博士論文

A new computational method to assist interpreting genomic variants of rare diseases based on disease-phenotype associations extracted from case reports

（症例報告由来の疾患-症状関連情報を用いた希少疾患ゲノムの変異解釈支援手法の開発）

藤原　豊史

Abstract

Recently, to speed up the differential diagnosis process based on symptoms and signs observed from a patient in the diagnosis of rare diseases, phenotype-driven differential diagnosis systems have been developed and implemented. The performance of those systems relies on the quantity and quality of underlying databases of disease–phenotype associations (DPAs). Although such databases are often developed by manual curation, they inherently suffer from limited coverage.

To address this problem, I propose a text mining approach to increase the coverage of DPA databases and consequently improve the performance of differential diagnosis systems. Our analysis showed that the text mining approach using one million case reports obtained from PubMed could increase the coverage of manual curated DPAs in Orphanet by 113.2%. I also present PubCaseFinder (https://pubcasefinder.dbcls.jp), a new phenotype-driven differential diagnosis system in a freely available web application. By utilizing automatically extracted DPAs from case reports in addition to manually curated DPAs, PubCaseFinder improves the performance of automated differential diagnosis. This approach is unique in designing a new phenotype-driven differential diagnosis system, and I believe that this approach shows a promising path for improving the performance of existing phenotype-driven differential diagnosis systems.

Moreover, PubCaseFinder helps clinicians search for relevant case reports using phenotype-based comparisons. For identifying new diseases and disease causative genes, it is necessary to collect multiple unrelated patients and case reports with common variants and similar phenotypes, but existing patient repositories and

matchmaking services lack methods to consult published case reports. To tackle this situation, I have provided API to facilitate searching for such case reports, and several patient repositories have employed our API.

Previous studies reported that case reports were an essential tool for extracting valuable information for rare diseases despite low certainty evidence due to its small samples. I, therefore, targeted the one million case reports included in PubMed, and this is, to our knowledge, the first demonstration that such an extensive collection of case reports was useful for tackling rare disease issues by using a text mining method. I will extend the collection of case reports to those of European in Europe PMC and those of Japanese in J-STAGE and will believe that further case reports will contribute more for tackling rare disease issues.

Acknowledgments

This thesis becomes a reality with the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I first deeply thank my supervisor, Professor Toshihisa Takagi for providing me with an opportunity to try this research. My dream, when I was a child, was to be a researcher in the field of rare diseases. Next, I would like to thank Associate Professor Yasunori Yamamoto and Associate Professor Jin-Dong Kim as my superiors in Database Center for Life Science. They always encouraged me greatly and provided me with valuable support and guidance to this thesis. Associate Professor Shoko Kawamoto also offered me with useful comments and delightful discussions about PubCaseFinder. Assistant Professor Toshiaki Katayama who organized international BioHackathon gave me an opportunity to collaborate with overseas researchers.

I would like to show my gratitude to Dr. Orion Buske who is a co-author of our paper published in American Journal of Human Genetics. He helped me develop MME API for PubCaseFinder and integrate PubCaseFinder with PhenomeCentral. Also, he and the Care4Rare Canada Consortium kindly shared the clinical cases on this research. I would also like to thank, Professor Kenjiro Kosaki, Dr. Tudor Groza, and Mr. Sadahiro Kumagai for integrating PubCaseFinder with IRUD Exchange which is the truly valuable repository for Japanese patients of rare diseases. Professor Soichi Ogishima provided me the Japanese Human Phenotype Ontology. Dr. Yuka Tateishi provided me with useful comments.

I would express a broad sense of gratitude to my parents to whom I owe my life for their constant love, encouragement, and moral support. They are the ultimate role models.

Most importantly, I wish to thank my wife Yukiko for her cheerful encouragement and my wonderful daughter Mirei who provides me eternal bliss. Thank you.

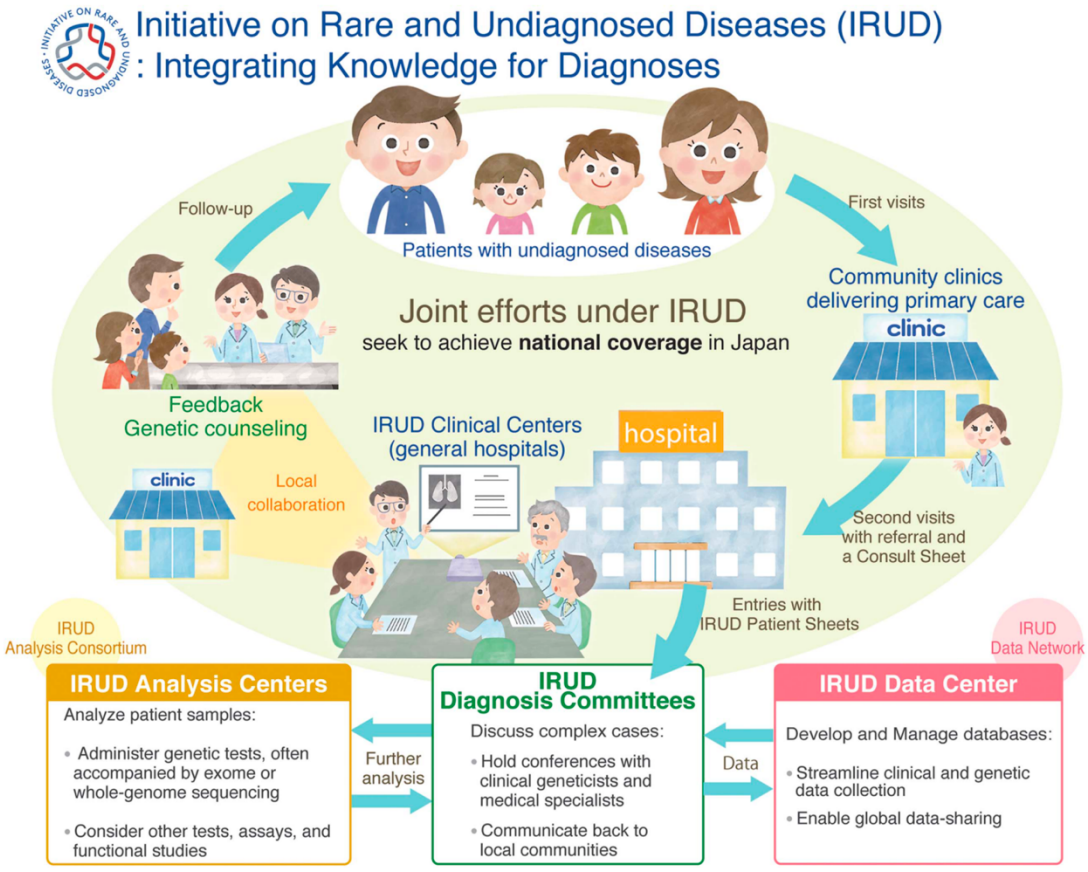
Contents

Chapter 1
Introduction

Undiagnosed patients of rare diseases

At present over 6,000 rare diseases have been identified, and ~80% of them are genetic in origin [1]. Unfortunately, a quarter of rare disease patients waited 5 to 30 years for a diagnosis, 40% of rare disease patients were misdiagnosed at first [2]. Such patients will likely lose opportunities such as optimization of clinical management and early intervention [3]. To tackle this situation, next-generation sequencing (NGS)-based analysis is being undertaken to identify candidate diseases for undiagnosed patients [4,5]. For example, the Japan Agency for Medical Research and Development (AMED) launched the Initiative on Rare and Undiagnosed Diseases (IRUD) in 2015 (Fig. 1.1), which conducted clinical whole-exome sequencing (CES) and clinical whole-genome sequencing (CGS) for undiagnosed patients on a nationwide scale [6]. IRUD have already analyzed more than 6,000 CES, and up to 40% of patients have been diagnosed. Figure 1.2 shows the process of interpreting the results of CES. At first, clinicians observe symptoms and signs in an undiagnosed patient, which are collectively called “phenotypes.” Second, more than tens of thousands of variants are identified by CES analysis. Third, the large number of variants such as synonymous ones, non-splice ones, high conserved ones, and common ones are filtered, and several genes which include the candidate variants are identified. Finally, the corresponding diseases are identified using disease-gene association databases such as Orphanet and OMIM. As a differential

diagnosis process, the set of phenotypes of the patient are compared for those of diseases, and then only diseases having highly phenotypic similarity remain as candidate diseases.



Adachi, T., Kawamura, K., Furusawa, Y., Nishizaki, Y., Imanishi, N., Umehara, S., Izumi, K., and Suematsu, M. (2017). Japan's initiative on rare and undiagnosed diseases (IRUD): Towards an end to the diagnostic odyssey. Eur. J. Hum. Genet. *25*, 1025–1028.

Figure 1.1: Process of the IRUD operation

There are three major functional units; 1) IRUD Diagnosis Committees, 2) IRUD Data Center, 3) IRUD Analysis Centers. These interact well with each other and are operated by principal research groups [6].

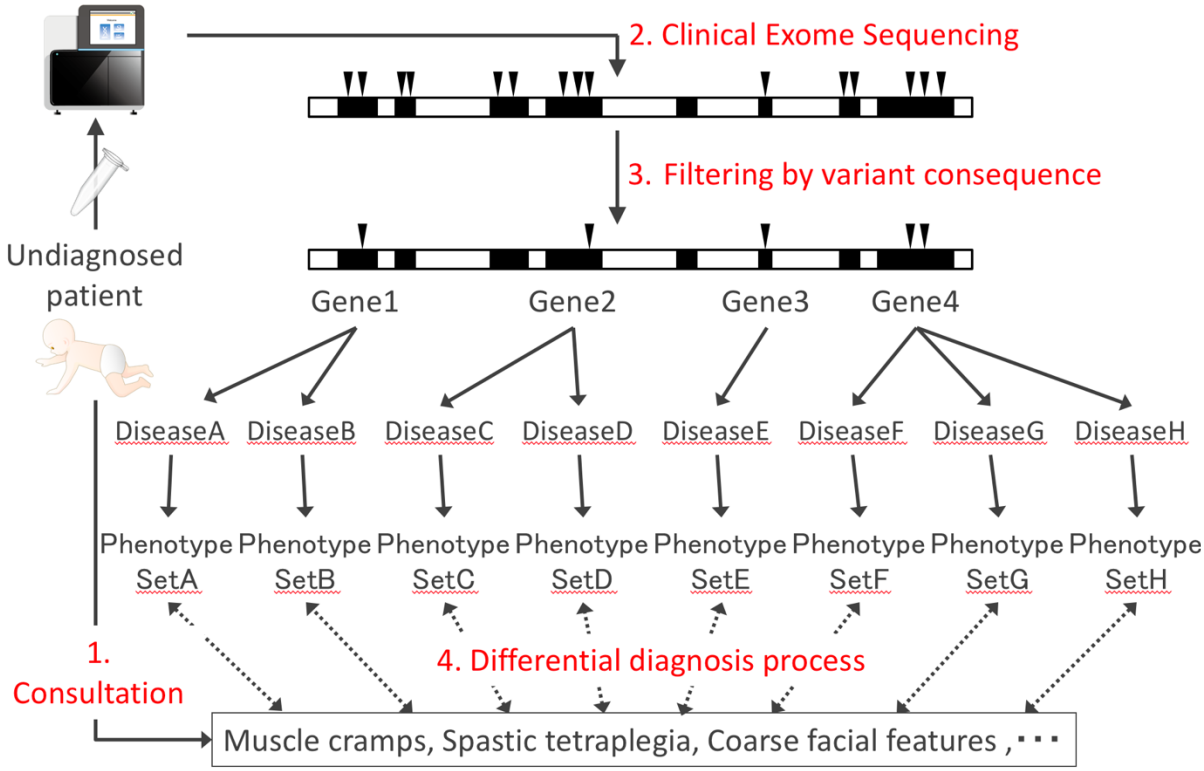


Figure 1.2: The process of interpreting the results of CES

Clinicians observe symptoms and signs in an undiagnosed patient (step 1), and more than tens of thousands of variants are identified by CES analysis (step 2). A large number of variants are filtered, and several genes which include the candidate variants are identified (step 3). The corresponding diseases are identified using disease-gene association databases such as Orphanet and OMIM. As the differential diagnosis process, the set of phenotypes of the patient are compared for those of diseases, and then only diseases with highly phenotypic similarity remain as candidate diseases (step 4).

Phenotype-driven differential diagnosis system

Even though the NGS-based analysis improves diagnostic rates [7,8], the differential diagnosis process (Fig. 1.2) is time-consuming [9]. At first, clinicians collect reported phenotypes from trusted medical sources (e.g., databases, textbooks, and papers) for each candidate disease and then check which diseases overlap regarding phenotype with the patient's phenotypes [10]. Recently, to speed up the process, phenotype-driven differential diagnosis systems such as Phenomizer [9], Phenolyzer [11], and FACE2GENE [12] have been implemented [13]. Phenomizer and Phenolyzer employ a semantic similarity computation method to compare the patient's phenotypes against a set of rare diseases associated with phenotypes. These systems use the Human Phenotype Ontology (HPO) for describing detailed and precise phenotypic abnormalities of a patient and rare diseases [13]. HPO has been curated by domain experts to provide a comprehensive vocabulary for describing phenotypic abnormalities that are widely seen in human genetic diseases. HPO is open source and consist of more than 12,000 classes which have hierarchical relationships (Fig. 1.3). More general terms are at the top, and more specific terms are below in the hierarchical relationships. Between each node and its parents has an “is-a” relationship (e.g., both “Abnormality of globe size” and “Aplasia/Hypoplasia affecting the eye” are a subtype of “Abnormality of the globe,” and “Microphthalmia” is a subtype of them). This semantic structure enables us to compute semantic similarity between phenotypic terms. Currently, Diverse groups such as international organizations for rare diseases, patient registries, biomedical resources, biomedical systems, and biomedical databases have employed HPO as a standard vocabulary for presenting

phenotypic abnormalities (Table. 1.1) [13]. This situation allows for the better interoperability of these groups through phenotypic terms of HPO.

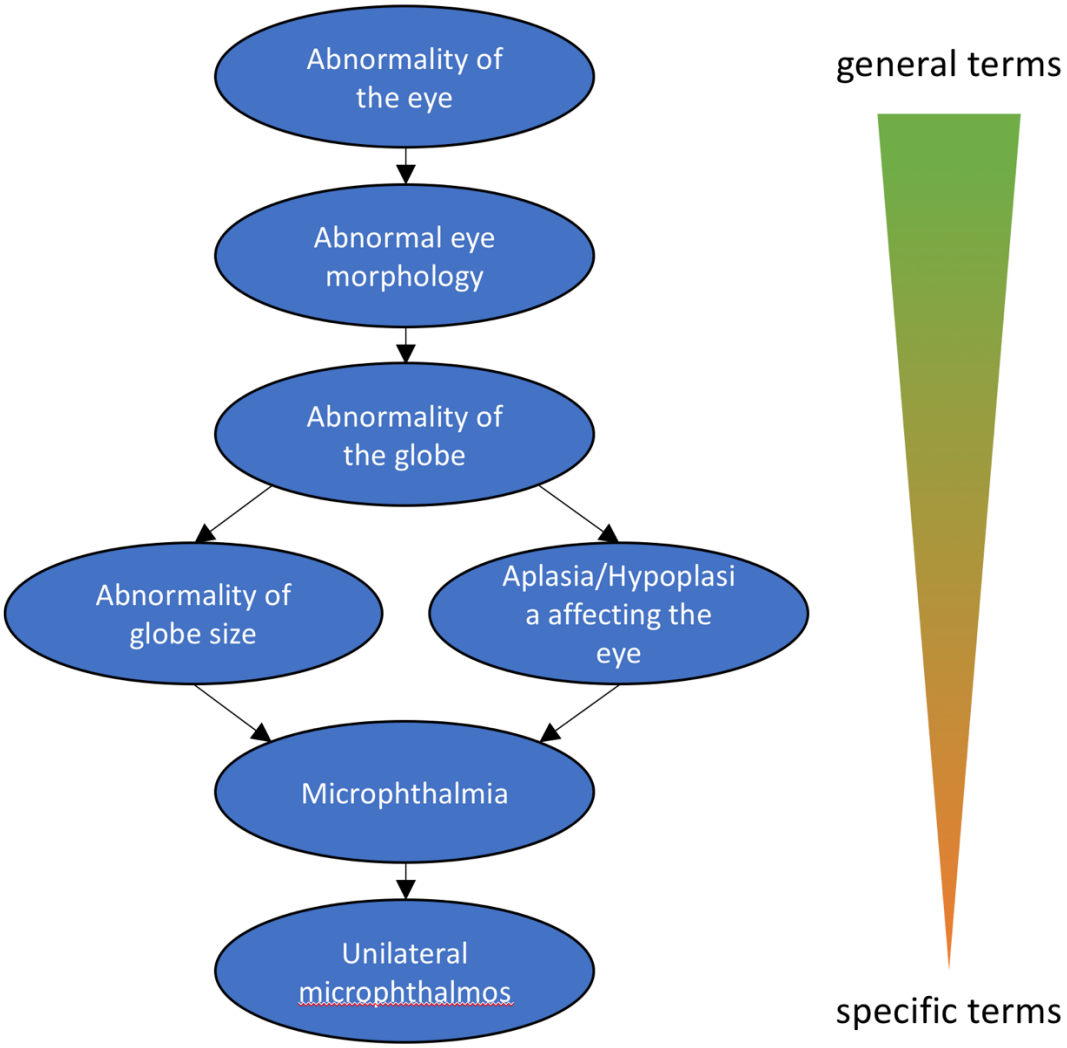


Figure 1.3: The structure of a part of the Human Phenotype Ontology
More general terms are at the top, and more specific terms are below in the hierarchical relationships. Between each node and its parents has an “is-a” relationship [13].

Table 1.1: Systems and Databases using HPO

These were reported in "Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res. 45, D865–D876" [13].

