ARTICLE

Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes

Jung Hoon Son,[1,9] Gangcai Xie,[1,2,3,9] Chi Yuan,[1] Lyudmila Ena,[1] Ziran Li,[1] Andrew Goldstein,[1] Lulin Huang,[2,3] Liwei Wang,[4] Feichen Shen,[4] Hongfang Liu,[4] Karla Mehl,[5] Emily E. Groopman,[5] Maddalena Marasa,[5] Krzysztof Kiryluk,[5] Ali G. Gharavi,[5] Wendy K. Chung,[6] George Hripcsak,[1] Carol Friedman,[1] Chunhua Weng,[1,10,*] and Kai Wang[1,2,3,7,8,10,*]

Integration of detailed phenotype information with genetic data is well established to facilitate accurate diagnosis of hereditary disorders. As a rich source of phenotype information, electronic health records (EHRs) promise to empower diagnostic variant interpretation. However, how to accurately and efficiently extract phenotypes from heterogeneous EHR narratives remains a challenge. Here, we present EHR-Phenolyzer, a high-throughput EHR framework for extracting and analyzing phenotypes. EHR-Phenolyzer extracts and normalizes Human Phenotype Ontology (HPO) concepts from EHR narratives and then prioritizes genes with causal variants on the basis of the HPO-coded phenotype manifestations. We assessed EHR-Phenolyzer on 28 pediatric individuals with confirmed diagnoses of monogenic diseases and found that the genes with causal variants were ranked among the top 100 genes selected by EHR-Phenolyzer for 16/28 individuals ($p < 2.2 \times 10^{-16}$), supporting the value of phenotype-driven gene prioritization in diagnostic sequence interpretation. To assess the generalizability, we replicated this finding on an independent EHR dataset of ten individuals with a positive diagnosis from a different institution. We then assessed the broader utility by examining two additional EHR datasets, including 31 individuals who were suspected of having a Mendelian disease and underwent different types of genetic testing and 20 individuals with positive diagnoses of specific Mendelian etiologies of chronic kidney disease from exome sequencing. Finally, through several retrospective case studies, we demonstrated how combined analyses of genotype data and deep phenotype data from EHRs can expedite genetic diagnoses. In summary, EHR-Phenolyzer leverages EHR narratives to automate phenotype-driven analysis of clinical exomes or genomes, facilitating the broader implementation of genomic medicine.

Introduction

Traditionally, the diagnostic workup of individuals with suspected monogenic disease has relied on sequential testing using a battery of genetic and biochemical studies, incurring substantial time and financial costs while the causal etiology remains elusive.[1,2] In addition, the diagnostic uncertainty, ambiguity regarding appropriate clinical management, and repeated medical evaluations during this "diagnostic odyssey" pose a weighty emotional and psychosocial burden on both affected individuals and their families.[3,4]

Since they were first reported to resolve a case with an undiagnosed genetic disease,[5] next-generation sequencing (NGS) methods, including whole-exome sequencing (WES) and whole-genome sequencing (WGS), have been quickly established as a scalable method for efficiently generating a molecular diagnosis.[6] The diagnostic yield of WES ranges from 25% to 51%[2,4,7–10] and has been shown to be cost effective when used as a first-line test.[2,4,6,11] However, the challenge of interpreting the vast amount of sequence data generated by genome-wide testing still hinders the broad clinical utilization of this technology.

The use of phenotype information has helped empower the discovery of genes with causal variants and enrich our understanding of disease pathogenesis.[12] Similarly, deep phenotyping can improve the efficiency of clinical WES analysis and increase diagnostic yield. Computational phenotype-based gene-prioritization tools, including Phevor,[13] Phen-Gen,[14] eXtasy,[15] PhenIX,[16] Exomiser,[17] Phenomizer,[18] and Phenolyzer,[19] have been demonstrated to aid NGS analysis pipelines[20–22] and improve diagnostic yields over those of undirected variant analysis alone.[23] All these tools require manual entry of an individual's clinical signs and symptoms (i.e., his or her phenotype) as input to identify a prioritized list of candidate genes. However, oftentimes only limited phenotype information about an individual is provided on a test requisition form.

All these gene-prioritization systems leverage the Human Phenotype Ontology (HPO), a powerful, robust ontology that enables computable representations of phenotype concepts with terms sourced from clinically oriented medical literature and gene-disease databases, such as Online Mendelian Inheritance in Man (OMIM).[24,25] Additional efforts such as the Monarch

[1]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA; [2]Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA; [3]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [4]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55901, USA; [5]Division of Nephrology, Department of Medicine, Columbia University, New York, NY 10032, USA; [6]Department of Pediatrics and Medicine, Columbia University, New York, NY 10032, USA; [7]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [8]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[9]These authors contributed equally to this work

[10]These authors contributed equally to this work

*Correspondence: cw2384@cumc.columbia.edu (C.W.), wangk@email.chop.edu (K.W.)

https://doi.org/10.1016/j.ajhg.2018.05.010.

Initiative[26] and PhenomeCentral[27] use high-quality crowd-sourced phenotype information to further enrich and refine computational abilities embedded in HPO.[28] Regarding this latter aspect, HPO serves as both a standardized terminology and a phenotype-genotype knowledge database.

Electronic health records (EHRs) are widely adopted and have the potential to serve as a rich, integrated source of phenotype information. Automatic extraction and recognition of phenotypes from EHR narratives can accelerate the adoption and utilization of phenotype-driven efforts to improve genomic diagnostics and gene discovery. Such automation is especially needed in the context of diagnostic sequencing, given that most clinical information is submitted as a copy of the free-text clinical evaluation note or as a short, relatively nonspecific clinical description (such as "congenital heart disease"). Moreover, the current proprietary nature of NGS informatics pipelines implemented in various laboratories impedes standardized processes for variant interpretation. This deficiency can be partially addressed via direct, systematic integration of phenotypes extracted from EHRs, therefore improving information synthesis at the time of interpretation.

As a first step toward these goals, we present EHR-Phenolyzer, an automated EHR-narrative-based phenotyping pipeline, to enable phenotype-based gene prioritization. Notably, the existing tools for phenotype-driven genetic analysis have been largely validated with either simulated phenotype data or WES data (e.g., altered VCF files).[22,29,30] In this study, our primary goal was to demonstrate the efficacy of EHR-derived deep phenotyping information in facilitating genetic diagnosis from WES data. Our secondary goal was to perform a comparative analysis of well-tested natural language processing (NLP) systems in parsing EHR narratives for phenotype extraction and normalization and to evaluate the ability of EHR-Phenolyzer to analyze real-world EHR data and prioritize candidate genes from WES of positively diagnosed individuals.

Material and Methods

In the following sections, we will introduce the four independent cohorts used in this study, the process of phenotype extraction by human experts or by NLP, the methods of performance evaluation, and information on accessing the software and data.

Individuals from Four Cohorts

This study was conducted in accordance with the Declaration of Helsinki and the national ethical guidelines and was approved by the institutional review boards of Columbia University and the Mayo Clinic. In total, we collected 28, 10, 46, and 20 affected individuals across cohorts 1, 2, 3, and 4, respectively, from two institutes. For cohorts 1 and 4, written informed consent was obtained from all individuals. For cohorts 2 (Mayo Clinic) and 3, which involved retrospective re-analysis of EHRs without further intervention, a waiver of consent in the study protocol was approved by the corresponding institutional review boards. For the several case studies with exome sequencing data, written informed consent was obtained. Our primary cohort at Columbia University consisted of 28 affected individuals, including 16 males and 12 females, with a mean age of 8.2 years (ranging from 0.2 to 53). Diagnostic testing for primary diagnoses was performed by various commercial and academic labs, including GeneDx, Ambry Genetics, Invitae, and Columbia University Medical Center's Personalized Genomic Medicine Laboratory (Table 1). The clinical records of these affected individuals were accessed and manually extracted as needed from our EHRs, Clinical Records On-Line Web Network (powered by Allscripts). We then manually de-identified extracted notes to remove all potentially identifiable information before using them in our study.

