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

Authors

Cong Liu, Casey N. Ta, Jim M. Havrilla, ...,
Wendy K. Chung, Kai Wang, Chunhua Weng

Correspondence

cw2384@cumc.columbia.edu

The open annotation for rare diseases (OARD) is a publicly accessible, data-driven resource containing summary statistics including phenotype frequencies and associations for rare diseases. It was derived from over 10 million individuals' electronic health records from two academic health institutions, which span wide age ranges and different disease subgroups.

Liu et al., 2022, The American Journal of Human Genetics *109*, 1591–1604
September 1, 2022 © 2022 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2022.08.002

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

Cong Liu,[1] Casey N. Ta,[1] Jim M. Havrilla,[2] Jordan G. Nestor,[4] Matthew E. Spotnitz,[1] Andrew S. Geneslaw,[3] Yu Hu,[2] Wendy K. Chung,[3] Kai Wang,[2] and Chunhua Weng[1,*]

Summary

Diagnosis for rare genetic diseases often relies on phenotype-driven methods, which hinge on the accuracy and completeness of the rare disease phenotypes in the underlying annotation knowledgebase. Existing knowledgebases are often manually curated with additional annotations found in published case reports. Despite their potential, real-world data such as electronic health records (EHRs) have not been fully exploited to derive rare disease annotations. Here, we present open annotation for rare diseases (OARD), a real-world-data-derived resource with annotation for rare-disease-related phenotypes. This resource is derived from the EHRs of two academic health institutions containing more than 10 million individuals spanning wide age ranges and different disease subgroups. By leveraging ontology mapping and advanced natural-language-processing (NLP) methods, OARD automatically and efficiently extracts concepts for both rare diseases and their phenotypic traits from billing codes and lab tests as well as over 100 million clinical narratives. The rare disease prevalence derived by OARD is highly correlated with those annotated in the original rare disease knowledgebase. By performing association analysis, we identified more than 1 million novel disease-phenotype association pairs that were previously missed by human annotation, and >60% were confirmed true associations via manual review of a list of sampled pairs. Compared to the manual curated annotation, OARD is 100% data driven and its pipeline can be shared across different institutions. By supporting privacy-preserving sharing of aggregated summary statistics, such as term frequencies and disease-phenotype associations, it fills an important gap to facilitate data-driven research in the rare disease community.

Introduction

Rare diseases collectively affect about 30 million Americans and cause significant morbidity and mortality.[1] Yet, they are difficult to diagnose because most clinicians are not familiar with them.[2,3] The diagnostic odyssey often includes numerous referrals and arduous tests, inflicting emotional and financial burdens on individuals and families,[4,5] who can suffer from lower quality of life, health deterioration, ineffective treatments, unnecessary procedures, and sometimes irreversible complications.[6] Recent advances in genomic-based diagnosis have facilitated diagnostic efficiency and efficacy and showed early promise to even identify new diseases in undiagnosed individuals.

Genomic-based diagnostic pipelines often involve prioritizing genes and variants on the basis of known clinical signs and symptoms. Some of the informatics systems that have been developed for this process include Exomiser,[7,8] Phevor,[9] GADO,[10] and ClinPhen[11] as well as our previously developed Phenolyzer,[12,13] EHR-Phenolyzer,[14] and Phen2Gene.[15] Efficient standardized' health information curation tools have also been developed, such as PhenoTips[16] and our previously developed Doc2Hpo[17] and PheNominal.[18] Although there are variations across these systems, they have a shared gene prioritization strategy based on a similarity score derived from disease-phenotype association knowledgebases. As a result of this shared approach, the ability for these systems to make accurate differential diagnoses relies largely on the quantity and quality of the underlying rare disease-phenotype annotation knowledge.

Currently, there are two main approaches to gather this knowledge: manual curation or automatic extraction from the literature. The most frequently used knowledgebase is the Human Phenotype Ontology (HPO),[19] which contains over 200,000 annotation relationships (i.e., edges) between ~16,000 standardized phenotypic concepts and ~12,000 rare diseases collected from OMIM,[20] Orphanet,[21] and DECIPHER.[22] The annotations in HPO were curated mainly by human experts with facilitation of natural language processing techniques to extract textual descriptions of each disease listed in OMIM, which is time consuming, labor intensive, and likely to miss annotations because of the large volume and rapid growth of rare disease knowledge. To provide a scalable way to increase the coverage of these annotations, text-mining from the literature has also been explored.[23] For example, Xu et al. developed a large-scale extraction approach to automatically extract disease-phenotype relationships from MEDLINE sentences.[24] Recently, PubCaseFinder used ConceptMapper to identify novel disease-phenotype associations from a large collection of case reports and increased the coverage of the association annotations by more than two times the original expert-curated Orphanet annotation.[23] By utilizing

[1]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA; [2]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [3]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY 10032, USA; [4]Division of Nephrology, Department of Medicine, Columbia University, New York, NY 10032, USA
*Correspondence: cw2384@cumc.columbia.edu
https://doi.org/10.1016/j.ajhg.2022.08.002.

semantic relations in phenotype ontologies, Kafkas et al. collected ICD10-HPO/MP associations from the public literature, original ontologies (e.g., UMLS [Unified Medical Language System]), and WikiData.[25]

Besides the aforementioned resources, electronic health records (EHRs) contain real-world evidence regarding the prevalence and associated information related to rare diseases and corresponding phenotypes, which can serve as a rich resource to enrich our current medical knowledge.[26] However, this resource has not been exploited yet for two main reasons. The first is the privacy concern, which is often magnified in the rare disease field as a result of the inherently low prevalence and identifiability of individuals with rare diseases. Second, rare diseases are not well documented in the EHR data. Commonly used billing codes (e.g., ICD) or clinical documentation codes (e.g., SNOMED) have a limited coverage of rare disease terms and their associated phenotypes.[27]

We previously developed Columbia Open Health Data (COHD), providing an open access to prevalence and co-occurrence statistics on conditions, drugs, procedures, and demographics derived from structured EHRs from Columbia University Irving Medical Center (CUIMC).[26] Because the architecture design of COHD does not store individual-level data, it provides a secure framework to share the rare disease knowledge derived from real-world EHRs. However, because COHD only analyzes structured EHR data from a single institution, it is underpowered and may be influenced by inherent biases of the specific coding process adopted in a single institution. To provide a sharable resource with a better phenotypic annotation coverage to the rare disease research community and complement the current existing HPO-Jax-based diagnosis pipeline with novel disease-phenotype associations that have not been reported in the current knowledgebase, we developed open annotation for rare diseases (OARD), which consists of the rare disease knowledge synthesized from multiple datasets on the basis of both structured and unstructured electronic health records from two major academic medical institutions: CUIMC and Children's Hospital of Philadelphia (CHOP).

Material and methods

Data sources

This study is approved by the institutional review boards (IRBs) at Columbia University and the Children's Hospital of Philadelphia (CHOP). Study does not involve recruitment procedures and a waiver of informed consent and assent is obtained at both institutions. There were three original data sources. (1) CUIMC/OMOP: we analyzed structured data from CUIMC's Observational Medical Outcomes Partnership (OMOP) database. The OMOP database was derived from longitudinal electronic health records including inpatient and outpatient data from 1985 to 2021. CUIMC's clinical data warehouse (CDW) was converted to OMOP Common Data Model[28] v5.3 in September 2021. CUIMC and NewYork-Presbyterian (NYP) Hospital serve New York, NY and the surrounding area. (2) CUIMC/Notes: we analyzed the unstructured notes from CUIMC's CDW including inpatient and outpatient data spanning from 1985 to 2020. Over 100 million notes were indexed along with relevant metadata (e.g., encounter date, provider, etc.) with the Solr technology. (3) CHOP/Notes: we analyzed the natural-language-processed clinical notes from CHOP's electronic health records spanning from 2000 to 2020. Over 15 million medical notes relevant to office visits were processed with the cTAKES-based NLP pipeline[29] to extract relevant clinical concepts and their contexts.

Standardized concepts derivation

Genomic-based diagnostic pipelines rely on standardized phenotypic and disease concepts such as HPO, MONDO, OMIM, and Orphanet as the input. In order to make the resource generalizable and easily adopted by those phenotype-based diagnosis tools, we used two types of standardized concepts from our original data sources with different approaches: (1) Human Phenotype Ontology (HPO) and (2) Mondo Disease Ontology (MONDO), which is a logic-based structure for unifying multiple disease resources. For HPO concepts, we only included those under the sub-ontology "phenotypic abnormality" (HP: 0000118). For MONDO concepts, we only included those under the subclass "rare" (MONDO: 0021136), which includes the diseases defined in Orphanet Rare Disease Ontology (ORDO) and Genetic and Rare Diseases Information Center (GARD).

For the CUIMC/OMOP dataset, we derived the standardized concepts by using the predefined cross-ontology mapping provided by the COHD application programming interface (API). Briefly, we extracted all rows from the OMOP condition_occurrence table to provide individuals' observed conditions. The OMOP IDs were then mapped to the corresponding HPO ID or MONDO ID via the COHD API, leveraging services established in the NCATS Biomedical Data Translator project to normalize Biolink entities.[30] Additional details are provided in the supplemental methods. In addition, we extracted all rows from the OMOP measurement table to provide phenotypes obtained by labs. The OMOP IDs and measurement results were then mapped to LOINC (short for Logical Observation Identifiers, Names, and Codes) IDs with the OMOP vocabulary and further mapped to the corresponding HPO IDs on the basis of the LOINC2HPO annotation.[31]

For CUIMC/Notes, we derived a context-aware keyword-searching approach by only considering keywords occurring within non-negated and non-family-member-related context to efficiently identify the standardized concepts. We first queried all note types and identified a list of relevant note types (e.g., notes titled "visit," "letter," "clinical summary," "surgical path," etc.) by ranking the phenotype term frequencies (Table S1). Then for each standard concept (i.e., HPO or MONDO), we queried for the concept within identified relevant note types by using a complex Solr query. We included both the concept's syntactic variations and synonyms to construct the Solr query, except when these terms were negated or family-member-related mentions. For example, to identify individuals with "muscular dystrophy (HP: 0003560)," we performed a Solr query to search for "muscular dystrophy" and its synonyms but excluded search hits that were negated or mentioned family members within a ten-word window. More details about the query formulation can be found in the supplemental methods.

For CHOP/Notes, we utilized the natural-language-processing toolkit cTAKES[29] to process only "office-visit"-related notes and identify the UMLS-based concept unique identifiers (CUIs). The UMLS-based CUIs were then mapped to corresponding HPO and MONDO concepts. We used the NegEx algorithm[32] to exclude those hits that were negated or not relevant to the person in the corresponding note.

Aggregate summary statistics

To share the real-world-derived knowledge securely, we adopted a similar backend architecture as described in COHD to only store the aggregated summary statistics. Briefly, there are two types of statistics involved—prevalence and co-occurrence frequencies of concepts. The prevalence of a standard concept C was defined as the ratio between the number of unique individuals observed with concept C in a dataset (or a subset as described below) and the total number of unique individuals observed in the same dataset/subset. The co-occurrence frequency of two standard concepts C_1 and C_2 were defined as the ratio between the number of unique individuals observed with both concepts C_1 and C_2 in a dataset/subset and the total number of unique individuals observed in the same dataset. To facilitate further association analysis, we also stored the total number of the unique individuals in a dataset/subset, and only individuals with at least one concept extracted were counted.

Association analysis between diseases and phenotypic concepts

Besides providing the public access to the prevalence and co-occurrence frequencies information, OARD also provides different perspectives on quantifying associations on the basis of a real-world diagnosis scenario. For example, a user might be interested in the most frequently associated rare diseases given one or more phenotypic terms. The OARD is able to provide a ranked association list based on the Chi-squared, relative frequency, and observed-expected frequency ratio statistics calculated on the fly as described in the original COHD paper (see supplemental methods for details).[26] In addition, we also calculated the Jaccard index to measure the correlation between pairs of concepts as follows:

$$r_{C_i,C_j} = \frac{N_{c_i,c_j}}{\left(N_{c_i} + N_{c_j} - N_{c_i,c_j}\right)} \qquad \text{(Equation 1)}$$

where N_{c_i,c_j} is the number of unique individuals observed with both concepts c_i an c_j and N_{c_i} and N_{c_j} are the counts of unique individuals with concepts c_i and c_j, respectively.

Subset stratification and hierarchical presentation

Often in clinical practice, the prevalence and co-occurrence information in an overall population is not enough for differential diagnosis. Therefore, we derived a number of subsets to reflect the concept prevalence and co-occurrence in different populations. We first generated four age strata based on when the phenotype or diagnosed disease was identified: (1) neonatal and early life, from birth to 2 years old; (2) childhood, from 3 years old to 11 years old; (3) adolescence, from 12 years old to 17 years old; and (4) adulthood, >18 years old. We did not further divide the adult group because the timestamp for the phenotypes and diseases in the EHR rarely accurately reflects the true ages of onset for adults,[33] which is especially problematic for genetic disorders, as opposed to conditions associated with old age. Second, we generated a number of strata based on disease subgroups. We classified the population into different disease systems on the basis of their observed phenotype terms. For example, the musculoskeletal subset included individuals who have at least one phenotype observed under the HPO “abnormality of the musculoskeletal system” subontology (HP: 0033127). In addition, we included a genetic subset as individuals have a genetic finding in their medical records defined as any condition under SNOMED code “4025367: genetic findings.” More details of the subset stratification can be found in Table S2. Furthermore, alternative to the individual concept prevalence, we derived another concept prevalence presentation on the basis of the concept group. A concept group is defined as a concept plus all of its descendants defined in the HPO or MONDO ontologies. Prevalence of the concept group was derived as the count of the total unique individuals with any concepts in that concept group. We also derived co-occurrence between a pair of concept groups by using a similar approach.

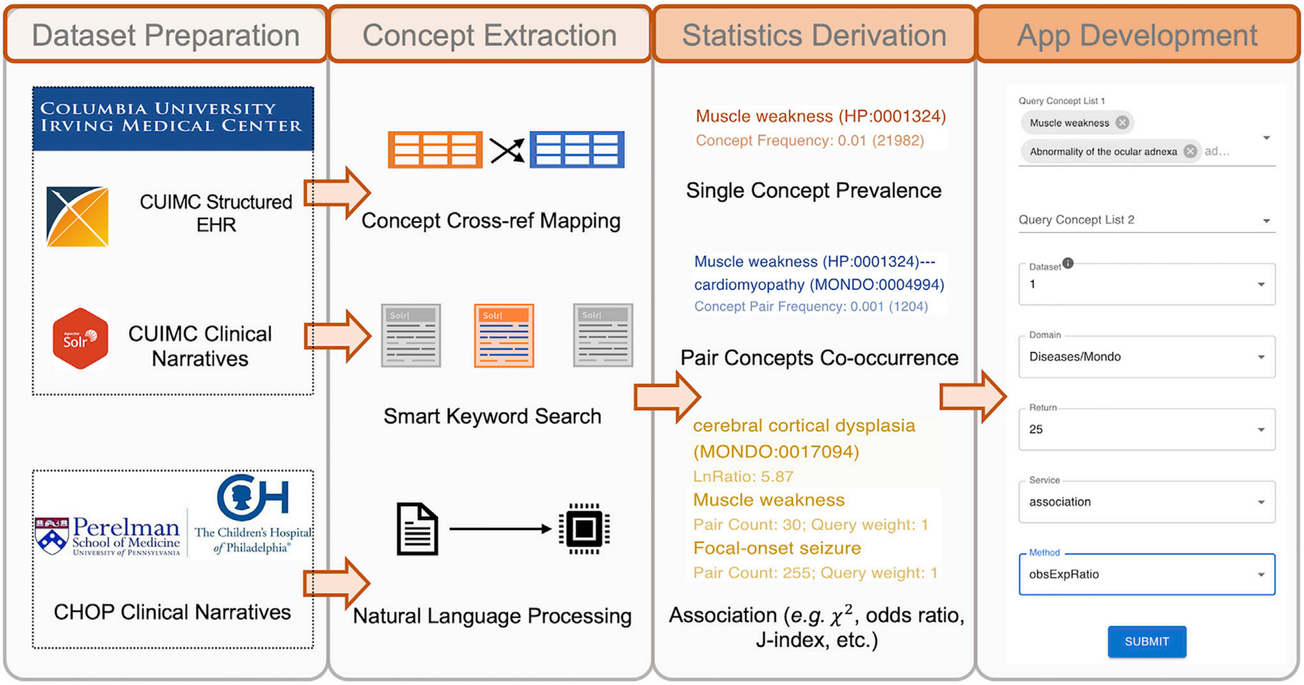
Evaluation of novel identified disease-phenotype associations

To estimate the accuracy of novel identified disease-phenotype association, we invited clinicians (J.G.N., A.S.G., and M.E.S.) for a manual review process. Because rare diseases require extensive knowledge and clinicians did not think they had expertise in >67% of pairs randomly sampled from all disease domains included in OARD, each clinician was then asked to independently select up to five rare diseases for which they felt confident to perform manual review. For each rare disease they selected, we then randomly sampled five novel identified disease-phenotype associations from each quantile (50%–60%, 60%–70%, 70%–80%, 80%–90%, and 90%–100%) according to the odds ratio calculated for the disease-phenotype pairs identified for that given disease in the CUIMC/Notes dataset. For each sampled pair, clinicians were asked to assign a confidence grade (0 = definitely wrong; 1 = likely to be correct; 2 = very likely to be correct; and 3 = definitely correct; NA = do not have expertise) on the basis of their clinical experience. A similar review approach was adopted to evaluate a Duchenne muscular dystrophy (DMD [MIM: 310200]) example, and a clinician (W.K.C.) who is an expert in DMD reviewed novel DMD-phenotype associations and determined whether a predicted phenotype is clinically associated with DMD.

Results

An overview of OARD and API availability

Figure 1 depicts the overall architecture for the OARD design. To protect individuals’ privacy, concepts and pairs of concepts with counts ≤ 10 were excluded. The OARD is publicly available at https://rare.cohd.io/. The size (i.e., total number of individuals, total number of unique concepts, and total number of concept pairs) of datasets and their stratifications were summarized in Figure 2A. For the CHOP/Notes dataset, only the unstratified dataset was obtained. Figure 2B shows the size of the dataset in each sub-disease strata. In general, the note-derived datasets are much larger than the structured-data-derived dataset in terms of the number of unique concepts and concept pairs, although both structured and unstructured datasets have a similar number of underlying individuals. The API endpoints, which are publicly available, are grouped into four resources, metadata, vocabulary,


Dataset Preparation
Concept Extraction
Statistics Derivation
App Development
Columbia University Irving Medical Center
CUIMC Structured EHR
CUIMC Clinical Narratives
Perelman School of Medicine University of Pennsylvania
The Children's Hospital of Philadelphia
CHOP Clinical Narratives
Concept Cross-ref Mapping
Solr
Smart Keyword Search
Natural Language Processing
Muscle weakness (HP:0001324)
Concept Frequency: 0.01 (21982)
Single Concept Prevalence
Muscle weakness (HP:0001324)---
cardiomyopathy (MONDO:0004994)
Concept Pair Frequency: 0.001 (1204)
Pair Concepts Co-occurrence
cerebral cortical dysplasia
(MONDO:0017094)
LnRatio: 5.87
Muscle weakness
Pair Count: 30; Query weight: 1
Focal-onset seizure
Pair Count: 255; Query weight: 1
Association (*e.g.* χ^2, odds ratio, J-index, etc.)
SUBMIT

Query Concept List 1: Muscle weakness; Abnormality of the ocular adnexa; ad...
Query Concept List 2
Dataset: 1
Domain: Diseases/Mondo
Return: 25
Service: association
Method: obsExpRatio

Figure 1. Overview of the OARD pipeline, data source, and architecture
Briefly, we first extracted individual levels of HPO and MONDO concepts from two institutions, CUIMC and CHOP, EHR databases. For each dataset, we also generated stratified subsets (as described in the material and methods) when it was applicable. Only the resulting frequency and co-occurrence data (without individual-level data) were stored in a MySQL database. The MySQL database was disconnected from the institutional internal network before it was served to the public via the RESTful web API (https://rare.cohd.io/). The association analysis is performed per query, and the results are returned on the fly.

frequency, and association, on the basis of their functionality. Table S3 lists the API endpoints and their descriptions. An API documentation is developed at the https://rare.cohd.io/api.

