Using Mondo to assemble rare disease cohorts in OMOP

Bryan Laraway, MS, Eric Hurwitz, PhD, Blake Byer, Daniel Korn, PhD, Megan Pearson, BS, Evan Connelly, BS, Sabrina Toro, PhD, Melissa Haendel, PhD
University of North Carolina at Chapel Hill

Background

The first step in working with electronic health record (EHR) data, including OMOP datasets, is evaluating whether cohorts of interest have a sufficient sample size for meaningful research and analysis. While creating an OMOP concept set for even a single disease, such as 'heart failure', can pose a significant challenge, identifying rare disease populations in an OMOP instance can present an even greater challenge. This is due to (1) the lack of coding systems for rare diseases in the source, (2) small numbers of patients for any given rare disease, (3) heterogeneous or missing mappings between rare diseases and OMOP concepts, and (4) lack of concept IDs in OMOP for most rare diseases. It has been estimated that ~80% of rare diseases do not have an ICD code, and this has resulted in additional bottlenecks to rare disease diagnosis and research, so much so that patient advocacy groups have created guides to help patients get their disease a code.(1,2)

Here, we propose a simple solution to assembling rare disease cohorts, including the ability to subset rare disease cohorts by rare disease authority, such as the National Organization for Rare Disorders (NORD) or Orphanet. The Mondo disease ontology is the result of ~5 years of reconciliation of rare disease concepts in partnership with several rare disease resources and coordinated by the Monarch Initiative (3). Mondo computationally and conceptually aligns rare disease terminologies to generate a coherent merged ontology that can point users to equivalent concepts in the sources, negating the need to navigate the incredibly complex landscape of rare disease resources with poorly provisioned mappings and often non-overlapping content.(4)

We present the process of mapping from Mondo's precise rare disease terminology to terminologies in OMOP's source vocabularies, and how these mappings can be used to "roll up" patient cohorts to intermediate rare disease concepts for querying within and across OMOP instances.

Methods

We utilized the Knowledge Graph Exchange (KGX) format version of Mondo (available from Knowledge Graph Hub, https://kghub.org/, version 2025-04-01)(5), which presents the ontology as flattened 'edges' and 'nodes' TSV files. Mondo is also available in other formats (OWL, OBO, JSON) to fit user needs, including a subset that includes only rare diseases and an international edition, available at https://mondo.monarchinitiative.org/.(6)

After first filtering on 'biolink:Disease' concepts, we utilized NetworkX(7) to select Mondo terms which are descendants of 'human disease' (MONDO:0700096)(**Figure 1**). We also removed all obsolete terms. We utilized Mondo's exact matches to other terminologies and filtered these mappings to terminologies included as OMOP vocabularies (ICD10-CM, MeSH, and SNOMED), generating one row per mapped terminology code. The terminology code and associated vocabulary were used to link to the OMOP Concept table, and were linked to the Concept_relationship table to identify standard OMOP concepts.

The final resulting Mondo-to-OMOP table can be utilized to query the Condition Occurrence table in any OMOP instance to quickly identify cohorts with the corresponding (rare) disease. Additionally, the hierarchical nature of Mondo allows for the rolling up of these cohorts to higher level disease classes. Conversely, a Mondo term could be used to identify the most relevant OMOP concept(s) for a precisely defined rare disease. Example code for generating the Mondo-to-OMOP table will be made available on GitHub.(8)

To illustrate the utility of Mondo in cohort identification, including the ability to roll up to higher-level disease classes, we connected the Mondo-to-OMOP table to the N3C OMOP instance(9) to identify patient cohorts that can be used for further analysis of COVID outcomes.

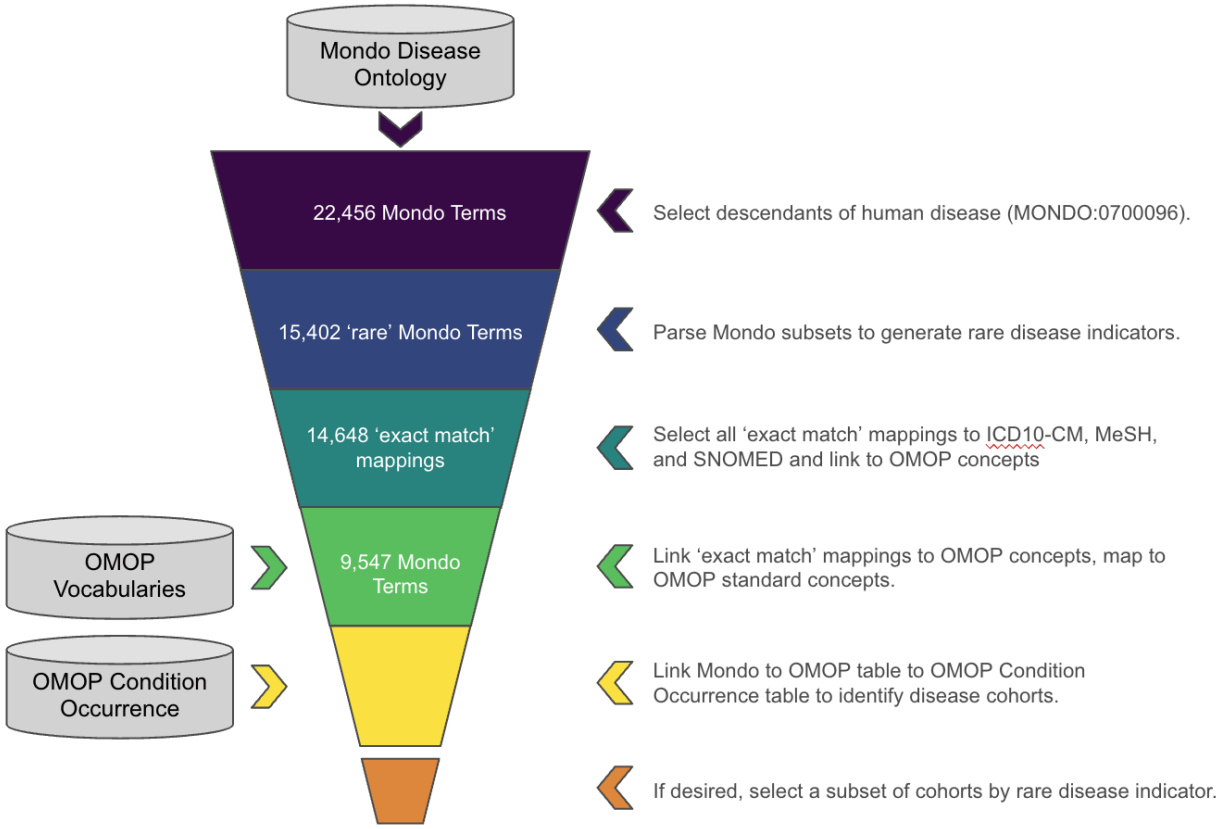

Figure 1. Data flow diagram for connecting the Mondo Disease Ontology to an OMOP instance.

Results

After performing the filtering process described above, 22,456 Mondo terms remained (including non-terminal concepts)(**Table 1**). Selecting for Mondo terms with mappings to SNOMED, ICD10-CM, or MeSH resulted in 14,648 mappings to OMOP concepts, where 9,547 Mondo terms are mapped onto 10,175 distinct standard OMOP concepts within the Condition domain (33 additional mappings are available for 32 Mondo terms when including other OMOP domains). These mappings were composed of 8,792 SNOMED codes, 4,802 MeSH codes, and 1,054 ICD10-CM codes. Using the rare disease indicators in the Mondo-to-OMOP table we can select subsets of rare diseases by rare disease authority. While

6,139 rare disease terms could be mapped to OMOP, over 9,000 rare diseases in Mondo are not mappable to any OMOP concept (**Figure 2**).

Table 1: Counts of Mondo terms by rare disease designation, and the counts of Mondo terms with available mappings to codes currently included in OMOP vocabularies. Inferred counts are based on term parentage.

	Mondo term count	Mondo term count with OMOP mapping
Mondo human disease terms	**22,456**	**9,547 (42.5% of all human Mondo terms)**
Any rare designation	**15,402**	**6139 (39.9% rare Mondo terms)**
GARD rare	**15,013**	**6055 (40.3% rare Mondo terms)**
NORD rare	**6885**	**2603 (37.8% rare Mondo terms)**
Orphanet rare	**6164**	**3966 (64.3% rare Mondo terms)**
Inferred rare	**272**	**56 (20.6% rare Mondo terms)**
Mondo rare	**12**	**5 (41.7% rare Mondo terms)**