To assess the cross-site validity of our automated pipeline for extracting phenotype concepts, we applied identical methods by using the EHR-derived records from the Mayo Clinic. This was to ensure that our pipeline was not biased toward the lexical similarities attributed to the note-writing styles of individuals and the practice settings of a given institution. In addition to showing the external validity of the pipeline, we aimed to evaluate whether the resultant pipeline could be deployed as created across institutions. Thus, we evaluated ten affected individuals with confirmed WES results (each with a positive finding) from the Mayo Clinic by using the same pipeline without modifications.

To further validate the clinical utility of the pipeline and to assess the real-world use of clinical genetics notes, we analyzed an independent set of clinical notes on 46 pediatric individuals seen by a genetic counselor at hospitals affiliated with Columbia University. Unlike the primary cohort, which included only individuals with positive (diagnostic) results from exome sequencing, this set of clinical notes together with the corresponding molecular pathology reports could be informative on the real-world use of clinical phenotype information in the context of various genetic testing techniques for a typical hospital in outpatient settings. For this set of 46 individuals, only 31 underwent genetic diagnostic testing, and 11 of this smaller set obtained positive results via various genetic assays.

Additionally, to evaluate whether our computational methods can facilitate the detection of specific genetic subtypes for a broad, clinically heterogeneous category of disease, we analyzed a fourth set of 20 individuals with chronic kidney disease (CKD). These individuals were drawn from an exome sequencing study of adults with CKD of unknown cause or familial nephropathy or hypertension[31] and represented those with diagnostic exome sequencing results for various monogenic etiologies of CKD.

Extraction of Phenotype Concepts from EHRs by Human Experts

In the current study, we used two different procedures to compile phenotype concepts from EHRs by human experts. These two procedures were (1) heuristic chart review and (2) targeted review of genetics notes. In brief, trained clinical experts extracted HPO concepts from two EHR data sources; the detailed procedures are described as follows.

Initially, we requested a domain expert to heuristically extract HPO concepts with the assistance of an HPO browser by using routine review of clinical charts. The expert could access any clinical note, lab and imaging results, and reports to help perform this task and noted the document source for each identified concept during this process. A second domain expert reviewed these concepts to come up with an agreed set of concepts selected for

Table 1. List of Affected Individuals at the Primary Site and Number of Phenotype Terms Extracted by Different Methods

					No. of Phenotype Terms			
Individual	**Sex**	**Age (Years)**	**Phenotype Examples**	**Primary Genetic Finding (Gene)**	**Expert All Note**	**Expert Single Note**	**MetaMap Single Note**	**MedLEE Single Note**
1	female	10–17	seizures, developmental regression	*SNAP25* (MIM: 600322)	10	22	19	34
2	female	10–17	skeletal dysplasia, short statue	*COL10A1* (MIM: 120110)	12	16	18	26
3	female	10–17	myopia, hypocalcemia	*ARID1B* (MIM: 614556)	14	8	22	22
4	female	10–17	tremor, atrial septal defect	*SCN1A* (MIM: 182389)	4	5	9	11
5	male	10–17	epilepsy, microcephaly	*CDKL5* (MIM: 300203)	9	7	17	17
6	male	4–9	strabismus, cognitive impairment	*MYH10* (MIM: 160776)	14	16	28	23
7	male	18+	atrial cardiomyopathy, dilated cardiomyopathy	*LMNA* (MIM: 150330)	3	4	9	11
8	female	4–9	absent speech, encephalopathy	*ALG13* (MIM: 300776)	7	4	10	9
9	female	4–9	open mouth, protruding tongue	*EHMT1* (MIM: 607001)	10	15	15	17
10	male	10–17	hypertonia, global developmental delay	*SLC1A4* (MIM: 600229)	3	4	7	8
11	male	4–9	overgrowth, hearing impairment	*MAN2B1* (MIM: 609458)	7	12	15	16
12	male	18+	visual loss, short stature	*MYO7A* (MIM: 276903)	3	4	17	13
13	female	10–17	lower-limb asymmetry, difficulty running	*TCF4* (MIM: 602272)	12	15	37	36
14	male	0–3	heart murmur, poor weight gain	*ARID1B* (MIM: 614556)	7	10	25	24
15	male	0–3	myopia, mild microcephaly	*EHMT1* (MIM: 607001)	6	8	19	24
16	male	0–3	large head, noisy breathing	*PTEN* (MIM: 601728)	10	15	23	25
17	male	0–3	clubbing of toes, depressed nasal bridge	*ATRX* (MIM: 300032, 300504)	15	17	21	26
18	female	4–9	hypertension, low birth weight	*TKT* (MIM: 606781)	7	8	21	24
19	male	0–3	oral aversion, muscle hypotonia	*PLA2G4A* (MIM: 600522)	9	13	10	19
20	female	4–9	depression, abnormal facial shape	*DDX3X* (MIM: 300160)	10	12	22	14
21	female	4–9	fever, dysmorphic facies	*HNRNPH2* (MIM: 300610)	6	15	18	22
22	male	0–3	hypovolemia, abnormal T-wave	*PTPN11* (MIM: 176876)	6	13	18	19
23	female	4–9	scarring, fragile skin	*COL7A1* (MIM: 120120)	2	17	14	15
24	male	0–3	cupped ear, highly arched eyebrow	*KMT2D* (MIM: 602113)	10	12	13	15
25	female	0–3	large forehead, difficulty running	*SHH* (MIM: 600725)	9	13	35	25
26	male	4–9	muscular hypotonia, heterotaxy	*NAA15* (MIM: 608000)	10	8	11	18
27	male	0–3	premature birth, prelonged neonatal jaundice	*CDKL5* (MIM: 300203)	3	4	6	10
28	male	0–3	myopathy, hypoplasia of penis	*POMT1* (MIM: 607423)	5	12	14	20

each affected individual. These HPO terms were referred to as "heuristic chart review" terms.

During the above procedure, we noticed that the majority of the expert-extracted HPO concepts were sourced from consultation notes authored by a clinical geneticist or genetic counselor. The genetics team typically performs a clinical evaluation before genetic testing in order to select the appropriate genetic test and ensure that affected individuals are thoroughly informed about the risks, benefits, and limitations associated with genetic testing. The resulting documentation of the encounter contains rich and descriptive but unstructured phenotype information. Therefore, this note could serve as the major source of an individual's phenotype information in the context of genomic medicine. For each individual, we selected a genetic evaluation note that (1) corresponded to the aforementioned encounter available from the EHR and (2) we believed could reliably represent the individual's clinical manifestation, or phenotype. A clinical expert then created a manually curated, gold-standard list of HPO concept terms that could be extracted from this note. These HPO terms were referred to as "single genetics note" terms.

Automated Recognition of HPO Concepts via NLP

We developed and evaluated our pipeline by using two well-regarded NLP tools, MedLEE[32,33] and MetaMap,[34,35] that can be used for extracting phenotype concepts from genetics counseling notes. In this subsection, we will describe the pre-processing of the clinical notes and the detailed configuration of the two NLP systems.

Pre-processing of Clinical Notes

We selected the most recent clinical genetic consultation notes before the WES-confirmed genetic diagnoses under the assumption that they were more complete and accurate than older consultation notes. In the primary cohort, of 28 individuals from Columbia University, four had genetic evaluation notes, which included information regarding the diagnostic genetic findings, because a prior diagnostic workup and/or sequencing from another institution or laboratory had become available by the time of their evaluation visit. For these individuals, the evaluation note included the documentation of genetic test results and a short description of the genetic diagnosis. To prevent such text from biasing our phenotyping, we manually removed these portions of the note before applying the NLP parsing.

