Mondo: Unifying diseases for the world, by the world

Nicole A Vasilevsky, University of Colorado Anschutz Medical Campus; Nicolas A Matentzoglu, Semanticly Ltd; Sabrina Toro, University of Colorado Anschutz Medical Campus; Joseph E Flack IV, Johns Hopkins University; Harshad Hegde, Lawrence Berkeley National Lab; Deepak R Unni, Lawrence Berkeley National Laboratory; Gioconda F Alyea, GARD; Joanna S Amberger, Johns Hopkins University School of Medicine; Larry Babb, Broad Institute; James P Balhoff, Renaissance Computing Institute, University of North Carolina; Taylor I Bingaman, Geisinger Autism & Developmental Medicine Institute; Gully A Burns, Chan Zuckerberg Initiative Foundation; Orion J Buske, PhenoTips; Tiffany J Callahan, Columbia University; Leigh C Carmody, Jackson Laboratory for Genomic Medicine; Paula Carrio Cordo, University of Zurich; Lauren E Chan, Oregon State University; George S Chang, National Cancer Institute, EVS Program; Sean L Christiaens, Ontoforce NV; Michel Dumontier, Institute of Data Science; Laura E Failla, National Institute of Allergy and Infectious Diseases, NIH; May J Flowers, Invitae; H. Alpha Garrett, Jr. MD, National Cancer Institute, EVS Program; Jennifer L Goldstein, University of North Carolina at Chapel Hill; Dylan Gration, Western Australian Register of Developmental Anomalies; Tudor Groza, European Molecular Biology Laboratory, European Bioinformatics Institute; Marc Hanauer, INSERM, Orphanet US14; Nomi L Harris, Lawrence Berkeley National Laboratory; Jason A Hilton, Stanford University; Daniel S Himmelstein, Related Sciences; Charles Tapley Hoyt, Harvard Medical School; Megan S Kane, National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health; Sebastian Köhler, Ada Health GmbH; David LAGORCE, INSERM, Orphanet US14; Abbe Lai, Boston Children's Hospital; Martin Larralde, European Molecular Biology Laboratory; Antonia Lock, EMBL-EBI; Irene López Santiago, Open Targets; Donna R Maglott, NCBI/NLM/NIH; Adriana J Malheiro, NIH / NLM / NCBI; Birgit H M Meldal, was: European Molecular Biology Laboratories - European Bioinformatics Institute; Monica C Munoz-Torres, University of Colorado Anschutz Medical Campus; Tristan H Nelson, Geisinger; Frank W Nicholas, The University of Sydney, Sydney School of Veterinary Science; David Ochoa, EMBL-EBI; Daniel P Olson, Critical Path Institue; Tudor I Oprea, University of New Mexico Health Sciences Center; David Osumi-Sutherland, European Bioinformatics Institute (EMBL/EBI); Helen Parkinson, European Bioinformatics Institute, EMBL-EBI; Zoë May Pendlington, European Bioinformatics Institute (EMBL-EBI); Ana Rath, INSERM, Orphanet US14; Heidi L Rehm, Massachusetts General Hospital and Broad Institute of MIT and Harvard; Lyubov Remennik, National Cancer Institute, EVS Program; Erin R Riggs, Geisinger; Paola Roncaglia, EMBL-EBI; Justyne E Ross, University of North Carolina at Chapel Hill; Marion F Shadbolt, Australian BioCommons; Kent A Shefchek, Helix; Morgan N Similuk, National Institute of Allergy and Infectious Diseases, NIH; Nicholas Sioutos, National Cancer Institute, EVS Program; Damian Smedley, Queen Mary University of London; Rachel Sparks, National Institute of Allergy and Infectious Diseases, NIH; Ray Stefancsik, European Bioinformatics Institute, EMBL-EBI; Ralf Stephan; Andrea L Storm, National Center for Advancing Translational Sciences, National Institutes of Health / Genetic and Rare Diseases Information Center; Doron Stupp, Hebrew University of Jerusalem; Gregory S Stupp, The Scripps Research Institute; Jagadish Chandrabose Sundaramurthi, The Jackson Laboratory for Genomic Medicine; Imke Tammen, The University of Sydney, Sydney School of Veterinary Science; Darin Tay; Courtney L Thaxton, University of North Carolina, Chapel Hill; Eloise Valasek, Jewish General Hospital; Jordi Valls-Margarit, MedBioInformatics Solutions; Alex H Wagner, Nationwide Children's Hospital; Danielle Welter, University of Luxembourg; Patricia L Whetzel, European Bioinformatics Institute (EMBL-EBI); Lori L Whiteman, National Cancer Institute, EVS Program; Valerie Wood, University of Cambridge, Department of Biochemistry; Colleen H Xu, The Scripps Research Institute; Andreas Zankl, The University of Sydney; Xingmin Aaron Zhang, The Jackson Laboratory for Genomic Medicine; Christopher G Chute, Johns Hopkins University; Peter N Robinson, The Jackson Laboratory; Christopher J Mungall, Lawrence Berkeley National Laboratory; Ada Hamosh, Johns Hopkins University; Melissa A Haendel, University of Colorado Anschutz Medical Campus

