PubCaseFinder: A Case-Report-Based, Phenotype-Driven Differential-Diagnosis System for Rare Diseases

Toyofumi Fujiwara,[1,2,*] Yasunori Yamamoto,[1] Jin-Dong Kim,[1] Orion Buske,[3] and Toshihisa Takagi[4]

Recently, to speed up the differential-diagnosis process based on symptoms and signs observed from an affected individual in the diagnosis of rare diseases, researchers have developed and implemented phenotype-driven differential-diagnosis systems. 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, we 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 a text-mining approach using one million case reports obtained from PubMed could increase the coverage of manually curated DPAs in Orphanet by 125.6%. We also present PubCaseFinder (see Web Resources), 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. Moreover, PubCaseFinder helps clinicians search for relevant case reports by using phenotype-based comparisons and confirm the results with detailed contextual information

Introduction

At present more than 6,000 rare diseases have been identified, and ~80% of them are genetic in origin.[1] Unfortunately, up to 50% of individuals affected by rare diseases never receive a diagnosis,[2] and such affected individuals will most likely lose opportunities such as optimization of clinical management and early intervention.[3] To tackle this situation, researchers are undertaking next-generation sequencing (NGS)-based analysis to identify candidate diseases for undiagnosed individuals.[4,5] After analysis, clinicians rank candidate diseases through a differential-diagnosis process based on symptoms and signs, collectively called "phenotypes," in the affected individual.[6]

Even though the analysis and process improve diagnostic rates,[7,8] the differential-diagnosis process is time-consuming.[6] At first, clinicians collect reported phenotypes from trusted medical sources (e.g., Orphanet and papers) for each candidate disease and then check which disease phenotypes overlap with the affected individual's phenotypes.[9] Recently, to speed up the process, phenotype-driven differential-diagnosis systems such as Phenomizer,[6] Phenolyzer,[10] and FACE2GENE[11] have been implemented.[12] Phenomizer and Phenolyzer employ a semantic similarity-computation method to compare the affected individual's phenotypes against a set of rare diseases associated with phenotypes and against a set of genes associated with phenotypes, respectively. FACE2GENE detects an affected individual's phenotypes from a face image and calculates a similarity score for each genetic disease via a deep-learning method. These systems provide a ranked list of diseases or genes on the basis of the similarity score, and the top-listed diseases represent the most likely differential diagnosis.

The performance of these systems is greatly influenced by the quantity and quality of underlying databases of disease-phenotype associations (DPAs). Currently, there are two well-known DPAs sources, whose focus is each slightly different from the other: Orphanet (Vasant et al. [2014]. ISMB 2014) and the Human Phenotype Ontology (HPO) consortium.[12] Orphanet provides DPAs for the rare diseases that are defined in Orphanet Rare Disease Ontology (ORDO) (Vasant et al. [2014]. ISMB 2014), and the HPO consortium mainly provides DPAs for the genetic diseases that are defined in OMIM. Note that databases that rely on manual curation inherently show a limited coverage.[13] 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.[14] Second, there are still many unknown phenotypes in rare diseases because phenotypic spectrums for many rare diseases are still under investigation.[15] For example, Elisabet et al.[16] quantified many atypical phenotypes of inherited kidney diseases caused by various genetic, epigenetic, and environmental factors. Sawyer et al.[2] diagnosed 105 undiagnosed rare-disease-affected individuals by using whole-exome sequencing and showed that

[1]Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Kashiwa-shi, Chiba-ken 277-0871, Japan; [2]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, Chiba-ken 277-8561, Japan; [3]Centre for Computational Medicine, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada; [4]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo-to, 113-0032, Japan
*Correspondence: fujiwara@dbcls.rois.ac.jp
https://doi.org/10.1016/j.ajhg.2018.08.003.

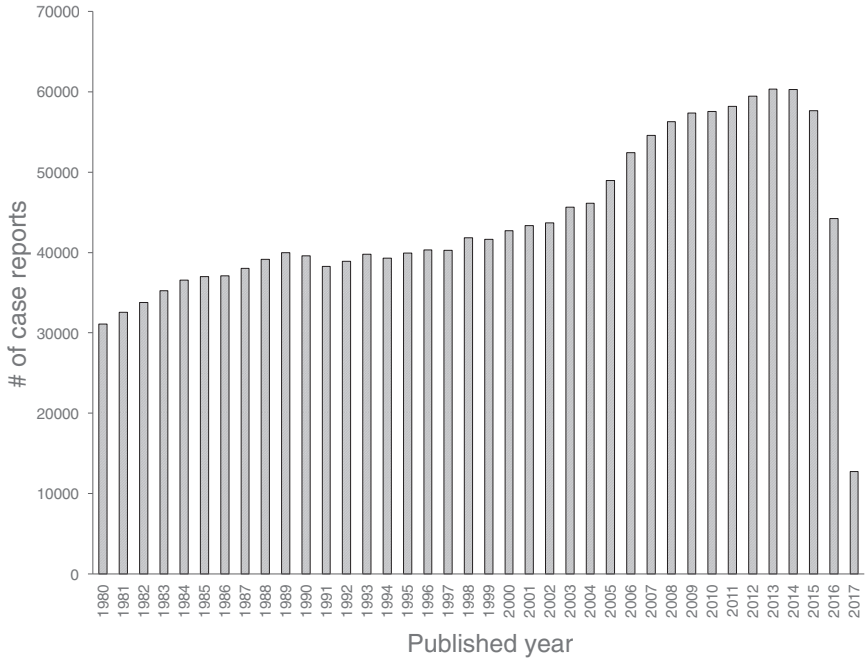

Figure 1. Distribution of the Number of Case Reports Published per Year in PubMed from 1980 to 2017

26 individuals presented with 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.[3,16]

Improving the performance of phenotype-driven differential-diagnosis systems and thus improving our ability to diagnose rare diseases will require overcoming the limited coverage of DPA databases. To address this problem, in this study we empirically explore one question on 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, we extract DPAs from case reports in PubMed by using a text-mining approach and compare those with DPAs from Orphanet. We focus on case reports because these are an important tool for quickly expanding the growing body of clinical knowledge on rare diseases,[17] and case reports often deal with previously undescribed and atypical phenotypes.[18] For example, with respect to cerebrotendinous xanthomatosis, Taboada et al.[13] automatically extracted DPAs from case reports in PubMed and obtained 11 new DPAs that did not appear in manually curated DPAs. Second, we develop a new phenotype-driven differential-diagnosis system called PubCaseFinder and demonstrate that automatically extracted DPAs without manual screening can obviously 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.

