



Published in final edited form as:

Stud Health Technol Inform. 2025 May 15; 327: 123–127. doi:10.3233/SHTI250286.

Identifying Phenotypes for Earlier Diagnosis of Rare Diseases

Casey N. TA^{a,1}, Cong LIU^{a,b}, Chunhua WENG^a

^aDepartment of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, U.S.A.

^bDivision of Genetics and Genomics, Department of Pediatrics, Boston Children's Hospital, Boston, MA, U.S.A.

Abstract

Rare diseases, while individually rare, cumulatively affect a large population, and patients often undergo long and arduous diagnostic odysseys. Toward the goal of supporting earlier diagnosis of rare diseases, we developed generalizable methods of extracting rare diseases and phenotypes from structured electronic health records and clinical notes. We analyzed the distributions of the age of onset of phenotypes per disease to identify disease-phenotype associations, producing a dataset with over 500 thousand associations covering 2300 rare diseases. Disease-phenotype associations are characterized by disease prevalence and mean age of onset of the phenotype to aid phenotype selection according to the priorities of the clinical decision support task.

Keywords

Rare Diseases; Phenotypes; Electronic Health Records; Clinical Notes

1. Introduction

The European Union regulation on orphan medicinal products defines rare disease as conditions that affect fewer than 5 people per 10,000, and in the United States, the Orphan Drug Act defines a rare disease as a condition that affects fewer than 200,000 people [1,2]. Although each rare disease affects a relatively small number of people, there are over 7,000 rare diseases that, in total, affect over 30 million people in the US. Due to the substantial number of rare diseases and their infrequency within the population, few providers are experts in diagnosing each rare disease. Consequently, patients with rare diseases often endure long diagnostic odysseys (median 9 months and average 4.7 years between symptom onset and confirmed diagnosis), with 22% consulting eight or more healthcare professionals, 73% were misdiagnosed at least once, and many experienced other factors contributing to diagnostic delays [3].

Recent approaches promise to improve diagnosis of rare diseases by integrating knowledge about the relationships between diseases, genes, variants, and phenotypes to improve accuracy and sensitivity of candidate disease and gene prioritization [4–7]. To aid in the

¹Corresponding Author: Casey N. Ta; ct2865@cumc.columbia.edu.

diagnosis of rare diseases, we developed a generalizable approach of mining electronic health records (EHR) to identifying phenotypes associated with rare diseases. The approach evaluates the age of onset of the phenotypes within the rare disease cohort and compares it to that of the general patient cohort to identify distinctive age of onset distributions. Additional data regarding phenotype prevalence and mean age of onset are provided to help prioritize the disease-phenotype relationships.

2. Methods

2.1 Data Sources

We analyzed the longitudinal EHR of 6.8 million patients from New York-Presbyterian/Columbia University Irving Medical Center's (NYP/CUIMC) clinical data warehouse (CDW). NYP/CUIMC is a tertiary medical center providing inpatient and outpatient care to the diverse population of New York City and surrounding areas. Structured clinical data from the CDW were transformed into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) V5.3 format. Unstructured clinical notes, including note metadata, were provided in Apache Avro format and loaded into an Apache Solr database for fast indexing and querying. This study received approval from the Columbia University's Institutional Review Board #AAAR3954.

2.2. Data Preprocessing

The following data preprocessing steps were described in detail in a previous publication [6] and are summarized briefly here. Diseases and phenotypes were extracted from both structured EHR data and clinical notes and normalized using Mondo Disease Ontology (MONDO) and Human Phenotypes Ontology (HPO) concept identifiers, respectively [8,9]. Only rare diseases (subclasses of MONDO:0021136) and phenotypic abnormalities (subclasses of HP:0000118) were included.

Structured EHR data were processed as follows. Condition concept identifiers and age of occurrence were extracted from the OMOP CONDITION_OCCURRENCE table. Condition concepts were mapped from OMOP standard concept identifiers to MONDO disease and HPO phenotype identifiers using cross-ontology mappings provided by Columbia Open Health Data (COHD) and the Biomedical Data Translator's Node Normalization service [10,11]. Additionally, phenotypes were assessed from the OMOP MEASUREMENT table using LOINC2HPO to analyze laboratory results and store abnormal findings as HPO identifiers [12].