Additional Pre-processing of Clinical Notes

For MetaMap, we removed the "review of systems" section (if present) from the evaluation notes because many of these sections contained un-parsable, template-based structured tables that became corrupted or lost during the extraction of EHR data to plain text. In addition, these sections typically contained tandem repeats of negated concepts (i.e., "no lymphadenopathy" or "no murmurs"), which add little value to the recognition of phenotype concepts. Because HPO concepts aim to represent mostly pertinent positive findings and only prominently salient negative findings (i.e., "absent speech"), we believe that the removal of this section was warranted. For MedLEE, such pre-processing was not necessary because the build-in section-detection methods were used to systematical delineate the sections via XML parsing.

Configuration of NLP Systems

For MetaMap, we used a local installation of MetaMap by using the latest supported version of the Unified Medical Language System (UMLS; 2016AA release). Starting from the UMLS 2015AB release, the entire HPO database had been integrated into UMLS,[25] which enabled us to make the configuration to restrict our output to HPO concepts (command-line parameter "-R 'HPO'"). In addition, our review of the expert-selected phenotypes revealed that the HPO phenotype concepts frequently belonged to a limited number of UMLS semantic types. In order to prevent an excessive number of non-relevant terms from being mapped, we chose seven UMLS semantic types that effectively represented the larger class of expert-curated HPO concepts selected. These included "congenital abnormality" (T019), "genetic function" (T045), "laboratory procedure" (T059), "laboratory or test result" (T034), "pathologic function" (T046), "disease or syndrome" (T047), and "finding" (T033). Specifically, the options "-I -p -J -K -8 –conj cgab,genf,lbpr,lbtr,patf,dsyn,fndg -R 'HPO'" were used in our application of MetaMap.

For MedLEE, the NLP engine's lexicon was loaded with HPO terms and synonyms available via UMLS (version 2017AA) for this task. The text files were processed, outputting an XML file with tagged tokens regarding information in the clinical note section, token information, HPO concept(s) identified, and certainty and negation information. A Python script using an XML-parsing library (lxml) was used to extract all HPO concepts. We were able to exclude the concepts found in the "review of systems" section without pre-processing.

Configurations of all NLP tools were set to allow for multiple suggestions for a given text phrase as we were performing semantic concept recognition. The scripts for recognition of HPO concepts and output parsing for each NLP tool are accessible at the EHR-Phenolyzer GitHub repository. The output of each tool is a list of HPO concepts (via HPO concept IDs and/or preferred terms) for each given clinical note input as plain text. To handle multiple instances for each concept within a given note, we selected only unique HPO concepts.

Performance Evaluation for Relevant Concept Recognition via NLP

In this subsection, we will describe two types of evaluation methods used in this study: one based on Phenolyzer analysis of the phenotype terms and another based on a comparison with expert-compiled terms.

Currently, there is no standard method for evaluating NLP performance in extracting ontology-based concepts, including evaluating accuracy, relevance, and appropriate granularity of the extracted concepts. Further complicating this evaluation is that for different tasks, relevance is a task-dependent concept. To recognize phenotypes pertinent to genomic testing, we evaluated the performance by using two methods: (1) indirect measurement of the surrogate performance benchmark via Phenolyzer ranking and (2) comparison between the NLP-generated term list and the expert-curated list.

First, these lists of recognized HPO concepts were input into a computational phenotype-based gene-prioritization tool named Phenolyzer, whose performance has been superior to that of other similar tools.[19] Although several other phenotype-based tools (e.g., Exomiser, PhenIX, and Phevor) are reported in the literature and are actively available, they are designed for different purposes and have additional requirements, such as concurrent analysis of VCF files, or produce output that lacks comparable gene ranking. Subsequently, we chose Phenolyzer as our main evaluation tool.

Second, we also approximated the well-established evaluation method for NLP systems by considering the expert-curated list of concepts as a gold standard. We explicitly acknowledge that manual extraction of phenotype terms is subjective and that different experts might generate different expert-curated lists, potentially compromising reproducibility; a reliable way is to measure inter-rater agreement among multiple domain experts and use the consensus-based concept sets as the gold standard. After the NLP tools output a list of HPO concepts (IDs and terms), we manually matched up semantically similar terms to calculate and compare precision, recall, and F-scores between the HPO concepts recognized by each NLP tool and the manually extracted HPO concepts from the same note.

Performance Evaluation of Gene Prioritization

We evaluated the performance of gene prioritization by using Phenolyzer and Phenomizer, which can both accept HPO terms as input and generate a ranked gene list as output. For Phenolyzer, we used the command-line tools available at the Phenolyzer GitHub repository (version v.0.2.0). We used the "-f -p -ph -logistic -addon DB_DISGENET_GENE_DISEASE_SCORE, DB_GAD_GENE_DISEASE_SCORE -addon_weight 0.25" argument in the command-line tool to ensure consistency with the Phenolyzer web server.

For analysis with Phenomizer, we used the web server available at the Phenomizer website because a command-line tool is not publicly available. For each individual, we manually entered HPO terms into the web interface for analysis. The "any" mode of inheritance was selected for the diagnosis, and if the number of input HPO terms was larger than five, we added the "symmetric" mode into the analysis. After Phenomizer generated results in the web interface, we manually downloaded the raw text output file for further processing by a custom Python script to get the gene rankings. Genes were first ranked by their Phenomizer p values in ascending order, and those with identical p values were further ranked by their Phenomizer scores. If one gene occurred multiple times, then the smallest p value was considered. If two genes had the same p values and scores, then the ranking order was randomly determined.

We analyzed the fourth independent cohort containing 20 individuals with CKD to evaluate whether EHR phenotypes can help classify disease subtypes. First, we applied EHR-Phenolyzer on the medical notes to generate HPO terms, and then we used a hierarchical clustering method to study the categorization of individuals with CKD. In the clustering analysis, we used "complete linkage" as the agglomeration method and "Euclidean distance" to calculate the distance between any two individuals. Only individuals with diagnostic genes ranked within the top 50 and with phenotype terms found in at least two individuals but not all were used in the clustering analysis.

Statistical Analysis

In this study, R language[36] was used for statistical computing. We used the exact binomial test to calculate the significance of the gene prioritization and the paired two-sided t test to compare the ranking efficiency of the gene prioritization among different methodologies. In the comparison between expert-rated genes and NLP-rated genes, Pearson's correlation coefficients and the related p values were calculated by R functions cor and cor.test, respectively.

Data Availability

The original clinical notes used for the current study are available from the corresponding authors upon reasonable requests and institutional approvals. The processed results generated or analyzed during this study are included in this published article (and its Supplemental Data files). The external site's clinical notes that support the findings of this study are available from the Mayo Clinic under restrictions. The processed results generated during this study are included in this published article (and its Supplemental Data files).

Results

Selection of Data Source: Comprehensive Chart Review versus Targeted Review of Genetic Notes

We experimented with two methods of selecting EHR data for phenotyping: (1) comprehensive chart review (reviewing the EHRs of each person and synthesizing phenotype concepts from various clinical notes, laboratory tests, imaging results, and pathology reports) and (2) targeted review of genetic notes (retrieving the most recent medical genetic consultation note before WES and synthesizing the phenotypes from the note). The latter examines a much smaller subset of phenotype concepts than the first approach but has the clear advantage of being more efficient and more likely to be fully automatable on EHRs.