Abstract

There are thousands of distinct disease entities and concepts, each of which are known by different and sometimes contradictory names. The lack of a unified system for managing these entities poses a major challenge for both machines and humans that need to harmonize information to better predict causes and treatments for disease. The Mondo Disease Ontology is an open, community-driven ontology that integrates key medical and biomedical terminologies, supporting disease data integration to improve diagnosis, treatment, and translational research. Mondo records the sources of all data and is continually updated, making it suitable for research and clinical applications that require up-to-date disease knowledge.

Introduction

In the past decade, there have been major advances in computational approaches to disease diagnosis and care management. However, the reference data on which these tools depend are not only heterogeneous and disaggregated, but also growing and ever changing. Standard

terminologies such as the Human Phenotype Ontology [1], the Online Mendelian Inheritance in Man [2], and Orphanet [3] have helped standardize medical terminology. However, reconciling the very many terminologies used to name diseases and their inherent meaning has continued to be challenging, making knowledge and data integration challenging. It is critical to develop an unambiguous resource for disease name reconciliation such that evidence can be accurately gathered on individuals with these diseases and inform their diagnosis, care, and treatment. This allows related resources such as gene, variant, and infectious agent resources to be interoperable and contribute to the ongoing building of medical knowledge bases. Dozens of terminological disease resources used for research and clinical applications exist [4,5], including for Mendelian diseases, common diseases, rare diseases, cancer, and infectious diseases, with others being more comprehensive and broad [2,3,6–8]. However, scope and classification are just the beginning of the ways these resources differ: additional differences include disease naming conventions, synonym encoding, cross references, and more. As a result, each terminology has different strengths and weaknesses. These resources partially overlap, often significantly [9,10]. The correspondence (mapping) among individual concepts is often accomplished through text-matching; for example, the terms 'muscular pseudohypertrophy-hypothyroidism syndrome' and 'B-cell immunodeficiency-limb anomaly-urogenital malformation syndrome' both have the exact synonym 'Hoffman syndrome', but they are entirely different diseases. Human declared mappings are represented as "cross-references", but the relationship between the two terms can be non-exact, incorrect, out of date, or otherwise not clearly defined. For example, the concept DOID:8923 'skin melanoma' cross references both OMIM:608035 'melanoma, cutaneous malignant, susceptibility to, 4' and OMIM:612263 'melanoma, cutaneous malignant, susceptibility to, 7' [11], which are two different types of susceptibility rather than a neoplasm. Further, some disciplines of medicine are not well covered by terminologies, for example pharmacogenetics. Therefore, the resulting integration across disease resources is often incomplete, inconsistent, and unreliable for diagnostics or research.