Phenotype coverage analysis

We investigated the coverage of phenotype concepts in the OARD database. In total, 7,046 out of 16,040 (43.9%) HPO phenotype concepts can be found in at least ten unique individuals in one of the three data sources (9,972 if requiring at least one unique individual). As shown in Figure 3A, both notes-derived datasets have much higher phenotype concept coverage (37.2% for CUIMC and 36.3% for CHOP) compared to the dataset derived from structured databases (4.0%), highlighting the need to include unstructured data for deriving rare disease annotation. We further investigated the patterns that might affect the coverage of the HPO terms. Figure 3B summarizes the coverage of the concepts in different subontologies. The highest coverage was found consistently across three datasets in "constitutional symptom." Despite using LOINC2HPO to retrieve lab-based phenotypes, lower concept coverage was observed in "abnormality of metabolism/homeostasis," which often contains lab and abnormal cellular phenotypes. Besides this subontology, limb anomaly is another subontology with relatively low coverage. Figure 3C shows the distribution of the HPO concept coverage according to the granularity levels. As expected, a more granular term is less likely to map to a person. However, we found a low match rate for concepts in the highest ontology levels for the structured-data-derived dataset, indicating that these are abstract concepts uncommonly used for billing purposes. Because natural-language-processing techniques were also involved in generating the two datasets, we also investigated the syntax pattern of the HPO concepts and summarized their coverage distribution. Figure 3D shows that the coverage dropped exponentially with the number of words (or tokens) involved in composing the concept name. If more than four words were involved in comprising the concept, the coverage dropped to less than 10% in any dataset. Furthermore, Figure 3E shows the coverage dropped slightly as the average word length increased; however, the effect is not as large as that observed for the number of words. Figure 3F shows the concept names with more "verb" tokens have lower coverage while nouns increased the coverage rate. We also investigated how the design pattern of the HPO affects the coverage. Figure 3G shows concepts involved in cell ontology (CL)[34] and chemical entities (CHEBI)[35] have lower coverage, which is consistent with the low coverage rate we found for the lab- and cellular-related phenotypes. In sum, leveraging the unstructured notes did dramatically increase the phenotype coverage compared to using structured codes only, although the coverage varies among different semantic and syntactic designs of the HPO concepts.

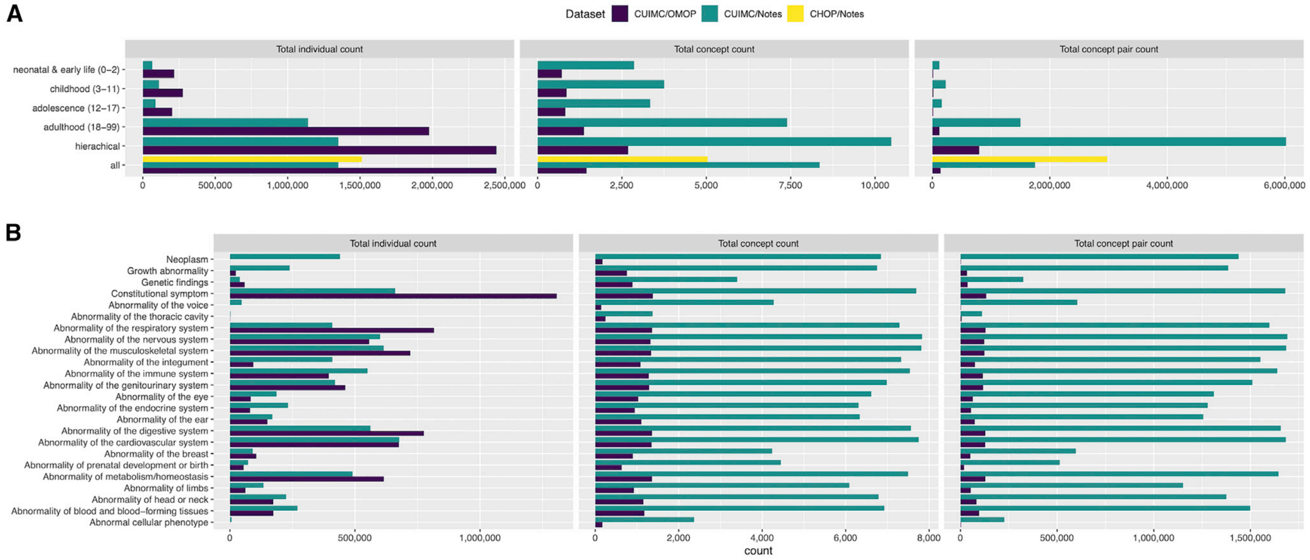


Figure 2. The size of each dataset and their subset

(A) The total unique individual count, unique concept count, and unique concept pair count found in each dataset. For the CUIMC-related dataset, the counts are also shown for the hierarchical represented dataset and subset from different ages.

(B) The total unique individual count, unique concept count, and unique concept pair counts found in different sub disease populations.

Rare diseases epidemiology reflected by EHR

The EHR has the potential to provide a low-cost means of estimating rare disease prevalence for epidemiologic research. Here, we present an analysis of epidemiologic data available in the OARD database. Among 7,619 unique rare disease concepts defined in MONDO, 2,531 (33%) have at least ten individuals identified in at least one of the three datasets. Figure 4A demonstrates the distribution of rare diseases according to the EHR prevalence (i.e., the percentage of individuals with this rare disease concept) estimated with the three datasets. Figure 4B shows the EHR-derived prevalence was correlated with the point prevalence annotation obtained in the Orphanet subset. For example, in the CUIMC/Notes dataset, the percentage of rare diseases with EHR prevalence less than 1/200,000 is increased along with the low Orphanet point prevalence annotation (i.e., 19.2% in "1–9/10,000," 29.3% in "1–9/100,000," 38.9% in "1–9/1,000,000," and 88.8% in "<1/1,000,000", respectively). Figure 4C further investigates the EHR prevalence in rare diseases with various inheritance patterns. As expected, "multigenic/multifactorial" and "not genetically inherited" rare diseases had a higher EHR prevalence, and diseases with "recessive" inheritance had a slightly lower prevalence than the "dominant" mode. Figure 4D shows the EHR prevalence for rare diseases was increased with older age of onset. We further compared the age of onset annotated with the EHR prevalence in the age-dependent subset derived with the CUIMC/Notes dataset. As shown in Figure S1, each subset has its highest disease prevalence observed in their corresponding onset age. For example, for the "adulthood" subset, the percentage of rare diseases with an EHR prevalence greater than 1/200,000 (green, cyan, and yellow) are 58.5% and 46.7% in the elderly- and adult-onset categories, which are the two largest among all other age-of-onset categories.

Comparison of disease-phenotype associations with expert-curated and literature-curated annotations

One of the most important features provided by OARD is its ability to identify associated pairs of concepts. We therefore compared the OARD results (using CUIMC/Notes as an example) against the HPO-MONDO concept pair annotations obtained from the HPO annotation file and public literature mining. In total, among 2,261,672 pairs of HPO-MONDO concept pairs in the CUIMC/Notes dataset, 34,932 of them were previously annotated in the HPO annotation files and 80,043 pairs were found in the literature by mining the PubCaseFinder database. By calculating the association with three different statistics, Chi-squared, odds ratio (OR) of observed counts and expected counts, and Jaccard index, Figure 5A shows the median statistics for a phenotype is significantly larger in the previously annotated pairs (median log scale for Chi-squared, 6.21; Jaccard index, −5.43; odds ratio, 5.03) than those pairs not annotated (median log scale for Chi-squared, 2.91; Jaccard index, −6.77; odds ratio, 2.56). For each phenotype, we conducted a Wilcoxon rank-sum test for pairs annotated (including expert curated and identified in literature) and pairs not annotated. Figure 5B shows that 67.8%, 50.6%, and 66.6% of annotated phenotype pairs have their Chi-squared, Jaccard index, and odds ratio statistics, respectively, ranked significantly higher (Wilcoxon rank-sum test; $p < 0.05$) than their not-annotated pairs. We then calculated the difference between the median association statistics among annotated and not-annotated pairs stratified by the provenance supporting the annotation (Figure 5C) and by the frequency of the pair according to