To evaluate whether targeted review of genetic notes alone is sufficient in practice, we compared the performance of gene prioritization by these two approaches on 28 affected individuals from Columbia University and for whom diagnostic mutations were identified by WES (Table 1). For each approach, we generated a list of phenotype terms, subjected them to Phenolyzer to generate a ranked gene list, and then examined where the gene with causal variants ranked. The rank performance for the two expert-based approaches is shown in Figure S1. We divided the results into bins of the top 10, 50, 100, 250, 500, and 1,000 genes to better compare the ranking performance of the gene prioritization on the basis of the two different expert methods. We found that the ranking performances were effectively identical between the two methodologies (paired t test $p = 0.44$ for testing differences in performance); more than 50% of confirmed genetic diagnosis occurred within the top 100 predicted candidate genes by Phenolyzer. Therefore, we can reliably use the latest genetic notes before diagnostic exome sequencing as the data source for gene ranking.

Performance Evaluation of NLP Tools in Extracting Phenotype Terms

An overview of our approach to evaluating the performance of different NLP tools is given in Figure 1. We first identified the types of EHR narratives that contain the documentation of phenotypes for genetic disorders, especially the notes authored by medical geneticists or genetic counselors, and then subjected the text to mature NLP systems to extract phenotype concepts and normalize them by using the HPO. Phenolyzer then analyzed these HPO terms to identify related genes with causal variants.

We adapted two different NLP systems, MedLEE and MetaMap, to process genetic notes from EHRs and extract and normalize phenotype concepts by using HPO, as illustrated in Figure 2. In general, both NLP systems tend to generate more terms (on average, 17.6 and 19.4 terms for MetaMap and MedLEE, respectively) than manual extraction by human experts (11.0 terms) (Table 2). The tabulated results based on the matching of terms to the expert-extracted list of HPO terms are shown in Table 3. MedLEE appears to have better concordance with the expert-compiled phenotype terms from the same clinical note. Furthermore, we also compared the phenotype-derived rankings of the genes carrying causal variants between those based on experts and those based on NLPs and found that they were highly correlated (Figure S5).

We next assessed Phenolyzer's ability to rank genes with causal variants by using phenotype terms compiled by experts or extracted by the NLP methods MetaMap and MedLEE. The ranking performances of these three methods are shown in Figure 3A. The NLP systems performed similarly to experts, although each NLP system

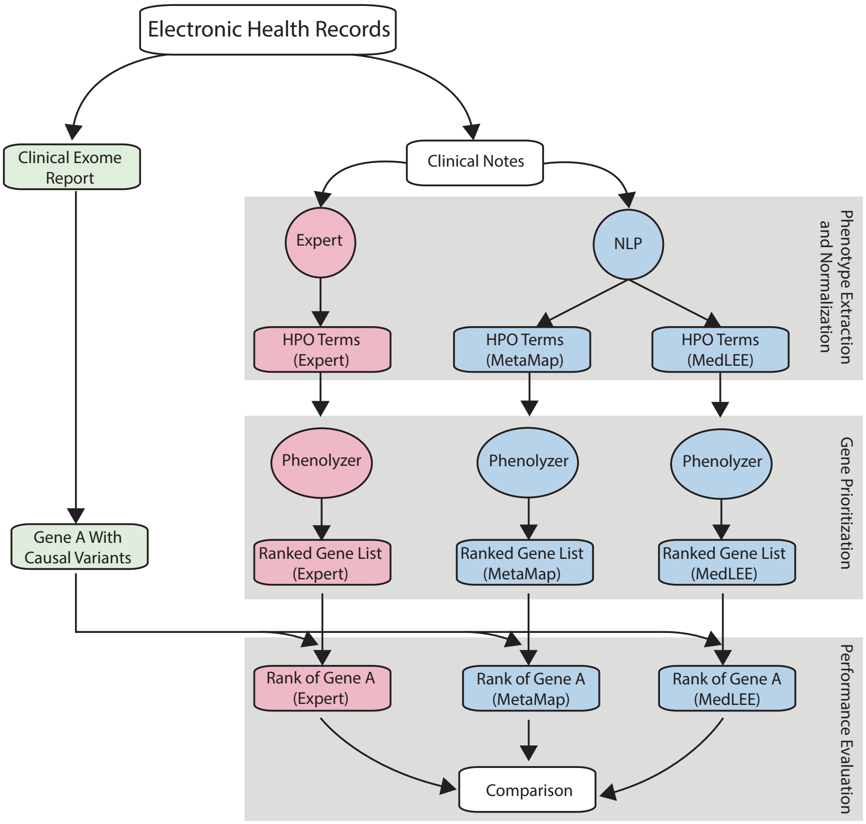


Figure 1. Overview of the Comparative Analysis for Evaluating Different NLP Tools

generated more terms than the experts did. The results showed that 39.3%–57.1% of gene candidates could be ranked within the top 100 genes and that 71.4%–75.0% of gene candidates could be ranked within the top 1,000 genes, both on the basis of only the phenotype concepts derived from the EHR. To evaluate different phenotype-based gene-prioritization tools, we also included gene-ranking results from Phenomizer on MetaMap-generated HPO terms. Our analysis demonstrated that Phenolyzer performs favorably against Phenomizer on the same set of HPO terms, most likely because Phenolyzer's gene-prioritization procedures incorporate multiple levels of prior biological knowledge. However, we acknowledge that Phenomizer was designed for disease diagnosis rather than gene prioritization, so it might not have performed optimally in our evaluation.

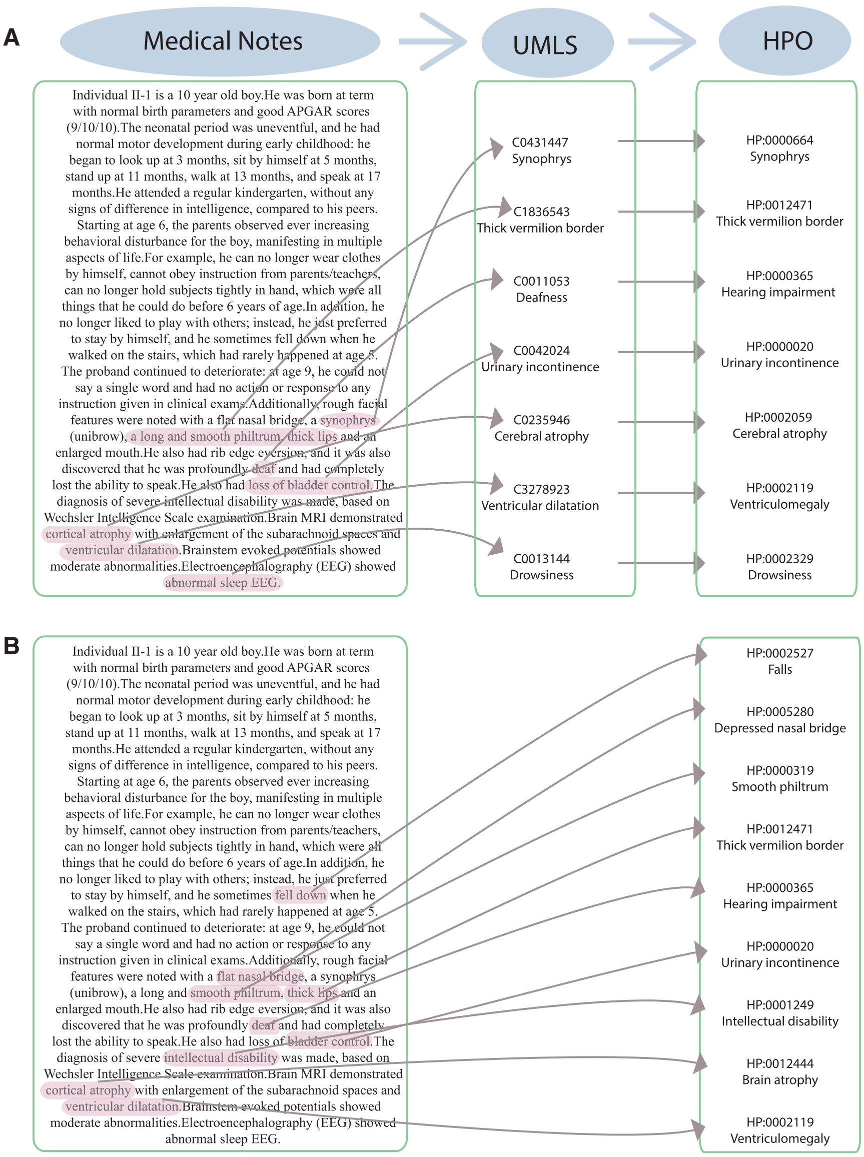
The fact that about 50% of diagnoses can be narrowed to the top 100 genes on the basis of only phenotype information documented in the EHR is remarkable, especially because this performance can be achieved by completely automated phenotype-concept-recognition methods (i.e., MetaMap or MedLEE). We believe that deep phenotypes from EHR data are valuable with the increasing adoption of genomics testing. Improving the prior probability of a diagnosis increases the positive predictive value of a test, although current genomic testing methods tend to forgo this step. Therefore, systematic integration of EHR-phenotype-based gene prioritization before variant interpretation can potentially improve workflow efficiency and help reach clinically valid results while improving diagnostic yield.