The figure of 7,500 rare diseases [12] is often quoted; however, in our systematic analysis across resources, we identified over 10,500 unique rare diseases [13]. Much of this heterogeneity results from a lack of consensus, both philosophically and practically, about how to classify diseases. Should diseases be classified based upon the anatomical structures they affect? Based on the doctor that first described them, such as 'Batten disease'? Or based upon their pathogenic mechanism (e.g. infectious, deficiency; hereditary, physiological)? What if two variants in the same gene give rise to different suites of phenotypic features; are those the same disease? The ClinGen "Lumping and splitting" group (https://clinicalgenome.org/working-groups/lumping-and-splitting/) has undertaken the development of curation rules to help inform such decisions, and Orphanet has set standard procedures as well[14] but the community still lacks a comprehensive, multiple classification of diseases that takes into account many other features such as treatment, onset, environmental factors, to name a few. Furthermore, standard clinical enterprise terminologies such as SNOMED-CT or ICD-11 are not released often enough to keep up to date with constantly-changing disease knowledge, and don't include rarer disease codes. Combined with slow code adoption and miscoding, it continues to be very challenging to identify patients with a given disease in Electronic Health Record systems. Furthermore, numerous clinical and research systems, such as laboratory variant pipelines and repositories such as ClinVar, require up-to-date disease information. At the time the report is written or the data submitted, the disease entities should have identifiers that can be reconciled across sources or over time as knowledge changes. Fundamentally, a mechanism is needed to computationally harmonize disease classifications in order to best take advantage of our collective disease knowledge and heterogeneous data assets. This requires a modern, granular, and interoperable approach to support improved coding that can take into account the community-developed, dynamic knowledge about diseases. Here we introduce the Mondo resource, which provides a sustainable

and fully-provenanced approach to integrating disease concepts from numerous sources across disease categories with the goal of better supporting precision medicine, diagnostics, and mechanistic disease research [15].

Terminology systems (or coding systems) often come in the form of taxonomies (simple classifications) or ontologies (conceptual domain models). The utilization of an ontology for biomedical knowledge representation enables data integration and navigation of large amounts of heterogeneous data. Additionally, an ontology encodes hierarchical and other relationships and definitions, which supports modern computational methods. Mondo includes multiple parentage, meaning concepts can be classified in multiple ways, which allows for more sophisticated querying and analytics **(Figure 1)**. For example, adult Refsum disease is a type of 'neurometabolic disease' and 'phytanoyl-CoA hydroxylase deficiency'. The new computationally friendly ICD11 incorporates a pragmatic mechanism for post-coordinating terms and concepts to accommodate the granular detail of complex clinical contents.[16,17]

Mondo currently harmonizes knowledge from 17 disease resources, collectively representing approximately 90,000 source concepts, and merges them into 22,157 distinct disease concepts. These resources were selected based on their scope, strengths, and usage (https://mondo.monarchinitiative.org/pages/sources/).