Occurrences of diseases and phenotypes were recognized from clinical notes by querying the notes in Apache Solr. We performed keyword searches for each MONDO and HPO concept, leveraging primary labels and synonyms referenced by MONDO and HPO concepts. To improve the relevancy of the detected concepts, we excluded search results where the hits were negated or within a ten-word context window to mentioned family members. Additionally, we only included notes with "visit", "letter", "summary", or "surgical path" in the note title, as previous evaluations showed these note types contained the most relevant information.

2.3. Disease-Phenotype Analysis

To identify age-sensitive disease-phenotype relationships, we merged the diseases and phenotypes extracted from both structured EHR data and clinical notes into a single dataset. For each disease, we defined the disease cohort as all patients with at least one occurrence of the disease identifier. For each patient in the disease cohort, we identified the first observations of each phenotype and recorded the age at occurrence. If there were at least ten patients in the disease cohort recorded with the phenotype and the prevalence of the phenotype within the cohort is greater than in the general patient cohort, we collected the age of first occurrences of the phenotype across the disease cohort to create disease-phenotype age of onset distributions. Similarly, we collected the phenotype age of onset across all patients to characterize the phenotype age of onset in the general patient cohort. To identify phenotypes associated with the disease, we compared each disease-specific phenotype age distribution to the general cohort phenotype age distribution using the Kolmogorov-Smirnoff test and selected the phenotypes with $p\text{-value} < 1 \times 10^{-10}$, adjusted for multiple hypotheses. The mean age of onset (MAO) was calculated per disease-phenotype age distribution to quantify how early the phenotype typically appears within the disease cohort.

3. Results

3.1 Data Summary

We analyzed EHR records of 6.8 million patients using NYP/CUIMC's clinical data warehouse, including structured clinical data in OMOP format and 100 million unstructured clinical notes. In total, 4461 rare diseases and 8871 phenotypes were observed from 2,811,977 patients (Table 1). Following statistical testing and selection criteria, 535,229 disease-specific phenotype age distributions were produced, covering 2303 diseases. If we only include phenotypes where the mean age of onset is lower in the disease cohort than mean age of onset of the disease, 212,904 disease-phenotype relationships were identified. If we only include phenotypes where the prevalence of the phenotype within the disease cohort is greater than 0.1, 0.25, 0.5, and 0.75, we identified 117,008, 41,181, 9,820, and 2,330 disease-phenotype relationships, respectively. The number of phenotypes associated with each disease were (median [first-quartile, third-quartile]): 79 [17, 283].

3.2 Example: Kennedy Disease

To provide examples, we look at the results for one use case: Kennedy disease, also known as bulbospinal muscular atrophy. Kennedy disease is a rare, slowly progressive adult-onset disease where patients present with weakness and wasting of muscles (e.g., in the face, mouth, throat, and limbs), endocrinological disturbances (e.g., elevated sexual hormones, gynecomastia, and reduced fertility), and sensory disturbances (e.g., abnormal sensory nerve action potentials). 84 patients were identified with Kennedy disease and had a mean age of onset at 53.8 years. 14 phenotypes were found to be associated with Kennedy Disease with a mean age of onset earlier than disease mean age of onset: Skeletal muscle atrophy, Fiber type grouping, Anxiety, Tongue fasciculations, Paresthesia, Muscle fibrillation, Tongue atrophy, Gynecomastia, Myopathy, Rimmed vacuoles, Headache, Neck pain, Proximal muscle weakness, Spinal muscular atrophy. Figure 1 shows the age of onset

distributions of the top eight phenotypes sorted by a) prevalence within the disease cohort, b) p-value, and c) mean age of onset among phenotypes with disease prevalence > 0.15. Some of the phenotypes with the earliest onset and minimally 15% prevalence included gynecomastia (HP:0000771, p-value=4.71e-34, prevalence=0.155, MAO=46.5 years), spinal muscular atrophy (HP:0007269, p-value=9.52e-24, prevalence=0.214, MAO=47.8 years), and skeletal muscle atrophy (HP:0003202, p-value=1.41e-18, prevalence=0.536, MAO=49.9 years).

4. Discussion

Disease-phenotype associations can be ranked by p-value, disease prevalence, or mean age of onset to identify the phenotypes most suitable for a given task. For example, to design a screening model with earliest sensitivity, we could select phenotypes with earlier mean age of onset and high disease prevalence. Phenotypes with MAO earlier than disease onset may potentially lead to earlier diagnoses. To produce a model with greater specificity, we may choose phenotypes with smaller P-values that differentiate the disease vs the general patient cohort. We provided examples from Kennedy disease, selecting only phenotypes with mean age of onset prior to disease onset (early detection task) and observed that identified phenotypes corresponded with known disease presentations. The methods are generalizable and portable to other institutions with OMOP and do not require specialized hardware to run.