External Validation of Automated Phenotype Description and Gene Prioritization

We applied the same pipeline by using the clinical notes written by genetic counselors from the Mayo Clinic. Information on ten affected individuals, together with confirmed genetic diagnoses in the genes cystic fibrosis transmembrane conductance regulator (*CFTR* [MIM: 602421]), peripheral myelin protein 22 (*PMP22* [MIM: 601097]), DM1 protein kinase (*DMPK* [MIM: 605377]), dynamin 1(*DNM1* [MIM: 602377]), coagulation factor VIII (*F8* [MIM: 300841]), fibrillin 1 (*FBN1* [MIM: 134797]), KAT8 regulatory NSL complex subunit 1(*KANSL1* [MIM: 612452]), NPC intracellular cholesterol transporter 1 (*NPC1* [MIM: 607623]), sodium voltage-gated channel alpha subunit 1 (*SCN1A* [MIM: 182389]), and SOS Ras/Rac guanine nucleotide exchange factor 1 (*SOS1* [MIM: 182530]), was provided. The ranking results are shown in Figure 3B. The results are comparable to the ranking performance obtained at our institution. Therefore, the analysis on the secondary-site validation data confirmed that the EHR-Phenolyzer approach can be used in different institutions with diverse sets of informatics infrastructure as long as an automated procedure for extracting clinical notes can be implemented in each site.

Additional Analysis on the Use of Phenotype Features in Real-World Genetic Diagnosis

To examine how clinical phenotypes are currently used in real-world settings to facilitate genetic diagnosis of people with rare monogenic diseases, we examined EHR data on 46 affected individuals, all of whom were assessed by a medical geneticist or genetic counselor at Columbia University-affiliated hospitals in an outpatient setting. This set of clinical notes, together with the corresponding molecular pathology reports, should be highly informative on the real-world use of clinical phenotype information in the context of genetic testing. We found that 15 of 46 affected individuals did not undergo diagnostic genetic testing (Figure 4A), the reasons for which were lack of known reimbursable tests (n = 7), lack of insurance (n = 2), refusal by family members (n = 2), lack of testing records in EHRs (n = 1), and other undescribed reasons (n = 3). Among



Individual II-1 is a 10 year old boy.He was born at term with normal birth parameters and good APGAR scores (9/10/10).The neonatal period was uneventful, and he had normal motor development during early childhood: he began to look up at 3 months, sit by himself at 5 months, stand up at 11 months, walk at 13 months, and speak at 17 months.He attended a regular kindergarten, without any signs of difference in intelligence, compared to his peers. Starting at age 6, the parents observed ever increasing behavioral disturbance for the boy, manifesting in multiple aspects of life.For example, he can no longer wear clothes by himself, cannot obey instruction from parents/teachers, can no longer hold subjects tightly in hand, which were all things that he could do before 6 years of age.In addition, he no longer liked to play with others; instead, he just preferred to stay by himself, and he sometimes fell down when he walked on the stairs, which had rarely happened at age 5. The proband continued to deteriorate: at age 9, he could not say a single word and had no action or response to any instruction given in clinical exams.Additionally, rough facial features were noted with a flat nasal bridge, a synophrys (unibrow), a long and smooth philtrum, thick lips and an enlarged mouth.He also had rib edge eversion, and it was also discovered that he was profoundly deaf and had completely lost the ability to speak.He also had loss of bladder control.The diagnosis of severe intellectual disability was made, based on Wechsler Intelligence Scale examination.Brain MRI demonstrated cortical atrophy with enlargement of the subarachnoid spaces and ventricular dilatation.Brainstem evoked potentials showed moderate abnormalities.Electroencephalography (EEG) showed abnormal sleep EEG.

Individual II-1 is a 10 year old boy.He was born at term with normal birth parameters and good APGAR scores (9/10/10).The neonatal period was uneventful, and he had normal motor development during early childhood: he began to look up at 3 months, sit by himself at 5 months, stand up at 11 months, walk at 13 months, and speak at 17 months.He attended a regular kindergarten, without any signs of difference in intelligence, compared to his peers. Starting at age 6, the parents observed ever increasing behavioral disturbance for the boy, manifesting in multiple aspects of life.For example, he can no longer wear clothes by himself, cannot obey instruction from parents/teachers, can no longer hold subjects tightly in hand, which were all things that he could do before 6 years of age.In addition, he no longer liked to play with others; instead, he just preferred to stay by himself, and he sometimes fell down when he walked on the stairs, which had rarely happened at age 5. The proband continued to deteriorate: at age 9, he could not say a single word and had no action or response to any instruction given in clinical exams.Additionally, rough facial features were noted with a flat nasal bridge, a synophrys (unibrow), a long and smooth philtrum, thick lips and an enlarged mouth.He also had rib edge eversion, and it was also discovered that he was profoundly deaf and had completely lost the ability to speak.He also had loss of bladder control.The diagnosis of severe intellectual disability was made, based on Wechsler Intelligence Scale examination.Brain MRI demonstrated cortical atrophy with enlargement of the subarachnoid spaces and ventricular dilatation.Brainstem evoked potentials showed moderate abnormalities.Electroencephalography (EEG) showed abnormal sleep EEG.

Figure 2. Illustration of How NLPs Work to Extract Phenotype Terms from Natural Language in Clinical Notes
The same clinical note was analyzed by MetaMap (A) and MedLEE (B) for the generation of HPO terms.

the 31 affected individuals who underwent genetic testing, the genetic tests used (Figure 4B) were clinical microarray (n = 11), PCR (n = 2), single-gene Sanger sequencing (n = 5), targeted panel (n = 2), clinical exome (n = 9), and undescribed (n = 2). Diagnostic results were detected in 11 of the 31 (35.5%) affected individuals; 7 (63.6%) of these individuals had been diagnosed via clinical WES.

