The American Journal of Human Genetics, Volume *109*

Supplemental information

OARD: Open annotations for rare diseases and their phenotypes based on real-world data

Cong Liu, Casey N. Ta, Jim M. Havrilla, Jordan G. Nestor, Matthew E. Spotnitz, Andrew S. Geneslaw, Yu Hu, Wendy K. Chung, Kai Wang, and Chunhua Weng

Supplemental Figure

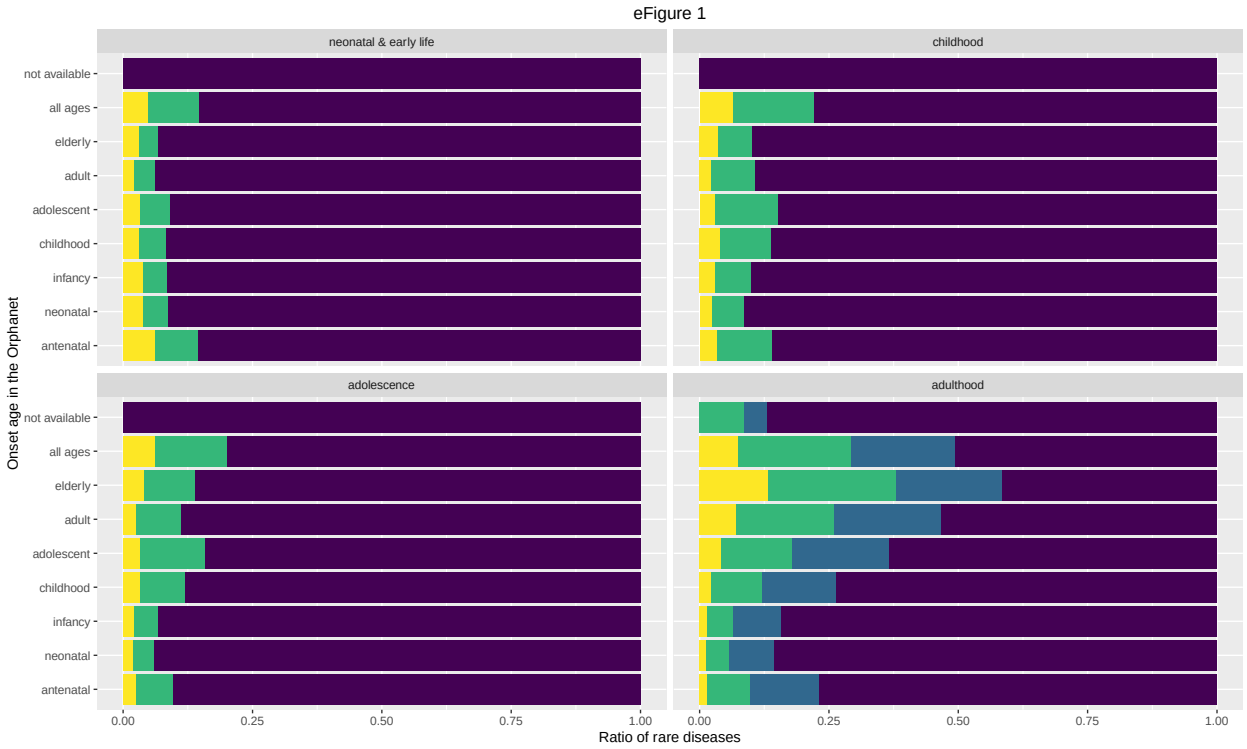

Figure S1 The prevalence of the rare disease concepts in the OARD database across different age groups. Y-axis is the onset annotation of the rare diseases in the Orphanet; Different colors represent the prevalence derived from the OARD database.

Supplemental Tables

Note title	# of individuals
Surgical Path Event	582,521
Follow-Up Visit	471,794
Letter	394,270
Initial Visit	377,863
Clinical Summary-RTF	325,959
Operative Note	325,886
12-Lead Electrocardiogram	305,887
Emergency Department Nursing Assessment Note	305,030
Consult Visit	268,538
Discharge Summary Note	265,582
Office Visit	252,210
Emergency Department Disposition Note	250,747
Patient Education	221,113
NYP Discharge Summary Note	221,097
Miscellaneous Nursing Note	205,409

Table S1. Distribution of the number of phenotypes in different note types found in CUIMC/Notes. Note title is assigned by clinical data warehouse operative team at CUIMC. # of individuals are the count of individuals with at least one of the phenotype concepts extracted from one of those encoded with a specific note title. Top 15 note titles were listed here. The “Sugical Path Event” is a specific note_title created by the CUIMC clinical data warehouse operative team to store the clinical manifest of a biospecimen sent to the pathology lab. We then defined the relevant notes as those with “visit”, “letter”, “summary”, and “surgical path” keywords in note_title.