We plan to continue developing these methods to produce a FAIR (findable, accessible, interoperable, reproducible) knowledge graph of disease-phenotype relationships. Additional evaluations are necessary to validate the associations produced by these methods as well as the utility of the association metrics and other methods (e.g., risk models) of identifying important phenotypes for diagnosing diseases and developing clinical decision support tools.

This study has several limitations. False-positives for diseases can occur from the structured EHR data (e.g., misdiagnosis, coding error, etc.) and from notes (e.g., false-positive named entity recognition (NER)). Since the cohort selection criteria only requires one occurrence of the disease, all phenotypes observed in these false-positive patients will falsely contribute towards the disease-phenotype age distribution. Conversely, not all phenotypes are treated within the care settings, and phenotypes missing from the EHR lead to false-negatives. With the goal of producing a publicly sharable knowledgebase, we imposed a minimum count of 10 people per disease-phenotype pair, which limits our ability to identify associations among the rarest diseases. Over the last few years, large language models (LLMs) have been established as the state of the art for most natural language processing tasks, including NER in clinical notes, however, we did not evaluate the use of LLMs in this study due to the cost and time required to process millions of clinical notes.

5. Conclusions

We analyzed structured EHR records and clinical notes from NYP/CUIMC's clinical data warehouse and found 500 thousand disease-phenotype associations for 2300 rare diseases. We looked at Kennedy disease as an example and observed that evaluating the phenotypes

by different association metrics could allow us to choose different phenotypes according to priorities, such as for developing clinical decision support tools for early screening of rare diseases. This research was supported by the National Human Genome Research Institute (NHGRI) and National Center For Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) under Award Numbers R01HG013031, R01HG012655, and OT2TR003434. This research leveraged Biomedical Data Translator services, which were funded by NCATS OT2TR003449.

References

- [1]. Commissioner O of the. Rare Diseases at FDA [Internet]. FDA. FDA; 2024 [cited 2024 Oct 7]. Available from: <https://www.fda.gov/patients/rare-diseases-fda>.
- [2]. Orphan medicinal products - European Commission [Internet]. 2024 [cited 2024 Oct 7]. Available from: https://health.ec.europa.eu/medicinal-products/orphan-medicinal-products_en.
- [3]. Faye F, Crocione C, Anido de Peña R, et al. Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a Rare Barometer retrospective patient survey. *Eur J Hum Genet*. 2024;32(9):1116–1126. [PubMed: 38755315]
- [4]. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015;10(12):2004–2015. [PubMed: 26562621]
- [5]. Birgmeier J, Haeussler M, Deisseroth CA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci Transl Med*. 2020;12(544):eaau9113. [PubMed: 32434849]
- [6]. Liu C, Ta CN, Havrilla JM, et al. OARD: Open annotations for rare diseases and their phenotypes based on real-world data. *Am J Hum Genet*. 2022;109(9):1591–1604. [PubMed: 35998640]
- [7]. Yang J, Shu L, Duan H, et al. A robust phenotype-driven likelihood ratio analysis approach assisting interpretable clinical diagnosis of rare diseases. *J Biomed Inform*. 2023;142:104372. [PubMed: 37105510]
- [8]. Vasilevsky NA, Matentzoglou NA, Toro S, et al. Mondo: Unifying diseases for the world, by the world [Internet]. medRxiv; 2022 [cited 2024 Oct 7]. p. 2022.04.13.22273750. Available from: <https://www.medrxiv.org/content/10.1101/2022.04.13.22273750v3>.
- [9]. G Ma, M N, C B, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res* [Internet]. 2024 [cited 2024 Oct 7];52(D1).
- [10]. Ta CN, Dumontier M, Hripcsak G, et al. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. *Sci Data*. 2018;5:180273. [PubMed: 30480666]
- [11]. Unni DR, Moxon SAT, Bada M, et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin Transl Sci*. 2022;15(8):1848–1855. [PubMed: 36125173]
- [12]. Zhang XA, Yates A, Vasilevsky N, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *Npj Digit Med*. 2019;2(1):1–9. [PubMed: 31304351]