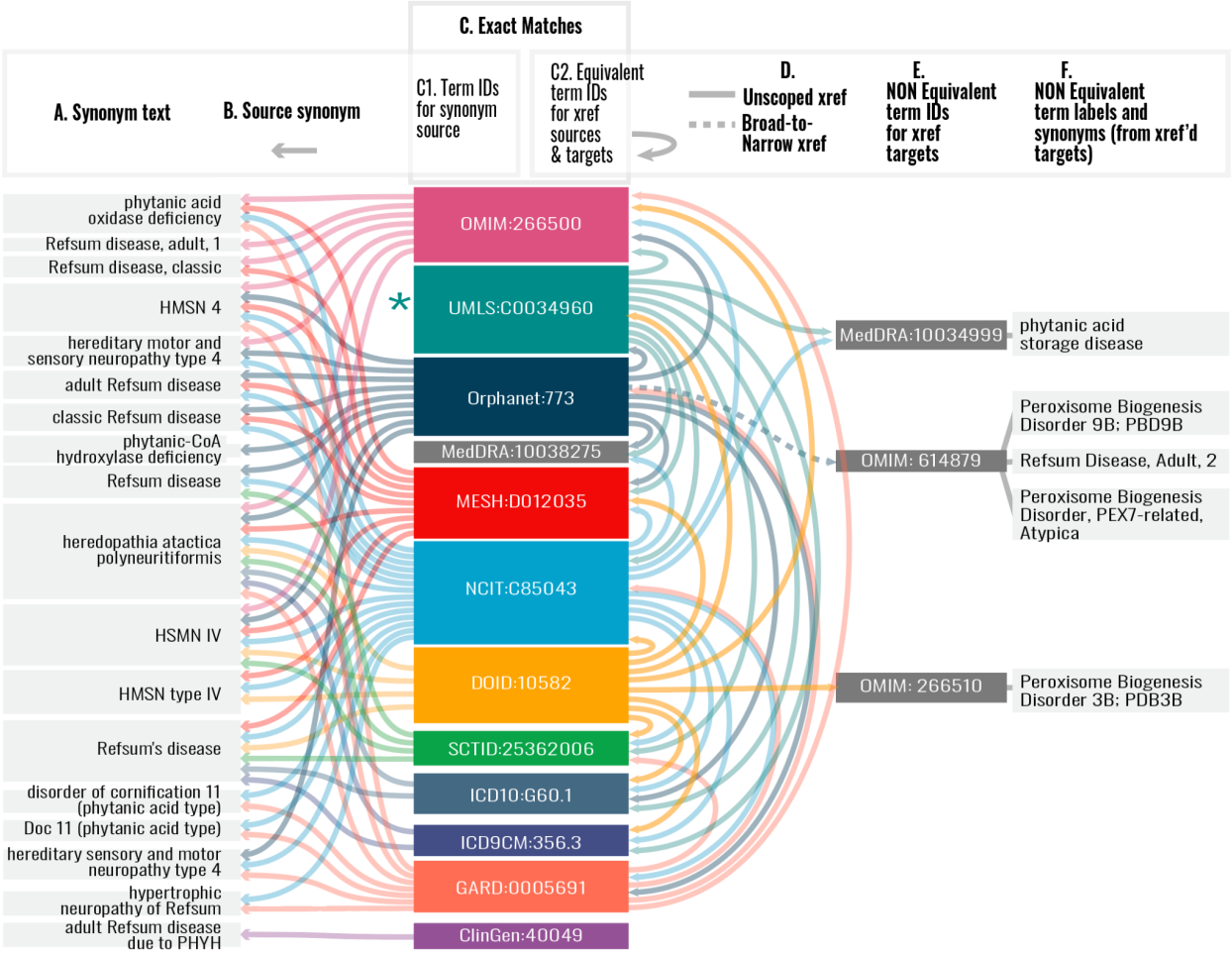

Figure 1: Aligning disease knowledge across sources: adult Refsum disease example. A-F. A Mondo term contains synonyms scoped as exact (shown), narrow, broad and related (not shown), database cross-

references (dbxref) to the source ontologies and terminologies. A. Exact synonyms for adult Refsum disease. B. The provenance for the source of exact synonyms is captured as a database cross-reference in the Mondo ontology file. C1. Representation of the identifiers (IDs) of synonym sources, which are also database cross-references for this Mondo term. C2. Representation of the mappings between source terms and other sources. For example, UMLS:C0034960 maps to OMIM:266500. D. A solid line represents an unscoped mapping (database cross-reference, i.e. the semantics of the mapping is not defined). A dotted line represents a broad (more general) to narrow (more specific) mapping. For example, Orphanet:773 is broader than OMIM:614879. E. Represents mappings between the source term to another external term that we reviewed and determined that they are not equivalent but there is no way in the source ontologies to determine that based on the information given in the source. F. Term labels for IDs shown in E. *UMLS pulls in the synonyms that are referenced by its cross-referenced neighbors (not shown). This is a subset of the mappings and does not reflect all of the mappings that exist in all of these sources.