dataset_id	clinical_site	source	subpopulation	subclass_category
1	CUIMC	OMOP	All	All
2	CUIMC	Notes	All	All
3	CHOP	Notes	All	All
10	CUIMC	OMOP	Age	Neonate and early life (0-2)
11	CUIMC	OMOP	Age	Childhood (3-11)
12	CUIMC	OMOP	Age	Adolescence (2-17)
13	CUIMC	OMOP	Age	Adulthood (18-99)
20	CUIMC	Notes	Age	Neonate and early life (0-2)
21	CUIMC	Notes	Age	Childhood (3-11)
22	CUIMC	Notes	Age	Adolescence (2-17)
23	CUIMC	Notes	Age	Adulthood (18-99)
100	CUIMC	OMOP	Specialist	Genetic
199	CUIMC	OMOP	All	Hierarchical
200	CUIMC	Notes	Specialist	Genetic
299	CUIMC	Notes	All	Hierarchical
190000119	CUIMC	OMOP	Specialist	Abnormality of the genitourinary system
190000152	CUIMC	OMOP	Specialist	Abnormality of head or neck
190000478	CUIMC	OMOP	Specialist	Abnormality of the eye
190000598	CUIMC	OMOP	Specialist	Abnormality of the ear
190000707	CUIMC	OMOP	Specialist	Abnormality of the nervous system
190000769	CUIMC	OMOP	Specialist	Abnormality of the breast
190000818	CUIMC	OMOP	Specialist	Abnormality of the endocrine system
190001197	CUIMC	OMOP	Specialist	Abnormality of prenatal development or birth
190001507	CUIMC	OMOP	Specialist	Growth abnormality
190001574	CUIMC	OMOP	Specialist	Abnormality of the integument
190001608	CUIMC	OMOP	Specialist	Abnormality of the voice
190001626	CUIMC	OMOP	Specialist	Abnormality of the cardiovascular system
190001871	CUIMC	OMOP	Specialist	Abnormality of blood and blood-forming tissues
190001939	CUIMC	OMOP	Specialist	Abnormality of metabolism/homeostasis
190002086	CUIMC	OMOP	Specialist	Abnormality of the respiratory system
190002664	CUIMC	OMOP	Specialist	Neoplasm
190002715	CUIMC	OMOP	Specialist	Abnormality of the immune system
190025031	CUIMC	OMOP	Specialist	Abnormality of the digestive system

dataset_id	clinical_site	source	subpopulation	subclass_category
190025142	CUIMC	OMOP	Specialist	Constitutional symptom
190025354	CUIMC	OMOP	Specialist	Abnormal cellular phenotype
190033127	CUIMC	OMOP	Specialist	Abnormality of the musculoskeletal system
190040064	CUIMC	OMOP	Specialist	Abnormality of limbs
190045027	CUIMC	OMOP	Specialist	Abnormality of the thoracic cavity
290000119	CUIMC	Notes	Specialist	Abnormality of the genitourinary system
290000152	CUIMC	Notes	Specialist	Abnormality of head or neck
290000478	CUIMC	Notes	Specialist	Abnormality of the eye
290000598	CUIMC	Notes	Specialist	Abnormality of the ear
290000707	CUIMC	Notes	Specialist	Abnormality of the nervous system
290000769	CUIMC	Notes	Specialist	Abnormality of the breast
290000818	CUIMC	Notes	Specialist	Abnormality of the endocrine system
290001197	CUIMC	Notes	Specialist	Abnormality of prenatal development or birth
290001507	CUIMC	Notes	Specialist	Growth abnormality
290001574	CUIMC	Notes	Specialist	Abnormality of the integument
290001608	CUIMC	Notes	Specialist	Abnormality of the voice
290001626	CUIMC	Notes	Specialist	Abnormality of the cardiovascular system
290001871	CUIMC	Notes	Specialist	Abnormality of blood and blood-forming tissues
290001939	CUIMC	Notes	Specialist	Abnormality of metabolism/homeostasis
290002086	CUIMC	Notes	Specialist	Abnormality of the respiratory system
290002664	CUIMC	Notes	Specialist	Neoplasm
290002715	CUIMC	Notes	Specialist	Abnormality of the immune system
290025031	CUIMC	Notes	Specialist	Abnormality of the digestive system
290025142	CUIMC	Notes	Specialist	Constitutional symptom
290025354	CUIMC	Notes	Specialist	Abnormal cellular phenotype
290033127	CUIMC	Notes	Specialist	Abnormality of the musculoskeletal system
290040064	CUIMC	Notes	Specialist	Abnormality of limbs
290045027	CUIMC	Notes	Specialist	Abnormality of the thoracic cavity

Table S2 Dataset/subset or hierarchical representation for different subpopulations. Dataset were derived from structured OMOP database or clinical

narratives (notes). Subset or hierarchical representation is currently not available for CHOP derived dataset.