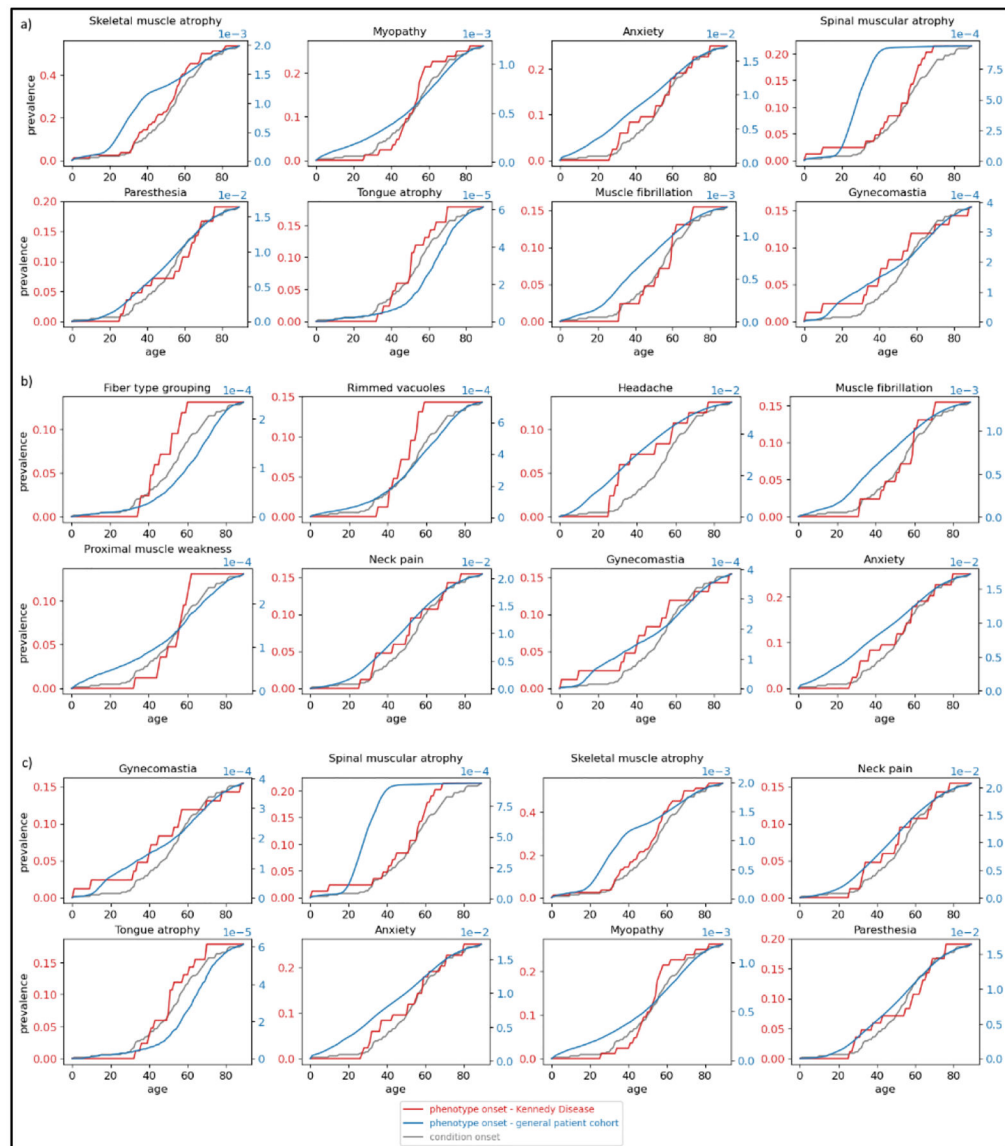


Figure 1.

The cumulative distributions of the age of onset for phenotypes in Kennedy Disease, sorted by a) disease prevalence, b) p-value, and c) mean age of onset. Red: age of onset of the phenotype in the disease; blue: age of onset of the phenotype in general patient cohort; gray: age on onset of the condition (max: 1.0).

Table 1.
Study sample characteristics. Q1: first-quartile; Q3: third-quartile; NA: not available

Characteristic	Data
Patients (N)	2,811,977
Age (median [Q1, Q3])	41.7 [23.6, 62.8] years
Sex (N (%))	female: 1,563,887 (55.6%) male: 1,246,836 (44.3%)
Ethnicity (N (%))	Not Hispanic or Latino: 717,696 (25.5%), Hispanic or Latino: 409,940 (14.6%), NA: 1,684,341 (59.9%)
Race (N (%))	White: 785,748 (27.9%), Black or African American: 238,381 (8.5%), Asian: 53,619 (1.9%), Other: 43,777 (1.6%), NA: 1,690,452 (60.1%)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript