenotypes related to a particular gene, or related to any genes involved in a particular biochemical pathway. This graph could additionally serve as a platform for identifying distinct phenotypic clusters, especially as data sources with more diverse sets of phenotypic observations such as electronic health records are incorporated.[23]

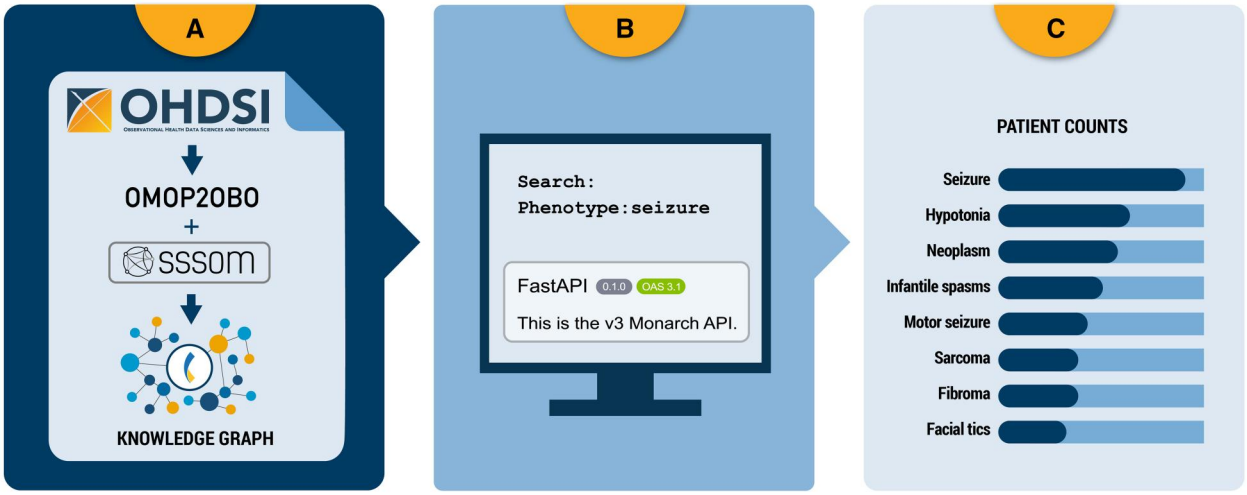


Figure 1. (A) Source data in the OMOP CDM is transformed into an extension of the Biolink model using OMOP2OBO mappings translated into a SSSOM-compliant format. The KG combines the mapped and translated patient-level data with biological association data present in the Monarch KG. (B) A semantic search application matches free-text queries to entities in the Monarch knowledge graph (genes, phenotypes, etc.). The Monarch API is used to resolve text searches to ontology terms, and to calculate semantic similarity between ontology terms. (C) Matching datasets as well as relevant patient counts are returned, for both the queried terms and other associated entities.

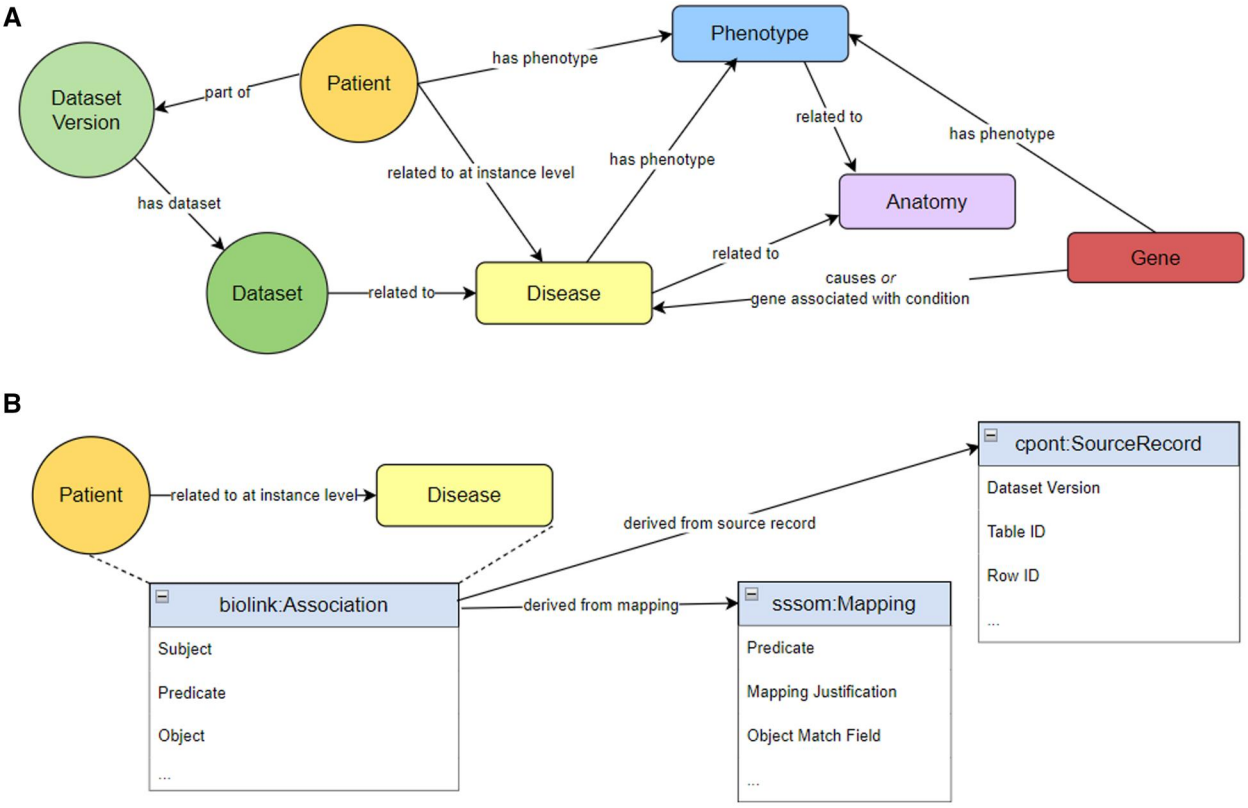
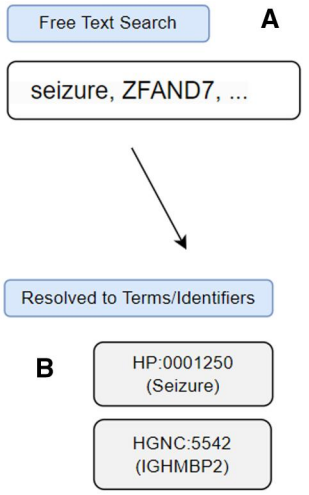