API endpoint	Description
/metadata/datasets ***parameter**: None*	Enumerates the datasets available in OARD
/metadata/domainCounts ***parameter**: dataset*	The number of concepts in each domain
/metadata/domainPairCounts ***parameter**: dataset, domain**	The number of pairs of concepts in each pair of domains
/metadata/patientCounts ***parameter**: dataset*	The number of patients in the dataset
/vocabulary/findConceptByName ***parameter**: q; domain**	Search for concepts by name and domain (optional: by domain)
/vocabulary/findConceptById ***parameter**: q;*	Search for concepts by pseudo OMOP ID
/vocabulary/findConceptByCode ***parameter**: q;*	Search for concepts by code (HPO or MONDO)
/vocabulary/findConceptByAny ***parameter**: q; domain**	Search for concepts by either name, code or pseudo OMOP ID (optional: by domain)
/frequencies/singleConceptFreq ***Parameter**: dataset; concept*	Clinical frequency of individual concepts
/frequencies/pairedConceptFreq ***Parameter**: dataset; concept1; concept2;*	Clinical frequency of a pair of concepts;
/frequencies/mostFrequency ***Parameter**: dataset; concept*; domain*,top_n**	Most frequent concepts (or concept pairs if q provided) (optional: by domain);
/association/chiSquare ***Parameter**: dataset; concept1; concept2*; domain*; top n*; ascending**	*Chi-squared analysis of paired concepts*
/association/obsExpRatio ***Parameter**: dataset; concept1; concept2*; domain*; top_n*; ascending**	Observed Count / Expected Count
/association/relativeFrequency ***Parameter**: dataset; concept1; concept2*; domain*; top_n*; ascending**	Relative frequency between pairs of concepts
/association/jaccardIndex	Jaccard Index between pairs of concepts

API endpoint	Description
***Parameter**: dataset; q1; q2*; domain*; top_n*; ascending**	

Table S3. API endpoints provided by OARD. * is the optional parameter. Here is a brief introduction of each parameter: *dataset*: which data source to extract the statistics; *domain*: “phenotypes” or “diseases” (by default, extract for all domains); *q*: query string for vocabulary search; *concept/concept1/concept2*: pseudo OMOP concept ID to extract frequencies and association statistics; *top_n*: only return top ranked records (by default, return all records); *ascending*: if rank negative association first (by default: false).

Supplemental Methods

Standardized concepts derivation

We used latest version of Human Phenotype Ontology (HPO)[1] and Mondo Disease Ontology (MONDO) ontology[2] to extract standardized concepts. We use the Python package owlready2 to access the ontology[3]. The HPO information acquired for HPO was up to 2021/12/15, and the MONDO information acquired for MONDO was updated 2022/01/15. The ontology IRI for MONDO is http://purl.obolibrary.org/obo/mondo.owl and the ontology IRI for HPO is http://purl.obolibrary.org/obo/hp.owl. For HPO IDs, we only includes the concepts under subclass 'Phenotypic abnormality' (HP:0000118), which is the root of the phenotypic abnormality subontology of the HPO. For MONDO IDs, we only included concepts under subclass 'rare' (MONDO:0021136), which includes the disease defined in Orphanet Rare Disease Ontology (ORDO)[4] and Genetic and Rare Diseases Information Center (GARD)[5].

In order to map the OMOP data to corresponding HPO or MONDO IDs, we leveraged the COHD API's omop_to_biolink API endpoint, which further references the Node Normalizer tool developed for the NCATS Biomedical Data Translator project to provide equivalence mappings between ontologies supported by the Biolink Model[6]. The Biolink Model supports both HPO and MONDO but does not support OMOP, thus a direct mapping from OMOP to HPO/MONDO was not possible using the Node Normalizer service directly. Since both the OMOP Standard Vocabulary and Biolink Model support SNOMED-CT, ICD10CM, ICD9CM, and MedDRA, we first mapped from the source OMOP concept IDs to these intermediate ontologies, then queried the Node Normalizer using these intermediate IDs to retrieve mappings to Biolink. If the preferred Biolink identifier selected by the Node Normalizer was an HPO or MONDO ID, then the mapping from OMOP to HPO/MONDO was created. The COHD API is publicly available at https://cohd.io/api. The source code for creating the OMOP-Biolink mappings is available at https://github.com/WengLab-InformaticsResearch/cohd_api/blob/master/cohd/biolink_mapper.py. The Translator Node Normalizer service is publicly available at https://nodenormalization-sri.renci.org/docs, and its code at https://github.com/TranslatorSRI/NodeNormalization.