Category	System / Database	Reference
Phenotype-driven differential diagnosis Phenomizer	Phenomizer	[9]
	BOQA	[14]
	FACE2GENE	[12]
	Phenolyzer	[11]
Phenotype-driven exome/genome analysis	Exomiser	[15]
	PhenIX	[16]
	Phevor	[17]
	PhenoVar	[18]
	eXtasy	[19]
	OMIMExplorer	[20]
	Phen-Gen	[21]
	Geno2MP	[22]
	Genomiser	[23]
	SimReg	[24]
Functional and network analysis	TopGene/ToppFunn	[25]
	WebGestalt	[26]
	SUPERFAMILY	[27]
	GREAT	[28]
	Random walk on heterogeneous network	[29]
	PANDA	[30]
	PREDICT	[31]
Clinical data management and analysis	Phenotips	[32]
	Patient Archive	[33]

	GENESIS (GEM.app)	[34]
Cross-species phenotype analysis	PhenoDigm	[35]
	MouseFinder	[36]
	Monarch	[37]
	PhenomeNet	[38]
	UberPheno	[39]
	MORPHIN	[40]
	PhenogramViz	[41]
Phenotype knowledge resources and databases	Orphanet	[42]
	MalaCards	[43]
	NIH genetic testing registry	[44]
	OMIM	[45]
	dcGO	[46]
	ClinVar	[47]
	GeneSetDB	[48]
	MSeqDR	[49]
	DIDA (digenic diseases database)	[50]
	Genetic and Rare Diseases (GARD) Information Center	[51]
Visualization	PhenoStacks	[52]
	PhenoBlocks	[53]
	DECIPHER (phenogram)	[54]
	phenogrid	[55]

FACE2GENE detects a patient's phenotypes from a face image and calculates a similarity score for hundreds of genetic diseases using a deep learning method (Fig. 1.4). It trained with over 26,000 patient cases form a rapidly growing phenotype-genotype database. Currently, FACE2GENE achieves 91% top-10-accuracy in identifying over 215 different genetic diseases and has outperformed clinical experts in their experiments using a large evaluation data set of patients' photos.

These phenotype-driven differential diagnosis systems provide a ranked list of diseases based on the similarity score, and the top-listed diseases represent the most likely differential diagnosis (Fig. 1.5).

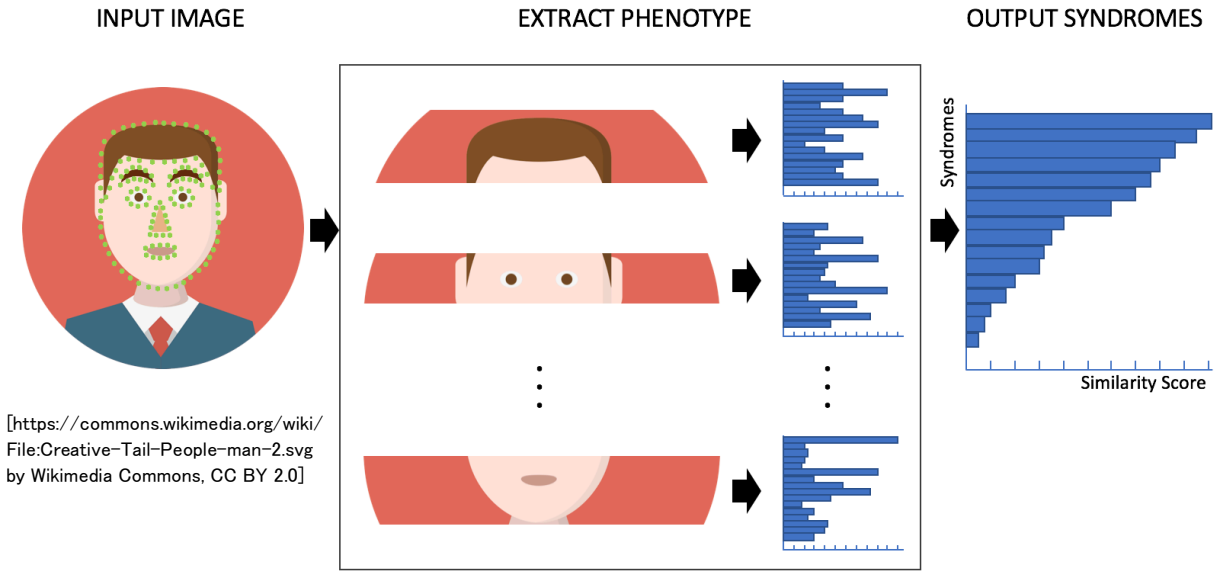


Figure 1.4: High-level flow from the input image to the output syndromes

At first, the input image is pre-processed to achieve a face detection, landmarks detection, and alignment. Second, the input image is cropped into facial regions. Third, each region outputs regional score based on phenotypic similarity with syndromes. Finally, FACE2GENE output syndromes, sorted by the aggregated similarity score of each regional phenotypic score [12].

p-value.	Disease Id.	Disease name.	Genes.
0.016	OMIM:312500	RETICULOENDOTHELIOSIS, X-LINKED	
0.0214	OMIM:616278	#616278 BILE ACID SYNTHESIS DEFECT, CONGENITAL, 5; CBAS5	ABCD3 (5825)
0.0214	OMIM:237800	%237800 HYPERBILIRUBINEMIA, SHUNT, PRIMARY; PSHB	
0.028	OMIM:206400	ANEMIA, NONSPHEROCYTIC HEMOLYTIC, POSSIBLY DUE TO DEFECT IN PORPHYRINMETABOLISM	
0.0294	OMIM:616649	#616649 SPHEROCYTOSIS, TYPE 2; SPH2;;SPHEROCYTOSIS, HEREDITARY, 2; HS2	SPTA1 (6708), SPTB (6710), EPB42 (2038), SLC4A1 (6521), ANK1 (286)
0.0294	OMIM:109270	+109270 SOLUTE CARRIER FAMILY 4 (ANION EXCHANGER), MEMBER 1; SLC4A1;;BAND 3 OF RED CELL MEMBRANE; BND3;;ERYTHROCYTE MEMBRANE PROTEIN BAND 3; EMPB3;;ERYTHROID PROTEIN BAND 3; EPB3;;ANION EXCHANGE PROTEIN 1; AE1ACANTHOCYTOSIS, ONE FORM OF, INCLUDED;;OVALOCYTOSIS, MALAYSIAN-MELANESIAN-FILIPINO TYPE, INCLUDED;;OVALOCYTOSIS, SOUTHEAST ASIAN, INCLUDED; SAO, INCLUDED;;ELLIPTOCYTOSIS 4, INCLUDED; EL4, INCLUDED;;ELLIPTOCYTOSIS, STOMATOCYTIC HEREDITARY, INCLUDED;;HE, STOMATOCYTIC, INCLUDED	SLC4A1 (6521)
0.0374	OMIM:182900	SPHEROCYTOSIS, HEREDITARY	SPTA1 (6708), SPTB (6710), EPB42 (2038), SLC4A1 (6521), ANK1 (286)
0.0374	OMIM:613977	#613977 CYANOSIS, TRANSIENT NEONATAL; TNCY	HBG2 (3048)
0.0374	ORPHANET:294	FETAL CYTOMEGALOVIRUS SYNDROME	
0.0481	OMIM:222800	DIPHOSPHOGLYCERATE MUTASE DEFICIENCY OF ERYTHROCYTE	BPGM (669)
0.0481	OMIM:235700	#235700 HEMOLYTIC ANEMIA, NONSPHEROCYTIC, DUE TO HEXOKINASE DEFICIENCY	HK1 (3098)
0.0501	OMIM:266200	PYRUVATE KINASE DEFICIENCY OF RED CELLS	PKLR (5313)
0.0501	ORPHANET:46532	HEREDITARY PERSISTENCE OF FETAL HEMOGLOBIN - BETA-THALASSEMIA	HBB (3043), KLF1 (10661), HBG1 (3047), HBG2 (3048)
0.0501	OMIM:273680	THANATOPHORIC DYSPLASIA, GLASGOW VARIANT	
0.0501	OMIM:300331	THROMBOCYTOSIS, FAMILIAL X-LINKED	MPL (4352), JAK2 (3717), THPO (7066)

Figure 1.5: Output image of Phenomizer in top-listed diseases [9]

Phenomizer reports the most likely differential diagnosis and provides p-value of each disease to consider that will result in specific diagnoses becoming significant.

Disease-phenotype associations

The quantity and quality of underlying databases of disease–phenotype associations (DPAs) greatly influence the performance of these systems. There are two representative sources of DPAs: Orphanet [56] and the Human Phenotype Ontology (HPO) consortium [13]. Orphanet provides DPAs for the rare diseases that are defined in ORDO, and the HPO consortium mainly provides DPAs for the genetic diseases that are defined in OMIM. For example, Orphanet delivers the 27 phenotypes associated with Fragile X syndrome (Fig. 1.6). Each phenotype is annotated with the frequency of occurrence in patients of Fragile X syndrome using Obligate (100%), Very frequently (99-80%), Frequent (79-30%), Occasional (29-5%), Very rare (4-1%), and Excluded (0%).

Note that databases which rely on manual curation inherently show a limited coverage [57]. In the case of Orphanet, more than half of the diseases (~60.5% of 6,268) are not associated with a phenotype. There are two main reasons for this limited coverage. First, the development of databases is based on the curation of papers by human experts, which is time-consuming and labor-intensive because of the large volume and rapid growth of life sciences papers [58]. Second, there are still many unknown phenotypes in rare diseases because phenotypic spectrums for many rare diseases are still under investigation [59]. For example, Elisabet et al. [60] quantified many atypical phenotypes of inherited kidney diseases caused by various genetic, epigenetic, and environmental factors (Fig. 1.7). Sawyer et al. [2] diagnosed 105 undiagnosed rare disease patients using whole-exome sequencing and showed that 26 patients presented atypical phenotypes of a known disease. With the rapid

adaptation of NGS-based diagnostics in clinical settings, phenotypic expansions of disease spectrums will become increasingly common [3,60].

HPO ID	Term	Frequency
HP:0000053	Macroorchidism	Very frequent (99-80%)
HP:0000389	Chronic otitis media	Very frequent (99-80%)
HP:0001388	Joint laxity	Very frequent (99-80%)
HP:0001763	Pes planus	Very frequent (99-80%)
HP:0002167	Neurological speech impairment	Very frequent (99-80%)
HP:0002342	Intellectual disability, moderate	Very frequent (99-80%)
HP:0003564	Folate-dependent fragile site at Xq28	Very frequent (99-80%)
HP:0000246	Sinusitis	Frequent (79-30%)
HP:0000256	Macrocephaly	Frequent (79-30%)
HP:0000275	Narrow face	Frequent (79-30%)
HP:0000276	Long face	Frequent (79-30%)
HP:0000303	Mandibular prognathia	Frequent (79-30%)
HP:0000388	Otitis media	Frequent (79-30%)
HP:0000411	Protruding ear	Frequent (79-30%)
HP:0001252	Muscular hypotonia	Frequent (79-30%)
HP:0002003	Large forehead	Frequent (79-30%)
HP:0002007	Frontal bossing	Frequent (79-30%)
HP:0002020	Gastroesophageal reflux	Frequent (79-30%)
HP:0007018	Attention deficit hyperactivity disorder	Frequent (79-30%)
HP:0000486	Strabismus	Occasional (29-5%)
HP:0000717	Autism	Occasional (29-5%)
HP:0000739	Anxiety	Occasional (29-5%)
HP:0001250	Seizures	Occasional (29-5%)
HP:0001634	Mitral valve prolapse	Occasional (29-5%)
HP:0002120	Cerebral cortical atrophy	Occasional (29-5%)
HP:0005111	Dilatation of the ascending aorta	Occasional (29-5%)
HP:0100716	Self-injurious behavior	Occasional (29-5%)

Figure 1.6: The phenotypes associated with Fragile X syndrome, which are provided by Orphanet

Each phenotype is annotated with the frequency of occurrence in patients of Fragile X syndrome.

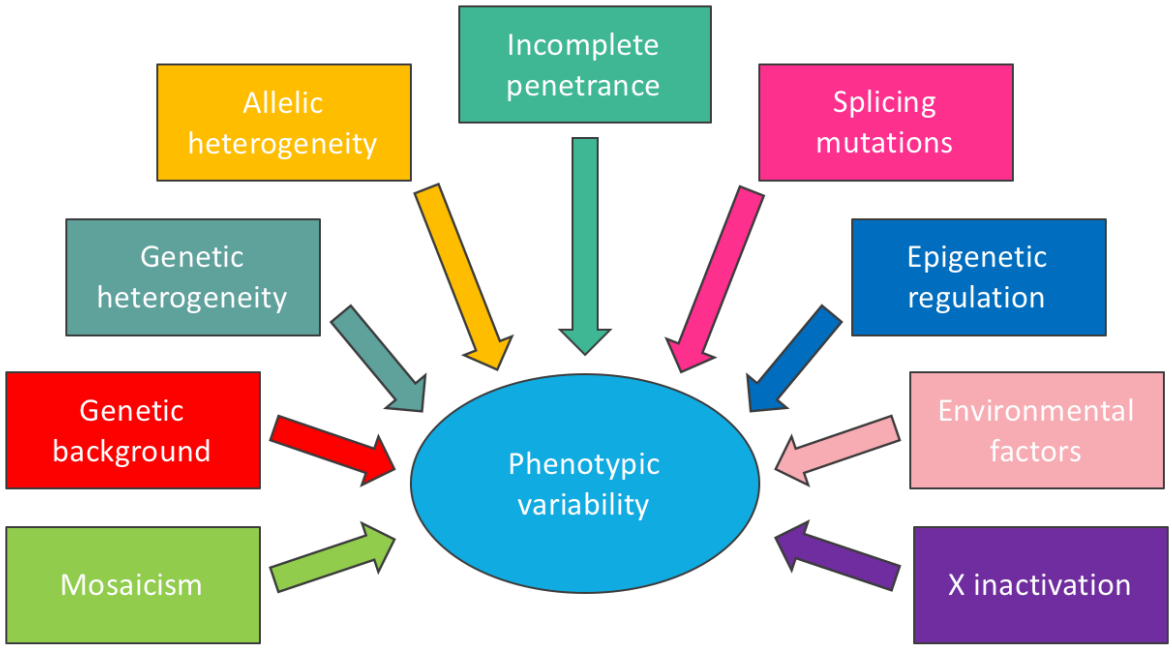


Figure 1.7: Possible explanations for phenotypic variability in inherited kidney diseases [60]

Problems need to be solved

To improve the performance of phenotype-driven differential diagnosis systems, the limited coverage problem of DPA databases needs to be overcome, which is an important problem for diagnosis of rare disease. In this study, to address the problem, I empirically explore one question in a large scale: Can automatically extracted DPAs from case reports contribute to improving the performance of phenotype-driven differential diagnosis systems for rare diseases?

First, I extract DPAs from case reports in PubMed using a text mining approach and compare those with DPAs from Orphanet. I focus on case reports as these are an essential tool for quickly expanding the growing body of clinical knowledge on rare diseases [61], and case reports often deal with previously undescribed and

atypical phenotypes [62]. For example, cencerning cerebrotendinous xanthomatosis, Taboada et al. [57] automatically extracted DPAs from case reports in PubMed and obtained 11 new DPAs that did not appear in manually curated DPAs. In addition, the total number of those case reports have been more than one million, and the number of PubMed-indexed case reports in each year has increased from 1980 (Fig. 1.8). This massive volume and rapid growth of case reports provide the possibilities to extract various DPAs. On the other hand, Orphanet also provides detailed descriptions of the rare diseases as unstructured data. As with case reports, these include diverse phenotypes for each rare disease, which are not included in the DPA database from Orphanet. However, these descriptions are not disclosed with an open license; it is not suitable for a text mining resource for extracting DPAs. Simple disease descriptions are disclosed with an free license, but these do not include diverse phenotypes for each rare disease.

Second, I develop a new phenotype-driven differential diagnosis system PubCaseFinder and demonstrate that automatically extracted DPAs without manual screening can contribute to improving the performance of automated differential diagnosis.

To the best of our knowledge, this is the first report on the potential of automatically extracted DPAs from one million case reports for improving the performance of phenotype-driven differential diagnosis systems for rare diseases. While existing phenotype-driven differential diagnosis systems use only curated DPAs retrieved from Orphanet and OMIM, I apply automatically extracted DPAs with curated DPAs to our system. Our approach is unique in designing a new phenotype-driven differential diagnosis system, and I believe that our approach shows a

promising path for improving the performance of existing phenotype-driven differential diagnosis systems.

Figure 1.8: Distribution of the number of case reports published per year in PubMed from 1980 to 2017

Chapter 2
Material and Methods

Collecting case reports

I used PubMed E-utilities to obtain an extensive collection of case reports [63]. To find an valid search query, our strategy was to use the publication type tag “Case Reports” that was manually tagged by human experts. A previous study has shown that PubMed includes many case reports that are not explicitly tagged as such [64]. I also chose to consider a paper as a case report if its title included “case report” or “case reports.” I used the following query to collect case reports and record titles and abstracts: “case reports” [Publication Type] OR “case reports” [ti] OR “case report” [ti]. I found that 1,895,021 PubMed entries were initially collected as case reports, among which only 1,083,283 had both titles and abstracts (as of July 20, 2017). Table 2.1 lists the top 20 journals (out of 7,649 containing case reports) ranked according to the number of case reports published.

Gagnier et al. [64] report the guidelines for writing case reports. According to the guidelines, titles should include the phenomenon of most significant interest (e.g., symptom, diagnosis, test, intervention). Moreover, abstracts should include case presentation such as the main symptoms of the patient, the primary clinical findings, the main diagnoses and interventions, and the leading outcomes. Thus, it is promising that I can extract the main and diverse DPAs from titles and abstracts of case reports.

Table 2.1: Top 20 journals ranked according to the number of case reports published

Journal title	Country of publication	Number of papers
BMJ case reports	England	9782
The Annals of thoracic surgery	Netherlands	6902
Internal medicine (Tokyo, Japan)	Japan	5697
Gan to kagaku ryoho. Cancer & chemotherapy	Japan	5618
Southern medical journal	United States	5144
Journal of pediatric surgery	United States	4776
Clinical nuclear medicine	United States	4479
Journal of neurosurgery	United States	4325
Chest	United States	4295
American journal of medical genetics	United States	4290
Urology	United States	4273
The Japanese journal of thoracic surgery	Japan	4116
Cancer	United States	4092
The Journal of laryngology and otology	England	4078
Neurosurgery	United States	4064
The Journal of urology	United States	3958
Neurology	United States	3880
Hinyokika kiyo. Acta urologica Japonica	Japan	3851
Journal of medical case reports	England	3845
Nederlands tijdschrift voor geneeskunde	Netherlands	3845

Identifying disease–phenotype associations

I extracted DPAs from our collection of case reports using a text mining approach (Fig. 2.1). At first, I annotated titles and abstracts of case reports with HPO terms and Orphanet Rare Disease Ontology (ORDO) terms using ConceptMapper [65] with HPO and ORDO. HPO, initially published in 2008, is a standardized vocabulary for describing phenotypic abnormalities [12]. ORDO, constructed by Orphanet and EBI, provides a standardized vocabulary for rare diseases extracted from papers and validated by international experts [13]. I downloaded the HPO file (releases/2017-06-30) supplied by the HPO consortium and the ORDO file (version 2.3) provided by Orphanet. HPO contains a set of 12,786 terms that were integrated with 9,473 textual definitions and 16,320 synonyms, and 16,443 is-a relationships were established between HPO terms. In this study, I use only the term HP:0000118 (Phenotypic abnormality) and all its descendants, whose total number is 12,485. Other terms such as “Mode of inheritance” and “Frequency” do not present signs and symptoms of patients. Table 2.2 shows all descendant HPO terms of the term HP:0000118 with the label, the disease definition, and the number of all its descendants.

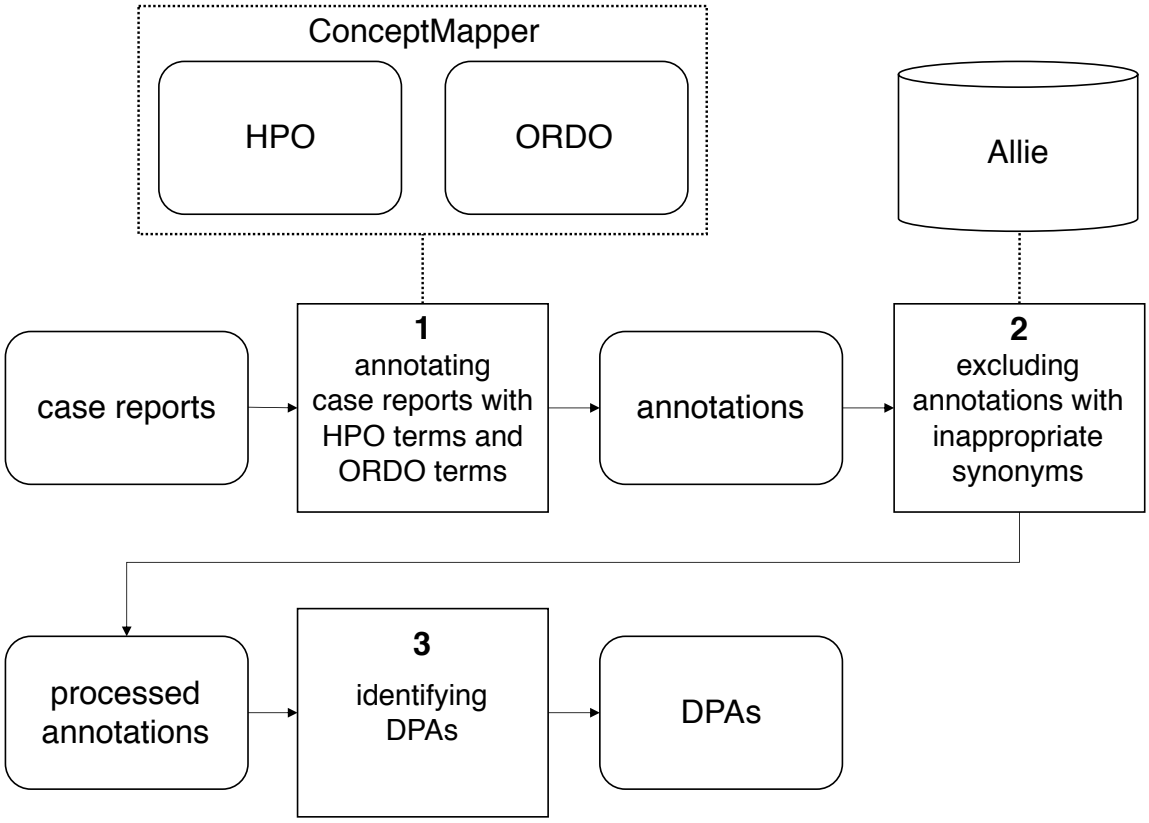


Figure 2.1. Process of identifying disease–phenotype associations (DPAs) from case reports

The set of titles and abstracts of case reports were annotated with HPO terms and ORDO terms using ConceptMapper with HPO and ORDO (step 1), and annotations with inappropriate synonyms were excluded using the Allie database (step 2). DPAs were identified in processed annotations (step 3).

Table 2.2: All descendant HPO terms of the term HP:0000118 as phenotypes

ID	Term (Label)	Definition	# of all descendant terms
HP:0000119	Abnormality of the genitourinary system	The presence of any abnormality of the genitourinary system.	814
HP:0000152	Abnormality of head or neck	An abnormality of head and neck.	1263
HP:0000478	Abnormality of the eye	Any abnormality of the eye, including location, spacing, and intraocular abnormalities.	904
HP:0000598	Abnormality of the ear	An abnormality of the ear.	281
HP:0000707	Abnormality of the nervous system	An abnormality of the nervous system.	1638
HP:0000769	Abnormality of the breast	An abnormality of the breast.	30
HP:0000818	Abnormality of the endocrine system	Ab abnormality of the endocrine system.	375
HP:0000924	Abnormality of the skeletal system	An abnormality of the skeletal system.	3591
HP:0001197	Abnormality of prenatal development or birth	An abnormality of the fetus or the birth of the fetus, excluding structural abnormalities.	132
HP:0001507	Growth abnormality	Growth abnormality.	84
HP:0001574	Abnormality of the integument	An abnormality of the integument, which consists of the skin and the superficial fascia.	836
HP:0001608	Abnormality of the voice	Abnormality of the voice.	17
HP:0001626	Abnormality of the cardiovascular system	Any abnormality of the cardiovascular system.	914
HP:0001871	Abnormality of blood and blood-forming tissues	An abnormality of the hematopoietic system.	544
HP:0001939	Abnormality of metabolism/homeostasis	Abnormality of metabolism/homeostasis.	861
HP:0002086	Abnormality of the respiratory system	An abnormality of the respiratory system, which include the airways, lungs, and the respiratory muscles.	357
HP:0002664	Neoplasm	An organ or organ-system abnormality that consists of uncontrolled autonomous cell-proliferation which can occur in any part of the body as a benign or malignant neoplasm (tumour).	541
HP:0002715	Abnormality of the immune system	An abnormality of the immune system	559

HP:0003011	Abnormality of the musculature	Abnormality originating in one or more muscles, i.e., of the set of muscles of body.	578
HP:0003549	Abnormality of connective tissue	Any abnormality of the soft tissues, including both connective tissue (tendons, ligaments, fascia, fibrous tissues, and fat).	201
HP:0025031	Abnormality of the digestive system	Abnormality of the digestive system.	559
HP:0025142	Constitutional symptom	A symptom or manifestation indicating a systemic or general effect of a disease and that may affect the general well-being or status of an individual.	63
HP:0025354	Abnormal cellular phenotype	An anomaly of cellular morphology or physiology.	4
HP:0040064	Abnormality of limbs	Abnormality of limbs.	2727
HP:0045027	Abnormality of the thoracic cavity	Abnormality of the thoracic cavity.	3
HP:0500014	Abnormal test result	Abnormal finding in a diagnostic test or assay.	36

ORDO contains a set of 13,321 terms integrated with 3,737 textual definitions and 20,542 synonyms and 15,973 is-a relationships and provides connections with other resources (e.g., OMIM and ICD10). In this study, I used 6,268 ORDO terms that are descendent terms of ORDO: 377788 (disease), ORDO: 377789 (malformation syndrome), ORDO: 377790 (biological anomaly), ORDO: 377791 (morphological anomaly), ORDO: 377792 (clinical syndrome), and ORDO: 377793 (particular clinical situation in a disease or syndrome) as rare diseases (Table 2.3).

For annotation, I used a dictionary-based system for recognizing concepts in the text. Christopher et al. [66] evaluated MetaMap [67], NCBO Annotator, and ConceptMapper on eight biomedical ontologies (Cell Type Ontology, Gene Ontology: Cellular Component, Gene Ontology: Molecular Function, Gene Ontology: Biological Process, Sequence Ontology, ChEBI, NCBI Taxonomy, Protein Ontology) using the Colorado Richly Annotated Full-Text Corpus (CRAFT). They examined over 1,000 combinations of parameters and concluded that ConceptMapper was the best-performing system, producing the highest F-measure for seven out of eight ontologies (Table. 2.4).

Table 2.3: Using all descendant ORDO terms of these terms as rare diseases

ID	Term (Label)	Definition	# of all descendant terms
ORDO: 377788	disease	An alteration of health status resulting from a physiopathological mechanism and having a homogeneous clinical presentation and evolution and homogeneous therapeutic possibilities. Excludes developmental anomalies.	3770
ORDO: 377789	malformation syndrome	A set of morphological anomalies resulting from a developmental anomaly involving more than one morphogenetic field regardless of the cause. Includes sequences and associations.	1702
ORDO: 377790	biological anomaly	An alteration of the normal values of biological products. Example : hypertransferrinemia.	7
ORDO: 377791	morphological anomaly	A set of anomalies resulting from a developmental anomaly involving only one morphogenetic field. Includes isolated anomalies and anatomical variants.	387
ORDO: 377792	clinical syndrome	A set of manifestations resulting from the alteration of a physiological state and that can be present in several diseases. Examples: nephrotic syndrome, hepatic failure.	32
ORDO: 377793	particular clinical situation in a disease or syndrome	A set of manifestations presenting as a subset of a disorder under particular circumstances.	30

Table 2.4: Best performance for eight ontologies

These results were reported in "Funk, C., Jr, W.B., Garcia, B., Roeder, C., Bada, M., Cohen, K.B., Hunter, L.E., and Verspoor, K. (2014). Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics *15*, 59" [66]. Bolded systems produced the highest F-measure.

Cell Type Ontology (CL)			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.32	0.76	0.20
MetaMap	0.69	0.61	0.80
ConceptMapper	0.83	0.88	0.78
Gene Ontology – Cellular Component (GO_CC)			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.40	0.75	0.27
MetaMap	0.70	0.67	0.73
ConceptMapper	0.77	0.92	0.66
Gene Ontology – Molecular Function (GO_MF)			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.08	0.47	0.15
MetaMap	0.09	0.09	0.34
ConceptMapper	0.14	0.44	0.29
Gene Ontology – Biological Process (GO_BP)			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.25	0.70	0.15
MetaMap	0.42	0.53	0.34
ConceptMapper	0.36	0.46	0.29
Sequence Ontology (SO)			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.44	0.63	0.33

MetaMap	0.50	0.47	0.54
ConceptMapper	0.56	0.56	0.57
ChEBI			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.56	0.7	0.46
MetaMap	0.42	0.36	0.50
ConceptMapper	0.56	0.55	0.56
NCBI Taxonomy			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.04	0.16	0.02
MetaMap	0.45	0.31	0.88
ConceptMapper	0.69	0.61	0.79
Protein Ontology (PRO)			
System	**F-measure**	**Precision**	**Recall**
NCBO Annotator	0.50	0.49	0.51
MetaMap	0.36	0.39	0.34
ConceptMapper	0.57	0.57	0.57

Besides, I evaluated MetaMap [67], NCBO Annotator, and ConceptMapper on HPO using HPO gold standard [13]. Although ConceptMapper produced lower F-measure than MetaMap (Table. 2.5), ConceptMapper was ~50 times faster than MetaMap and NCBO Annotator (Table. 2.6). Considering these results, I decided to use ConceptMapper for annotating published case reports. To conduct ConceptMapper with HPO and ORDO, I used ccp nlp-pipelines with default parameter sets [66].

Table 2.5: Best performance for HPO

Bolded system produced the highest F-measure.

System	F-measure	Precision	Recall
NCBO Annotator	0.51	0.54	0.47
MetaMap	0.56	0.51	0.61
ConceptMapper	0.52	0.52	0.51

Table 2.6: Best performance for HPO gold standard

Bolded system produced the fastest processing time to annotate 228 abstracts of HPO gold standard.

System	Processing time (sec)
NCBO Annotator	206.0
MetaMap	351.0
ConceptMapper	4.3

Many synonyms are present in HPO and ORDO, and some of them are abbreviations of labels. A previous study reported that 81.2% of abbreviations are ambiguous and have an average of 16.6 meanings [68]. Thus, there are instances where case reports annotated with synonyms that are abbreviations do not include their labels. For example, the label of ORDO: 103918 is “tropical pancreatitis,” and its synonym is “TCP.” A case report with PubMed ID 24472742 includes “TCP,” but it does not include “tropical pancreatitis” and instead includes “thrombocytopenia.” To exclude inappropriate annotations, I used the Allie database that deposits abbreviations generated by all titles and abstracts in PubMed (Table. 2.7). Annotations including synonyms were excluded if a case report did not include both the synonym and its label in the text.

Table 2.7: Examples of Allie’s outputs

Abbreviation	Long form
SPF	Specific pathogen-free
	S-phase fraction
	Sun protection factor
MAP	Mean arterial pressure
	Mitogen-activated protein
	Mean arterial blood pressure
	Microtubule-associated protein
BAC	Bacterial artificial chromosome
	Blood alcohol concentration
	Bronchioloalveolar carcinoma
	Benzalkonium chloride

Finally, using the processed annotations, I identified DPAs in all titles and abstracts of published case reports. Various approaches have been proposed to extract relations such as protein-protein interactions and disease-gene associations from biomedical text. I used the most straightforward approach to identify DPAs that are co-occurrences of an ORDO term and an HPO term within a sentence using the processed annotations [69]. Owing to the intrinsic complexity of the biomedical text, most of the cases using this approach work on the sentence-based level. Note that, this approach tends to be with high recall but be with low precision. I chose this approach due to acquiring diverse DPAs.

Development of PubCaseFinder

I developed PubCaseFinder, a new phenotype-driven differential diagnosis system using the DPAs extracted from the one million of case reports. PubCaseFinder is based on a DPA database where phenotypes are associated with diseases defined in Orphanet. Some of the DPAs are from Orphanet, whereas some originate from text mining results. The goal of the system was to help clinicians rank candidate diseases for a patient who is suspected to be a case of a rare disease. PubCaseFinder takes as input a set of HPO terms that describe the signs and symptoms of the patient. The case representation is then compared to diseases in the database. Note that each disease in the database is also represented by a set of HPO terms. Thus, the comparison is performed as a similarity computation between two sets of HPO terms. As a result, PubCaseFinder outputs a ranked list of candidate diseases according to the similarity score.

To calculate semantic similarity between two sets of HPO terms, several measures such as Resnik [70], Lin [71], Jiang-Conrath [72], simGIC [73], and GeneYenta [74] have been recommended (Table. 2.8). These measures are used for patient diagnosis [6], enrichment analysis of gene sets and disease sets [75,76], discovering causative genes of rare diseases [74], and other applications. Resnik, Lin, and Jiang-Conrath define semantic similarity between two HPO terms as the information content (IC) of the most informative common ancestor. simGIC is defined as the sum of IC of HPO terms shared by two sets of HPO terms, divided by the sum of IC of those HPO terms.

Table 2.8: Measures to calculate semantic similarity between two sets of HPO terms

The measures of Resnik, Lin, and Jiang-Conrath are used for comparing two HPO terms (*a, b*) using each equation. There are two variations in measuring the similarity between two sets of HPO terms (*P, Q*). The Avg is the average score, and the Max is the highest score in the scores for HPO terms in *P*. On the other hand, the measure of simGIC directly measures the similarity between two sets of HPO terms. g^P is the set of HPO terms in *P* and all ancestral HPO terms of them.

Measure	Equation	Variations	Reference
Resnik(*a,b*)	$\max_{t \in g^a \cap g^b} IC(t)$	Avg, Max	[70]
Lin(*a,b*)	$\frac{2 * Resnik(a,b)}{IC(a) + IC(b)}$	Avg, Max	[71]
Jiang-Conrath(*a,b*)	$\frac{1}{IC(a) + IC(b) - 2 * Resnik(a,b) + 1}$	Avg, Max	[72]
simGIC(*P,Q*)	$\frac{\sum_{t \in g^P \cap g^Q} IC(t)}{\sum_{t \in g^P \cup g^Q} IC(t)}$		[73]

PubCaseFinder uses GeneYenta (based on Resnik's measure) and constitutes a user-weighted matching algorithm that empowers researchers to leverage their expertise and knowledge to customize results. In clinical practice, a specific phenotype may be extremely prominent or severe; thus, this algorithm allows users to set a matching weight for each phenotype [77]. The Gene Yenta algorithm represents the similarity ranging from 0% for no phenotypic overlap to 100% for complete phenotypic overlap. The algorithm starts with determining the information content (IC_t) of each HPO term *t*. $P(t)$ is the probability of occurrence of an HPO term *t* in a set of case reports, and the IC_t of the HPO term *t* is defined as follows:

$$P(t) = \frac{|annot_t|}{|annot_{all}|},$$

$$IC_t = -\log P(t),$$

where $annot_{all}$ is the total number of annotations of all HPO terms in case reports and $annot_t$ is the total number of annotations of the HPO term *t* and all its descendants in case reports. That is, for the root node, $P(t)$ is 1 and IC_t is 0. There is an inverse relation between IC and the total number of annotations of an HPO term *t*. IC_t of the most informative common ancestor of the two HPO terms was assigned as the similarity sim_{terms} between two HPO terms. This is defined as follows:

$$sim_{terms}(t, t') = max_{a_t \in A_t \cap A_{t'}} IC_{a_t},$$

where A_t is the HPO term *t* and all ancestral HPO terms of *t*, and a_t is the HPO term of a intersection of A_t and $A_{t'}$. The similarity $sim_{case\_disease}$ between a case and a disease assesses the resemblance between their sets of HPO terms and is defined as follows:

$$sim_{case\_disease}(c, d) = \frac{\sum_{t \in T_c} R_t \times max_{t' \in T_d} sim_{terms}(t, t')}{\sum_{t \in T_c} R_t \times IC_t},$$

where R_t allows users to assign weights ranging from 1 to 5 that represent how important a term t is for the user. For this evaluation, I assigned 1 to R_t for any HPO terms, where c represents a case, and d represents a disease. T_c and T_d represent HPO terms for a case and disease, respectively. PubCaseFinder provides a ranked list of diseases according to $sim_{case\_disease}$, but the disease with the fewest T_d becomes highest ranking in the case of diseases with the same $sim_{case\_disease}$.

Chapter 3
Results

Identifying disease–phenotype associations from case reports

I annotated titles and abstracts of 1,083,283 case reports with HPO terms and ORDO terms and identified DPAs that are co-occurrences of an ORDO term and an HPO term within a sentence. As a result, 810,705 case reports were annotated with 6,380 HPO terms and 316,674 case reports were annotated with 3,788 ORDO terms. Using these annotations, I identified 70,011 DPAs consisting of 3,881 HPO terms and 3,072 ORDO terms. I also obtained 51,590 DPAs composed of 4,832 HPO terms and 2,478 ORDO terms from Orphanet. Figure 3.1 shows the overlap between the two sets of ORDO terms included in DPAs from case reports and Orphanet. I found that 1,483 ORDO terms were common to the two data sources and 1,589 ORDO terms included in DPAs from case reports were not found in DPAs from Orphanet.

Within the overlapping 1,483 ORDO terms, I compared 40,512 DPAs from case reports with 35,172 DPAs from Orphanet. I regarded ORDO terms as the same if their related HPO terms were located in the same, superordinate, or subordinate part of the ontology hierarchy. As a result, 11,593 DPAs were in common, and 28,919 new DPAs were added to 1,483 rare diseases included in DPAs from Orphanet. I also identified 29,499 DPAs for 1,589 rare diseases that are not associated with a

phenotype in Orphanet. In total, our text mining approach could identify 58,418 new DPAs and increase the coverage of DPAs in Orphanet by 113.2%.

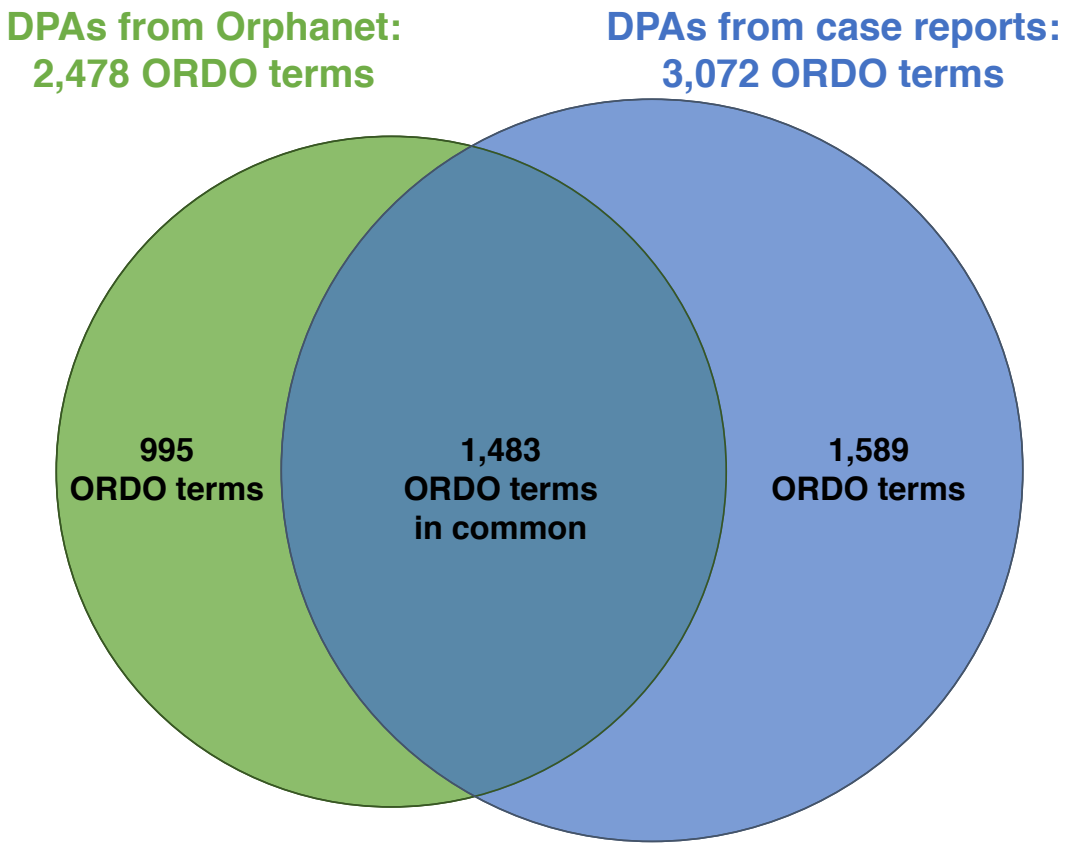


Figure 3.1: Overlap between two sets of ORDO terms found in disease–phenotype associations (DPAs) from Orphanet and case reports

An overview of PubCaseFinder

I implemented the algorithms described above in a web application called PubCaseFinder. By typing a patient's phenotype in the search box, candidate HPO terms are displayed (Fig. 3.2). This enables rapid entry of HPO terms because users select appropriate HPO terms from the list. Moreover, users can obtain detailed information about HPO terms such as definition, synonyms, superordinate concepts, and subordinate concepts (Fig. 3.3). The patient is then compared with all rare diseases in Orphanet based on phenotypic similarity, and Figure 3.4 shows the ranked list of rare diseases. Users can also narrow down the ranked list of rare diseases to specify the causative genes of rare diseases. The higher the phenotypic similarity, the higher the displayed probability as a candidate disease. In addition to comparing a patient's phenotypes with rare diseases, users can compare a patient's phenotypes against published case reports that are associated with their HPO terms in the same manner (Fig. 3.5). By the ranked lists of rare diseases and case reports, clinicians can discuss differential diagnoses for undiagnosed patients with suspected rare diseases. To confirm detailed contextual information on the presence of DPAs, PubCaseFinder shows the context in which a DPA appears (Fig. 3.6). To keep up with new DPAs that are continuously introduced in case reports, PubCaseFinder is equipped with an automatic update system.

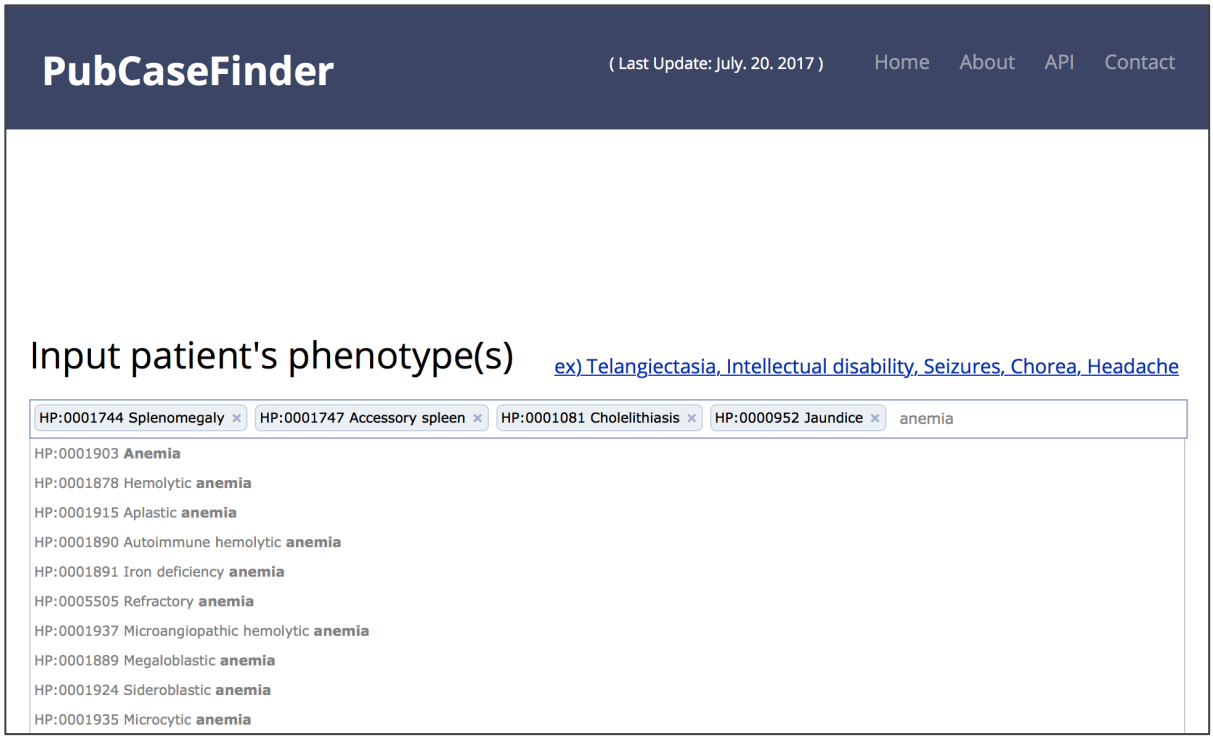


Figure 3.2: Textbox having autocompleted feature for helping user's rapid entry of HPO terms

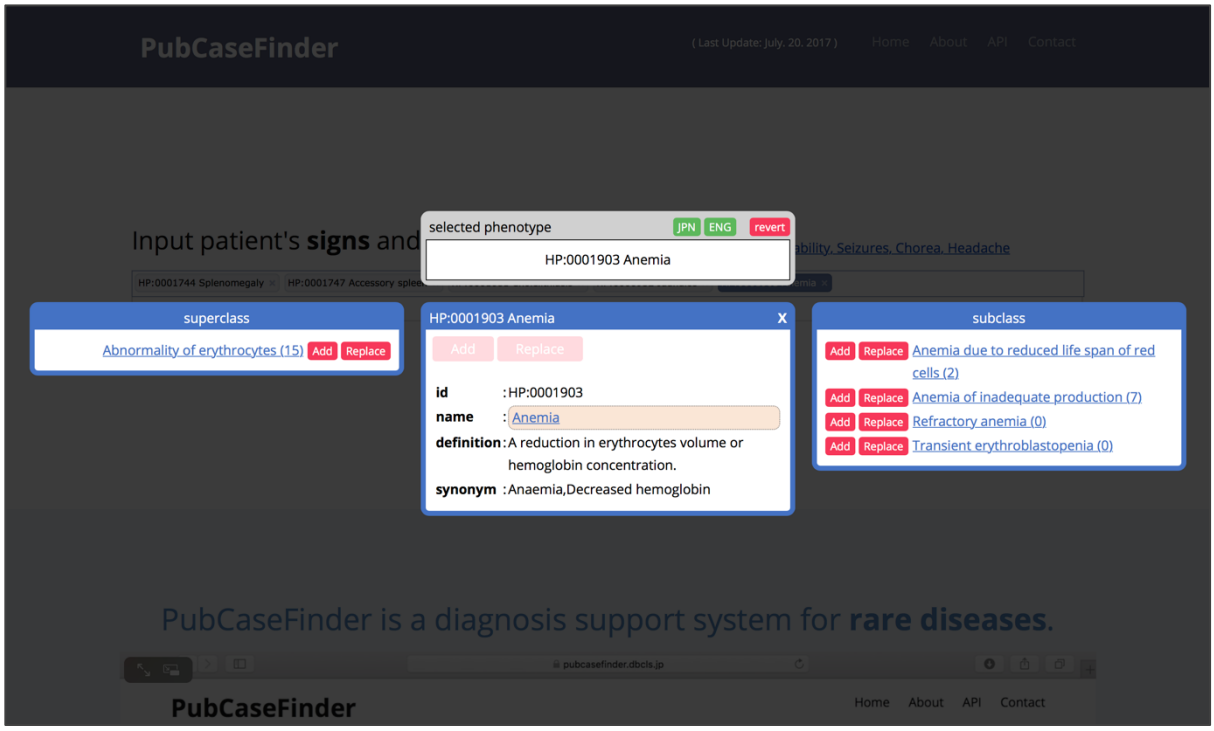


Figure 3.3: Presenting detailed information about HPO terms such as definition, synonyms, superordinate concepts, and subordinate concepts to facilitate the inputting of HPO terms

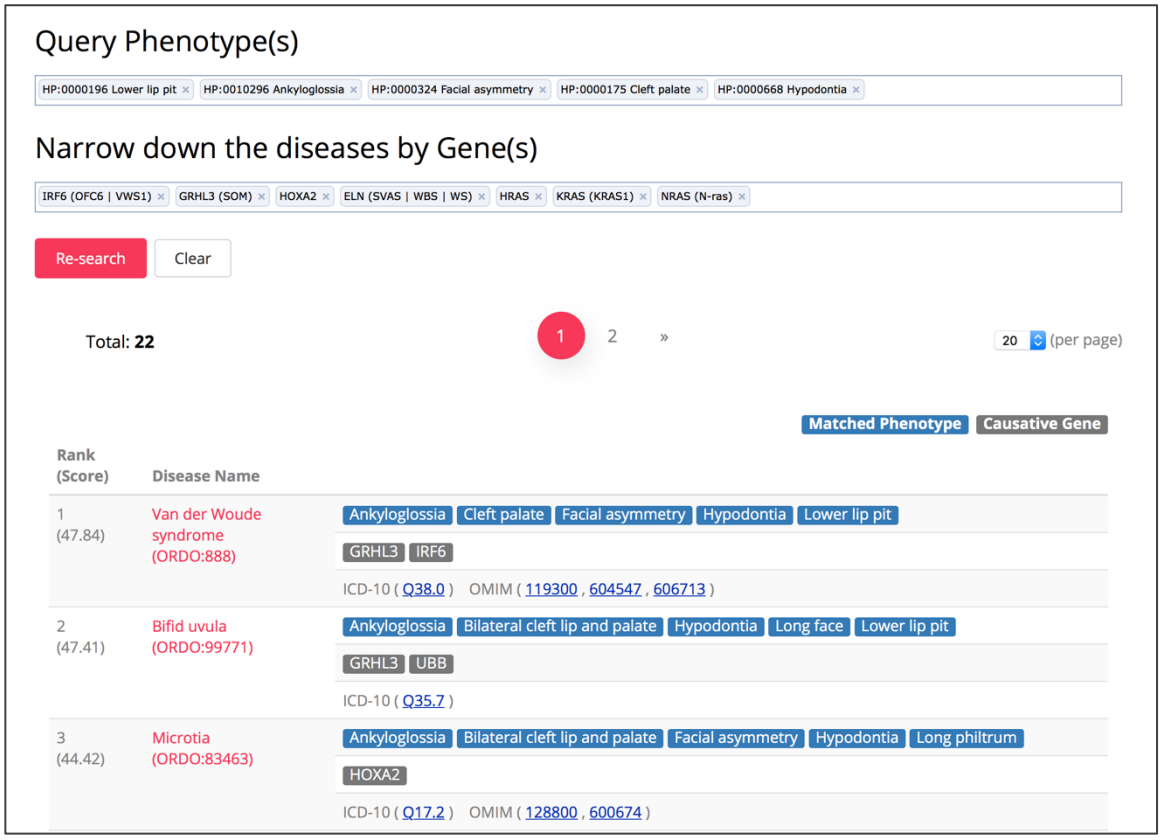


Figure 3.4: Providing a ranked list of rare diseases defined in Orphanet based on phenotypic similarity for supporting differential diagnosis of rare diseases

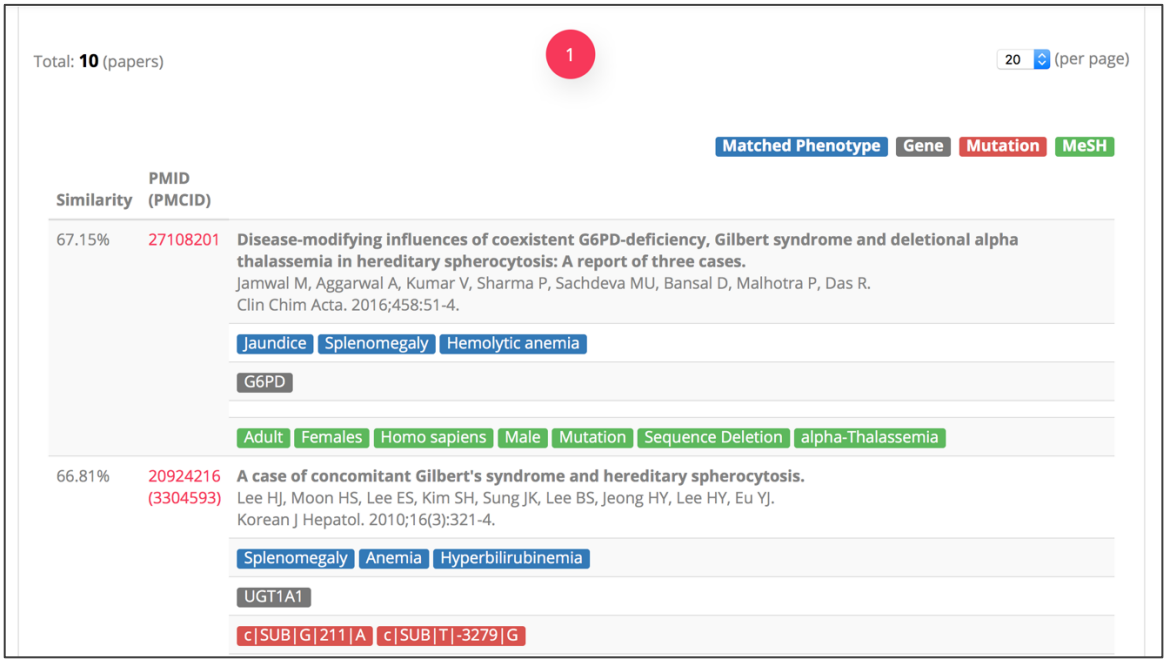


Figure 3.5: Providing a ranked list of published case reports in PubMed based on phenotypic similarity for supporting differential diagnosis of rare diseases

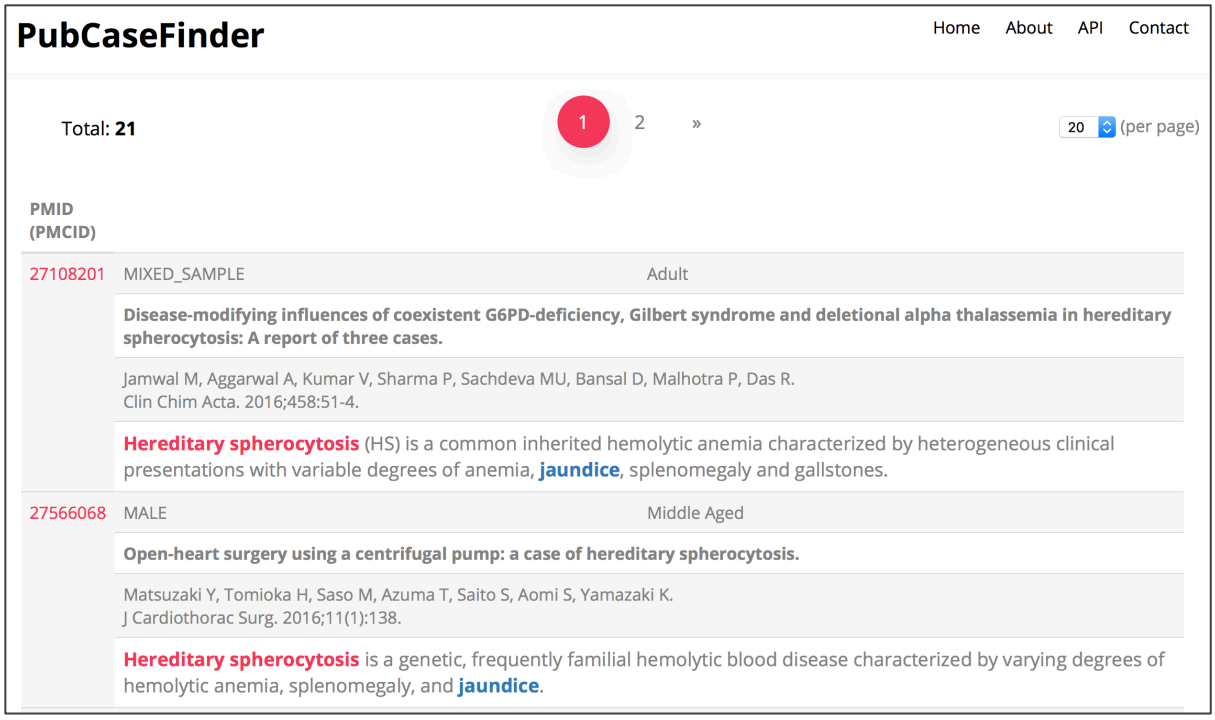


Figure 3.6: Providing detailed contextual information on the presence of disease-phenotype associations

To identify new causative genes and new diseases for rare diseases, the comparison of the exomes or genomes of unrelated patients and case reports with similar phenotypes is a promising method, but it is a nontrivial task. Such patients will likely be seen by different clinicians at different hospitals and different countries, and the clinician will often be unaware of other cases. Currently, to find such patients on a worldwide scale, many patient repositories and matchmaking services have been implemented all over the world. These repositories and services made possible to find such patients. However, those lack methods to consult published case reports. So, I integrated PubCaseFinder with available patient repositories and matchmaking services, namely, IRUD (Initiative on Rare and Undiagnosed

Diseases) Exchange [6]/Patient Archive [33] and PhenomeCentral [79] in BioHackathon2017. IRUD is actively engaged in the diagnosis of patients with suspected rare diseases in Japan, and IRUD Exchange is a customized system of the Patient Archive platform for IRUD. PhenomeCentral is a portal for phenotypic and genotypic matchmaking of patients with suspected rare genetic diseases. I developed a JSON-based REST endpoint to query PubCaseFinder using HPO terms and Ensemble gene IDs and to return ranked lists of rare diseases and case reports based on phenotypic similarity. I also developed the Matchmaker Exchange (MME) application programming interface (API) [79] as a secondary querying option for PubCaseFinder. Using the PubCaseFinder API and the MME API, I enabled the use of PubCaseFinder in both IRUD Exchange/Patient Archive and PhenomeCentral.

Performance evaluation of PubCaseFinder

To evaluate the performance of PubCaseFinder as a phenotype-driven differential diagnosis system, I collected 1,584 clinical cases from PhenomeCentral, which were registered by the Care4Rare Canada Consortium. It turned out only 243 cases out of them had both phenotypes and diagnoses, the former represented by HPO terms and the latter represented by MIM IDs. I used them as the test cases of our evaluation. All MIM IDs of the cases were converted to Orpha numbers using connections between MIM IDs and Orpha numbers in ORDO. To evaluate the effect of DPAs form case reports, I compared the performance of PubCaseFinder in three different settings: one with DPAs only from Orphanet (PubCaseFinder-O), one with DPAs only from case reports (PubCaseFinder-CR), and one with DPAs from both (PubCaseFinder-O/CR). For a reference purpose, I included Orphamizer (a

customized system of Phenomizer for Orphanet) in our comparison because it was the most similar system among phenotype-driven differential diagnosis systems, using DPAs from Orphanet and targeting the diseases defined in ORDO. For the evaluation, I compiled two exclusive sets of diseases as targets of differential diagnosis; one consisted of 2,323 diseases that were associated with phenotypes in Orphanet and consequently could be potentially solved by both PubCaseFinder and Orphamizer (Target-A), the other consisted of 1,589 diseases that were not associated with a phenotype in Orphanet (Target-B).

First, I evaluated the performance of PubCaseFinder (in the three different settings) and Orphamizer, when targeting target-A. Figure 3.7 shows the evaluation process. The 135 cases out of the 243 PhenomeCentral cases were used for this evaluation (Table. S1), as they had diagnoses which belonged to Target-A. The result of each run was obtained as a ranked list of diseases. They were represented in terms of “recall at ranks” (i.e., the fraction of cases where the correct diagnosis appeared in the top-listed diseases). Figure 3.8 shows the recall rates by PubCaseFinder-O and PubCaseFinder-O/CR (the recall numbers are shown in Table 3.1). The top 10 recall rate of PubCaseFinder-O/CR is 57% (Fig. 3.8), which means that there is a correct diagnosis in the top 10 of a ranked list of 2,323 diseases for about one in every two cases. All recall rates of PubCaseFinder-O/CR are higher than those of PubCaseFinder-O (Fig. 3.8). The top 50 recall rate of PubCaseFinder-O is lower than the top 20 recall rate of PubCaseFinder-O/CR, which means that even if a user checks the top 50 diseases of PubCaseFinder-O, the diagnostic rate is lower than when checking the top 20 diseases of PubCaseFinder-O/CR.

Figure 3.9 shows the recall rates by Orphamizer and PubCaseFinder-O/CR (the recall numbers are shown in Table 3.1). All recall rates of PubCaseFinder-O/CR are higher than those of Orphamizer (Fig. 3.9). The top 100 recall rate of Orphamizer is lower than the top 10 recall rate of PubCaseFinder-O/CR, which means that even if a user checks the top 100 diseases of Orphamizer, the diagnostic rate is lower than when checking the top 10 diseases of PubCaseFinder-O/CR. I also evaluated the statistical significance of a correct diagnosis appearing in the top 10 with a binomial test and found that the p-value of PubCaseFinder-O/CR was 4.01×10^{-144}, whereas those of PubCaseFinder-O and Orphamizer were 2.83×10^{-108} and 4.73×10^{-65}, respectively. Those results clearly show the potential of DPAs from case reports to improve the performance of phenotype-driven differential diagnosis systems.

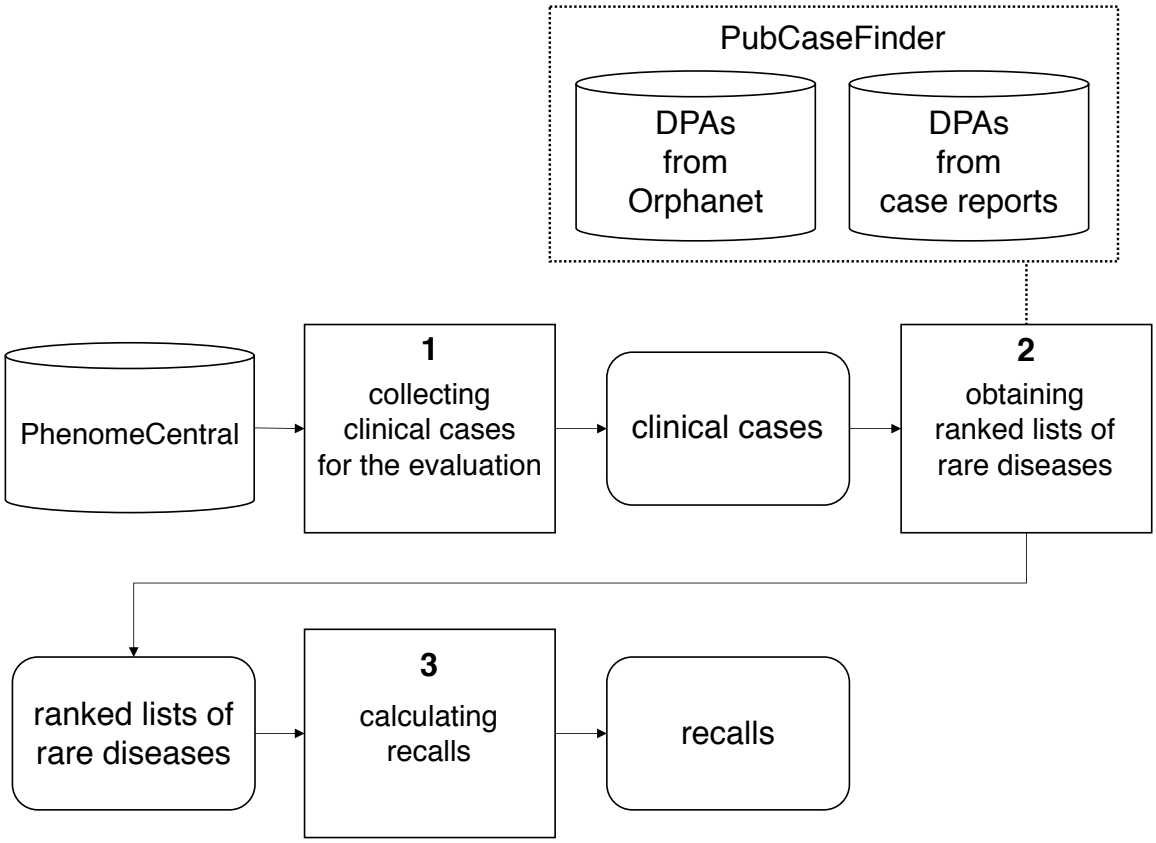


Figure 3.7: Performance evaluation of PubCaseFinder

Clinical cases of rare diseases were collected from PhenomeCentral (step 1), and a ranked list of rare diseases based on phenotypic similarity was obtained with PubCaseFinder for each clinical case (step 2). The performance of PubCaseFinder was evaluated using the “recall at ranks” metric (step 3).

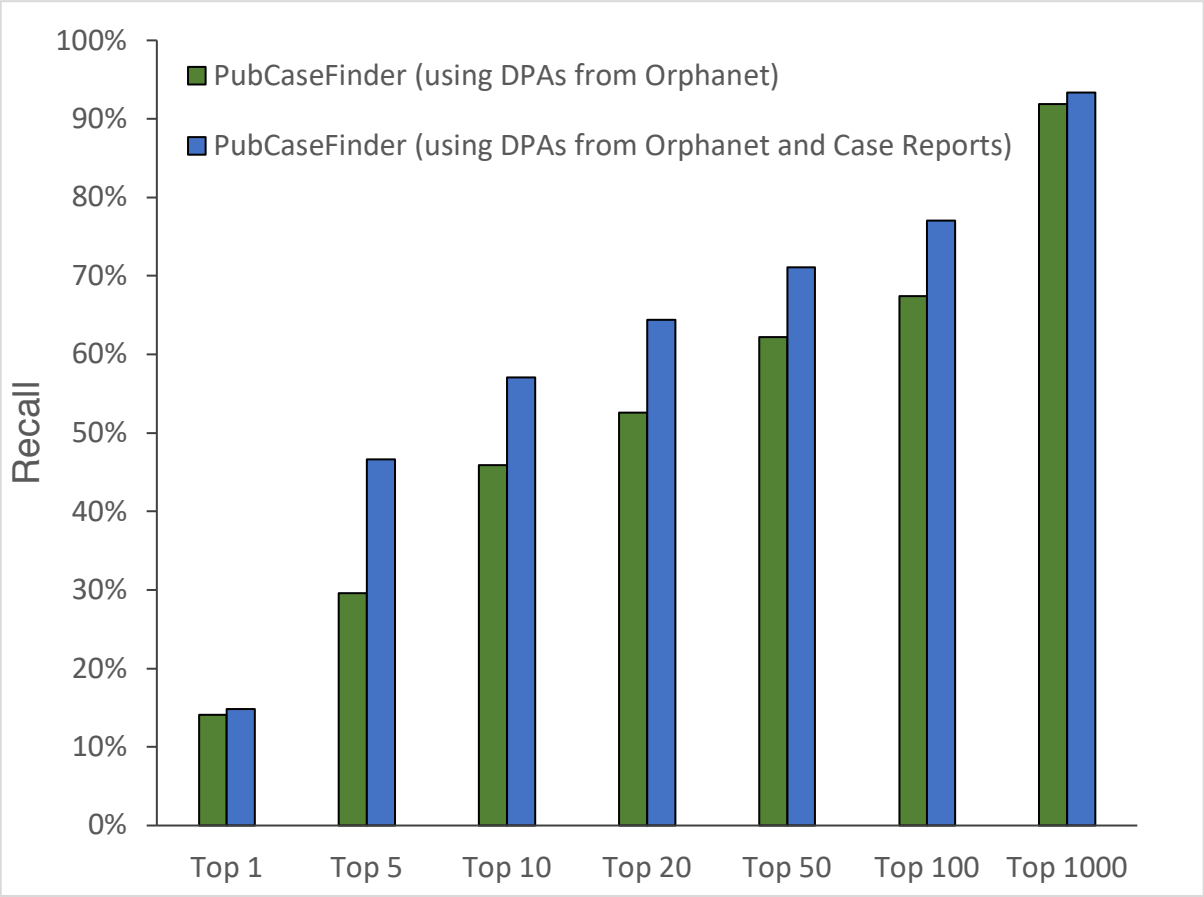


Figure 3.8: Performance comparison of PubCaseFinder-O and PubCaseFinder-O/CR

Recalls were calculated on the basis of ranked lists of 2,323 rare diseases for 135 clinical cases from PhenomeCentral.

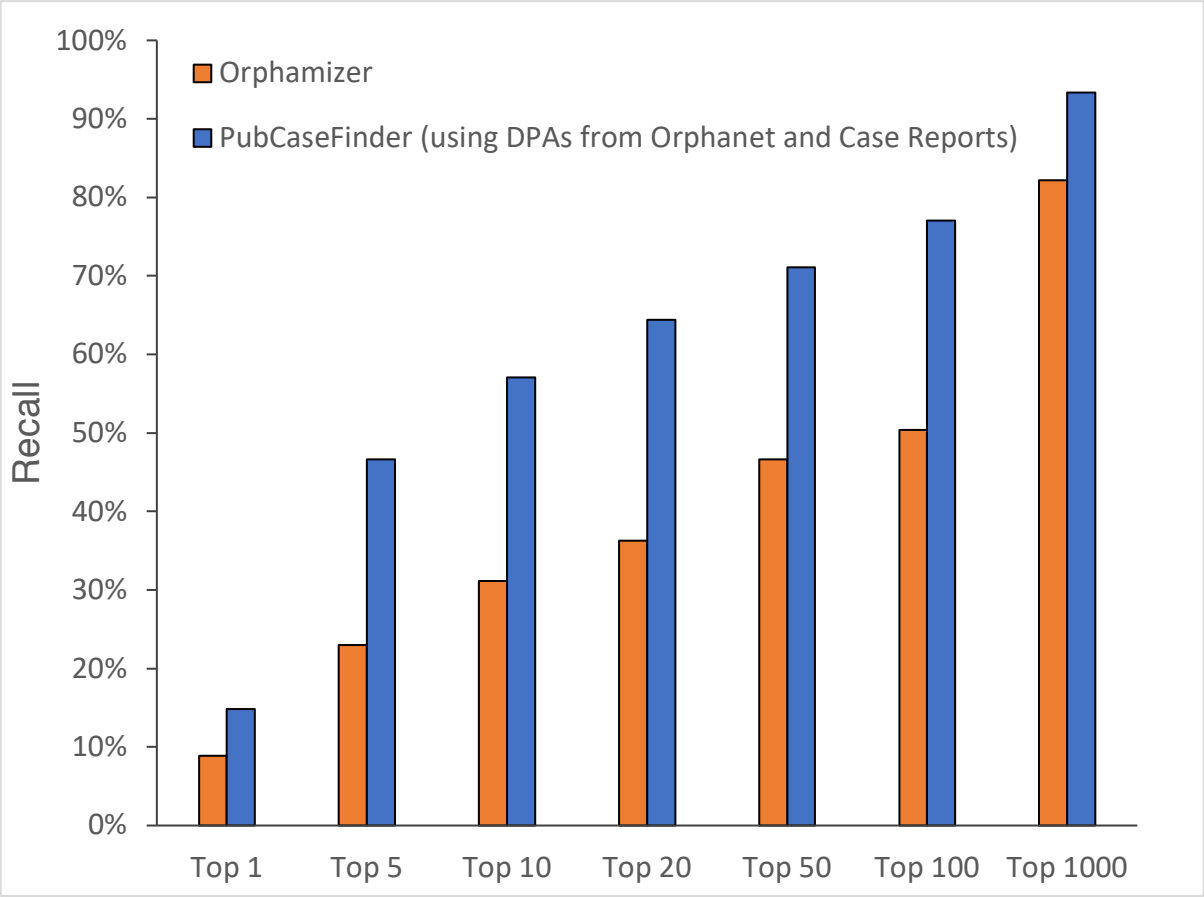


Figure 3.9: Performance comparison of Orphamizer and PubCaseFinder-O/CR

Recalls were calculated on the basis of ranked lists of 2,323 rare diseases for 135 clinical cases from PhenomeCentral.

Table 3.1: Recall numbers by PubCaseFinder in three different settings and by Orphamizer

Differential diagnosis system	Top 1 recall number (rate)	Top 5 recall number (rate)	Top 10 recall number (rate)	Top 20 recall number (rate)	Top 50 recall number (rate)	Top 100 recall number (rate)	Top 1000 recall number (rate)
Orphamizer	12 (8.9%)	31 (23.0%)	42 (31.1%)	49 (36.3%)	63 (46.7%)	68 (50.4%)	111 (82.2%)
PubCaseFinder (using DPAs from Orphanet)	19 (14.1%)	40 (29.6%)	62 (45.9%)	71 (52.6%)	84 (62.2%)	91 (67.4%)	124 (91.9%)
PubCaseFinder (using DPAs from Orphanet and Case Reports)	20 (14.8%)	63 (46.6%)	77 (57.0%)	87 (64.4%)	96 (71.1%)	104 (77.0%)	126 (93.3%)

Let us take a running example. A clinical case from PhenomeCentral had HP:0000657 (Oculomotor apraxia), HP:0001263 (Global developmental delay), and HP:0002066 (Gait ataxia), as the phenotypes, which were diagnosed with ORDO:2318 (Joubert syndrome with oculorenal defect). PubCaseFinder-O could place ORDO:2318 only at the 41st rank because the association between it and HP:0000657 was missing in the DPAs from Orphanet. However, the association existed in the DPAs from case reports, and PubCaseFinder-O/CR could place it at the 5th rank.

Second, I evaluated the performance of PubCaseFinder-CR when targeting target-B. I narrowed down 243 cases of PhenomeCentral to 59 cases (Table. S2) whose

diagnoses were part of the target-B. For the 59 cases, I obtained ranked lists of target-B using PubCaseFinder-CR and then calculated recalls on the basis of the results. PubCaseFinder-CR showed the recall number (rate), 2(3.4%)@1, 3(5.1%)@5, 5(8.5%)@10, 6(10.2%)@20, 13(22.0%)@50, 24(40.7%)@100, and 56(94.9%)@1000 (Fig. 3.10). I evaluated the statistical significance of a correct diagnosis appearing in the top 10 by using a binomial test and found a p-value of 3.72×10^{-5}. Although Figure 3.10 highlights the low recall rates of PubCaseFinder-CR, the *p*-value shows the potential of PubCaseFinder for differential diagnosis of rare diseases that were not associated with a phenotype in Orphanet. Note that the recall rates of PubCaseFinder-CR for target-B were lower than those of PubCaseFinder-CR for target-A even though they both exploited DPAs from case reports. From examining the number of associated DPAs from case reports for target-A and target-B, on average, each disease in target-A had 27.3 DPAs, while each disease in target-B had 18.6 DPAs. It is interpreted that the difference of the number of associated DPAs causes the difference in recall rates.

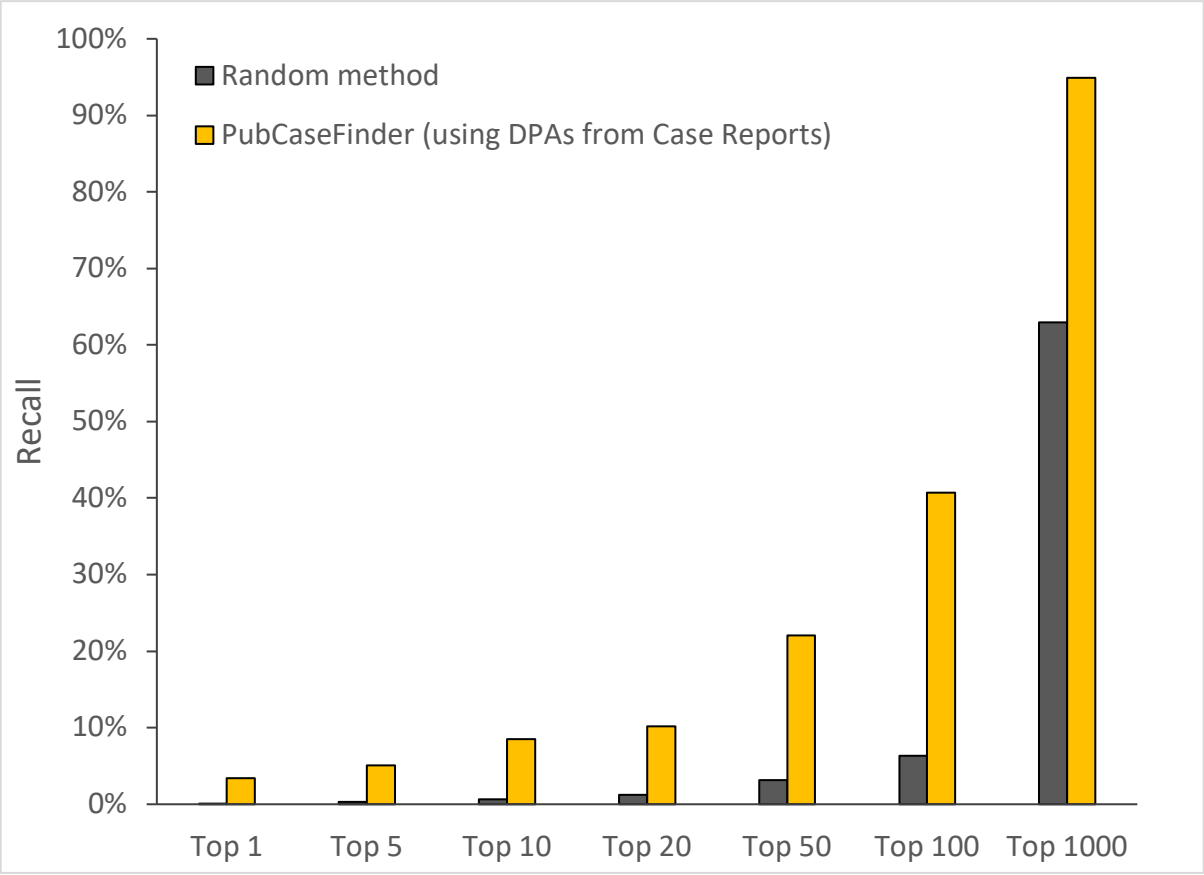


Figure 3.10: Performance comparison of a random method and PubCaseFinder-CR for rare diseases not included in disease–phenotype associations (DPAs) from Orphanet

Recalls were calculated on the basis of ranked lists of 1,589 rare diseases for 59 clinical cases from PhenomeCentral.

Filtering of unreliable disease-phenotype associations

In a previous study, Tudor et al. [80] also tried to extract DPAs for common diseases from papers in PubMed, and they suggested to ignore frequent low occurrences to filter out potentially noisy DPAs. This method is often used and is based on the hypothesis that if two entities are frequently mentioned together, it is likely that they are related [69]. However, I found that most DPAs identified in this study appeared in few case reports. Figure 3.11 shows the distribution of DPA numbers from case reports according to frequencies of occurrence in case reports. More than half of DPAs appeared in only one case report, and the ratio of DPAs that appeared in multiple case reports was only ~34.0%. Using the 135 clinical cases from PhenomeCentral, I calculated the top 10 recall rate of PubCaseFinder that exploits each set of DPAs filtered by their frequencies of occurrence in case reports (Fig. 3.11, all results are shown in Table S4). The top 10 recall rates gradually decreased from 57.0% to 49.6% when increasing the frequency of occurrence. Our results show that low-frequency DPAs from case reports include many DPAs that are informative for the differential diagnosis of rare diseases. I should therefore not filter out DPAs for rare diseases using frequencies of occurrence.

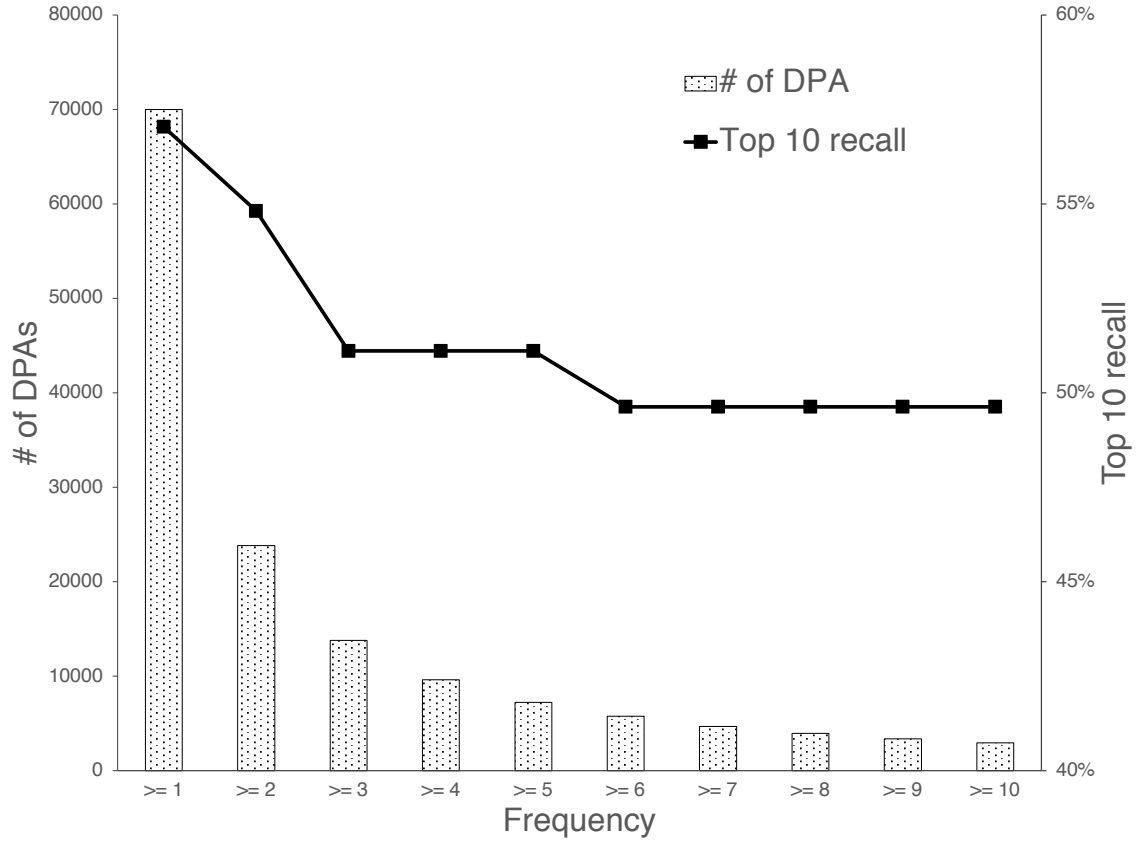


Figure 3.11: Distribution of numbers of disease–phenotype associations (DPAs) from case reports (bars) and Top 10 recall rates (solid line)

For each set of DPAs ordered according to the frequency of occurrence in case reports, the number of DPAs was counted, and the top 10 recall rate was calculated to evaluate the performance of PubCaseFinder using the set of DPAs.

Chapter 4
Discussion and Future Work

To speed up the differential diagnosis process based on symptoms and signs observed from patients, phenotype-driven differential diagnosis systems for rare diseases have been developed and implemented. The performance of these systems is influenced by the quantity and quality of underlying DPA databases. I found that the limited coverage of manually curated databases was a significant problem that hindered the further progress of automated differential diagnosis. To address the problem, I developed a text mining approach to extend the coverage of DPAs in manually curated databases like Orphanet. By applying the approach to a million case reports from PubMed, I could increase the coverage of DPAs from Orphanet more than two times. Based on the extended DPA database, I also developed PubCaseFinder, a new phenotype-driven differential diagnosis system. A series of experiments which was conducted using clinical cases from PhenomeCentral showed that the performance of phenotype-driven differential diagnosis could substantially be improved thanks to the extension of the DPA database. Previous studies reported that case reports were an essential tool for extracting valuable information for rare diseases in spite of low certainty evidence due to its small samples. I, therefore, targeted the one million case reports included in PubMed, and this is, to our knowledge, the first demonstration that such a extensive collection of case reports was useful for tackling rare disease issues by using a text mining method. I will extend the collection of case reports to those of European in Europe PMC [84] and those of Japanese in J-STAGE [85] and will

believe that further case reports will contribute more for tackling rare disease issues. In addition, I will also use all papers in PubMed instead of the collection of case reports to examine whether the performance will further improve.

Note that automatic text mining techniques are often regarded as assistive tools to help manual curation of databases, due to its potentially high chance of noisy results. Our automatically extracted DPAs using text mining techniques also included noisy results, but they included many new DPAs which were not obtained by manual curation of Orphanet. Figure 6 shows that the performance of PubCaseFinder was much low when using automatically extracted DPAs independently. However, I could regard them as useful supplementary information for manual curated DPAs since the performance was most high by using both in combination. Manual curation is the best approach for obtaining correct DPAs, but our proposed approach using text mining techniques is deemed practically useful because manual curation will take enormous time and cost, particularly considering the large volume and rapid growth of case reports.

For annotation with HPO terms and ORDO terms, I used ConceptMapper which was reported as a state-of-the-art concept recognition system among publicly available ones. Recently, Bio-LarK, which was also a concept recognizer specifically tailored to annotate HPO terms have become a publicly available system. A previous study showed that Bio-LarK was benchmarked using both the gold standard and the test suite corpora for HPO and outperformed other concept recognizers [81]. As our approach does not rely on a specific concept recognizer, I am planning to seek a further performance improvement by finding and adopting a more optimal concept recognizer.

Our experiment and discussion on filtering of unreliable DPAs suggest that a simple filtering method based on the frequency of occurrence will not work well for automated differential analysis of rare diseases, although it was reported to be useful for common diseases. I attribute the reason to the nature of data for rare diseases which are much less frequent than that of common diseases. Although the coverage of DPAs in Orphanet was improved with a text mining approach, unreliable DPAs (i.e., including negations in the sentence, which might represent a cause of false-positive association) were found. As a preliminary study, I performed an NLP tool negation-detection [39] against all sentences of case reports, which include DPAs and detected the sentences including negations. As a result, negated HPO terms or negated ORDO terms were detected in the 4.2% of sentences including DPAs (Table 4.1). As a future work, I am planning to exclude such negated sentences by manual curation and develop much more sophisticated filtering methods than simple, frequency-based filtering. Conversely, I will consider increasing DPAs which may include more noisy ones. Although DPAs are extracted within a sentence including both an ORDO term and HPO term in this study, I am planning to expand the searching region to two sentences or more. For example, in case of extracting DPAs within the entire abstract, more noisy results may increase, but the performance of PubCaseFinder may be improved using them.

In clinical practice, a specific phenotype may be extraordinarily prominent or severe; thus, I used the GeneYenta algorithm that allows users to set a matching weight for each phenotype. However, I always assigned the same weights to HPO terms in this evaluation in order to only evaluate the contribution of automatically extracted DPAs from case reports for improving the automated differential diagnosis.

As a future work, I am planning to modify the user interface of PubCaseFinder to make users set weights to HPO terms, which empowers users to leverage their expertise and knowledge to customize results. On the other hand, instead of assigning the weights to HPO terms of patients, it is possible to assign the weights to HPO terms for each disease. Even in such a case, it is considered that physician's expertise and knowledge can be reflected as the weights. Besides, to improve the accuracy of matching more, it is necessary to consider utilization other than phenotypes such as gender, age, family history, and medical history. When using a ranking system, top ranking results like Top 5 are essential. However, the scores of the top-ranked diseases in PubCaseFinder tend to be similar. In other words, only using the set of phenotypes alone, there is not much difference in ranking results. So, for example, I will discuss a new matching method that considers sex, age, family history, and medical history of a patient for calculating similarity scores. As a result, there is a possibility that the accuracy of top ranking results is improved.

I will integrate PubCaseFinder with TogoVar as future work. TogoVar is a database providing genomic variant information and the corresponding allele frequency information among ExAC [83], Japanese public variant database, and original Japanese dataset consisted from 183,884 samples genotyped by SNP-array and 125 whole exome sequence samples. Further, it provides biological information annotated by Variant Effect Predictor, publications registered in PubMed, and clinical data of ClinVar. Using integrated databases, for example, users will be able to search for case reports which will include variants with low allele frequency only in Japanese public variant database. In this study, I used only the phenotypes as input data for the evaluation. But, after integrating PubCaseFinder and TogoVar, I am

planning to evaluate my method using patient's variants in order to evaluate whether PubCaseFinder is effective in the process of genome analysis.

Table. 4.1 Examples of sentences including negations detected by the negation-detection tool

The disease names and symptoms are indicated in red and blue, respectively.

No	Sentence including a disease-phenotype association	Reference
1	Mucolipidosis III (ML-III), or pseudo-Hurler polydystrophy, is an autosomal recessive Hurler-like disorder without mucopolysaccharidiuria.	[86]
2	Van der Woude syndrome is an autosomal dominant disease characterized by lower lip pits with or without cleft lip and/or cleft palate.	[87]
3	The majority of the cases of nephroblastoma do not present with abdominal pain.	[88]
4	Developmental regression has not been reported in SPG56 patients.	[89]
5	This type of arthropathy has not been described in dermatomyositis or polymyositis.	[90]
6	Herein we report a case of a possible PHACE syndrome without hemangioma of the head but with a large segmental hemangioma of the trunk.	[91]
7	Systemic therapy for granulocytic sarcoma presenting without evidence of leukemia is reviewed.	[92]
8	Although there was a widely held belief that ALS does not cause cognitive impairment, cognitive function in patients with ALS has received more attention recently.	[93]
9	Our results indicate that hyperparathyroidism is not more prevalent in affected individuals with osteosarcoma than in the general population.	[94]

10	We report a rare case of toxocariasis with thoracic and pleural involvement without transient pulmonary infiltrates.	[95]

Web Resources

URLs for data presented herein are as follows:

Orphanet, http://www.orpha.net/consor/cgi-bin/index.php/

Phenomizer, http://compbio.charite.de/phenomizer/

Phenolyzer, http://phenolyzer.wglab.org

FACE2GENE, https://www.face2gene.com

Orphanet Rare Disease Ontology, http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php/

Human Phenotype Ontology consortium, http://human-phenotype-ontology.github.io

PubMed, https://www.ncbi.nlm.nih.gov/pubmed/

EBI, https://www.ebi.ac.uk

OMIM, https://www.omim.org

ICD10, http://www.who.int/classifications/icd/icdonlineversions/en/

MetaMap, https://metamap.nlm.nih.gov

NCBO Annotator, https://bioportal.bioontology.org/annotator

Ccp nlp-pipelines, https://github.com/UCDenver-ccp/ccp-nlp-pipelines

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

PubCaseFinder API, https://pubcasefinder.dbcls.jp/mme

Patient Archive, http://www.patientarchive.org

PhenomeCentral, https://www.phenomecentral.org

MME API, https://github.com/ga4gh/mme-apis

BioHackathon2017, http://2017.biohackathon.org

Care4Rare Canada Consortium, http://care4rare.ca

Orphamizer, http://compbio.charite.de/phenomizer_orphanet

negation-detection, https://github.com/gkotsis/negation-detection

Bibliography

1. Boat, T. F., & Field, M. J. (Eds.). (2011). Rare diseases and orphan products: Accelerating research and development (National Academies Press).

2. Sawyer, S.L., Hartley, T., Dyment, D.A., Beaulieu, C.L., Schwartzentruber, J., Smith, A., Bedford, H.M., Bernard, G., Bernier, F.P., Brais, B., et al. (2016). Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: Time to address gaps in care. Clin. Genet. *89*, 275–284.

3. Yu, H., and Zhang, V.W. (2015). Precision Medicine for Continuing Phenotype Expansion of Human Genetic Diseases. Biomed Res. Int. *2015*, 745043.

4. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P. A, Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N. Engl. J. Med. *369*, 1502–1511.

5. Stranneheim, H., and Wedell, A. (2016). Exome and genome sequencing: A revolution for the discovery and diagnosis of monogenic disorders. J. Intern. Med. *279*, 3–15.

6. Adachi, T., Kawamura, K., Furusawa, Y., Nishizaki, Y., Imanishi, N., Umehara, S., Izumi, K., and Suematsu, M. (2017). Japan's initiative on rare and undiagnosed diseases (IRUD): Towards an end to the diagnostic odyssey. Eur. J. Hum. Genet. *25*, 1025–1028.

7. Zhu, X., Petrovski, S., Xie, P., Ruzzo, E.K., Lu, Y.F., Melodi McSweeney, K., Ben-Zeev, B., Nissenkorn, A., Anikster, Y., Oz-Levi, D., et al. (2015). Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med.

17, 774–781.

8. Trujillano, D., Bertoli-Avella, A.M., Kumar Kandaswamy, K., Weiss, M.E., Köster, J., Marais, A., Paknia, O., Schröder, R., Garcia-Aznar, J.M., Werber, M., et al. (2017). Clinical exome sequencing: results from 2819 samples reflecting 1000 families. Eur. J. Hum. Genet. *25*, 176–182.

9. 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.

10. Fang, H., Wu, Y., Yang, H., Yoon, M., Jiménez-Barrón, L.T., Mittelman, D., Robison, R., Wang, K., and Lyon, G.J. (2017). Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine. BMC Med. Genomics *10*, 10.

11. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods *12*, 841–843.

12. Basel-Vanagaite, L., Wolf, L., Orin, M., Larizza, L., Gervasini, C., Krantz, I.D., and Deardoff, M.A. (2016). Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. Clin. Genet. *89*, 557–563.

13. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res. *45*, D865–D876.

14. Bauer, S., Köhler, S., Schulz, M.H. and Robinson, P.N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics. *28*, 2502–2508.

15. Robinson, P.N., Köhler, S., Oellrich, A. and Sanger Mouse Genetics ProjectSanger Mouse Genetics Project, Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D. *et al.* (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res., *24*, 340–348.

16. Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., ... & Øien, N. C. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science translational medicine*, *6*(252), 252ra123-252ra123.

17. Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., ... & Huff, C. D. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *The American Journal of Human Genetics*, *94*(4), 599-610.

18. Trakadis, Y. J., Buote, C., Therriault, J. F., Jacques, P. É., Larochelle, H., & Lévesque, S. (2014). PhenoVar: a phenotype-driven approach in clinical genomics for the diagnosis of polymalformative syndromes. *BMC medical genomics*, *7*(1), 22.

19. Sifrim, A., Popovic, D., Tranchevent, L. C., Ardeshirdavani, A., Sakai, R., Konings, P., ... & Moreau, Y. (2013). eXtasy: variant prioritization by genomic data fusion. *Nature methods*, *10*(11), 1083.

20. James, R. A., Campbell, I. M., Chen, E. S., Boone, P. M., Rao, M. A., Bainbridge, M. N., ... & Shaw, C. A. (2016). A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. *Genome medicine*, *8*(1), 13.

21. Javed, A., Agrawal, S., & Ng, P. C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature methods*, *11*(9), 935.

22. Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., ... & Akdemir, Z. H. C. (2015). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *The American Journal of Human Genetics*, *97*(2), 199-215.

23. Smedley, D., Schubach, M., Jacobsen, J. O., Köhler, S., Zemojtel, T., Spielmann, M., ... & Haendel, M. A. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *The American Journal of Human Genetics*, *99*(3), 595-606.

24. Greene, D., Richardson, S., Turro, E., & BioResource, N. I. H. R. (2016). Phenotype similarity regression for identifying the genetic determinants of rare diseases. *The American Journal of Human Genetics*, *98*(3), 490-499.

25. Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, *37*(suppl_2), W305-W311.

26. Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic acids research*, *41*(W1), W77-W83.

27. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., ... & Gough, J. (2008). SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic acids research*, *37*(suppl_1), D380-D386.

28. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., ... & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, *28*(5), 495.

29. Jiang, R. (2015). Walking on multiple disease-gene networks to prioritize candidate genes. *Journal of molecular cell biology*, *7*(3), 214-230.

30. Hart, S. N., Moore, R. M., Zimmermann, M. T., Oliver, G. R., Egan, J. B., Bryce, A. H., & Kocher, J. P. A. (2015). PANDA: pathway and annotation explorer for visualizing and interpreting gene-centric data. *PeerJ*, *3*, e970.

31. Gottlieb, A., Stein, G.Y., Ruppin, E. and Sharan, R. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol*.*, *7*, 496. 32. Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., ... & So, J. (2013). PhenoTips: patient phenotyping software for clinical and research use. *Human mutation*, *34*(8), 1057-1065.

33. Mcmurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., et al. (2016). Navigating the phenotype frontier: The Monarch Initiative. Genetics *203*, 1491–1495.

34. Gonzalez, M., Falk, M. J., Gai, X., Postrel, R., Schüle, R., & Zuchner, S. (2015). Innovative genomic collaboration using the GENESIS (GEM. app) platform. *Human mutation*, *36*(10), 950-956.

35. Smedley, D., Oellrich, A., Köhler, S., Ruef, B., Sanger Mouse Genetics Project, Westerfield, M., ... & Mungall, C. (2013). PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database*, *2013*, bat025.

36. Chen, C. K., Mungall, C. J., Gkoutos, G. V., Doelken, S. C., Köhler, S., Ruef, B. J., ... & Schofield, P. N. (2012). MouseFinder: candidate disease genes from mouse phenotype data. *Human mutation*, *33*(5), 858-866.

37. Mungall, C. J., Washington, N. L., Nguyen-Xuan, J., Condit, C., Smedley, D., Köhler, S., ... & Lewis, S. E. (2015). Use of model organism and disease databases

to support matchmaking for human disease gene discovery. *Human mutation*, *36*(10), 979-984.

38. Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, *39*(18), e119-e119.

39. Köhler, S., Doelken, S. C., Ruef, B. J., Bauer, S., Washington, N., Westerfield, M., ... & Robinson, P. N. (2013). Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, *2*.

40. Hwang, S., Kim, E., Yang, S., Marcotte, E. M., & Lee, I. (2014). MORPHIN: a web tool for human disease research by projecting model organism biology onto a human integrated gene network. *Nucleic acids research*, *42*(W1), W147-W153.

41. Köhler, S., Schoeneberg, U., Czeschik, J. C., Doelken, S. C., Hehir-Kwa, J. Y., Ibn-Salem, J., ... & Robinson, P. N. (2014). Clinical interpretation of CNVs with cross-species phenotype data. *Journal of medical genetics*, *51*(11), 766-772.

42. Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., & Ayme, S. (2012). Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Human mutation*, *33*(5), 803-808.

43. Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., ... & Lancet, D. (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database*, *2013*.

44. Rubinstein, W. S., Maglott, D. R., Lee, J. M., Kattman, B. L., Malheiro, A. J., Ovetsky, M., ... & Husain, N. (2012). The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive

information and improve transparency. *Nucleic acids research*, *41*(D1), D925-D935.

45. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014). OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, *43*(D1), D789-D798.

46. Fang, H., & Gough, J. (2012). DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic acids research*, *41*(D1), D536-D544.

47. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, *42*(D1), D980-D985.

48. Araki, H., Knapp, C., & Tsai, P. (2012). GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, *2*(1), 76-82.

49. Shen, L., Diroma, M. A., Gonzalez, M., Navarro-Gomez, D., Leipzig, J., Lott, M. T., ... & Chinnery, P. F. (2016). MSeqDR: a centralized knowledge repository and bioinformatics web resource to facilitate genomic investigations in mitochondrial disease. *Human mutation*, *37*(6), 540-548.

50. Gazzo, A. M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., ... & Lenaerts, T. (2015). DIDA: A curated and annotated digenic diseases database. *Nucleic acids research*, *44*(D1), D900-D907.

51. Morgan,T.,Schmidt,J.,Haakonsen,C.,Lewis,J.,DellaRocca,M., Morrison,S., Biesecker,B. and Kaphingst,K.A. (2014) Using the internet to seek information about genetic and rare diseases: a case study comparing data from 2006 and 2011. JMIR Res. Protoc., *3*, e10.

52. Glueck, M., Gvozdik, A., Chevalier, F., Khan, A., Brudno, M., & Wigdor, D. (2017). PhenoStacks: cross-sectional cohort phenotype comparison visualizations. *IEEE transactions on visualization and computer graphics*, *23*(1), 191-200.

53. Glueck, M., Hamilton, P., Chevalier, F., Breslav, S., Khan, A., Wigdor, D., & Brudno, M. (2016). PhenoBlocks: phenotype comparison visualizations. *IEEE transactions on visualization and computer graphics*, *22*(1), 101-110.

54. Chatzimichali, E. A., Brent, S., Hutton, B., Perrett, D., Wright, C. F., Bevan, A. P., ... & Swaminathan, G. J. (2015). Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. *Human mutation*, *36*(10), 941-949.

55. McMurry, J. A., Köhler, S., Washington, N. L., Balhoff, J. P., Borromeo, C., Brush, M., ... & Foster, E. (2016). Navigating the phenotype frontier: the monarch initiative. *Genetics*, *203*(4), 1491-1495.

56. 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. Phenoday@ISMB2014, http://phenoday2014.bio-lark.org/. 61.

57. Taboada, M., Rodriguez, H., Martinez, D., Pardo, M., and Sobrido, M.J. (2014). Automated semantic annotation of rare disease cases: a case study. Database *2014*, bau045.

58. Howe, D., and Yon, S. (2008). The future of biocuration. Nature *455*, 47–50.

59. Ramoni, R.B., Mulvihill, J.J., Adams, D.R., Allard, P., Ashley, E.A., Bernstein, J.A., Gahl, W.A., Hamid, R., Loscalzo, J., Mccray, A.T., et al. (2017). The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease.

Am. J. Hum. Genet. *100*, 185–192.

60. Ars, E., and Torra, R. (2017). Rare diseases, rare presentations: recognizing atypical inherited kidney disease phenotypes in the age of genomics. Clin. Kidney J. *10*, 586–593.

61. Carey, J.C. (2010). The Importance of Case Reports in Advancing Scientific Knowledge of Rare Diseases. Adv Exp Med Biol *686*, 77–86.

62. Sudhakaran, S., and Surani, S. (2014). "The Role of Case Reports in Clinical and Scientific Literature." Austin J Clin Case Rep *1*, 1–2.

63. Sayers, Eric. (2008). E-utilities quick start. Entrez Programming Utilities Help [Internet], https://www.ncbi.nlm.nih.gov/books/NBK25500/.

64. Gagnier, J.J., Kienle, G., Altman, D.G., Moher, D., Sox, H., and Riley, D. (2013). The CARE guidelines: Consensus-based clinical case reporting guideline development. Headache *53*, 1541–1547.

65. Tanenblatt, M., Coden, A., and Sominsky, I. (2010). The ConceptMapper Approach to Named Entity Recognition. LREC 546–551.

66. Funk, C., Jr, W.B., Garcia, B., Roeder, C., Bada, M., Cohen, K.B., Hunter, L.E., and Verspoor, K. (2014). Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics *15*, 59.

67. Aronson, A.R., and Lang, F. (2010). An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Informatics Assoc. *17*, 229–236.

68. Yamamoto, Y., Yamaguchi, A., Bono, H., and Takagi, T. (2011). Allie: A database and a search service of abbreviations and long forms. Database *2011*, 1–8.

69. Garten, Y., Coulet, A., and Altman, R.B. (2010). Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics

11, 1467–1489.

70. Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Roceedings 14th Int. Jt. Conf. Artif. Intell. - Vol. 1 - IJCAI'95 *1*, 6.

71. Lin, D. (1998). An Information-Theoretic Definition of Similarity. Proc. ICML 296–304.

72. Jiang, J.J., and Conrath, D.W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proc. Int. Conf. Res. Comput. Linguist. (ROCLING X).

73. Pesquita, C., Faria, D., Bastos, H., Falcão, A.O., and Couto, F.M. (2007). Evaluating GO-based semantic similarity measures. InProceedings of 10thAnnualBio-OntologiesMeeting *37*, 38.

74. Gottlieb, M.M., Arenillas, D.J., Maithripala, S., Maurer, Z.D., TarailoGraovac, M., Armstrong, L., Patel, M., van Karnebeek, C., and Wasserman, W.W. (2015). GeneYenta: A PhenotypeBased Rare Disease Case Matching Tool Based on Online Dating Algorithms for the Acceleration of Exome Interpretation. Hum. Mutat. *36*, 432–438.

75. Deng, Y., Gao, L., Wang, B., and Guo, X. (2015). HPOSim: An r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS One *10*, e0115692.

76. Hoehndorf, R., Gruenberger, M., Gkoutos, G. V, and Schofield, P.N. (2015). Similarity-based search of model organism, disease and drug effect phenotypes. J. Biomed. Semantics *6*, 6.

77. Buske, O.J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W.P., et al. (2015). PhenomeCentral:

A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. Hum. Mutat. *36*, 931–940.

78. Adachi, T., Kawamura, K., Furusawa, Y., Nishizaki, Y., Imanishi, N., Umehara, S., Izumi, K., and Suematsu, M. (2017). Japan's initiative on rare and undiagnosed diseases (IRUD): Towards an end to the diagnostic odyssey. Eur. J. Hum. Genet. *25*, 1025–1028.

79. Buske, O.J., Schiettecatte, F., Hutton, B., Dumitriu, S., Misyura, A., Huang, L., Hartley, T., Girdea, M., Sobreira, N., Mungall, C., et al. (2015). The Matchmaker Exchange API: Automating Patient Matching Through the Exchange of Structured Phenotypic and Genotypic Profiles. Hum. Mutat. *36*, 922–927.

80. Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., Schriml, L.M., Kibbe, W.A., Schofield, P.N., Beck, T., et al. (2015). The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am. J. Hum. Genet. *97*, 111–124.

81. Groza, T., Kohler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F.M., Baynam, G., Zankl, A., and Robinson, P.N. (2015). Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. Database *2015*, bav005-bav005.

82. Gkotsis, G., Velupillai, S., Oellrich, A., Dean, H., Liakata, M., and Dutta, R. (2016). Don't Let Notes Be Misunderstood : A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records. Proc. Third Work. Comput. Lingusitics Clin. Psychol. 95–105.

83. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature,

536(7616), 285.

84. Europe PMC Consortium. (2014). Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic acids research, 43(D1), D1042-D1048.

85. 1: Kudo S. [Public accessibility to the journal of Internal Medicine through J-STAGE]. Nihon Naika Gakkai Zasshi. 2006 Mar 10;95(3):564-9.

86. Chen, H. H., Lan, J. L., Shu, S. G., Chen, D. Y., & Lan, H. H. (2004). A mucolipidosis III patient presenting characteristic sonographic and magnetic resonance imaging findings of claw hand deformity. Journal of the Formosan Medical Association= Taiwan yi zhi, 103(9), 715-720.

87. Ziai, M. N., Benson, A. G., & Djalilian, H. R. (2005). Congenital lip pits and van der Woude syndrome. Journal of Craniofacial Surgery, 16(5), 930-932.

88. Li, A., Asch, M., Joseph Lasky III, L. S., & Lee, S. L. (2012). The surgical management of a stage III Wilms tumor presenting with perforated appendicitis. Journal of pediatric hematology/oncology, 34(5).