Figure 2. (A) The extension of the Biolink schema models datasets with patients related to phenotypes and diseases, which are additionally connected to other phenotypes, anatomical entities, and genes via associations from the Monarch KG. (B) Associations from patients to phenotypes and diseases are annotated with an extension of the Biolink Association class that includes a reference to the original source record and the SSSOM-compliant mapping used to transform the source object (a SNOMED concept) into a node in the Monarch KG (an HPO or Mondo term).

Semantic search application

In addition to the KG itself, a dashboard was developed to provide a semantic search interface for querying rare disease dataset metadata and obtaining aggregated patient counts based on associations with biological entities. The free-text search supports queries that are diseases, phenotypes, anatomical entities, or genes (Figure 3A). The query can be names, labels, identifiers, or a combination, eg, "respiratory system, IGHMBP2, 5542." The query can also specify the category of each element in the comma-separated list, eg,



C Matching Dataset Disease-Areas

Dataset	Query	Match	Overlapping Concept
Study 1	Seizure (HPO)	Related phenotype(s) of **Prader-Willi syndrome** (Mondo)	Seizure (HPO)
Study 2	Seizure (HPO)	Related phenotype(s) of **fibrous dysplasia** (Mondo)	Abnormal nervous system physiology (HPO)
Study 3	IGHMBP2 (HGNC)	Causal gene(s) of **SMARD1** (Mondo)	IGHMBP2 (HGNC)

D Matching Patient Conditions

Dataset	Query	Match	Overlapping Concept	Count
Study 4	Seizure (HPO)	**Seizure** (HPO)	Seizure (HPO)	N_1
Study 5	Seizure (HPO)	**Atypical behavior** (HPO)	Abnormal nervous system physiology (HPO)	N_2
Study 5	Seizure (HPO)	Related phenotype(s) of **cardiomyopathy** (Mondo)	Seizure (HPO)	N_3
Study 6	IGHMBP2 (HGNC)	Related gene(s) of **Scoliosis** (HPO)	IGHMBP2 (HGNC)	N_4

Figure 3. (A) Free-text search interface supports querying for one or more phenotypes, diseases, anatomical entities, or genes. In the example here, we show a query for a phenotype (seizure) and gene (ZFAND7, an alias of IGHMBP2). (B) Each query element is resolved to an OBO term or gene identifier using the Monarch search API. (C) Datasets with semantically similar disease areas to a queried disease or that are related to queried phenotypes, anatomical entities, or genes are returned. D. The counts of patients with conditions semantically similar (rows 1 and 2) or related to (rows 3 and 4) the queried entities are returned. The bolded text indicates the condition directly associated with a given patient or dataset.

Table 1. Entities present in the knowledge graph.

Entity	Terminology	Biolink category/categories	Distinct entity count
Dataset	Not applicable	Dataset	43
Dataset version	Not applicable	DatasetVersion	53
Patient	Not applicable	Case	1703
Disease	Mondo disease ontology (Mondo)	Disease	201
Phenotype	Human phenotype ontology (HPO)	PhenotypicFeature, PhenotypicQuality	3685
Anatomical entity	Uberon anatomy ontology (UBERON)	AnatomicalEntity, GrossAnatomicalStructure	56
Gene	HUGO gene nomenclature committee (HGNC)	Gene	4141

"anatomy: respiratory system, gene: IGHMBP2." The Monarch API was used to resolve queries to the best matching ontology term in the case of diseases, phenotypes, and anatomical entities, or identifier in the case of genes (Figure 3B). Datasets are returned where that dataset's disease-area is identical or semantically similar to a queried disease, or where the disease-area is associated with other biological entities that are semantically similar to the queried terms (Figure 3C). Similarly, datasets are also returned if the conditions of patients within that dataset are related or semantically similar to the queried terms, and, in this case, the patient counts for that specific condition within a given datasets are also returned (Figure 3D). In all cases, semantic similarity was calculated using the Jaccard similarity metric.

Datasets

The semantic search application is being developed as a containerized application, with plans to integrate this functionality more closely into the RDCA-DAP platform. The dataset-level metadata for each RDCA-DAP dataset was ingested into the KG, enabling discoverability through dataset-level disease area annotations. Five datasets focused on Friedreich's Ataxia, desmoid tumors, and rare epilepsies, respectively, have been mapped into the OMOP CDM and are compatible with the described pipeline and semantic search application. They comprise more than 1500 patients (Table 1).

Discussion

Representing patient conditions using nodes in the Monarch KG (OBO ontology terms) allows researchers to discover datasets with patient data based on connections from wider biological knowledge between those phenotypes and diseases and other biological entities. Integration with a resource like the Monarch KG has the additional advantage that the inferred connections will be automatically expanded and refined. When additional connections are added to the Monarch KG, including to more diverse entity types such as medical interventions, that information will be propagated into downstream discoverability tools such as the RDCA-DAP if they are leveraging these connections.

The approach described here was limited to a subset of the OMOP CDM (patient conditions), because that data incorporates phenotypes and diseases, which are foundational elements of the Monarch KG. However, disease mentions are also present in the OBSERVATION table (where "history of" records are represented), and phenotypes may be indirectly represented in the MEASUREMENT table (eg, a specific blood pressure value indicating the phenotype of hypertension). The inclusion of these additional concepts into the semantic model is possible through projects such as LOINC2HPO,[24] but would require additional specificity within the model, as the presence of a measurement alone does not imply a given phenotype. In addition, there are OBO ontologies such as OBI[25] and MAXO[26] that capture the measurement concepts themselves, and existing associations[27] to Mondo could be leveraged for interoperability with these ontologies.

Conclusion

Discoverability is a critical component in ensuring that rare disease datasets adhere to the FAIR data principles, with the aim of ensuring that researchers, patients, and other stakeholders can find all relevant data for their particular use case or question. Integrating patient data with an in-depth aggregation of biological knowledge like the Monarch KG ensures that the patient-level data contained within a dataset is searchable not only by the concepts directly contained within the data, but also by the context where each of those concepts exist within the framework of broader medical and biological knowledge. The method and search application described in this work demonstrate how CDMs such as the OMOP CDM and standardized semantic models such as SSSOM can be used to connect patient data to the Monarch KG in a richly semantic way that can be used to support enriched semantic search interfaces, paving the way for this increased discoverability on platforms such as the RDCA-DAP.

Acknowledgments

The authors would additionally like to acknowledge the contributions and assistance of James Overton and Tim Putman on this project.

Author contributions

Ian Braun (Conceptualization, Methodology, Visualization, Writing—original draft, Writing—review and editing), Emily Hartley (Conceptualization, Methodology, Writing—review and editing), Daniel Olson (Conceptualization, Methodology, Writing—review and editing), Nicolas Matentzoglu (Conceptualization, Methodology, Writing—review and editing), Kevin Schaper (Software), Ramona Walls (Conceptualization, Project administration, Supervision, Writing—review and editing), and Nicole Vasilevsky (Conceptualization, Project administration, Supervision, Writing—review and editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

Critical Path Institute is supported by the Food and Drug Administration (FDA) of the Department of Health and Human Services (HHS) and is 56% funded by the FDA/HHS, totaling $23 740 424, and 44% funded by non-government source(s), totaling $18 881 611. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the US Government.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The Monarch KG files in the N-Triples (.nt) format used in this project are available at https://data.monarchinitiative.org/monarch-kg/latest/index.html. The mappings referenced in this project are available at https://zenodo.org/records/7250177. Data access requests can be made through the RDCA-DAP portal available at: https://portal.rdca.c-path.org.