Material and Methods

Collecting Case Reports

We used PubMed E-utilities to obtain a large collection of case reports.[19] To find an effective search query, we used the publication-type tag "case reports," which was manually tagged by human experts. A previous study has shown that PubMed includes many case reports that are not explicitly tagged as such.[20] We also chose to consider a paper as a case report if its title included "case report" or "case reports." We 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]. We found that 1,895,021 PubMed entries were initially collected as case reports and that among these only 1,083,283 had both titles and abstracts (as of July 20, 2017). Figure 1 shows the growth of PubMed-indexed case reports published per year. The apparent decrease observed in recent years is attributed to the delay with PubMed indexing for many reasons, e.g., a manual tagging process. Table 1 lists the top 20 journals (out of 7,649 containing case reports) ranked according to the number of published case reports (all journals are shown in Table S1).

Identifying Disease-Phenotype Associations

We extracted DPAs from our collection of case reports by using a text-mining approach (Figure 2). At first, we annotated titles and abstracts of case reports with HPO terms and ORDO terms by using ConceptMapper (Tanenblatt et al. [2010]. LREC 546–551). HPO, initially published in 2008, has been curated by domain experts to provide a standardized vocabulary for describing phenotypic abnormalities that are widely seen in human genetic diseases.[12] ORDO, constructed by Orphanet and EBI, provides a standardized vocabulary for rare diseases extracted from papers and validated by international experts (Vasant et al. [2014]. ISMB 2014). We downloaded the HPO file (releases/2017-06-30) provided 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 (parent-child) relationships were established between HPO terms. ORDO contains a set of 13,321 terms integrated with 3,737 textual definitions, 20,542 synonyms, and 15,973 is-a relationships and provides connections with other resources (e.g., OMIM and ICD10). In this study, we 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.

For annotation, we used ConceptMapper, a dictionary-based system for recognizing concepts in text. Christopher et al.[21]

Table 1. Top 20 Journals Ranked According to the Number of Published Case Reports

Journal Title	Country of Publication	Number of Published Case Reports
BMJ Case Reports	England	9,782
The Annals of Thoracic Surgery	Netherlands	6,902
Internal Medicine (Tokyo, Japan)	Japan	5,697
Gan to Kagaku Ryoho (Cancer and Chemotherapy)	Japan	5,618
Southern Medical Journal	United States	5,144
Journal of Pediatric Surgery	United States	4,776
Clinical Nuclear Medicine	United States	4,479
Journal of Neurosurgery	United States	4,325
Chest	United States	4,295
American Journal of Medical Genetics	United States	4,290
Urology	United States	4,273
The Japanese Journal of Thoracic Surgery	Japan	4,116
Cancer	United States	4,092
The Journal of Laryngology and Otology	England	4,078
Neurosurgery	United States	4,064
The Journal of Urology	United States	3,958
Neurology	United States	3,880
Hinyokika Kiyo Acta Urologica Japonica	Japan	3,851
Journal of Medical Case Reports	England	3,845
Nederlands Tijdschrift voor Geneeskunde	Netherlands	3,845

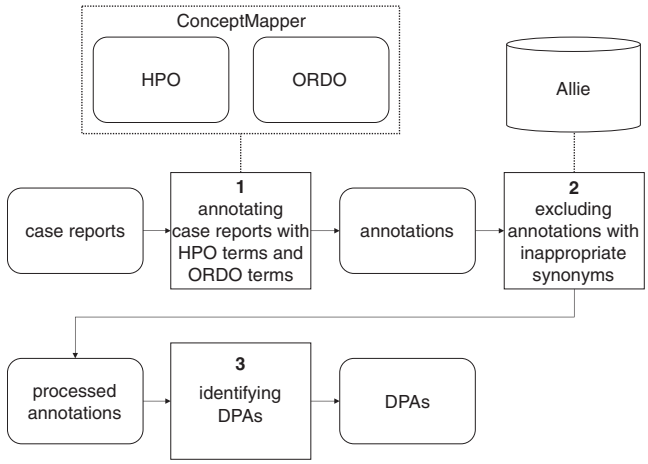

Figure 2. Process of Identifying Disease-Phenotype Associations from Case Reports
The set of titles and abstracts of case reports were annotated with HPO terms and ORDO terms using ConceptMapper with HPO and ORDO (step 1), and annotations with inappropriate synonyms were excluded using the Allie database (step 2). DPAs were identified in processed annotations (step 3).

evaluated MetaMap,[22] NCBO Annotator, and ConceptMapper on eight biomedical ontologies by using the Colorado Richly Annotated Full-Text Corpus. They examined more than 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. On the basis of our preliminary survey on the processing speed of the three systems, we confirmed that ConceptMapper was ~50 times faster than MetaMap and NCBO Annotator. In order to conduct ConceptMapper with HPO and ORDO, we used Colorado Computational Pharmacology (University of Colorado School of Medicine) natural language-processing (NLP) pipelines with default parameter sets.[21]

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.[23] 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 "thrombocytopaenia." To exclude inappropriate annotations, we used the Allie database that deposits abbreviations generated on the basis of all titles and abstracts in PubMed. Annotations including synonyms were excluded if a case report did not include both the synonym and its label in the text.

Finally, using the processed annotations, we identified DPAs in all titles and abstracts of case reports. Others have proposed various approaches, such as the co-occurrence approach, rule-based approach, and machine-learning approach, to extract relations such as protein-protein interactions and disease-gene associations from biomedical text.[24] We took the co-occurrence approach, which is simple but known to be successful for many previous works, such as AliBaba,[25] EBIMed,[26] iHOP,[27] and Pharmspresso.[28] This approach regarded two entities co-occurring in a textual unit of some defined size as having relations. Due to the intrinsic complexity of the biomedical text, most of the cases using this approach work on a sentence-based level.[25–28] Thus, we regarded co-occurrences of an ORDO term and an HPO term within a sentence as DPAs.

Development of PubCaseFinder

We developed PubCaseFinder, a new phenotype-driven differential-diagnosis system that uses the DPAs extracted from one million 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 an individual who is suspected have a rare disease. A case is represented by a set of HPO terms that describe the phenotypes of the individual. 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 shows a ranked list of candidate diseases according to the similarity score.

To calculate semantic similarity between two sets of HPO terms, various studies, such as those by Resnik,[29] Lin,[30] Jiang-Conrath,[31] simGIC,[32] and GeneYenta,[33] have recommended several different

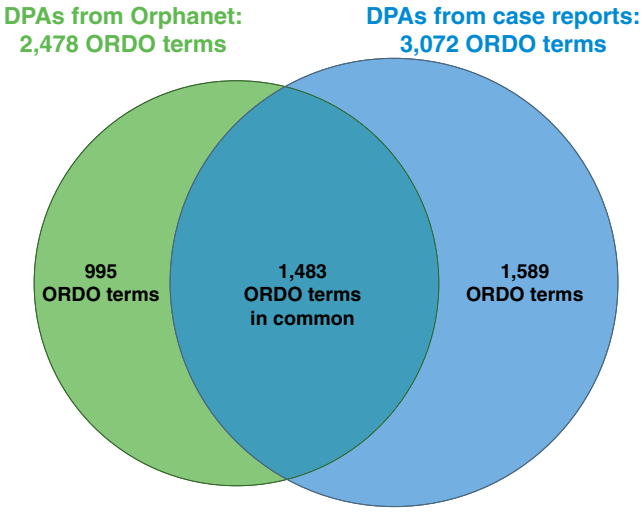

Figure 3. Overlap between Two Sets of ORDO Terms Found in Disease-Phenotype Associations from Orphanet and from Case Reports

measures. These measures are used for an affected individual's diagnosis,[6] enrichment analysis of gene sets and disease sets,[34,35] discovering causative genes of rare diseases,[33] 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.

PubCaseFinder uses GeneYenta (based on Resnik's measure), which is a user-weighted matching algorithm that sets a matching weight for each phenotype.[33] The 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 a 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 relationship 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 the intersection of A_t and $A_{t'}$. The similarity $sim_{case\_disease}$ between a case and a disease reflects 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 c represents a case, d represents a disease, and R_t allows users to assign weights ranging from 1 to 5 in accordance with how important a term t is for the user. For this evaluation, we assigned 1 to R_t for any HPO terms. T_c and T_d represent HPO terms for a case and for a disease, respectively. PubCaseFinder provides a ranked list of diseases according to $sim_{case\_disease}$, but the disease with the fewest T_d values becomes highest ranking in the case of diseases with the same $sim_{case\_disease}$.

Results

Identifying Disease-Phenotype Associations from Case Reports

We 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. From the annotated case reports, we extracted 70,011 DPAs consisting of 3,881 HPO terms and 3,072 ORDO terms. Note that we also obtained 51,590 DPAs from Orphanet, which consists of 4,832 HPO terms and 2,478 ORDO terms. Figure 3 shows the overlap between the two sets of ORDO terms included in DPAs from case reports and from Orphanet. We found that 1,483 ORDO terms were common to the two data sources and that 1,589 ORDO terms included in DPAs from case reports were not found in DPAs from Orphanet.

Within the overlapping 1,483 ORDO terms, we compared 40,512 DPAs from case reports with 35,172 DPAs from Orphanet. We regarded DPAs as the same if their related HPO terms were located in the same part of the ontology hierarchy. As a result, 5,217 DPAs were in common, and 35,295 new DPAs were added to 1,483 rare diseases included in DPAs from Orphanet. We 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 64,794 new DPAs and increase the coverage of DPAs in Orphanet by 125.6%.

An Overview of PubCaseFinder

We implemented the algorithms described above in a web application called PubCaseFinder (Figures 4A and 4B). In addition to comparing an affected individual's phenotypes with rare diseases, a user can compare an affected individual's phenotypes against published case reports that are associated with their HPO terms (Figure 4C). On the basis of the ranked lists of rare diseases and case reports, clinicians can discuss differential diagnoses for undiagnosed individuals with suspected rare diseases. Users can also narrow down the ranked list of rare diseases to specify the causative genes for rare diseases. PubCaseFinder shows the context in which a DPA appears to confirm detailed contextual information on the presence of DPAs (Figure 4D). To keep up with new DPAs that are

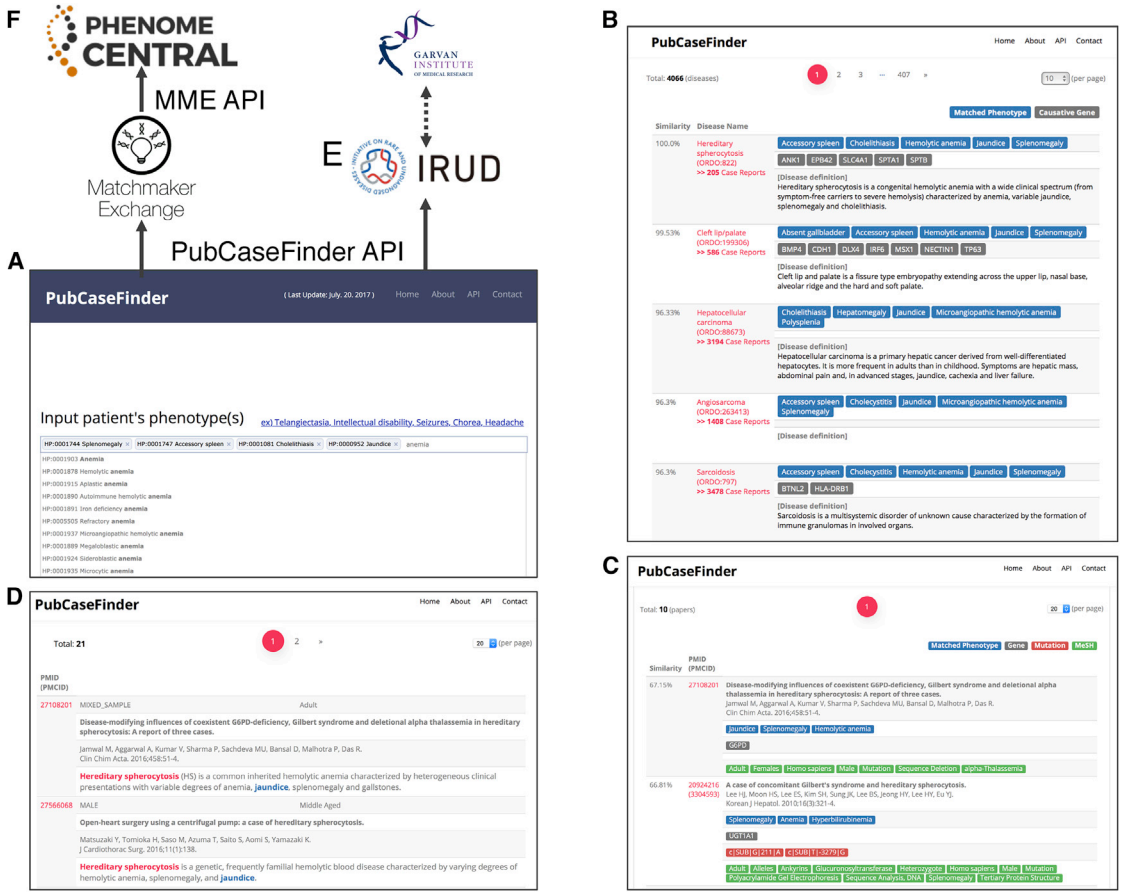

Figure 4. PubCaseFinder at a Glance and Integration of PubCaseFinder with IRUD Exchange and PhenomeCentral
Once a user types an affected individual's phenotype in the search box, PubCaseFinder displays candidate HPO terms. This enables rapid entry of HPO terms because users select appropriate HPO terms from the list (A). The affected individual is then compared with all rare diseases in Orphanet on the basis of phenotypic similarity, and the ranked list of rare diseases is shown (B). The higher the phenotypic similarity, the higher the displayed probability as a candidate disease. Users can also obtain a ranked list of published case reports in the same manner (C). The context in which a DPA appears is useful for confirming detailed contextual information on the presence of DPAs (D). This figure also shows the integration of PubCaseFinder with IRUD Exchange (a customized system of Patient Archive) (E) and PhenomeCentral (F) via the PubCaseFinder application programming interface (API). The PubCaseFinder API is also developed as the Matchmaker Exchange (MME) API.

continuously introduced in case reports, we equipped PubCaseFinder with an automatic update system.

Currently, existing data-sharing and matchmaking services for affected individuals lack methods for consulting published case reports, so we integrated PubCaseFinder with available services, namely IRUD (Initiative on Rare and Undiagnosed Diseases) Exchange,[36] which is a customized system of Patient Archive[37] (Figure 4E), and PhenomeCentral[38] (Figure 4F) in BioHackathon 2017. IRUD is actively engaged in the diagnosis of individuals 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 individuals with suspected rare genetic diseases. We developed a JSON-based REST endpoint to query PubCaseFinder by using HPO terms and Ensemble gene IDs and to return ranked lists of rare diseases and case reports based on phenotypic similarity. We also developed the Matchmaker Exchange (MME) application programming interface (API)[39] as a secondary querying option for PubCaseFinder. Using the PubCaseFinder API and the MME API, we enabled the use of PubCaseFinder in both IRUD Exchange (from Patient Archive) and PhenomeCentral.

Performance Evaluation of PubCaseFinder

To evaluate the performance of PubCaseFinder as a phenotype-driven differential-diagnosis system, we collected 1,584 PhenomeCentral clinical cases, 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. We used them as the test cases of our evaluation. We converted all MIM IDs of the cases to Orpha numbers by using connections between MIM IDs and Orpha numbers in ORDO. To evaluate the effect of DPAs form case reports, we compared the performance of PubCaseFinder in three different settings: one

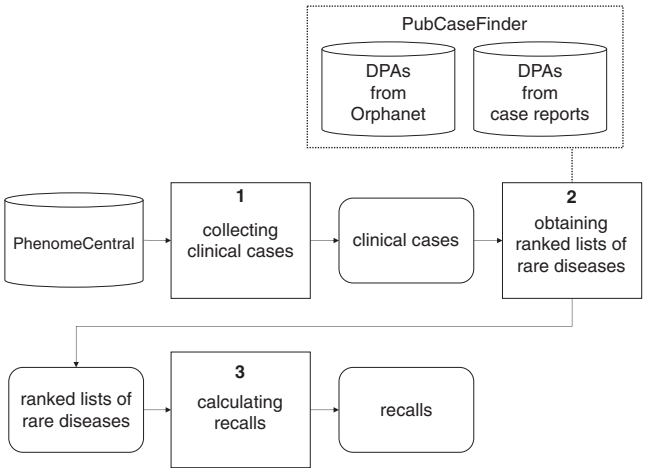

Figure 5. Performance Evaluation of PubCaseFinder
Clinical cases of rare diseases were collected from PhenomeCentral (step 1), and a ranked list of rare diseases based on phenotypic similarity was obtained with PubCaseFinder for each clinical case (step 2). The performance of PubCaseFinder was evaluated via the "recall at ranks" metric (step 3).

with DPAs from Orphanet only (PubCaseFinder-O), one with DPAs from case reports only (PubCaseFinder-CR), and one with DPAs from both (PubCaseFinder-O and -CR). For reference purposes, we included Phenomizer and Orphamizer (a customized system of Phenomizer for Orphanet) in our comparison because they were the most comparative systems among phenotype-driven differential-diagnosis systems. For the evaluation, we 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 potentially be targeted with PubCaseFinder, Phenomizer, and Orphamizer (target A), and the other consisted of 1,589 diseases that were not associated with a phenotype in Orphanet (target B).

First, we evaluated the performance of PubCaseFinder in the three different settings when targeting target A. Figure 5 shows the evaluation process. The 135 cases (all PhenomeCentral IDs are shown in Table S2) out of the 243 PhenomeCentral cases were used for this evaluation, because they had diagnoses that belonged to target A. The result of each run was obtained as a ranked list of diseases (all results are shown in Table S3). 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 6 shows the recall rates by PubCaseFinder in the three settings (the recall numbers are shown in Table 2). The top-10 recall rate of PubCaseFinder-O and -CR is 57% (Figure 6), 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 and -CR are higher than those of PubCaseFinder-O and PubCaseFinder-CR (Figure 6). The top 50 recall rate of PubCaseFinder-O is lower than the top 20 recall rate of PubCaseFinder-O and -CR, which means that even if a user checks the top 50 diseases of PubCaseFinder-O, the diagnostic rate is lower than when that user checks the top 20 diseases of PubCaseFinder-O and -CR. We also evaluated the statistical significance of the appearance of a correct diagnosis in the top 10 with a binomial test and found that the p value of PubCaseFinder-O and -CR was 4.01×10^{-144}, whereas those of PubCaseFinder-O and PubCaseFinder-CR were 2.83×10^{-108} and 4.89×10^{-33}, respectively. Those results clearly show the potential of DPAs from case reports to improve the performance of phenotype-driven differential-diagnosis systems.

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 and -CR could place it at the 5th rank.

Moreover, we compared the performance of PubCaseFinder-O and -CR to that of Phenomizer and Orphamizer by using the 135 cases with target A (all results are shown in Table S3). Figure 7 shows the recall rates of the three systems (the recall numbers are shown in Table 2). Note, however, that the comparison should be taken only for a reference purpose for at least two reasons. First, Phenomizer is based on the OMIM ID system, whereas the other two are based on the ORDO ID system, and the two ID systems are not directly comparable to each other. Second, the development stages of the three systems are all different: Phenomizer is a mature system, PubCaseFinder is entering into a production stage, and Orphamizer is at a development stage.

Second, we evaluated the performance of PubCaseFinder-CR when targeting target B. We narrowed down 243 cases of PhenomeCentral to 59 cases whose diagnoses were part of the target B. For the 59 cases (all PhenomeCentral IDs are shown in Table S2), we obtained ranked lists of target B by using PubCaseFinder-CR (all results are shown in Table S4) and then calculated recalls on the basis of the results. PubCaseFinder-CR showed the recall number (rate) as follows: 2 (3.4%) in the top 1, 3 (5.1%) in the top 5, 5 (8.5%) in the top 10, 6 (10.2%) in the top 20, 13 (22.0%) in the top 50, and 24 (40.7%) in the top 100 (Figure 8). We evaluated the statistical significance of the appearance of a correct diagnosis in the top 10 by using a binomial test and found a p value of 3.72×10^{-5}. Although Figure 8 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

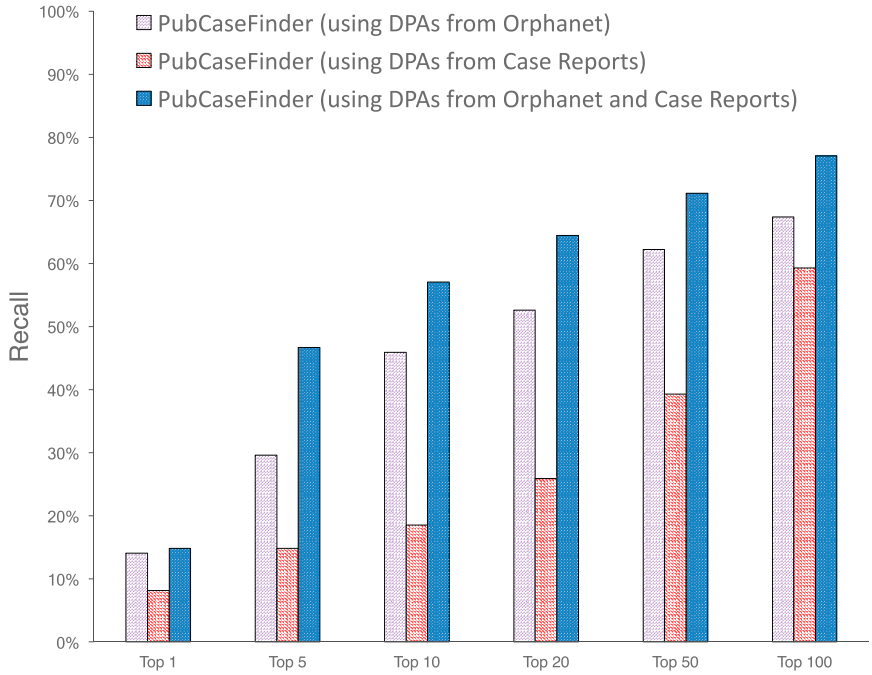

Figure 6. Performance Comparison of Three Different Settings of PubCaseFinder
Recalls were calculated on the basis of ranked lists of 2,323 rare diseases for 135 clinical cases from PhenomeCentral.

both exploited DPAs from case reports. Examining the number of associated DPAs from case reports for target A and target B revealed that, on average, each disease in target-A had 27.3 DPAs, whereas each disease in target B had 18.6 DPAs. Our interpretation is that the difference in recall rates is caused by the difference of the number of associated DPAs.

Filtering of Unreliable Disease-Phenotype Associations

In a previous study, Tudor et al.[40] also tried to extract DPAs for common diseases from papers in PubMed, and they suggested ignoring low frequency 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.[24] However, we found that most DPAs identified in this study appeared in few case reports. Figure 9 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, we calculated the top-10 recall rate of PubCaseFinder that exploits each set of DPAs filtered by their frequencies of occurrence in case reports (Figure 9; all results are shown in Table S5). The top-10 recall rates gradually decreased from 57.0% to 49.6% when the frequency of DPA occurrence increased. That is, this indicates that not filtering out DPAs by their frequencies of occurrence improves the performance of PubCaseFinder. We therefore chose not use frequencies of occurrence to filter out DPAs from case reports.

Discussion

To speed up the differential-diagnosis process on the basis of symptoms and signs observed from affected individuals, researchers have developed and implemented phenotype-driven differential-diagnosis systems for rare diseases. The performance of these systems is influenced by the quantity and quality of underlying DPA databases. We found that the limited coverage of manually curated databases was a major problem that hindered the further progress of automated differential diagnosis. To address the problem, we developed a text-mining approach to extend the coverage of DPAs in manually curated databases such as Orphanet. By applying the approach to a million case reports from PubMed, we could increase the coverage of DPAs from Orphanet more than two times. By using this extended

Table 2. Recall Numbers by Phenomizer, Orphamizer, and Three Different Settings of PubCaseFinder

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)
PubCaseFinder (with DPAs from Orphanet)	19 (14.1%)	40 (29.6%)	62 (45.9%)	71 (52.6%)	84 (62.2%)	91 (67.4%)
PubCaseFinder (with DPAs from Case Reports)	11 (8.1%)	20 (14.8%)	25 (18.5%)	35 (25.9%)	53 (39.3%)	80 (59.3%)
PubCaseFinder (with DPAs from Orphanet and Case Reports)	20 (14.8%)	63 (46.6%)	77 (57.0%)	87 (64.4%)	96 (71.1%)	104 (77.0%)
Phenomizer	22 (16.3%)	46 (34.1%)	63 (46.7%)	84 (62.2%)	99 (73.3%)	111 (82.2%)
Orphamizer	12 (8.9%)	31 (23.0%)	42 (31.1%)	49 (36.3%)	63 (46.7%)	68 (50.4%)

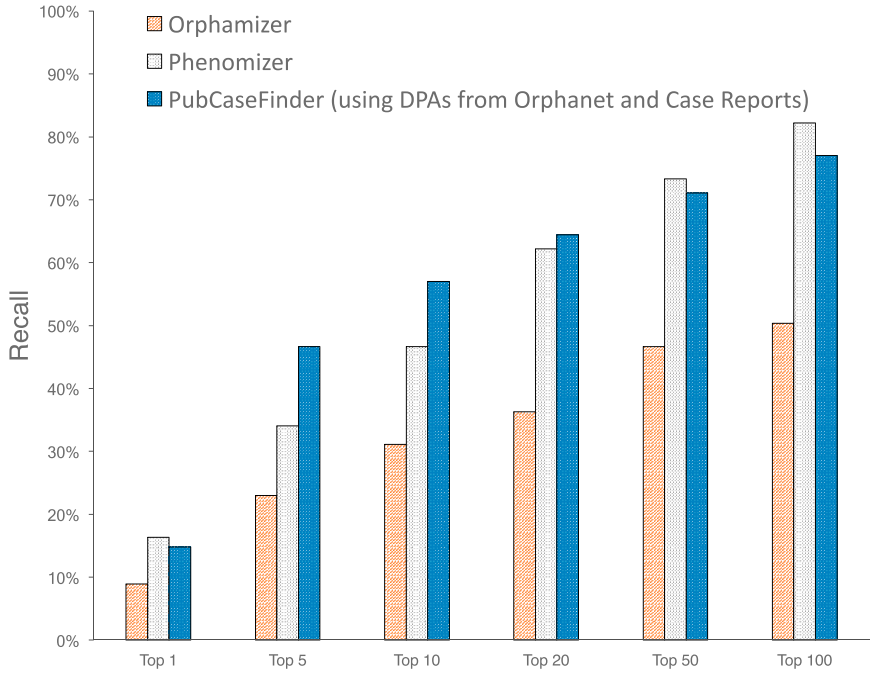

Figure 7. Performance Comparison of PubCaseFinder (using DPAs from Orphanet and Case Reports), Phenomizer, and Orphamizer
Recalls were calculated on the basis of ranked lists of 2,323 rare diseases for 135 clinical cases from PhenomeCentral.

DPA database, we also developed PubCaseFinder, a new phenotype-driven differential-diagnosis system. A series of experiments that was conducted with clinical cases from PhenomeCentral showed that the performance of phenotype-driven differential diagnosis could be substantially improved thanks to the extension of the DPA database.

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

For annotation with HPO terms and ORDO terms, we used ConceptMapper, which was reported as a state-of-the-art

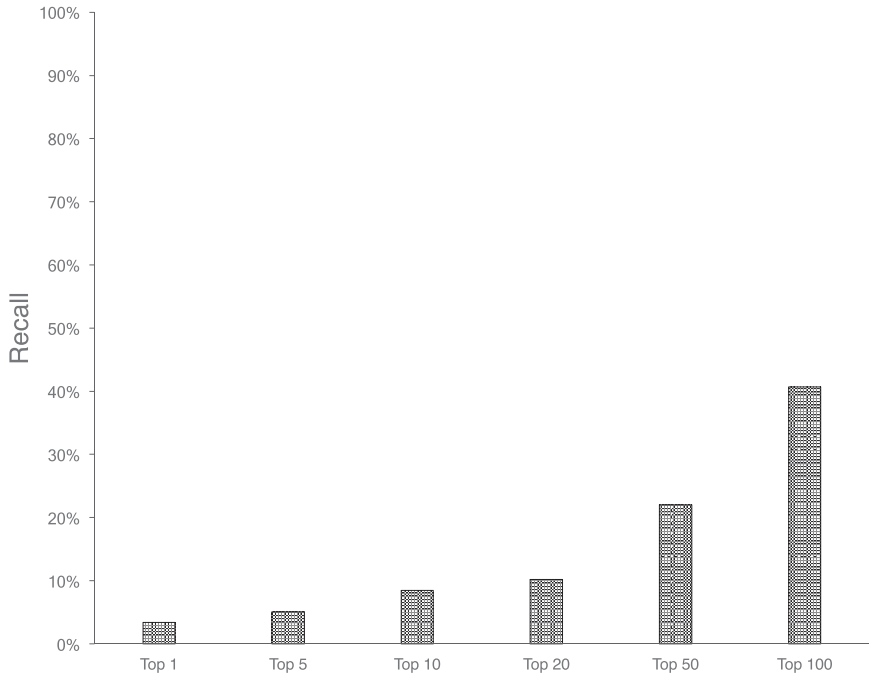

Figure 8. Recalls of PubCaseFinder for Rare Diseases Not Included in Disease-Phenotype Associations from Orphanet
Recalls were calculated on the basis of the ranked lists of 1,589 rare diseases via PubCaseFinder for 59 clinical cases from PhenomeCentral.

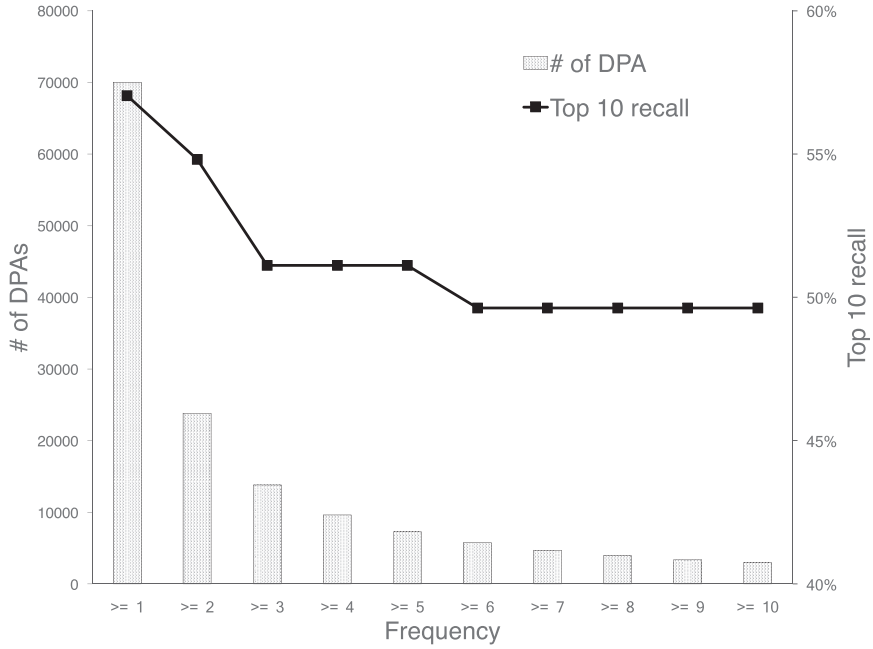

Figure 9. Distribution of Numbers of Disease-Phenotype Associations from Case Reports and Top-10 Recall Rates
For each set of DPAs ordered according to the frequency of occurrence in case reports, we counted the number of DPAs and calculated the top-10 recall rate to evaluate the performance of PubCaseFinder from the set of DPAs. Bars indicate case reports, and a solid line indicates top 10 recall rates.

concept-recognition system among publicly available ones. Recently, Bio-LarK, which was also a concept recognizer specifically tailored to annotating HPO terms, has become a publicly available system. A previous study showed that Bio-LarK was benchmarked with both the gold standard and the test suite corpora for HPO and outperformed other concept recognizers.[41] Because our approach does not rely on a specific concept recognizer, we are planning to seek a further performance improvement by finding and adopting a more optimal concept recognizer.

In clinical practice, a specific phenotype can be extremely prominent or severe; thus, we used the GeneYenta algorithm, which allows users to set a matching weight for each phenotype.[33] However, we 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. In future work, we are planning to modify the user interface of PubCaseFinder to make users set weights to HPO terms; setting weights will empower users to leverage their expertise and knowledge to customize results.

Our experiment and discussion on the 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 effective for common diseases. We attribute this to the fact that data for rare diseases are by nature much less available than those of common diseases. As a future work, we are planning to develop a much more sophisticated filtering method, such as using a negation detector.[42]

Supplemental Data

Supplemental Data include five tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.08.003.

Acknowledgments

This work was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST). We thank Prof. Kenjiro Kosaki, Dr. Shoko Kawamoto, Dr. Tudor Groza, Dr. Yuka Tateishi, and Mr. Sadahiro Kumagai for their help. We also thank the Care4Rare Canada Consortium for sharing the clinical cases.

Declaration of Interests

The authors declare no competing interests.

Received: April 12, 2018
Accepted: August 1, 2018
Published: August 30, 2018

Web Resources

BioHackathon2017, http://2017.biohackathon.org
Care4Rare Canada Consortium, http://care4rare.ca
CCP NLP pipelines, https://github.com/UCDenver-ccp/ccp-nlp-pipelines
EBI, https://www.ebi.ac.uk
FACE2GENE, https://www.face2gene.com
Human Phenotype Ontology consortium, http://human-phenotype-ontology.github.io
ICD10, http://www.who.int/classifications/icd/icdonlineversions/en/
MetaMap, https://metamap.nlm.nih.gov
MME API, https://github.com/ga4gh/mme-apis
NCBO Annotator, https://bioportal.bioontology.org/annotator
negation-detection, https://github.com/gkotsis/negation-detection
OMIM, https://www.omim.org
Orphamizer, http://compbio.charite.de/phenomizer_orphanet
Orphanet, http://www.orpha.net/consor/cgi-bin/index.php/
Orphanet Rare Disease Ontology, http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php/
Patient Archive, http://www.patientarchive.org

Phenolyzer, http://phenolyzer.wglab.org
Phenomizer, http://compbio.charite.de/phenomizer/
PubMed, https://www.ncbi.nlm.nih.gov/pubmed/
PubCaseFinder, https://pubcasefinder.dbcls.jp/
PubCaseFinder API, https://pubcasefinder.dbcls.jp/mme
PhenomeCentral, https://www.phenomecentral.org

References

1. Boat T.F. and 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.; FORGE Canada Consortium; and Care4Rare Canada Consortium (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. 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.
7. Zhu, X., Petrovski, S., Xie, P., Ruzzo, E.K., Lu, Y.F., McSweeney, K.M., 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. 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.
10. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat. Methods *12*, 841–843.
11. 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.
12. 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* (D1), D865–D876.
13. Taboada, M., Rodríguez, H., Martínez, D., Pardo, M., and Sobrido, M.J. (2014). Automated semantic annotation of rare disease cases: A case study. Database (Oxford) *2014*, bau045.
14. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., et al. (2008). Big data: The future of biocuration. Nature *455*, 47–50.
15. 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.; Undiagnosed Diseases Network (2017). The Undiagnosed Diseases Network: Accelerating discovery about health and disease. Am. J. Hum. Genet. *100*, 185–192.
16. 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.
17. Carey, J.C. (2010). The importance of case reports in advancing scientific knowledge of rare diseases. Adv. Exp. Med. Biol. *686*, 77–86.
18. Sudhakaran, S., and Surani, S. (2014). The role of case reports in clinical and scientific literature. Austin J. Clin. Case Rep. *1*, 1–2.
19. Sayers, E. (2008). E-utilities quick start. In Entrez Programming Utilities Help [Internet] (National Center for Biotechnology Information). https://www.ncbi.nlm.nih.gov/books/NBK25500/.
20. Gagnier, J.J., Kienle, G., Altman, D.G., Moher, D., Sox, H., Riley, D.; and CARE Group (2013). The CARE guidelines: consensus-based clinical case reporting guideline development. Headache *53*, 1541–1547.
21. Funk, C., Jr., Baumgartner, W., Jr., 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.
22. Aronson, A.R., and Lang, F.M. (2010). An overview of MetaMap: Historical perspective and recent advances. J. Am. Med. Inform. Assoc. *17*, 229–236.
23. Yamamoto, Y., Yamaguchi, A., Bono, H., and Takagi, T. (2011). Allie: A database and a search service of abbreviations and long forms. Database (Oxford) *2011*, bar013.
24. Garten, Y., Coulet, A., and Altman, R.B. (2010). Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics *11*, 1467–1489.
25. Hoffmann, Robert, and Valencia, Alfonso (2005). Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics *21* (*Suppl 2*), ii252–ii258.
26. Garten, Y., and Altman, R.B. (2009). Pharmspresso: A text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics *10* (*Suppl 2*), S6.
27. Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., and Leser, U. (2006). AliBaba: PubMed as a graph. Bioinformatics *22*, 2444–2445.
28. Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., and Stoehr, P. (2007). EBIMed–text crunching to gather facts for proteins from Medline. Bioinformatics *23*, e237–e244.
29. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Conference on Artificial Intelligence *1*, 6.
30. Lin, D. (1998). An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, 296–304.
31. 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).

32. Pesquita, C., Faria, D., Bastos, H., Falcão, A.O., and Couto, F.M. (2007). Evaluating GO-based semantic similarity measures. Proceedings of 10th Annual Bio-Ontologies Meeting *37*, 38.
33. Gottlieb, M.M., Arenillas, D.J., Maithripala, S., Maurer, Z.D., Tarailo Graovac, M., Armstrong, L., Patel, M., van Karnebeek, C., and Wasserman, W.W. (2015). GeneYenta: A phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation. Hum. Mutat. *36*, 432–438.
34. 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.
35. 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.
36. 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.
37. McMurry, J.A., Köhler, S., Washington, N.L., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2016). Navigating the phenotype frontier: The Monarch Initiative. Genetics *203*, 1491–1495.
38. 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.
39. Buske, O.J., Schiettecatte, F., Hutton, B., Dumitriu, S., Misyura, A., Huang, L., Hartley, T., Girdea, M., Sobreira, N., Mungall, C., and Brudno, M. (2015). The Matchmaker Exchange API: Automating patient matching through the exchange of structured phenotypic and genotypic profiles. Hum. Mutat. *36*, 922–927.
40. 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.
41. Groza, T., Köhler, 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 (Oxford) *2015*, bav005–bav005.
42. 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. Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology, 95–105.