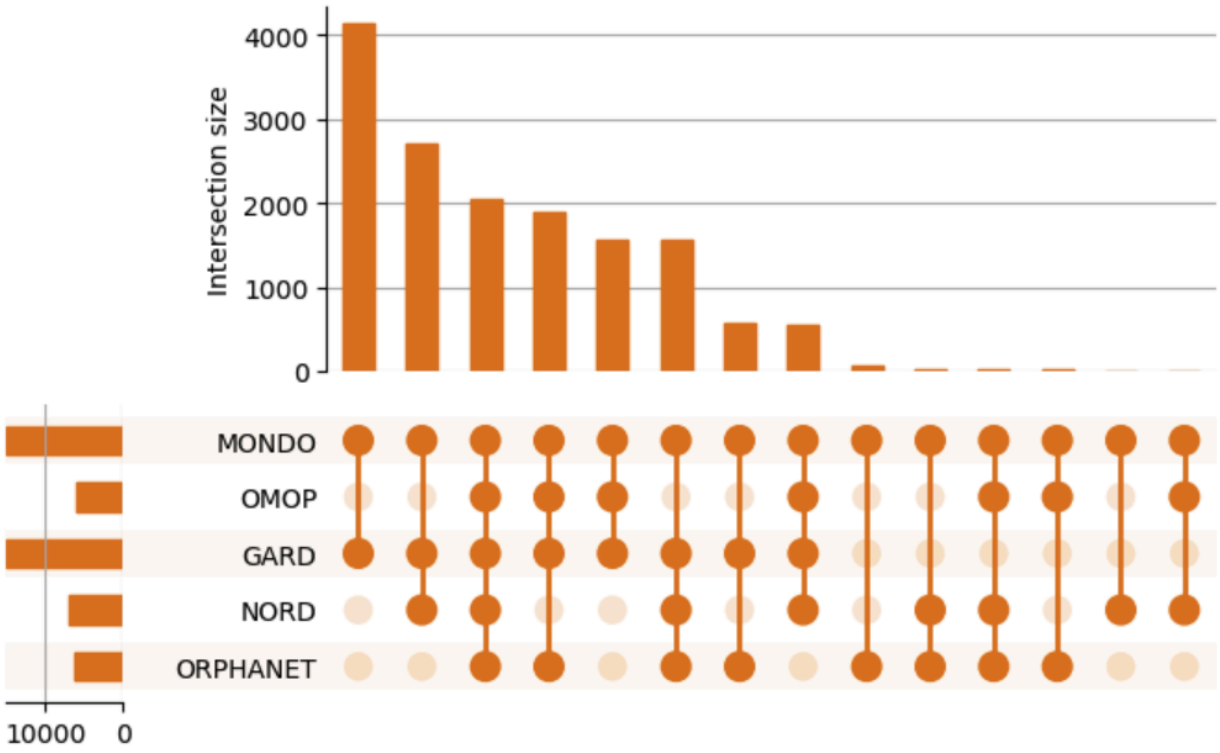

Figure 2. Mondo rare disease terms and their presence in different knowledge sources

(GARD/NORD/Orphanet) and mappings to OMOP concepts (OMOP). The left horizontal bars indicate the total number of Mondo terms connected to each resource. Of the 15,402 Mondo rare disease terms, only 6,139 terms could be mapped to OMOP using an OMOP vocabulary (ICD10-CM, SNOMED, MeSH).

Upon connecting our Mondo-to-OMOP mapping table to N3C's data enclave, we were able to identify 3,335 rare disease cohorts for further study of outcomes of SARS-CoV-2 infection. While over half (1,623) of these rare disease cohorts were too small for further analysis (<20 patients), these cohorts could be rolled up to higher-level disease classes for analysis, as shown below (**Figure 3**).