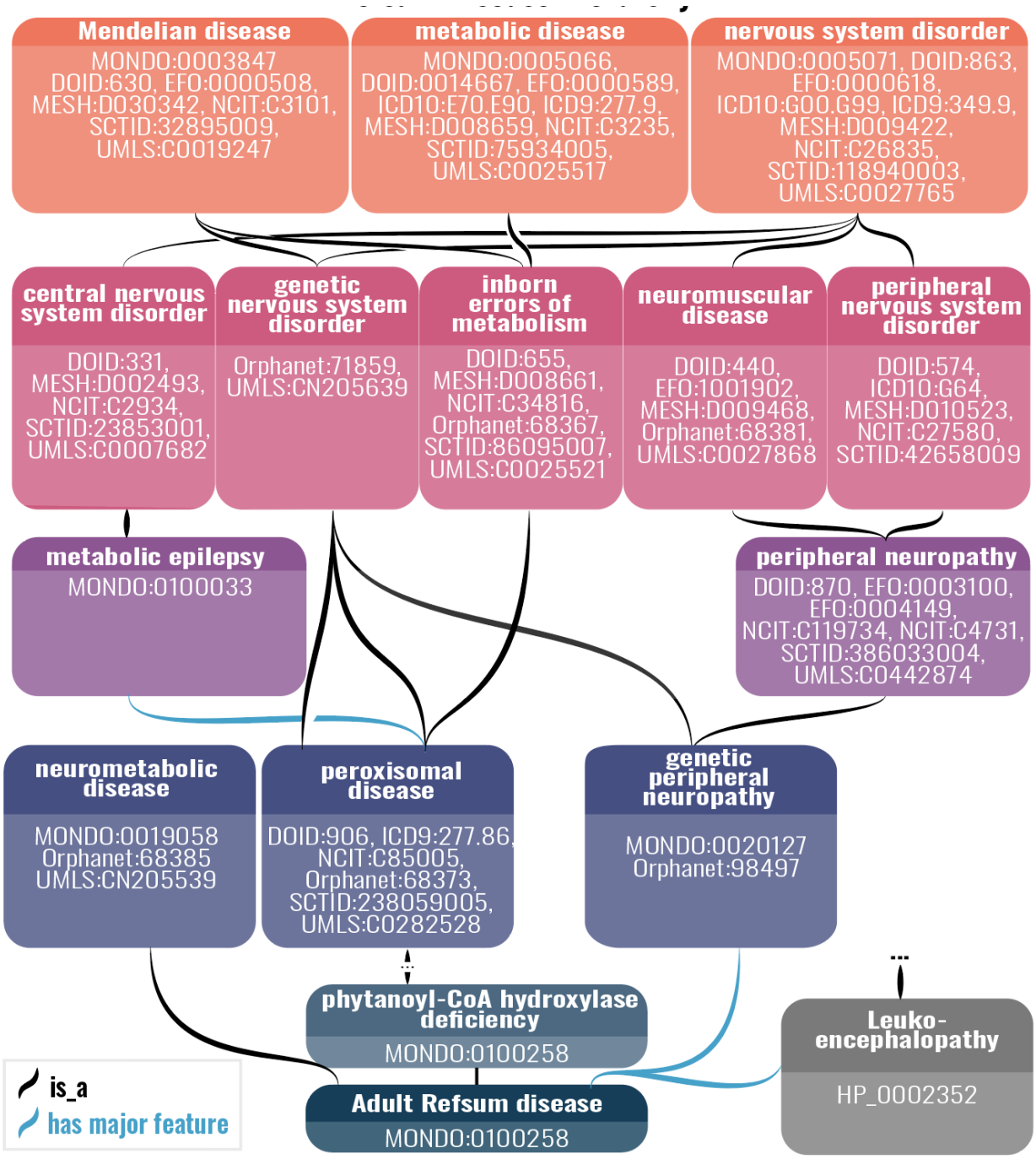

Figure 2: Hierarchical classification of adult Refsum disease. Mondo terms are classified in a hierarchy and can have multiple parentage, i.e., a class can have more than one parent term. Example classification of adult Refsum disease. Relationships between terms can be defined in a subclass of relationship (is a), or via additional relationships, such as 'has major feature', where a phenotype or associated disease is a feature of that disease. Each of these parent classes is similarly complex with dbxrefs hailing from 10 terminologies.

The Human Genome Nomenclature Committee (HGNC) standardizes human gene naming, but there is no comparable global standard for reconciling the heterogeneity of disease naming systems nor their semantic interoperability. Disease names vary — not only by language and

region — but also over time due to social norms and improved understanding of underlying pathogenic mechanisms; moreover, different stakeholders that speak the same language may still prefer different names for different reasons. As a consequence, it is vital to have reliable disease identifiers which durably refer to the same concept over time — accommodating both changes and preferences. Mondo functions as a broker for disease nomenclature; the disease names and respective identifiers are a handle (i.e. stable reference), whereby synonyms, related knowledge and definitions can evolve over time with full provenance. Mondo supports multiple synonyms and synonym types as well as annotating which labels are preferred by which groups. In Mondo, synonyms are classified as exact, broad, narrow, and related [18]. Mondo aims to accommodate all community requests and prioritizes community and medical expert recommendations for naming. More details about disease naming in Mondo is available here (https://mondo.monarchinitiative.org/pages/disease-naming/).