In order to identify the HPO concepts and MONDO concepts for each individual, we created a context-awared query for each of the concept. The query is defined as "%concept_string% NOT '%negation_trigger% %concept_string%'~10 NOT '%family_trigger% muscular dystrophy'~10". The concept string with "%%" can be replaced by concept name or synonyms for that query concept. And similarly the negation and family triggers with "%%" can be also replaced by a set of context triggers. Those triggers are predefined with the aim to reduce false positives and reduce the timing for the query operation. All identified documents were returned for a HPO query and then preprocessed to extract the individual patient ID and encounter time related to the specific document.

Association analysis

The OARD is able to provide a ranked association list based on the Chi-squared, relative frequency, observed-expected frequency ratio, Jaccard-index calculated using concept

prevalence, concept pair co-occurrence and total patient count. The Jaccard-index definition can be found in the main content. We listed the calculation formular for the other three summary statistics as below.

The concept prevalence is defined as

$$P_H^c = \frac{|T_H^c|}{|T_H|}$$

Where P_H^c is the prevalence of concept C in dataset H. T_H^c is the set of unique patients in dataset H observed with concept C , and T_H is the set of unique inpatient visits of patients. $|S|$ represents the number of elements in set S.

The concept co-occurrence frequency is defined as

$$P_H^{c_1,c_2} = \frac{\left|T_H^{c_1} \cap T_H^{c_2}\right|}{|T_H|}$$

Where $P_H^{c_1,c_2}$ is the co-occurrence frequency of concept $C_1$and C_2 in dataset H.

The relative frequency indicates how frequently concept C_1 occurs among patients who have concept C_2. This is similar to the conditional probability of C_1 given C_2. Relative frequency is calculated as:

$$F_H(C_1|C_2) = \frac{\left|T_H^{c_1} \cap T_H^{c_2}\right|}{\left|T_H^{c_2}\right|}$$

where $F_H(C_1|C_2)$ is the relative frequency of concept C_1 among patients observed with concept C_2 in dataset H.

The observed-expected frequency ratio quantifies the strength of the dependence between two concepts. The natural logarithm of observed-expected frequency ratio (log ratio for short) is calculated as:

$$LR_H(C_1, C_2) = log\frac{\left|T_H^{c_1} \cap T_H^{c_2}\right| \times |T_H|}{\left|T_H^{c_1}\right| \times \left|T_H^{c_2}\right|}$$

where $LR_H(C_1, C_2)$ is the log ratio of concepts C_1 and C_2 in dataset H.

The chi-squared analysis is informative of the dependence between two concepts. However, this analysis becomes very sensitive with large population sizes, such that statistically significant results may not be scientifically significant.

$$\chi_H^2(C_1, C_2) = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Where

$O_{ij} = \left|T_H^{c_1} \cap T_H^{c_2}\right|^{i\times j} \times \left|\left(T_H - T_H^{c_1}\right) \cap T_H^{c_2}\right|^{(1-i)\times j} \times \left|T_H^{c_1} \cap \left(T_H - T_H^{c_2}\right)\right|^{i\times(1-j)} \times \left|\left(T_H - T_H^{c_1}\right) \cap \left(T_H - T_H^{c_2}\right)\right|^{(1-i)\times(1-j)}$

and

$$E_{ij} = \frac{\left(\left|T_H^{c_1}\right| \times \left|T_H^{c_2}\right|\right)^{i \times j}}{|T_H|} \times \frac{\left(\left|T_H - T_H^{c_1}\right| \times \left|T_H^{c_2}\right|\right)^{(1-i) \times j}}{|T_H|} \times \frac{\left(\left|T_H - T_H^{c_2}\right| \times \left|T_H^{c_1}\right|\right)^{(1-j) \times i}}{|T_H|} \times \frac{\left(\left|T_H - T_H^{c_1}\right| \times \left|T_H - T_H^{c_1}\right|\right)^{(1-i) \times (1-j)}}{|T_H|}$$

Reference:

1. Kohler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res *47*, D1018-D1027. 10.1093/nar/gky1105.
2. Vasilevsky, N., Essaid, S., Matentzoglu, N., Harris, N.L., Haendel, M., Robinson, P., and Mungall, C.J. (2020). Mondo Disease Ontology: harmonizing disease concepts across the world. (CEUR-WS).
3. Jean-Baptiste, L. (2021). Ontologies with Python: Programming OWL 2.0 Ontologies with Python and Owlready2 (Springer).
4. Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P.N., Parkinson, H., and Rath, A. (2014). Ordo: an ontology connecting rare disease, epidemiology and genetic data. (researchgate. net).
5. Zhu, Q., Nguyen, D.-T., Grishagin, I., Southall, N., Sid, E., and Pariser, A. (2020). An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). Journal of Biomedical Semantics *11*, 1-13.
6. Consortium, B.D.T. (2019). Toward a universal biomedical data translator. Clinical and translational science *12*, 86.