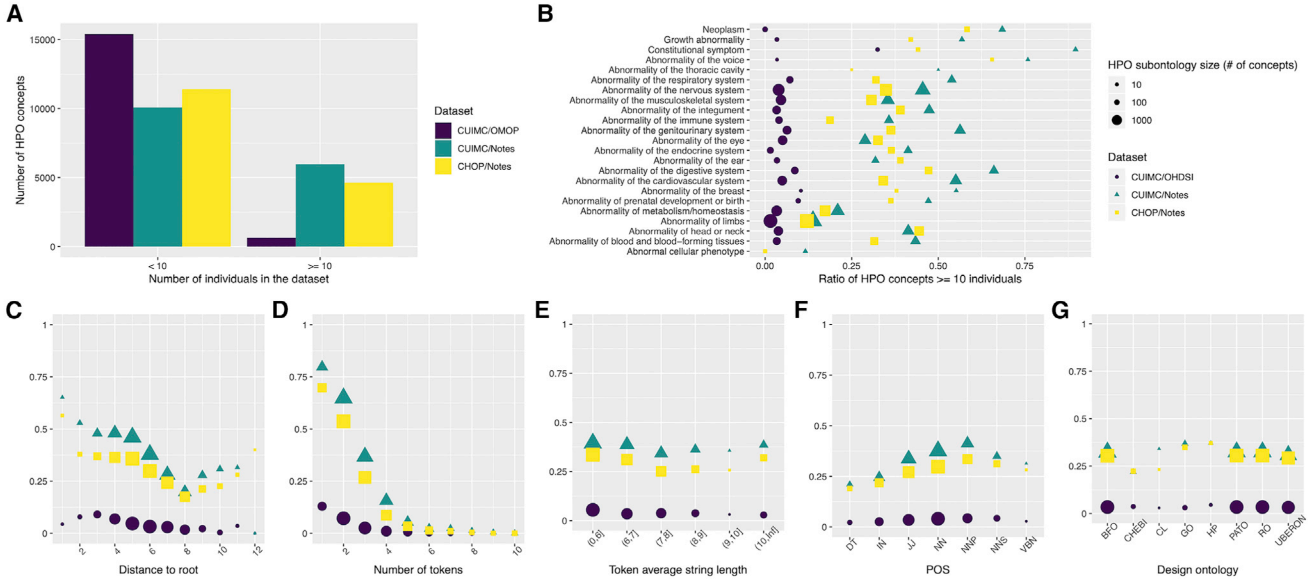


Figure 3. The coverage of phenotype concepts in the OARD database
(A) Overall coverage for each dataset.
(B) Coverage in different phenotype subontology.
(C) Coverage for concepts with different distances to the root (i.e., phenotypic abnormality).
(D) Coverage for concepts with a different number of tokens in the concept string.
(E) Coverage for concepts with different average token lengths.
(F) Coverage for concepts with different composition of grammatical properties. If a concept contains multiple POSs, it will be counted multiple times. POS, part of speech; DT, determiner; IN, preposition/subordinating conjunction; JJ, adjective; NN, noun; NNP, proper noun; NNS, noun plural; VBN, verb.
(G) Coverage for concepts with different ontologies involved in its design patterns. If a concept contains multiple ontologies in its design pattern, it will be counted multiple times. BFO, Basic Formal Ontology; CHEBI, Chemical Entities of Biological Interest; CL, Cell Ontology; GO, Gene Ontology; HP, Human Phenotype Ontology; PATO, Phenotype And Trait Ontology; RO, Relation Ontology; UBERON, Uber Anatomy Ontology.

the annotation (Figure 5D). While the difference of statistics between annotated and not-annotated pairs is not obviously affected by the annotation provenance, it is larger in the more frequently annotated pairs. Table 1 shows the number of phenotype-disease concept pairs in different association quantiles, which is defined as the quantiles of the statistics distribution of the HPO-MONDO concept pairs for a given disease. Our results showed the original HPO-annotated concept association pairs are significantly enriched in the high confident categories, indicating those statistics derived from the EHR can serve as a proxy to measure the association strength between two concepts and potentially identify previously unknown associated phenotype-disease pairs.

Manual review of novel disease-phenotype pairs

In total, 271 novel disease-phenotype pairs have been sampled and graded and received confidence scores from clinicians (Table S4). As shown in Figure 6A, among those pairs, >65% of them are found to be at least "likely to be associated" (i.e., precision). Figure 6B shows the mean of clinicians' confidence across different ORs (log scale). A significant positive correlation was observed between ORs (log scale) quantiles for the given disease and clinicians' confidence (Pearson correlation $r^2 = 0.2$; p value $= 0.005$) for that novel identified disease-phenotype pairs. On average, the phenotypes falling into the top quantiles for that disease evaluated received a higher score. We also investigated the distribution of absolute ORs (log scale) among novel disease-phenotype pairs with different confidence levels as shown in Figure 6C. There is a clear trend that with an increasing OR (log scale), clinician confidence in a true association also increased (Pearson correlation $r^2 = 0.2$; p value < 0.001).

Example: Duchenne muscular dystrophy

To illustrate the utility of the OARD database, we will take Duchenne muscular dystrophy (DMD), which is a severe type of muscular dystrophy, as an example. DMD is a lethal X-linked recessive muscle dystrophy caused by different mutations including mostly frame-shifting deletions and duplications and rare point mutations in the *DMD*.[36] Commonly associated phenotypes include muscle weakness, progressive loss of skeletal muscle mass, and later-onset cardiomyopathy, which ultimately leads to cardiac and respiratory failure and premature death in DMD-affected individuals. In the original HPO annotation, there are 16 phenotype concepts annotated with DMD. In the OARD (CUIMC/Notes dataset), there are 211 phenotype concepts identified. One clinician (W.K.C.), who has extensive knowledge of DMD, manually reviewed the list and determined whether the phenotypes identified are

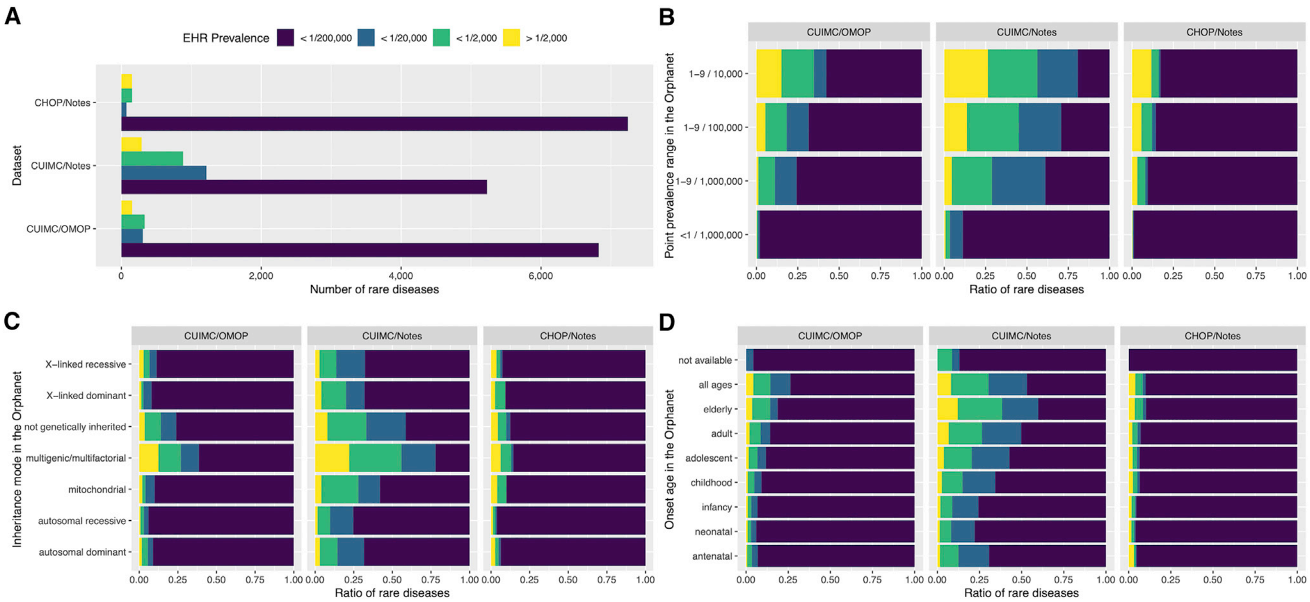


Figure 4. The prevalence of the rare disease concepts in the OARD database
(A) The overall prevalence for each dataset.
(B) Distribution of the EHR observed rare disease concepts (represented by different colors) across different annotated point estimation categories. Only those who have a point prevalence annotated in the Orphanet databases (different colors) are included in this figure.
(C) Distribution of the EHR observed rare disease concepts across different annotated inherit modes. Only those who have an inherit mode annotated in the Orphanet databases (different colors) are included in this figure.
(D) Distribution of the EHR observed rare disease concepts across different annotated onset ages.

associated with DMD (Table S5). Figure 7A showed the precision increased with the ORs (log scale). Among those with ORs larger than 5, the precision is 91.7% (11/12). While some concepts such as "Gower sign" (ORs [log scale]: 7.56), "calf muscle hypertrophy" (ORs [log scale]: 7.76), and "pelvic girdle muscle weakness" (ORs [log scale]: 6.20) are well-known phenotypes associated with DMD, OARD is able to identify some phenotypes, including "elevated hepatic transaminase" (ORs [log scale]: 2.41),"autism" (ORs [log scale]: 2.20), and "osteopenia" (ORs [log scale]: 1.55), that are completely missed in the original HPO annotation and provides much more granular phenotype annotations with detailed statistics for some phenotypes. For example, in the original HPO annotation, "respiratory insufficiency; HP: 0002093" is annotated with DMD, while in OARD, a more granular term, "respiratory insufficiency due to muscle weakness; HP: 0002747," was provided. To provide an example of how OARD can identify those phenotypes facilitating the differential diagnosis, we selected another disease, "juvenile dermatomyositis" (JDM [ORPHA: 93672]), which can be misdiagnosed as muscular dystrophy,[37] and added those phenotypes identified in JDM for review. Figure 7B showed there are 70 phenotypes overlapped between the two diseases, 19 phenotypes that are uniquely identified in JDM, and 141 that phenotypes uniquely identified in DMD. Figure 7C showed 64.5% of the phenotypes found uniquely in DMD, 17.1% of the overlapping phenotypes, and 15.8% of the phenotypes found uniquely in JDM are determined to be true positive phenotypes associated with DMD, indicating the unique annotated phenotypes can serve as distinguishing features for differential diagnosis between DMD and JDM.