To understand how phenotype information is used in current clinical practice to assist in genetic diagnosis, we manually examined the genetic diagnostic reports for each of the 31 affected individuals (Figure 4C). These diagnostic reports were generally provided as scanned PDF files from the following clinical labs: Ambry Genetics (n = 4), GeneDx (n = 12), Columbia University Personalized Genomic Medicine Laboratory Hospital lab (n = 3), Integrated Genetics (n = 5), LabCorp (n = 4), Mayo Clinic (n = 1), and unspecified (n = 2). We found that 19 (61%) of the 31 diagnostic reports contained no indication of a clinical phenotypes, suggesting that clinical phenotypes were either not provided to diagnostic labs or not used by diagnostic labs in making a diagnosis (we acknowledge that this situation is quite common for targeted tests but less likely for WES). Among the 12 genetic diagnostic reports with information about the indication for testing, the indication was most commonly listed in an unstructured sentence or paragraph format (8/12 [67%]); in the others, it was listed simply as ICD codes (3/12 [25%]) or as the single general term “diagnostic” (1/12 [8%]). We compared the indication with clinical phenotypes inferred by MetaMap or MedLEE from clinical notes in EHRs (Table S1). With the exception of one individual for whom we do not have detailed notes by the genetic counselor, the clinical phenotypes from EHRs were consistently more comprehensive and detailed than those provided in the indication, which could improve the diagnostic yield for clinical labs.

For the 11 individuals with positive results from genetic diagnostic testing, we next examined whether deep phenotypes from EHRs can facilitate prioritization of candidate genes, similarly to what we had done on the primary and secondary cohorts described above. We found that the genes with causal variants were ranked among the top 100 or top 1,000 genes for over 50% or 91%, respectively, of the affected individuals (Figure 4D), again suggesting that EHR-derived phenotype information could greatly increase the efficiency of genetic diagnosis. Furthermore, similar to previous observations, we also found that Phenolyzer outperformed Phenomizer on this set of affected individuals, justifying the use of computational tools specifically designed for phenotype-driven gene prioritization.

Analysis of Specific Disease Subtypes within a Broad Disease Category

We next investigated whether EHR-Phenolyzer can be useful for discerning specific genetic forms of a broader category of disease with CKD as a model. Discerning hereditary versus acquired etiologies of CKD oftentimes has a

Table 2. Comparisons of Manual Extraction and Two NLP Tools in Terms of the Total and Correct HPO Concepts Recognized from Clinical Notes

	Expert Manual Extraction		NLP	
Phenotype Extraction Methods	**Heuristic Chart Review**	**Single Genetics Note**	**MetaMap**	**MedLEE**
Total no. of HPO concepts selected	223	309	493	543
Mean no. of concepts per individual	8.0	11.0	17.6	19.4
Median no. of concepts per individual	8	12	17.5	19.0
Semantically correctly matched concepts	–	–	199 (64%)	222 (72%)
Mean no. of correctly matched concepts per individual note	–	–	7.1	7.9

The following abbreviations are used: HPO, Human Phenotype Ontology; and NLP, natural language processing.

substantial impact on clinical prognosis and management; however, the two can be indistinguishable by traditional diagnostics alone.[37] Because many hereditary nephropathies display substantial genetic and phenotypic heterogeneity,[38] gene panels or genome-wide testing can help diagnose individuals with a suspected monogenic renal disease. We evaluated the EHRs of a set of 20 individuals with CKD and confirmed genetic diagnosis.[31] We found that EHR-Phenolyzer (based on either MedLEE or MetaMap) worked especially well for this set of individuals in that it ranked the genes with causal variants within the top ten for nearly half of them (Figure S3A). This observation can be attributed to two reasons: (1) given that these individuals were recruited from a large academic referral center for renal disease, many were already well characterized and had been diagnosed by traditional methods (e.g., kidney biopsy for Alport syndrome), so genetic testing served as a merely confirmatory test; and (2) the specificity of the kidney-related phenotypes listed in these individuals' EHRs would also restrict the number of candidate genes. We additionally performed a hierarchical clustering on this set of individuals on the basis of the presence or absence of specific phenotype terms. For the 13 individuals with diagnostic genes ranked within the top 50 by EHR-Phenolyzer, we found that the individuals with the same genes with causal variants, such as the two individuals with uromodulin (*UMOD* [MIM: 191845]) mutations and the four individuals with collagen type IV alpha 5 chain (*COL4A5* [MIM: 303630]) mutations, tended to be clustered together according to the phenotype terms (Figure S3B). Nevertheless, there were also scenarios in which affected individuals with the same diagnostic genes had quite distinct phenotypes from each other (such as the individuals with *COL4A4* [MIM: 120131] mutations), which suggests that EHR-Phenolyzer can tolerate some noise in the phenotype-extraction procedure, supporting its utility for genetic diseases that have clinically heterogeneous presentations.

Detailed Examination of Selected Case Studies

To understand the degree to which or the contexts in which the methods work, we performed a detailed examination of several illustrative cases. We analyzed a 15-year-old female with multiple organ-system anomalies, including intellectual disability and skeletal dysplasia. Clinical exome sequencing identified collagen type X alpha 1 chain (*COL10A1* [MIM: 120110]) as the gene with causal variants, yielding a molecular diagnosis of Schmid-type metaphyseal chondrodysplasia (MCDS [MIM: 156500]). MCDS is caused by heterozygous mutations in *COL10A1* and is characterized by short stature and bowing of the long bones.[39] For this individual, 15, 25, and 18 phenotype terms were compiled by experts, MedLEE, and MetaMap, respectively (Figure 5), but only five terms (spondylometaphyseal dysplasia, skeletal dysplasia, short stature, intellectual disability, and global developmental delay) were shared by all three methods. Nevertheless, this gene was ranked as #4 by Phenolyzer on all three sets of terms separately, suggesting that Phenolyzer can tolerate inaccuracies in phenotype terms and upweight highly specific terms in its scoring scheme. This example clearly demonstrates that as long as a core set of highly informative phenotype terms can be identified from EHR narratives, good ranking performance can be achieved, even if extra less-relevant terms are also included.

We further analyzed an individual for whom expert-compiled terms and MedLEE-generated terms worked much better than the terms generated by MetaMap. The affected individual is a 13-year-old female with generalized seizures and a mutation in *SCN1A*. *SCN1A* encodes a voltage-gated sodium channel essential for the generation and propagation of action potentials and is associated with four Mendelian phenotypes in OMIM, including generalized epilepsy with febrile seizures plus type 2 (MIM: 604403), early infantile epileptic encephalopathy (MIM: 607208), familial febrile seizures 3A (MIM: 604403), and familial hemiplegic migraine 3 (MIM: 609634). Surprisingly, although expert-compiled terms and MedLEE-compiled terms are generally quite broad, this gene ranked as #1 and #18 on the basis of these terms, respectively (Figure S4). In comparison, MetaMap generated more specific phenotype terms such as "pneumonia" and "hepatic encephalopathy" (which are unrelated to *SCN1A*), as well as candidate disease diagnosis "autism spectrum disorders," but *SCN1A* was not ranked within the top 100 genes.

The above analyses highlight that EHR narratives typically contain concepts that can include both pertinent

Table 3. Evaluation of Two NLP Tools with Expert-Extracted HPO Terms as the Gold Standard

		NLP tools	
Data Source	**Measure**	**MetaMap**	**MedLEE**
Primary site (n = 28)	precision	0.40	0.41
	recall	0.64	0.72
	F-score	0.50	0.52
Secondary site (n = 10)	precision	0.60	0.51
	recall	0.58	0.68
	F-score	0.59	0.58

The following abbreviation is used: NLP, natural language processing.

and irrelevant signs, symptoms, clinical descriptions, and clinical histories with variable levels of confidence or relevance. Thus, despite the limitations of NLP systems, the clinical information contained within the note can be extracted with the assistance of computationally enabled ontologies such as HPO and tools such as Phenolyzer. In a purely hypothetical example, the two phenotype concepts "intellectual disability" and "generalized seizure" would ideally strengthen the confidence of the computational representation of the disorder "seizure disorder" given these semantically and ontologically related concepts, improving the confidence score of finding seizure-disorder-related genes. Less-relevant concepts identified for the same individual can be regarded as peripheral to the main genetic etiology in computational phenotype-based gene prioritization. Thus, a robust relevance metric is critical for filtering out irrelevant concepts.

Combined Analysis of Phenotype and Genotype Data Expedite Genetic Diagnosis

Our analysis above focused on phenotype-driven prioritization of genes and demonstrated that genes with causal variants can be ranked much higher than other genes with the use of phenotype information extracted from EHRs. To further demonstrate the applicability of this method in real-world settings to facilitate the identification of disease-causing variants, we analyzed several previously published cases,[40,41] for which we performed a combined analysis of genotype data (VCF files) and clinical descriptions from the methods sections of the published manuscripts. We observed that the clinical descriptions in scientific manuscripts were professionally edited and could be of higher quality than typical EHR narratives, but extracting HPO terms from the public case reports poses challenges similar to those faced in EHR settings.

The first case study was of an individual diagnosed with KBG syndrome,[40] which is a rare autosomal-dominant genetic condition characterized by intellectual disability, seizures, and distinct facial, hand, and skeletal features. We previously identified a *de novo*, single-nucleotide insertion in ankyrin repeat domain 11(*ANKRD11* [MIM: 611192]) as the disease-causing variant through the analysis of trio WES data. In the current study, we did not use parental information to infer *de novo* variants and instead analyzed the exome data of the proband only. We identified all missense, nonsense, stop-gain, frameshift, and splice variants with an allele frequency $< 1 \times 10^{-5}$ in gnomAD,[42] a publicly available allele-frequency database of 123,136 exomes. There were 459 prioritized variants in this list (typically <100 variants are identified after filtering, but these exome data were generated on the Ion Torrent platform, resulting in a large number of potential false-positive calls). Using phenotype terms derived from the EHR-Phenolyzer pipeline with the MetaMap engine (Table S2), our method ranked *ANKRD11* as #6. After we compared the overlap between the Phenolyzer list and the prioritized variant list, *ANKRD11* was ranked first, providing a molecular diagnosis of KBG syndrome even without parental information (Figure 6). We replicated this result by using the EHR-Phenolyzer pipeline with MedLEE as the NLP engine, yielding identical results.

The second case study was focused on a sibling pair (brother and sister) both affected by progressive cognitive decline starting from 6 years of age. We previously identified compound-heterozygous mutations in N-acetyl-alpha-glucosaminidase (*NAGLU* [MIM: 609701]), leading to a genetic diagnosis of Sanfilippo syndrome (mucopolysaccharidosis IIIB).[41] Biochemical tests confirmed the complete loss of activity of alpha-N-acetylglucosaminidase (encoded by *NAGLU*) in both individuals. In the current study, we did not filter for shared variants between the siblings and instead analyzed each individual's exome separately. We used an allele-frequency threshold of 0.01 to account for the possibility that causal variants for recessive conditions could be observed in public databases with a relatively high allele frequency. For the sister, using phenotype terms derived from the EHR-Phenolyzer pipeline with MetaMap engine (Table S3), our method ranked *NAGLU* as #42 among all human genes. After we compared the overlap between this list and the prioritized list of 885 variants, *NAGLU* was ranked as #1 for the observed phenotypes. For the brother, *NAGLU* was ranked as #201, and the intersection between this list and the prioritized list of 892 variants increased the rank to #1. Therefore, in both cases, we were able to easily identify the gene with causal variant and yield a molecular diagnosis through combined analysis of genotypes and phenotypes. Similar results were obtained with the EHR-Phenolyzer pipeline with MedLEE as the NLP engine, confirming that the combination of EHR-Phenolyzer and exome data can often significantly expedite molecular diagnosis of monogenic disorders.

Discussion

In this study, we evaluated the clinical validity of automated extraction of HPO concepts from EHR narratives for computational phenotype-driven gene prioritization and demonstrated that the proposed method greatly

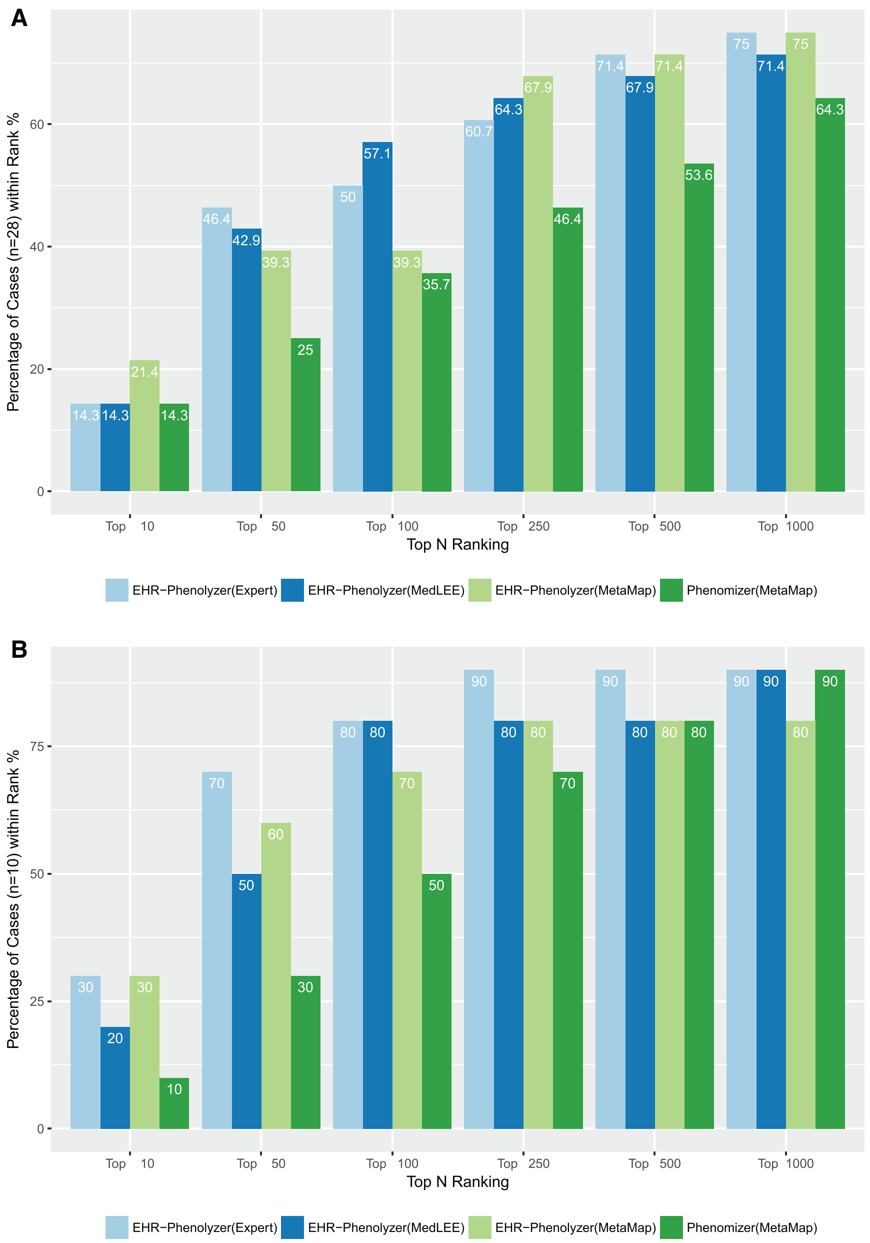


Figure 3. Comparison of Four Methods of Ranking Genes with Causal Variants
28 individuals in the primary site (A) and ten individuals in the secondary site (B). For each individual, three methods were used to extract phenotype terms and then used in Phenolyzer or Phenomizer to find a ranked list of candidate genes. The MedLEE approach achieved the best performance in ranking the genes with causal variants within the top 100 of all genes in both datasets.

facilitates the interpretation of clinical exome sequencing data. The EHR-Phenolyzer framework operates in two steps: the first uses NLP-driven HPO-concept recognition through either the publicly available tool MetaMap or the proprietary tool MedLEE, and the second utilizes the open-source computational phenotype tool Phenolyzer for gene prioritization. Finally, through retrospective case studies, we demonstrated how combined analyses of genotype and phenotype data from EHRs can expedite genetic diagnoses by using clinical exomes. We conclude that EHR-Phenolyzer enables comprehensive utilization of deep phenotypes in EHR narratives, allows for phenotype-driven ordering and analysis of clinical exome tests, and facilitates the implementation of genomic medicine.