89. Kariminejad, A., Schöls, L., Schüle, R., Tonekaboni, S. H., Abolhassani, A., Fadaee, M., ... & Gleeson, J. G. (2016). CYP2U1 mutations in two Iranian patients with activity induced dystonia, motor regression and spastic paraplegia. european journal of paediatric neurology, 20(5), 782-787.

90. Bradley, J. D. (1986). Jaccoud's arthropathy in adult dermatomyositis. Clinical and experimental rheumatology, 4(3), 273-276.

91. Huther, M., Gronier, C., & Lipsker, D. (2015, October). Infantile segmental hemangioma without facial involvement: A cutaneous marker of vascular malformations such as in PHACE syndrome?. In Annales de dermatologie et de

venereologie (Vol. 142, No. 10, pp. 563-566).

92. Beck, T. M., Day, J. C., Smith, C. E., & Eddy, H. E. (1984). Granulocytic sarcoma treated as an acute leukemia: report of a case. Cancer, 53(8), 1764-1766.

93. Abe, K., Fujimura, H., Toyooka, K., Sakoda, S., Yorifuji, S., & Yanagihara, T. (1997). Cognitive function in amyotrophic lateral sclerosis. Journal of the neurological sciences, 148(1), 95-100.

94. Jimenez, C., Yang, Y., Kim, H. W., Al-Sagier, F., Berry, D. A., El-Naggar, A. K., ... & Gagel, R. F. (2005). Primary hyperparathyroidism and osteosarcoma: examination of a large cohort identifies three cases of fibroblastic osteosarcoma. Journal of Bone and Mineral Research, 20(9), 1562-1568.

95. Sakai, K., Hirasawa, Y., & Hashimoto, A. (2002). A case of toxocariasis with eosinophil-rich pleural effusion. Nihon Kokyuki Gakkai zasshi= the journal of the Japanese Respiratory Society, 40(6), 494-498.

Supplemental Data

Table S1: 135 clinical cases from PhenomeCentral

No	MIM ID	ORDO ID	Disease Name	# of annotated HPO terms
1	MIM:212840	ORDO:1173	Cerebellar ataxia-hypogonadism syndrome	2
2	MIM:212840	ORDO:1173	Cerebellar ataxia-hypogonadism syndrome	2
3	MIM:301835	ORDO:1187	Lethal ataxia with deafness and optic atrophy	30
4	MIM:301835	ORDO:1187	Lethal ataxia with deafness and optic atrophy	38
5	MIM:165199	ORDO:1215	Autosomal dominant optic atrophy plus syndrome	9
6	MIM:117650	ORDO:1393	Cerebro-costo-mandibular syndrome	6
7	MIM:117650	ORDO:1393	Cerebro-costo-mandibular syndrome	2
8	MIM:117650	ORDO:1393	Cerebro-costo-mandibular syndrome	7
9	MIM:117650	ORDO:1393	Cerebro-costo-mandibular syndrome	6
10	MIM:117650	ORDO:1393	Cerebro-costo-mandibular syndrome	4
11	MIM:117650	ORDO:1393	Cerebro-costo-mandibular syndrome	9
12	MIM:135900	ORDO:1465	Coffin-Siris syndrome	8
13	MIM:135900	ORDO:1465	Coffin-Siris syndrome	7
14	MIM:612794	ORDO:1478	Interatrial communication	1
15	MIM:612794	ORDO:1478	Interatrial communication	1
16	MIM:204500	ORDO:168491	Late infantile neuronal ceroid lipofuscinosis	4
17	MIM:608156	ORDO:178303	8q22.1 microdeletion syndrome	9
18	MIM:608156	ORDO:178303	8q22.1 microdeletion syndrome	13
19	MIM:608156	ORDO:178303	8q22.1 microdeletion syndrome	17
20	MIM:608156	ORDO:178303	8q22.1 microdeletion syndrome	14
21	MIM:608156	ORDO:178303	8q22.1 microdeletion syndrome	7
22	MIM:133540	ORDO:191	Cockayne syndrome	15
23	MIM:133540	ORDO:191	Cockayne syndrome	6
24	MIM:136140	ORDO:2044	Floating-Harbor syndrome	16
25	MIM:136140	ORDO:2044	Floating-Harbor syndrome	17
26	MIM:136140	ORDO:2044	Floating-Harbor syndrome	19

27	MIM:136140	ORDO:2044	Floating-Harbor syndrome	23
28	MIM:136140	ORDO:2044	Floating-Harbor syndrome	15
29	MIM:136140	ORDO:2044	Floating-Harbor syndrome	17
30	MIM:136140	ORDO:2044	Floating-Harbor syndrome	16
31	MIM:136140	ORDO:2044	Floating-Harbor syndrome	19
32	MIM:136140	ORDO:2044	Floating-Harbor syndrome	21
33	MIM:136140	ORDO:2044	Floating-Harbor syndrome	16
34	MIM:136140	ORDO:2044	Floating-Harbor syndrome	17
35	MIM:136140	ORDO:2044	Floating-Harbor syndrome	22
36	MIM:136140	ORDO:2044	Floating-Harbor syndrome	20
37	MIM:219000	ORDO:2052	Fraser syndrome	5
38	MIM:206200	ORDO:209981	IRIDA syndrome	6
39	MIM:206200	ORDO:209981	IRIDA syndrome	5
40	MIM:234100	ORDO:2108	Hallermann-Streiff syndrome	19
41	MIM:236100	ORDO:2162	Holoprosencephaly	10
42	MIM:236100	ORDO:2162	Holoprosencephaly	5
43	MIM:243150	ORDO:2300	Multiple intestinal atresia	3
44	MIM:243150	ORDO:2300	Multiple intestinal atresia	2
45	MIM:243150	ORDO:2300	Multiple intestinal atresia	2
46	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	5
47	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	3
48	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	3
49	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	1
50	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	3
51	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	3
52	MIM:612285	ORDO:2318	Joubert syndrome with oculorenal defect	4
53	MIM:147920	ORDO:2322	Kabuki syndrome	3
54	MIM:223370	ORDO:235	Dubowitz syndrome	12
55	MIM:223370	ORDO:235	Dubowitz syndrome	11
56	MIM:223370	ORDO:235	Dubowitz syndrome	7
57	MIM:612965	ORDO:242	46,XY complete gonadal dysgenesis	1
58	MIM:612965	ORDO:242	46,XY complete gonadal dysgenesis	1
59	MIM:613807	ORDO:244	Primary ciliary dyskinesia	5
60	MIM:154400	ORDO:245	Nager syndrome	8
61	MIM:154400	ORDO:245	Nager syndrome	4
62	MIM:154400	ORDO:245	Nager syndrome	5
63	MIM:154400	ORDO:245	Nager syndrome	2
64	MIM:154400	ORDO:245	Nager syndrome	2
65	MIM:154400	ORDO:245	Nager syndrome	12
66	MIM:154400	ORDO:245	Nager syndrome	8
67	MIM:201200	ORDO:2500	Acrogeria	18
68	MIM:600118	ORDO:2510	Micro syndrome	6
69	MIM:159950	ORDO:2590	Spinal muscular atrophy-progressive myoclonic epilepsy syndrome	4

70	MIM:252010	ORDO:2609	Isolated complex I deficiency	13
71	MIM:276820	ORDO:2879	Phocomelia, Schinzel type	10
72	MIM:276820	ORDO:2879	Phocomelia, Schinzel type	8
73	MIM:173800	ORDO:2911	Poland syndrome	3
74	MIM:173800	ORDO:2911	Poland syndrome	5
75	MIM:135100	ORDO:337	Fibrodysplasia ossificans progressiva	1
76	MIM:232400	ORDO:366	Glycogen storage disease due to glycogen debranching enzyme deficiency	3
77	MIM:232400	ORDO:366	Glycogen storage disease due to glycogen debranching enzyme deficiency	4
78	MIM:615630	ORDO:474	Jeune syndrome	27
79	MIM:614615	ORDO:475	Joubert syndrome	3
80	MIM:614615	ORDO:475	Joubert syndrome	4
81	MIM:614615	ORDO:475	Joubert syndrome	3
82	MIM:614615	ORDO:475	Joubert syndrome	4
83	MIM:614615	ORDO:475	Joubert syndrome	4
84	MIM:614615	ORDO:475	Joubert syndrome	3
85	MIM:614615	ORDO:475	Joubert syndrome	5
86	MIM:614615	ORDO:475	Joubert syndrome	4
87	MIM:614615	ORDO:475	Joubert syndrome	3
88	MIM:614970	ORDO:475	Joubert syndrome	7
89	MIM:604168	ORDO:48431	Congenital cataracts-facial dysmorphism-neuropathy syndrome	21
90	MIM:615846	ORDO:51	Aicardi-Gouti $\bar{\tau}$ δ ι res syndrome	4
91	MIM:157900	ORDO:570	Moebius syndrome	32
92	MIM:309900	ORDO:580	Mucopolysaccharidosis type 2	5
93	MIM:309900	ORDO:580	Mucopolysaccharidosis type 2	8
94	MIM:602771	ORDO:598	Multiminicore myopathy	6
95	MIM:158810	ORDO:610	Bethlem myopathy	3
96	MIM:158810	ORDO:610	Bethlem myopathy	10
97	MIM:608553	ORDO:65	Leber congenital amaurosis	4
98	MIM:602440	ORDO:65684	Monomelic amyotrophy	3
99	MIM:609734	ORDO:71526	Obesity due to pro-opiomelanocortin deficiency	7
100	MIM:609734	ORDO:71526	Obesity due to pro-opiomelanocortin deficiency	7
101	MIM:105830	ORDO:72	Angelman syndrome	12
102	MIM:190350	ORDO:77258	Trichorhinophalangeal syndrome type 1 and 3	20
103	MIM:312750	ORDO:778	Rett syndrome	9
104	MIM:180849	ORDO:783	Rubinstein-Taybi syndrome	23
105	MIM:613794	ORDO:791	Retinitis pigmentosa	1
106	MIM:613794	ORDO:791	Retinitis pigmentosa	1
107	MIM:610536	ORDO:79113	Mandibulofacial dysostosis-microcephaly syndrome	14
108	MIM:610536	ORDO:79113	Mandibulofacial dysostosis-microcephaly syndrome	13
109	MIM:610536	ORDO:79113	Mandibulofacial dysostosis-microcephaly syndrome	19

110	MIM:610536	ORDO:79113	Mandibulofacial dysostosis-microcephaly syndrome	12
111	MIM:610536	ORDO:79113	Mandibulofacial dysostosis-microcephaly syndrome	11
112	MIM:610536	ORDO:79113	Mandibulofacial dysostosis-microcephaly syndrome	6
113	MIM:606574	ORDO:79435	Oculocutaneous albinism type 4	25
114	MIM:248200	ORDO:827	Stargardt disease	11
115	MIM:615938	ORDO:83473	Megalencephaly-polymicrogyria-postaxial polydactyly-hydrocephalus syndrome	7
116	MIM:615938	ORDO:83473	Megalencephaly-polymicrogyria-postaxial polydactyly-hydrocephalus syndrome	6
117	MIM:614381	ORDO:88637	Hypomyelination-hypogonadotropic hypogonadism-hypodontia syndrome	2
118	MIM:614381	ORDO:88637	Hypomyelination-hypogonadotropic hypogonadism-hypodontia syndrome	1
119	MIM:610743	ORDO:88644	Autosomal recessive ataxia, Beauce type	4
120	MIM:313400	ORDO:93284	Spondyloepiphyseal dysplasia tarda	3
121	MIM:250220	ORDO:93317	Spondylometaphyseal dysplasia, Sedaghatian type	18
122	MIM:250220	ORDO:93317	Spondylometaphyseal dysplasia, Sedaghatian type	15
123	MIM:271510	ORDO:93357	SPONASTRIME dysplasia	7
124	MIM:102500	ORDO:955	Acroosteolysis dominant type	7
125	MIM:102500	ORDO:955	Acroosteolysis dominant type	6
126	MIM:102500	ORDO:955	Acroosteolysis dominant type	3
127	MIM:102500	ORDO:955	Acroosteolysis dominant type	11
128	MIM:102500	ORDO:955	Acroosteolysis dominant type	8
129	MIM:102500	ORDO:955	Acroosteolysis dominant type	10
130	MIM:102500	ORDO:955	Acroosteolysis dominant type	6
131	MIM:102500	ORDO:955	Acroosteolysis dominant type	4
132	MIM:102500	ORDO:955	Acroosteolysis dominant type	6
133	MIM:614707	ORDO:97229	Riboflavin transporter deficiency	13
134	MIM:614707	ORDO:97229	Riboflavin transporter deficiency	8
135	MIM:300068	ORDO:99429	Complete androgen insensitivity syndrome	1

Table S2: 59 clinical cases from PhenomeCentral

PhenomeCentral ID	MIM ID	ORDO ID	Disease Name	# of annotated HPO terms
1	MIM:604187	ORDO:100991	Autosomal dominant spastic paraplegia type 10	3
2	MIM:604187	ORDO:100991	Autosomal dominant spastic paraplegia type 10	3
3	MIM:610246	ORDO:101109	Spinocerebellar ataxia type 28	3
4	MIM:610246	ORDO:101109	Spinocerebellar ataxia type 28	8
5	MIM:601338	ORDO:1171	Cerebellar ataxia-areflexia-pes cavus-optic atrophy-sensorineural hearing loss syndrome	11
6	MIM:601338	ORDO:1171	Cerebellar ataxia-areflexia-pes cavus-optic atrophy-sensorineural hearing loss syndrome	12
7	MIM:245570	ORDO:163721	Rolandic epilepsy-speech dyspraxia syndrome	10
8	MIM:245570	ORDO:163721	Rolandic epilepsy-speech dyspraxia syndrome	2
9	MIM:611523	ORDO:166073	Pontocerebellar hypoplasia type 6	12
10	MIM:611523	ORDO:166073	Pontocerebellar hypoplasia type 6	11
11	MIM:610125	ORDO:178364	Syndromic microphthalmia type 5	4
12	MIM:610125	ORDO:178364	Syndromic microphthalmia type 5	3
13	MIM:612541	ORDO:178503	Dursun syndrome	21
14	MIM:235510	ORDO:2136	Hennekam syndrome	9
15	MIM:182212	ORDO:2462	Shprintzen-Goldberg syndrome	10
16	MIM:130720	ORDO:2789	Lateral meningocele syndrome	31
17	MIM:260600	ORDO:280293	Pelizaeus-Merzbacher-like disease due to AIMP1 mutation	3
18	MIM:260600	ORDO:280293	Pelizaeus-Merzbacher-like disease due to AIMP1 mutation	2
19	MIM:233400	ORDO:2855	Perrault syndrome	14
20	MIM:233400	ORDO:2855	Perrault syndrome	7
21	MIM:233400	ORDO:2855	Perrault syndrome	6
22	MIM:233400	ORDO:2855	Perrault syndrome	6
23	MIM:614261	ORDO:294016	Microcephaly-capillary malformation syndrome	14
24	MIM:614261	ORDO:294016	Microcephaly-capillary malformation syndrome	11
25	MIM:614261	ORDO:294016	Microcephaly-capillary malformation syndrome	11
26	MIM:614261	ORDO:294016	Microcephaly-capillary malformation syndrome	12

27	MIM:614261	ORDO:294016	Microcephaly-capillary malformation syndrome	6
28	MIM:180700	ORDO:3107	Autosomal dominant Robinow syndrome	5
29	MIM:180700	ORDO:3107	Autosomal dominant Robinow syndrome	15
30	MIM:312830	ORDO:3134	SCARF syndrome	21
31	MIM:312830	ORDO:3134	SCARF syndrome	10
32	MIM:604213	ORDO:314597	Chudley-McCullough syndrome	5
33	MIM:604213	ORDO:314597	Chudley-McCullough syndrome	5
34	MIM:604213	ORDO:314597	Chudley-McCullough syndrome	3
35	MIM:604213	ORDO:314597	Chudley-McCullough syndrome	5
36	MIM:604213	ORDO:314597	Chudley-McCullough syndrome	3
37	MIM:604213	ORDO:314597	Chudley-McCullough syndrome	4
38	MIM:269880	ORDO:3163	SHORT syndrome	8
39	MIM:269880	ORDO:3163	SHORT syndrome	9
40	MIM:605130	ORDO:319182	Wiedemann-Steiner syndrome	14
41	MIM:275400	ORDO:3363	Trichomegaly-retina pigmentary degeneration-dwarfism syndrome	8
42	MIM:613477	ORDO:3451	West syndrome	6
43	MIM:614736	ORDO:361	Familial glucocorticoid deficiency	8
44	MIM:614736	ORDO:361	Familial glucocorticoid deficiency	3
45	MIM:615290	ORDO:363454	Autosomal dominant childhood-onset proximal spinal muscular atrophy with contractures	8
46	MIM:615630	ORDO:474	Jeune syndrome	27
47	MIM:608931	ORDO:590	Congenital myasthenic syndrome	8
48	MIM:602771	ORDO:598	Multiminicore myopathy	6
49	MIM:606002	ORDO:64753	Spinocerebellar ataxia with axonal neuropathy type 2	5
50	MIM:601675	ORDO:670	PIBIDS syndrome	7
51	MIM:609015	ORDO:746	Mitochondrial trifunctional protein deficiency	16
52	MIM:212065	ORDO:79318	PMM2-CDG	4
53	MIM:300539	ORDO:93606	Nephrogenic syndrome of inappropriate antidiuresis	4
54	MIM:300539	ORDO:93606	Nephrogenic syndrome of inappropriate antidiuresis	11
55	MIM:108500	ORDO:97	Familial paroxysmal ataxia	12
56	MIM:108500	ORDO:97	Familial paroxysmal ataxia	13
57	MIM:270550	ORDO:98	Autosomal recessive spastic ataxia of Charlevoix-Saguenay	5

58	MIM:270550	ORDO:98	Autosomal recessive spastic ataxia of Charlevoix-Saguenay	4
59	MIM:128230	ORDO:98808	Autosomal dominant dopa-responsive dystonia	3