Discussion

Our study provides a real-world-data-driven resource using the EHR to generate this knowledge in an automated and scalable fashion to enrich the current phenotype-disease knowledge graph. Prior to our study, we found the EHR has rarely been explored to generate disease-phenotype associations on a large scale. In the eRAM study, the MIMIC-III dataset has been leveraged to identify disease-phenotype associations.[38] Unlike MIMIC-III dataset,[39] our study is based on multiple datasets generated by two large academic institutions that covering more than 10 million individuals, providing associations identified in much more diverse clinical environments, spanning a wide age range and different disease subgroups. Similar to our study, Shen et al. have applied a data-driven approach to enrich existing rare disease resources by mining phenotype-disease associations from the EHR. However, their resource of the disease-phenotype association is not publicly shareable.[40] We provide OARD with the aim to facilitate the knowledge dissemination of the EHR-based disease-phenotype annotation and enrich the current disease-phenotype knowledge graph, specifically for rare genetic disorders. To the best of our knowledge, OARD is the largest publicly available open resource of rare diseases and their

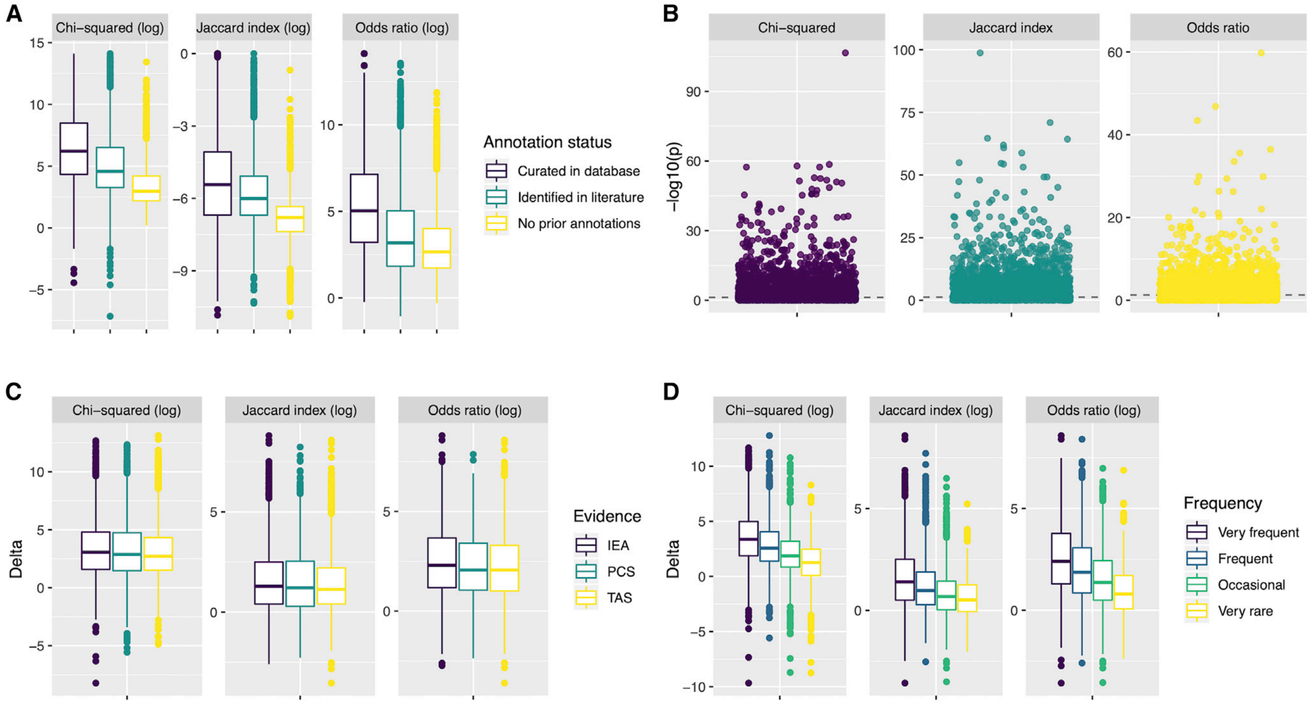


Figure 5. Comparison of association statistics between annotated pairs and not-annotated pairs
(A) Boxplot of the Chi-squared, Jaccard index, and odds ratio distribution (log scale) calculated in the OARD for annotated disease-phenotype pairs and not-annotated disease-phenotype pairs.
(B) Distribution of the p value of the concepts obtained from the Wilcoxon test by comparing its annotated pairs and not-annotated pairs. The dashed line is p value = 0.05.
(C) Distribution of the mean difference of the statistics between annotated pair and not-annotated pair of a given disease for different evidence levels. IEA, annotations that have been extracted by parsing the clinical features sections of the OMIM; PCS, annotations that have been extracted from articles in the medical literature; TAS, annotations that have a "traceable author statement", usually reviews or disease entries (e.g., OMIM) that only refers to the original publication.
(D) Distribution of the mean difference of the statistics between annotated pair and not-annotated pair of a given disease for different levels of frequency that a phenotype was found in the given diseases. Very rare, in 1% to 4% of the cases; occasional: in 5% to 29% of the cases; frequent, in 30% to 79% of the cases; very frequent, in 80% to 99% of the cases.

phenotype annotations curated from millions of records extracted from two large academic institutions. OARD also provides fine-grained statistics, including frequency and association, derived from the underlying real-world data, which is more than the rough estimation and annotation generated by human experts, filling an important gap in the rare disease community to enable data-driven disease-phenotype research. However, considering only less than half of the rare diseases were annotated in the OARD, it is important to emphasize OARD should not be used alone to replace the current HPO-Jax-annotation-based diagnostic pipeline. Rather, it should be treated as another resource to complement the current knowledge graph with the ability to provide non-binary (summary statistics) relationship annotations.

Novel disease-phenotype association identified

Existing expert-curated annotations and literature-mined associations showed a higher association strength (e.g., odds ratio) in OARD compared to unknown annotations, demonstrating OARD-identified associations align well with previously expert-curated knowledge. By sampling novel identified paired concepts, our evaluation based on a sampled set showed >65% precision for those novel identified annotations in the top 50% quantiles (Figure 6A), which in total corresponded to >1 million disease-phenotype associations not previously annotated by humans or found in literature. If we use OR-based thresholds, the precision of novel identified concept pairs are 72%, 78%, 80%, and 87% for corresponding cutoffs at 2, 3, 4, and 5, respectively. We did not select a cut-off specifically for the users to determine the novel disease-phenotype associations. In practice, the edges connecting nodes (i.e., association between pair-wised diseases and phenotypes or phenotypes and phenotypes) can be treated as probabilistic values rather than deterministic binary values in implementing the differential diagnostic pipeline. By adopting a privacy-conserved infrastructure, we made the association statistics available by querying the publicly available API or browsing a web app interface. Researchers can easily leverage those provided statistics to either enrich the current disease-phenotype knowledge graph on the basis of their selected cut-off or train a co-occurrence-based embedding (e.g., GloVe) for rare disease relevant concepts.[41] To note, beside the overall population-derived statistics, those associations were also available for each

Table 1. The number of previously annotated disease-phenotype concept pairs captured by OARD based on CUIMC/Notes dataset in different quantiles of association statistic

Quantile	Odds ratio	Jaccard index	χ^2
0%–10%	1,395 (0.5%)	1,353 (0.5%)	1,299 (0.5%)
10%–20%	1,914 (0.7%)	1,816 (0.7%)	1,412 (0.5%)
20%–30%	2,472 (0.9%)	2,329 (0.9%)	1,604 (0.6%)
30%–40%	2,826 (1.1%)	2,594 (1.0%)	1,964 (0.8%)
40%–50%	3,359 (1.3%)	3,032 (1.2%)	2,279 (0.9%)
50%–60%	3,740 (1.4%)	3,405 (1.3%)	2,773 (1.1%)
60%–70%	3,974 (1.5%)	3,679 (1.4%)	3,216 (1.2%)
70%–80%	4,160 (1.6%)	3,814 (1.5%)	3,975 (1.5%)
80%–90%	4,378 (1.7%)	4,202 (1.6%)	5,136 (2.0%)
90%–100%	6,623 (2.5%)	8,617 (3.3%)	11,183 (4.2%)

The concept pairs are ordered according to the association statistic and divided into different quantiles. Only disease concepts with at least ten disease-phenotype pairs were included. Percentages in the parentheses represent the percentage of HPO-original-annotated or literature-reported pairs in this quantile.

subset. Therefore, researchers who focused on a specific domain could flexibility select the corresponding dataset to enhance the knowledge for their research domain.

Building differential diagnosis algorithm using summary statistics only

Besides the pair-wise association provided by mining the EHR, OARD is also capable of aiding with differential diagnosis by inputting multiple phenotypes. Unlike other differential diagnosis systems,[13,42] OARD does not simply assume all the phenotypes observed are independent. Instead, by borrowing the "linkage disequilibrium (LD)"[43] concept widely used in developing polygenetic risk prediction models,[44–47] OARD assign less weights to those query phenotypes carrying redundant information. Considering an extreme situation when two identical phenotypes were provided in the query, OARD assures each (same) phenotype can only contribute half of the association to the final score, which makes the final score identical to the results when only one phenotype is provided. In the future, we believe more advanced methods, such as Bayesian-based approaches such as LDpred[25,45] and PRS-CS,[46] can be adopted to develop a phenotypic risk score[48] for rare disease diagnosis purposes. On the other hand, by leveraging the "relative frequency" of a phenotype in a given disease provided by OARD, we could potentially build an interactive system (similar to our previously developed DQueST[49]) that would be helpful for the phenotype assessment and differential diagnosis by providing recommendations on the assessment of those high "relative frequency" phenotypes in a suspected rare disease. While the focus of this manuscript is not to develop a phenotypic-driven risk algorithm, OARD provides a fundamental knowledgebase for such algorithms by sharing pair-wise association statistics, which is similar to many genome-wide association knowledgebase, such as GWAS catalog.[50] We encourage the broader research community to explore this resource and develop more advanced diagnostic algorithms. Furthermore, despite the latest advances in data standardization (e.g., PhenoPacket[51]) for individual-level data exchange, regulatory barriers still exist for most researchers to get individual-level data access. OARD enables the development of federated machine learning approaches,[52–54] which do not require individual-level data access. This is extremely important for rare disease research because the sample size is often limited in a single institution. By aggregating the summary-level statistics from multiple institutions, novel knowledge related to rare diseases can be discovered.