Articulating similarities between concepts across a set of ontologies or terminologies is challenging and often unreliable due to the prevailing use of purely automated approaches (such as text matching). Such methods lack context and can match concepts incorrectly; moreover a lack of declared rules, provenance, and versioning for these mappings makes them difficult to use for computational purposes. Mondo contains precise semantic mappings between source ontologies and terminologies, such as between OMIM, ICD10-CM, Orphanet, the National Cancer Institute Thesaurus (NCIt), and many others [19]. A computational strategy that predicts equivalency based on a variety of features - such as labels, synonyms, cross-references (including existing semantics such as those provisioned by Orphanet), graph structure, and priors that indicate classification features specific to each source - was first applied to generate a set of mappings between concepts [20]. The output of this computational equivalency assessment was reviewed by dedicated curation and technical teams and by the Mondo user community. Introduction of new concepts and subsequent refinements to the hierarchy and mappings are then carried out as needed.

Mondo leverages a wealth of expert knowledge and authoritative terminologies to create a resource that is optimized for computational use in diagnostic, clinical, and research applications. Released on a monthly cycle, Mondo is iteratively developed to meet the evolving needs and input of a diverse, global community of contributors. There are currently more than 100 clinical and domain expert contributors from over 25 institutions that help evolve the resource, including ClinGen [21], OMIM [22], GARD [23], Orphanet [24], and others. Mondo also has a rich community of users that have implemented Mondo in a variety of settings, including incorporation into standards, such as in the Global Alliance for Genomics and Health (GA4GH) [25] and ISO standards such as Phenopackets [26], in the HL7 Terminology Authority [27], use in tools and data management programs, as well as in databases such as ClinGen [21], MedGen [28], Gabriella Miller Kids First [29], Pharos [30], and many others. A full list of users (https://mondo.monarchinitiative.org/pages/users/) and contributors (https://mondo.monarchinitiative.org/pages/contributors/) are available.

Mondo is more than just a source of robust and reproducible mappings between disease terminologies; Mondo also includes n-of-1, rare diseases, environmentally-influenced diseases, and complex genetic diseases that may not be documented in other sources can be added to Mondo and partitioned out for different uses. By integrating knowledge fully provenanced from the many existing and ever-evolving disease resources, acquired through years of work by researchers, clinicians, terminologists, and scientists from around the world, Mondo aims to make unified, comprehensive disease knowledge readily accessible to the scientific community and grow its value through logical connections across resources.

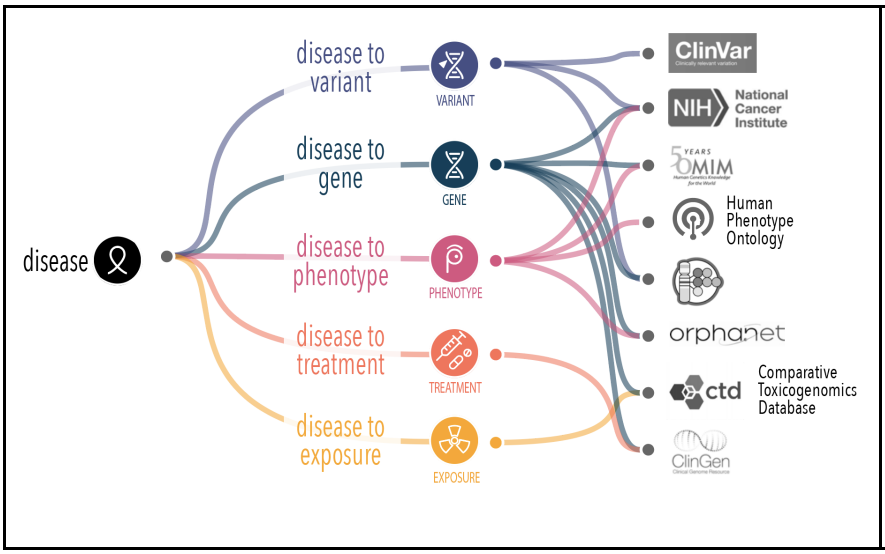

Figure 3: Different communities capture different disease relationships, at different levels of granularity and using different vocabularies. In order to form a complete picture of knowledge about a given disease, we need an authoritative handle to robustly and reproducibly collate disease features.

Figure 3. Mondo supports alignment of different disease attributes that are captured in different sources.

Table 1: Summary statistics across all Mondo concepts. (Version at: https://github.com/monarch-initiative/mondo/releases/tag/v2022-03-01)

Disease term feature	Count
Total number of terms	22,157
Database cross references	104,479
Term definition	15,443
Exact synonyms	66,247
Related synonyms	30,661
Narrow (more specific) synonym	2,214

Table 2: Disease concept statistics for select disease categories. Note that these groupings are overlapping. (Version at: https://github.com/monarch-initiative/mondo/releases/tag/v2022-03-01)

Disease type	Count (Concepts)
Rare disease	10,443
Infectious disease	1,240
Cancer (including neoplasms)	4,298
Mendelian diseases	11,380

Acknowledgements

Mondo is generously supported by the NIH National Human Genome Research Institute Phenomics First Resource, **NIH-NHGRI** # 1 RM1 HG010860–01, a Center of Excellence in Genomic Science; and a **NIH Office of the Director** Grant #5R24OD011883 for the Monarch Initiative. Thank you to Damien Goutte-Gattat for assistance in mining GitHub for our list of contributors.

References

1. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49: D1207–D1217.

2. Amberger JS, Hamosh A. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. Curr Protoc Bioinformatics. 2017;58: 1.2.1–1.2.12.

3. Slebodnik M. Orphanet: The Portal for Rare Diseases and Orphan Drugs2009384Orphanet: The Portal for Rare Diseases and Orphan Drugs. Paris: Institute National de la Santé et de la Recherche Médicale (INSERM) Last visited June 2009. Gratis URL: www.orpha.net/. Reference Reviews. 2009. pp. 45–46. doi:10.1108/09504120911003492

4. Haendel MA, McMurry JA, Relevo R, Mungall CJ, Robinson PN, Chute CG. A Census of Disease Ontologies. Annu Rev Biomed Data Sci. 2018;1: 305–331.