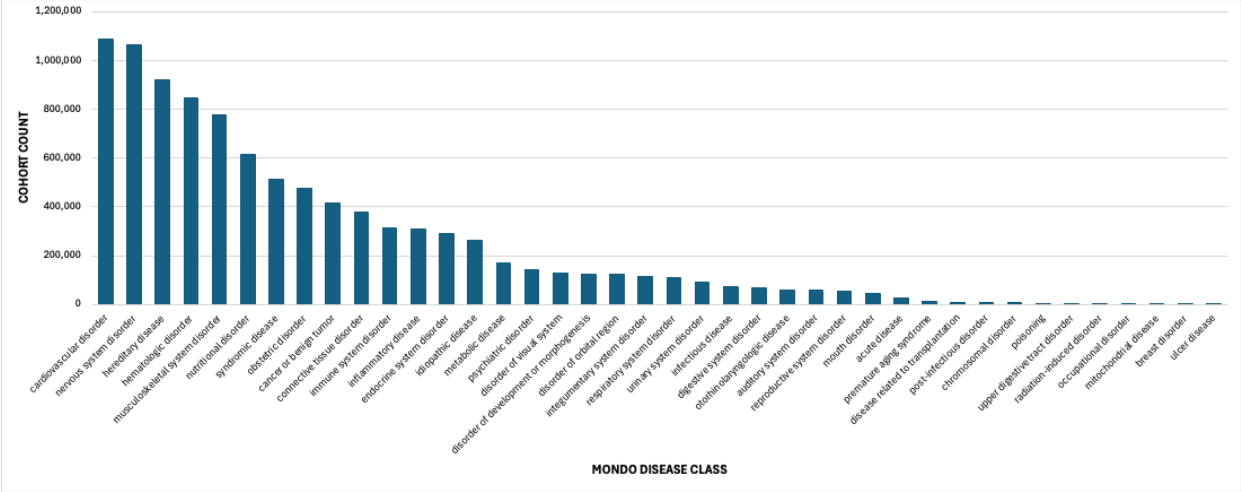

Figure 3. Using Mondo to identify and roll up patient cohorts to higher-level disease classes

Conclusion

We describe a simple and rigorous process for utilizing the Mondo disease ontology together with OMOP to identify rare disease cohorts, with applications for all diseases. More than 9K rare diseases cannot be directly mapped to an OMOP concept currently due to the lack of inclusion of rare diseases in OMOP and its source vocabularies. Most recently, IMO has incorporated Mondo as an EHR interface terminology used directly by clinicians, and can now provision more precise rare disease concepts for a subset of Mondo rare disease terms during the clinical encounter.(10) Other efforts are underway to integrate Mondo with ICD and into EHR systems in partnership with WHO and CDC. However, while we wait for more accurate rare disease codes to percolate into the OHDSI ecosystem, the use of the Mondo mapping and roll-up strategy presented here can support the identification of rare disease cohorts across OMOP instances, improving research and encouraging further investigation for specific rare diseases.

We strongly recommend incorporating Mondo as a first-class vocabulary in the OMOP Standardized Vocabulary collection and look forward to partnering with the new rare disease OHDSI working group.

Acknowledgements

Supported by the NCATS Center for Data to Health #U24TR002306, NHGRI Phenomics First award #RM1HG010860, and The Monarch Initiative NIH Office of Director #5R24OD011883.

References

1. McMurry J, Sizer A, Allaway R, Boerkoel C, Chen AJ, Colquitt J, et al. Critical bottlenecks in Rare Disease research and care: A community perspective [Internet]. Zenodo; 2025. Available from: https://doi.org/10.5281/zenodo.14907384

2. EveryLife Foundation for Rare Diseases [Internet]. 2021 [cited 2025 Jun 30]. ICD Code Roadmap. Available from: https://everylifefoundation.org/icd-code-roadmap/

3. Vasilevsky NA, Matentzoglu NA, Toro S, Flack JE, Hegde H, Unni DR, et al. Mondo: Unifying diseases for the world, by the world [Internet]. medRxiv. 2022 [cited 2025 Jun 26]. p. 2022.04.13.22273750. Available from: https://www.medrxiv.org/content/10.1101/2022.04.13.22273750v1.abstract

4. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nature reviews Drug discovery. 2020 Feb;19(2):77.

5. Caufield JH, Putman T, Schaper K, Unni DR, Hegde H, Callahan TJ, et al. KG-Hub-building and exchanging biological knowledge graphs. Bioinformatics [Internet]. 2023 Jul 1;39(7). Available from: http://dx.doi.org/10.1093/bioinformatics/btad418

6. Mondo Disease Ontology [Internet]. [cited 2025 Jun 30]. Available from: http://mondo.monarchinitiative.org/

7. Exploring Network Structure, Dynamics, and Function Using NetworkX. 2008.

8. mondo2omop: This repository will provide example code for creating a Mondo to OMOP mapping table that can be used to connect Mondo to an OMOP instance and generate Mondo patient cohorts [Internet]. Github; [cited 2025 Jul 1]. Available from: https://github.com/monarch-initiative/mondo2omop

9. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021 Mar 1;28(3):427–43.

10. Patrick M. IMO Health. 2025 [cited 2025 Jul 1]. Mondo 101: Your guide to clinical terminology for rare diseases. Available from: https://www.imohealth.com/resources/mondo-101-your-guide-to-clinical-terminology-for-rare-diseases/