A semi-portable pipeline

Our pipeline to derive phenotype and rare-disease associations from the structured data is completely portable to institutions that have converted their databases into the OMOP Common Data Model.[28] Unfortunately, as shown in our results, while the datasets derived on the basis of on standardized codes (i.e., CUIMC/OMOP) have a similar number of individuals included, the coverage of the phenotypic concepts were significantly lower than clinical narrative derived datasets (i.e., CUIMC/Notes, CHOP/Notes). This is consistent with previous findings that rare-disease-related concepts have lower coverage in standard terminology.[27] While many natural-language-processing pipelines have been developed to recognize rare-disease-related concepts (e.g., HPO concepts),[17,29,55] it is notorious for its lack of portability and requiring great efforts to configure those pipelines with institution specific customization.[56] As an alternative, we explored a context-aware keyword search for concept extraction in this study. Given most sites have their clinical notes already indexed with some indexing technology (e.g., Solr), it is straightforward to identify individuals with a combination of queries consisting of concept string and the relevant context. A similar strategy has been previously used to identify relevant medical concept correlations in an unstructured nephrology database TBase.[57] We believe this approach is more generalizable, easier to customize, and provides a faster solution when dealing with huge amounts of clinical narratives (e.g., ~100 million) while the number of query concepts is relatively small (<100,000). However, future comparison studies are needed to evaluate its impact on the efficiency, accuracy, and portability of the concept recognition pipeline.

Bottleneck in concept recognition

Despite leveraging different methods, the correlation between phenotype concept coverage and various concept properties (e.g., token numbers, subontology, parts of speech [POSs], design ontology) is fairly consistent among the results obtained via NLP and context-aware keyword searches, even though they derived from two institutions. For example, the number of tokens involved in the concepts (rather than the average string length) can largely

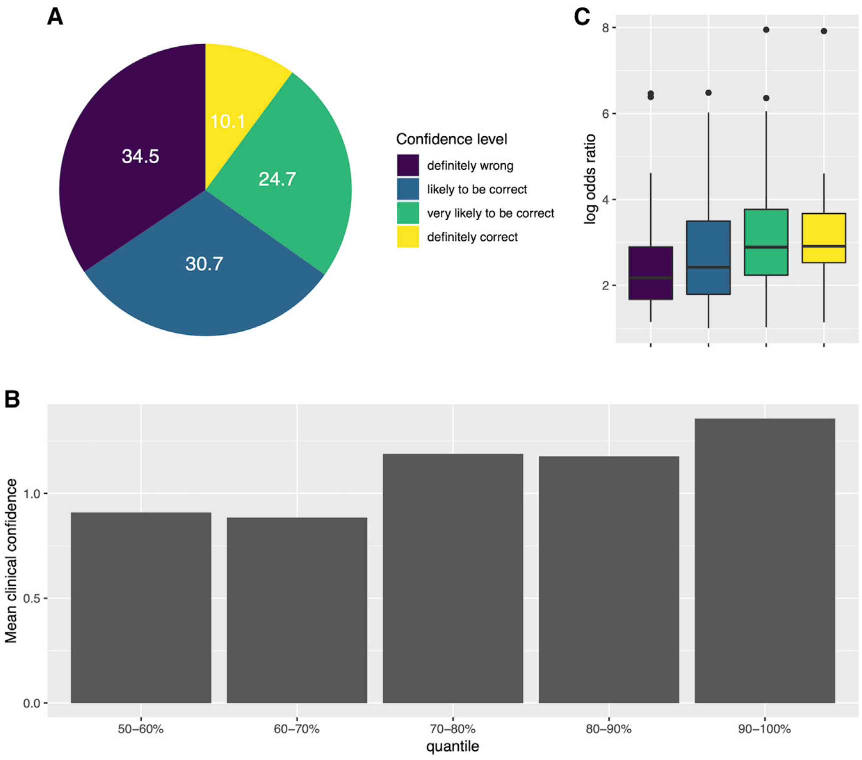


Figure 6. Manual review of a sampled list of novel identified disease-phenotype pairs
Quantiles were derived from the odds ratio distribution of the disease-phenotype pairs for a given disease. Clinician-assigned confidence scores were defined as 0 = definitely wrong; 1 = likely to be correct; 2 = very likely to be correct; and 3 = definitely correct.
(A) The distribution of clinician-assigned confidence scores.
(B) Mean clinical confidence score in each quantile based on the odds ratio for a given disease.
(C) The distribution of log odds ratio in different clinician-assigned confidence scores.

affect the retrieval results. Once the number of tokens in a concept is larger than five, it is likely that no individuals can be returned either by querying that concept in the note or by processing documents to recognize that concept. Also, if concepts are composed of determiner, preposition/subordinating conjunction, adjective, the chance of the concept mapping is lower. This indicates that a major bottleneck in concept recognition is the concept mapping between the narrative language and concept definition on the ontology. The most commonly used concept mapping can be treated as a document retrieval procedure, which is often based on a Lucene index strategy.[58] Advanced language model based approaches should be considered in the future to improve this mapping procedure.[59] Meanwhile, the design of the ontology should also take this aspect into consideration so that the challenges of the concept recognition can be eased in the future. Though our resource only contains phenotypes identified from "visit"-related notes, it does not necessarily mean phenotypes identified by other procedures (e.g., radiography) are not included. If physicians believe those phenotypes are important to note, those can still be presented in the "office visit"-related notes as a summarization of individuals' phenotypes. For example, "pulmonary nodule (HP: 0033608)," which is typically observed by chest radiography or computer tomography imaging, does show up in ~5,800 individuals in our results.

Rare disease epidemiology reflected using the EHR

In order to inform public policy, it is important to have an evidence-based estimation of disease prevalence. Currently, the best estimation of the point prevalence is available via the epidemiological data in the Orphanet database.[60] Our EHR-derived rare-disease concept prevalence showed a high correlation with the point prevalence obtained from the Orphanet, indicating the potential of using the EHR to support rare disease epidemiology research. However, the reported prevalence is a "concept" prevalence. While the trend is consistent between these two estimates of prevalence, it might not be appropriate to equate the "concept prevalence" as a "disease prevalence." In fact, identifying rare disease individuals from the EHR itself remains a challenging phenotyping problem. Among many ongoing phenotyping efforts for specific rare disorders,[61–63] few can be scaled up to provide a systematic and reliable approach to estimate the prevalence for any rare diseases. Future studies are needed to develop more generalizable and scalable approaches for rare disease prevalence estimation with EHR data.

Limitations and future work

There are a few limitations. First, although using the NLP and information retrieval techniques greatly improve the concept coverage, this is by no means a validated phenotyping algorithm. Therefore, the absolute frequencies and cooccurrences provided by OARD should be interpreted with caution. Although a subset of the associations identified by the OARD were partially evaluated, there may still be false positives caused by systematic errors. Second, unlike the genetic data reflecting the underlying genetic profiles, the data derived from the EHR is actually a measure of the health care process. Given the current OARD data are derived from two academic institutions, users should be aware that the health care processes can be different among different types of medical settings (i.e., community, academic, tertiary care settings). Third, though we identify the commonalities of the NLP-based and information-retrieval-based pipelines in rare disease concepts recognition, we were unable to

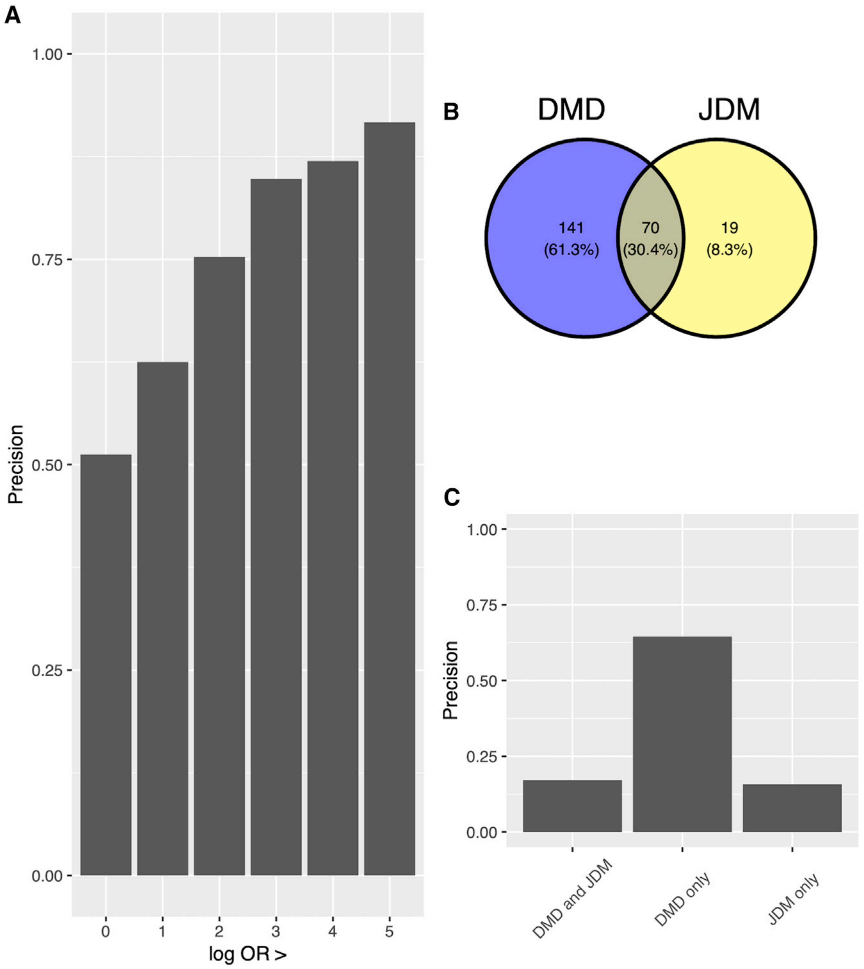


Figure 7. Manual review of OARD-identified DMD-associated phenotypes
(A) Precision according to different ORs (log-scale) thresholds.
(B) A Venn diagram showing the number of phenotypes identified for DMD and JDM.
(C) Precision for phenotypes having different association with DMD and JDM. DMD, Duchenne muscular dystrophy; JDM, juvenile dermatomyositis.

compare the differences between these two methods given it is fully confounded by the two-institution design. Finally, our manual evaluation is only limited to a small set of identified phenotype-disease associations, and the limited number of the rare disease specialists forced us to adopt an expert-oriented evaluation approach, which might cause bias, and the accuracy reported from our evaluation may vary across different rare diseases.

Data and code availability

OARD web app is publicly available at http://rare.cohd.io. The source code for OARD API can be found at https://github.com/WengLab-InformaticsResearch/cohd-rare and the source code for OARD React web application can be found at https://github.com/WengLab-InformaticsResearch/oard-react.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.08.002.

Acknowledgments

This study was supported by National Library of Medicine/National Human Genomic Research Institute grant R01LM012895 and National Center for Advancing Translational Sciences of the National Institutes of Health under award number OT2TR003434. We thank Dr. Natarajan Karthik and the CUIMC DBMI operational team for providing access to the CUIMC clinical data warehouse and clinical narratives. We also thank the CHOP Arcus team for technical assistance in compiling and analyzing the cTAKES processed clinical notes.

Declaration of interests

The authors declare no competing interests.

Received: May 3, 2022
Accepted: August 1, 2022
Published: August 22, 2022

Web resources

Columbia Open Health Data, https://cohd.io/

HPO JAX browser, https://hpo.jax.org

MONDO ontology, https://www.ebi.ac.uk/ols/ontologies/mondo

OARD GitHub, https://github.com/WengLab-InformaticsResearch/oard-react

OARD SMART API documents, http://rare.cohd.io/api

OARD web app, http://rare.cohd.io

ORPHANET browser, https://www.orpha.net/consor/cgi-bin/index.php

Pubcasefinder, https://pubcasefinder.dbcls.jp/

References

1. Griggs, R.C., Batshaw, M., Dunkle, M., Gopal-Srivastava, R., Kaye, E., Krischer, J., Nguyen, T., Paulus, K., Merkel, P.A.; and Rare Diseases Clinical Research Network (2009). Clinical research for rare disease: opportunities, challenges, and solutions. Mol. Genet. Metab. *96*, 20–26.
2. Anderson, M., Elliott, E.J., and Zurynski, Y.A. (2013). Australian families living with rare disease: experiences of diagnosis, health services use and needs for psychosocial support. Orphanet J. Rare Dis. *8*, 22.
3. Zurynski, Y., Frith, K., Leonard, H., and Elliott, E. (2008). Rare childhood diseases: how should we respond? Arch. Dis. Child. *93*, 1071–1074.
4. Adams, L.S., Miller, J.L., and Grady, P.A. (2016). The spectrum of caregiving in palliative care for serious, advanced, rare diseases: key issues and research directions. J. Palliat. Med. *19*, 698–705.
5. Engel, P., Bagal, S., Broback, M., and Boice, N. (2013). Physician and patient perceptions regarding physician training in rare diseases: the need for stronger educational initiatives for physicians. J Rare Dis *1*, 1–14.
6. Bogart, K.R., and Irvin, V.L. (2017). Health-related quality of life among adults with diverse rare disorders. Orphanet J. Rare Dis. *12*, 177. https://doi.org/10.1186/s13023-017-0730-1.
7. Cipriani, V., Pontikos, N., Arno, G., Sergouniotis, P.I., Lenassi, E., Thawong, P., Danis, D., Michaelides, M., Webster, A.R., Moore, A.T., et al. (2020). An improved phenotype-driven tool for rare mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. Genes *11*, E460. https://doi.org/10.3390/genes11040460.
8. Smedley, D., Jacobsen, J.O.B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat. Protoc. *10*, 2004–2015. https://doi.org/10.1038/nprot.2015.124.
9. Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am. J. Hum. Genet. *94*, 599–610. https://doi.org/10.1016/j.ajhg.2014.03.010.
10. Deelen, P., van Dam, S., Herkert, J.C., Karjalainen, J.M., Brugge, H., Abbott, K.M., van Diemen, C.C., van der Zwaag, P.A., Gerkes, E.H., Zonneveld-Huijssoon, E., et al. (2019). Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat. Commun. *10*, 2837.
11. Deisseroth, C.A., Birgmeier, J., Bodle, E.E., Kohler, J.N., Matalon, D.R., Nazarenko, Y., Genetti, C.A., Brownstein, C.A., Schmitz-Abe, K., Schoch, K., et al. (2019). ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet. Med. *21*, 1585–1593.
12. Xin, H., Changchen, W., Lei, L., Meirong, Y., Ye, Z., and Bo, P. (2019). The phenolyzer suite: prioritizing the candidate genes involved in microtia. Ann. Otol. Rhinol. Laryngol. *128*, 556–562. https://doi.org/10.1177/0003489419840052.
13. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat. Methods *12*, 841–843. https://doi.org/10.1038/nmeth.3484.
14. Son, J.H., Xie, G., Yuan, C., Ena, L., Li, Z., Goldstein, A., Huang, L., Wang, L., Shen, F., Liu, H., et al. (2018). Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. Am. J. Hum. Genet. *103*, 58–73. https://doi.org/10.1016/j.ajhg.2018.05.010.
15. Zhao, M., Havrilla, J.M., Fang, L., Chen, Y., Peng, J., Liu, C., Wu, C., Sarmady, M., Botas, P., Isla, J., et al. (2020). Phen2-Gene: rapid phenotype-driven gene prioritization for rare diseases. NAR Genom. Bioinform. *2*, lqaa032. https://doi.org/10.1093/nargab/lqaa032.
16. Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K.M., Chénier, S., Chitayat, D., Faghfoury, H., Meyn, M.S., Ray, P.N., et al. (2013). Pheno tips: patient phenotyping software for clinical and research use. Hum. Mutat. *34*, 1057–1065.
17. Liu, C., Peres Kury, F.S., Li, Z., Ta, C., Wang, K., and Weng, C. (2019). Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res. *47*, W566–W570. https://doi.org/10.1093/nar/gkz386.
18. Havrilla, J.M., Singaravelu, A., Driscoll, D.M., Leonard, M., Helbig, I., Medne, L., Wang, K., Krantz, I., and Desai, B.R. (2021). PheNominal: An EHR-Integrated Web Application for Structured Deep Phenotyping at the Point of Care.
19. Köhler, 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. https://doi.org/10.1093/nar/gky1105.
20. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798. https://doi.org/10.1093/nar/gku1205.
21. Pavan, S., Rommel, K., Mateo Marquina, M.E., Höhn, S., Lanneau, V., and Rath, A. (2017). Clinical practice guidelines for rare diseases: The Orphanet Database. PLoS One *12*, e0170365. https://doi.org/10.1371/journal.pone.0170365.
22. Bragin, E., Chatzimichali, E.A., Wright, C.F., Hurles, M.E., Firth, H.V., Bevan, A.P., and Swaminathan, G.J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. *42*, D993–D1000. https://doi.org/10.1093/nar/gkt937.
23. Fujiwara, T., Yamamoto, Y., Kim, J.D., Buske, O., and Takagi, T. (2018). PubCaseFinder: a case-report-based, phenotype-driven differential-diagnosis system for rare diseases. Am. J. Hum. Genet. *103*, 389–399. https://doi.org/10.1016/j.ajhg.2018.08.003.
24. Xu, R., Li, L., and Wang, Q. (2013). Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. Bioinformatics *29*, 2186–2194. https://doi.org/10.1093/bioinformatics/btt359.
25. Kafkas, Ş., Althubaiti, S., Gkoutos, G.V., Hoehndorf, R., and Schofield, P.N. (2021). Linking common human diseases to their phenotypes; development of a resource for human phenomics. J. Biomed. Semantics *12*, 17. https://doi.org/10.1186/s13326-021-00249-x.

26. Ta, C.N., Dumontier, M., Hripcsak, G., Tatonetti, N.P., and Weng, C. (2018). Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. Sci. Data *5*, 180273. https://doi.org/10.1038/sdata.2018.273.

27. Fung, K.W., Richesson, R., and Bodenreider, O. (2014). Coverage of rare disease names in standard terminologies and implications for patients, providers, and research. AMIA Annu. Symp. Proc. *2014*, 564–572.

28. Hripcsak, G., Duke, J.D., Shah, N.H., Reich, C.G., Huser, V., Schuemie, M.J., Suchard, M.A., Park, R.W., Wong, I.C.K., Rijnbeek, P.R., et al. (2015). Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud. Health Technol. Inform. *216*, 574–578.

29. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., and Chute, C.G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. *17*, 507–513. https://doi.org/10.1136/jamia.2009.001560.

30. Unni, D.R., Moxon, S.A., Bada, M., Brush, M., Bruskiewich, R., Caufield, J.H., Clemons, P.A., Dancik, V., Dumontier, M., and Fecho, K. (2022). Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clinical and Translational Science.

31. Zhang, X.A., Yates, A., Vasilevsky, N., Gourdine, J.P., Callahan, T.J., Carmody, L.C., Danis, D., Joachimiak, M.P., Ravanmehr, V., Pfaff, E.R., et al. (2019). Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit. Med. *2*, 32. https://doi.org/10.1038/s41746-019-0110-4.

32. Chapman, W., Dowling, J., and Chu, D. (2007). ConText: An Algorithm for Identifying Contextual Features from Clinical Text, pp. 81–88.

33. Hribar, M.R., Read-Brown, S., Goldstein, I.H., Reznick, L.G., Lombardi, L., Parikh, M., Chamberlain, W., and Chiang, M.F. (2018). Secondary use of electronic health record data for clinical workflow analysis. J. Am. Med. Inform. Assoc. *25*, 40–46. https://doi.org/10.1093/jamia/ocx098.

34. Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., et al. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J. Biomed. Semantics *7*, 44. https://doi.org/10.1186/s13326-016-0088-7.