5. Vasilevsky N. The landscape of disease and phenotype ontologies. Zenodo; 2022. doi:10.5281/ZENODO.6299898

6. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40: 30–43.

7. Bello SM, Shimoyama M, Mitraka E, Laulederkind SJF, Smith CL, Eppig JT, et al. Disease Ontology: improving and unifying disease annotations across species. Dis Model Mech. 2018;11. doi:10.1242/dmm.032839

8. Rogers FB. Medical subject headings. Bull Med Libr Assoc. 1963;51: 114–116.

9. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007;14: 687–696.

10. Kamdar MR, Tudorache T, Musen MA. A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies. Semant Web. 2017;8: 853–871.

11. Human Disease Ontology release v2022-03-02. [cited 11 Apr 2022]. Available: http://purl.obolibrary.org/obo/doid/releases/2022-03-02/doid.owl

12. FAQs about rare diseases. [cited 26 Feb 2022]. Available: https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases

13. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare

diseases are there? Nat Rev Drug Discov. 2020;19: 77–78.

14. Procedural document: Orphanet nomenclature and classification of rare diseases. Orphanet; 2020 Mar. Available: http://www.orpha.net/orphacom/cahiers/docs/GB/eproc_disease_inventory_R1_Nom_Dis_EP_04.pdf

15. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. N Engl J Med. 2018;379: 1452–1462.

16. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak. 2021;21: 206.

17. Chute CG. The rendering of human phenotype and rare diseases in ICD-11. J Inherit Metab Dis. 2018. doi:10.1007/s10545-018-0172-5

18. Entities - Mondo documentation. [cited 12 Apr 2022]. Available: https://mondo.readthedocs.io/en/latest/editors-guide/f-entities/

19. Sources. 26 Feb 2022 [cited 11 Apr 2022]. Available: https://mondo.monarchinitiative.org/pages/sources/

20. Mungall CJ, Koehler S, Robinson P, Holmes I, Haendel M. k-BOOM: A Bayesian approach to ontology structure inference, with applications in disease ontology construction. bioRxiv. 2019. p. 048843. doi:10.1101/048843

21. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015;372: 2235–2242.

22. McKusick V. Online Mendelian Inheritance in Man, OMIM™. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. World Wide Web URL: https://omim org. 2009.

23. About GARD. [cited 26 Feb 2022]. Available: https://rarediseases.info.nih.gov/about-gard

24. Maiella S, Rath A, Angin C, Mousson F, Kremp O. Orphanet et son réseau : où trouver une information validée sur les maladies rares. Rev Neurol . 2013;169: S3–S8.

25. Thorogood A, Rehm HL, Goodhand P, Page AJH, Joly Y, Baudis M, et al. International federation of genomic medicine databases using GA4GH standards. Cell Genom. 2021;1. doi:10.1016/j.xgen.2021.100032

26. Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Callahan TJ, et al. The GA4GH Phenopacket schema: A computable representation of clinical data for precision medicine. medRxiv. 2021 [cited 26 Feb 2022]. doi:10.5167/uzh-210475

27. HL7.TERMINOLOGY\HL7 terminology home page - FHIR v4.0.1. [cited 26 Feb 2022]. Available: https://build.fhir.org/ig/HL7/UTG/

28. Home - MedGen - NCBI. [cited 26 Feb 2022]. Available: https://www.ncbi.nlm.nih.gov/medgen/

29. Working together to put kids first. [cited 27 Feb 2022]. Available: https://kidsfirstdrc.org/

30. Pharos: Disease List. [cited 27 Feb 2022]. Available: https://pharos.nih.gov/diseases