References

1. Haendel M, Vasilevsky N, Unni D, et al. How many rare diseases are there? *Nat Rev Drug Discov*. 2020;19:77-78. https://doi.org/10.1038/d41573-019-00180-y
2. Putman TE, Schaper K, Matentzoglu N, et al. The monarch initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Res*. 2024;52:D938-D949. https://doi.org/10.1093/nar/gkad1082
3. Wilkinson MD, Dumontier M, Aalbersberg I, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
4. RDCA-DAP Portal. Accessed October 18, 2024. https://portal.rdca.c-path.org/
5. Overview—Monarch Ingest Documentation. Accessed October 18, 2024. https://monarch-initiative.github.io/monarch-ingest/Sources/
6. Unni DR, Moxon SAT, Bada M, et al.; Biomedical Data Translator Consortium. Biolink model: a universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical Translational Sci*. 2022;15:1848-1855. https://doi.org/10.1111/cts.13302
7. Jackson R, Matentzoglu N, Overton JA, et al. OBO foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database (Oxford)*. 2021;2021:baab069. https://doi.org/10.1093/database/baab069
8. Monarch API. Accessed October 18, 2024. https://api.monarchinitiative.org/v3/docs
9. O'Neil ST, Schaper K, Elsarboukh G, et al. Phenomics assistant: an interface for LLM-based biomedical knowledge graph exploration. bioRxiv. Published Online First: February 2, 2024. https://doi.org/10.1101/2024.01.31.578275
10. Reich C, Ostropolets A, Ryan P, et al. OHDSI standardized vocabularies—a large-scale centralized reference ontology for international data harmonization. *J Am Med Inf Assoc*. 2024;31:583-590. https://doi.org/10.1093/jamia/ocad247
11. Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012;19:54-60. https://doi.org/10.1136/amiajnl-2011-000376
12. OMOP Common Data Model. Accessed October 18, 2024. https://ohdsi.github.io/CommonDataModel/
13. SNOMED CT. Accessed December 16, 2024. https://www.nlm.nih.gov/healthit/snomedct/index.html
14. Callahan TJ, Stefanski AL, Wyrwa JM, et al. Ontologizing health systems data at scale: making translational discovery a reality. *NPJ Digit Med*. 2023;6:89. https://doi.org/10.1038/s41746-023-00830-x
15. Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021. *Nucleic Acids Res*. 2021;49:D1207-D1217. https://doi.org/10.1093/nar/gkaa1043
16. Vasilevsky NA, Matentzoglu NA, Toro S, et al. Mondo: unifying diseases for the world, by the world. medRxiv. Published Online First: April 16, 2022. https://doi.org/10.1101/2022.04.13.22273750
17. linkml/linkml: Linked Open Data Modeling Language. Accessed October 18, 2024. https://github.com/linkml/linkml
18. Matentzoglu N, Balhoff JP, Bello SM, et al. A simple standard for sharing ontological mappings (SSSOM). *Database (Oxford)*. 2022;2022:baac035. https://doi.org/10.1093/database/baac035
19. Callahan TJ, Wyrwa JM, Vasilevsky NA, et al. OMOP2OBO Condition Occurrence Mappings. 2022. Accessed October 18, 2024. https://zenodo.org/record/7250177
20. mapping-commons/semantic-mapping-vocabulary. 2024. Accessed October 18, 2024. https://github.com/mapping-commons/semantic-mapping-vocabulary
21. SKOS Simple Knowledge Organization System Reference. Accessed October 18, 2024. https://www.w3.org/TR/skos-reference/
22. mapping-commons/mapping-commons.github.io: Repo for the user facing documentation of mapping-commons. Accessed October 18, 2024. https://github.com/mapping-commons/mapping-commons.github.io
23. Reese JT, Blau H, Casiraghi E, et al.; RECOVER Consortium. Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. *eBioMedicine*. 2023;87:104413. https://doi.org/10.1016/j.ebiom.2022.104413
24. Zhang XA, Yates A, Vasilevsky N, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *NPJ Digit Med*. 2019;2:32. https://doi.org/10.1038/s41746-019-0110-4
25. Bandrowski A, Brinkman R, Brochhausen M, et al. The ontology for biomedical investigations. *PLoS One*. 2016;11:e0154556. https://doi.org/10.1371/journal.pone.0154556
26. monarch-initiative/MAxO. 2024. Accessed October 18, 2024. https://github.com/monarch-initiative/MAxO
27. monarch-initiative/MAxO—Releases including Mondo annotations. GitHub. Accessed October 18, 2024. https://github.com/monarch-initiative/MAxO/releases

JAMIA Open, 2025, 8, 1–6
https://doi.org/10.1093/jamiaopen/ooaf001
Brief Communications