35. Hastings, J., Glauer, M., Memariani, A., Neuhaus, F., and Mossakowski, T. (2021). Learning chemistry: exploring the suitability of machine learning for the task of structure-based chemical ontology classification. J. Cheminform. *13*, 23. https://doi.org/10.1186/s13321-021-00500-8.

36. Salmaninejad, A., Jafari Abarghan, Y., Bozorg Qomi, S., Bayat, H., Yousefi, M., Azhdari, S., Talebi, S., and Mojarrad, M. (2021). Common therapeutic advances for Duchenne muscular dystrophy (DMD). Int. J. Neurosci. *131*, 370–389.

37. Cheng, Z., Min, X., Yiming, S., and Cheng, Z. (2016). A Case of Misdiagnosed Juvenile Dermatomyositis. J. Neurol. Neuromedicine *1*, 45–47.

38. Jia, J., An, Z., Ming, Y., Guo, Y., Li, W., Liang, Y., Guo, D., Li, X., Tai, J., Chen, G., et al. (2018). eRAM: encyclopedia of rare disease annotations for precision medicine. Nucleic Acids Res. *46*, D937–D943. https://doi.org/10.1093/nar/gkx1062.

39. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., and Mark, R.G. (2016). MIMIC-III, a freely accessible critical care database. Sci. Data *3*, 160035. https://doi.org/10.1038/sdata.2016.35.

40. Shen, F., Zhao, Y., Wang, L., Mojarad, M.R., Wang, Y., Liu, S., and Liu, H. (2019). Rare disease knowledge enrichment through a data-driven approach. BMC Med. Inform. Decis. Mak. *19*, 32. https://doi.org/10.1186/s12911-019-0752-9.

41. Lee, J., Liu, C., Kim, J.H., Butler, A., Shang, N., Pang, C., Natarajan, K., Ryan, P., Ta, C., and Weng, C. (2021). Comparative effectiveness of medical concept embedding for feature engineering in phenotyping. JAMIA open *4*, ooab028.

42. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. *85*, 457–464. https://doi.org/10.1016/j.ajhg.2009.09.003.

43. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., and Lander, E.S. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204. https://doi.org/10.1038/35075590.

44. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. *97*, 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001.

45. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.

46. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. https://doi.org/10.1038/s41467-019-09718-5.

47. Liu, C., Zeinomar, N., Chung, W.K., Kiryluk, K., Gharavi, A.G., Hripcsak, G., Crew, K.D., Shang, N., Khan, A., Fasel, D., et al. (2021). Generalizability of Polygenic Risk Scores for Breast Cancer Among Women With European, African, and Latinx Ancestry. JAMA Netw. Open *4*, e2119084.

48. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science *359*, 1233–1239. https://doi.org/10.1126/science.aal4043.

49. Liu, C., Yuan, C., Butler, A.M., Carvajal, R.D., Li, Z.R., Ta, C.N., and Weng, C. (2019). DQueST: dynamic questionnaire for search of clinical trials. J. Am. Med. Inform. Assoc. *26*, 1333–1343.

50. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901. https://doi.org/10.1093/nar/gkw1133.

51. Jacobsen, J.O.B., Baudis, M., Baynam, G.S., Beckmann, J.S., Beltran, S., Buske, O.J., Callahan, T.J., Chute, C.G., Courtot, M., Danis, D., et al. (2022). The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat. Biotechnol. *40*, 817–820.

52. Hossen, M.N., Panneerselvam, V., Koundal, D., Ahmed, K., Bui, F.M., and Ibrahim, S.M. (2022). Federated machine learning for detection of skin diseases and enhancement

of Internet of Medical Things (IoMT) security. IEEE J. Biomed. Health Inform., 1. https://doi.org/10.1109/JBHI.2022.3149288.

53. Vaid, A., Jaladanki, S.K., Xu, J., Teng, S., Kumar, A., Lee, S., Somani, S., Paranjpe, I., De Freitas, J.K., Wanyan, T., et al. (2021). Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. JMIR Med. Inform. *9*, e24207. https://doi.org/10.2196/24207.

54. Zerka, F., Barakat, S., Walsh, S., Bogowicz, M., Leijenaar, R.T.H., Jochems, A., Miraglio, B., Townend, D., and Lambin, P. (2020). Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO Clin. Cancer Inform. *4*, 184–200. https://doi.org/10.1200/CCI.19.00047.

55. Luo, L., Yan, S., Lai, P.T., Veltri, D., Oler, A., Xirasagar, S., Ghosh, R., Similuk, M., Robinson, P.N., and Lu, Z. (2021). PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. Bioinformatics *37*, 1884–1890. https://doi.org/10.1093/bioinformatics/btab019.

56. Shang, N., Liu, C., Rasmussen, L.V., Ta, C.N., Caroll, R.J., Benoit, B., Lingren, T., Dikilitas, O., Mentch, F.D., Carrell, D.S., et al. (2019). Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. J. Biomed. Inform. *99*, 103293. https://doi.org/10.1016/j.jbi.2019.103293.

57. Schmidt, D., Budde, K., Sonntag, D., Profitlich, H.-J., Ihle, M., and Staeck, O. (2017). A novel tool for the identification of correlations in medical data by faceted search. Comput. Biol. Med. *85*, 98–105.

58. Zhang, Y., and Li, J.-l. (2009). Research and Improvement of Search Engine Based on Lucene (IEEE), pp. 270–273.

59. Luo, L., Yan, S., Lai, P.-T., Veltri, D., Oler, A., Xirasagar, S., Ghosh, R., Similuk, M., Robinson, P.N., and Lu, Z. (2021). PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. Bioinformatics *37*, 1884–1890.

60. Nguengang Wakap, S., Lambert, D.M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur. J. Hum. Genet. *28*, 165–173. https://doi.org/10.1038/s41431-019-0508-0.

61. Sun, A.Z., Shu, Y.-H., Harrison, T.N., Hever, A., Jacobsen, S.J., O'Shaughnessy, M.M., and Sim, J.J. (2020). Identifying patients with rare disease using electronic health record data: The Kaiser Permanente Southern California membranous nephropathy cohort. Perm. J. *24*.

62. Cohen, A.M., Chamberlin, S., Deloughery, T., Nguyen, M., Bedrick, S., Meninger, S., Ko, J.J., Amin, J.J., Wei, A.J., and Hersh, W. (2020). Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. PLoS One *15*, e0235574.

63. Khare, R., Kappelman, M.D., Samson, C., Pyrzanowski, J., Darwar, R.A., Forrest, C.B., Bailey, C.C., Margolis, P., Dempsey, A.; and And the PEDSnet Computable Phenotype Working Group (2020). Development and evaluation of an EHR-based computable phenotype for identification of pediatric Crohn's disease patients in a national pediatric learning health system. Learn. Health Syst. *4*, e10243.

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

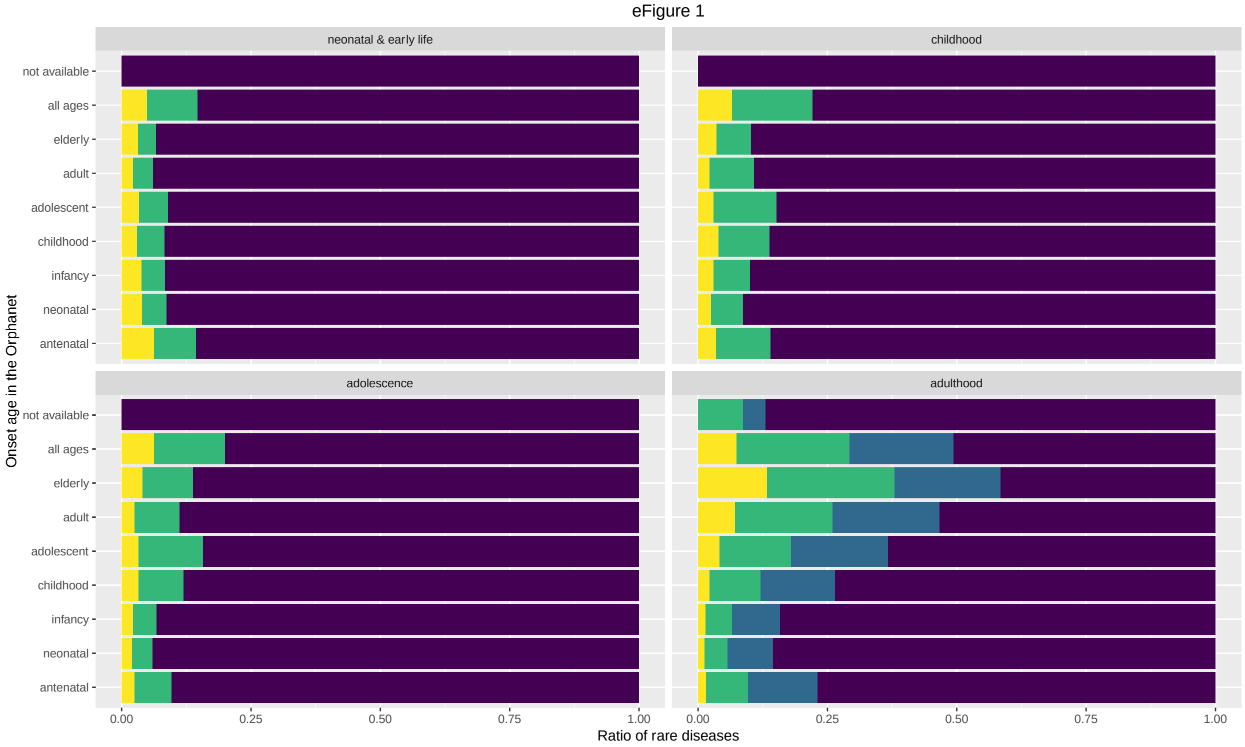


Figure S1 The prevalence of the rare disease concepts in the OARD database across different age groups. Y-axis is the onset annotation of the rare diseases in the Orphanet; Different colors represent the prevalence derived from the OARD database.

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

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

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.