Before using NLP systems for phenotype recognition, we first examined whether expert-curated HPO concepts can be derived from EHR narratives to perform a phenotype-driven gene prioritization task. Upon our discovery that the majority of the curated phenotype concepts were sourced from a single document – the latest genetics evaluation note – a document authored by a member of the clinical genetics team, we performed a focused HPO term recognition of this single clinical document. We show that the performance of expert extraction of the "heuristic chart review" and the "single genetics note" (Figure S1) are largely comparable for the purposes of computational phenotype-based gene prioritization task. However, the latter approach has the clear advantage of being able to be automated through NLP approaches, and therefore serves as the foundation of EHR-Phenolyzer framework for phenotype extraction from EHRs.

Current Issues with Extraction of HPO Concepts

We have shown that, when used with computational phenotype-based gene-ranking tools, automatic recognition of HPO concepts by two NLP systems, MetaMap and MedLEE, achieves a gene-ranking performance comparable to that of expert curation of phenotype terms from EHR narratives, despite the seemingly modest traditional NLP performance. In our internal review of the extracted HPO concepts for each tool, we recognized that many "non-phenotype" concepts, such as "left (HP:0012834)," were also recognized by MedLEE and that this most likely affected the NLP evaluation metrics shown in Table 2. MetaMap was able to filter out these non-phenotype concepts by using the UMLS "semantic types" filter that we applied. This is one of the inherent limitations of HPO: the expansion of HPO creates modifiers represented as both pre-coordinated concepts and post-coordinated concepts, which can be represented by the integration of multiple smaller concepts. For

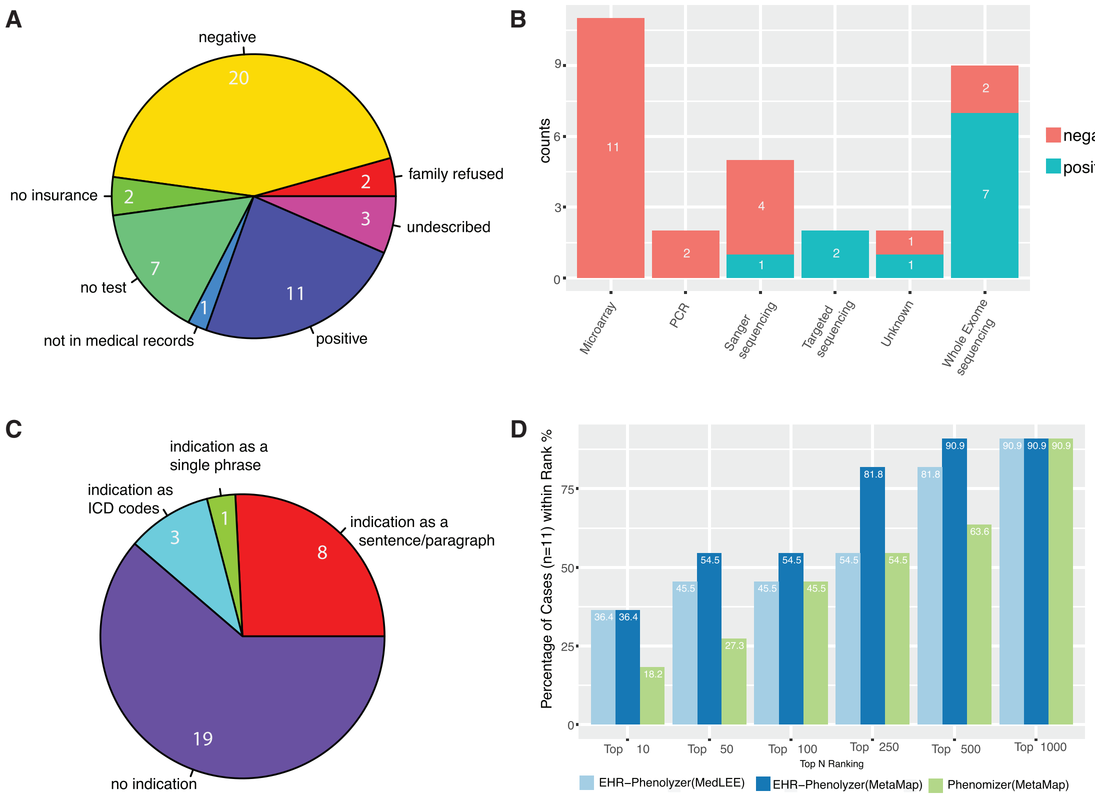


Figure 4. Detailed Analysis of Genetic Counselors' Notes and Genetic Diagnostic Reports on 46 Affected Individuals from Cohort 3
(A) A breakdown of the affected individuals according to diagnostic genetic testing.
(B) The distribution of various genetic tests that were used on this cohort.
(C) The distribution of the types of phenotype information used in genetic diagnosis.
(D) Performance of EHR-Phenolyzer in ranking the genes with causal variants among all candidate genes.

example, MetaMap recognizes "severe global developmental delay (HP:0011344)" as a pre-coordinated concept, but MedLEE identifies "severe" as a modifier of the separate concept "global developmental delay (HP:0001263)." This is a well-known redundancy issue recognized by the contributors of HPO, and it demonstrates the ongoing work needed to continue improving HPO.[25]

Improving Performance and Limitations by Restricting to OMIM Genes

In our previous analysis, we examined the ranking of genes with causal variants among the ~20,000 human genes. However, in practice, clinical diagnostic labs might examine only the subset of genes known to be associated with monogenic disorders, which would make gene prioritization somewhat easier. To gain a deeper understanding of the performance of the EHR-Phenolyzer approach in clinical settings, we assessed how our approach can rank genes among a selected list of about 5,000 OMIM genes that are known to be associated with Mendelian diseases rather than among all 20,000 genes. Our results showed that restricting the analysis to OMIM genes further improved the performance of EHR-Phenolyzer in detecting genes with causal variants (Figure S2). However, we also note that two positive diagnoses were made on myosin heavy chain 10 (*MYH10* [MIM: 160776]) and N(alpha)-acetyltransferase 15, NatA auxiliary subunit (*NAA15* [MIM: 608000]), which had not yet been documented in OMIM as being associated with a Mendelian phenotype, suggesting that expanded analysis could still be warranted if OMIM-restricted analyses do not yield positive results. *MYH10* and *NAA15* were both discovered recently from several sequencing studies on congenital heart disease and developmental disorders.[43–45]

Limitations of the Current Study

The current study has several limitations. First, our evaluation was done retrospectively, whereas ideally we would apply our pipeline prospectively to investigate whether it offers higher diagnostic yields, decreases the time for genomic data analysis, and improves reproducibility. We also did not directly compare Phenolyzer results against the pipelines that have already integrated computational phenotyping for annotations (e.g., Exomiser and Phevor) because these other tools require more than phenotype terms in HPO formats. We are currently designing prospective studies that assess a large number of clinical cases of suspected monogenetic disease to formally quantify the impact of EHR-Phenolyzer on a healthcare system to facilitate the implementation of genomic medicine in a streamlined, efficient, and scalable manner.

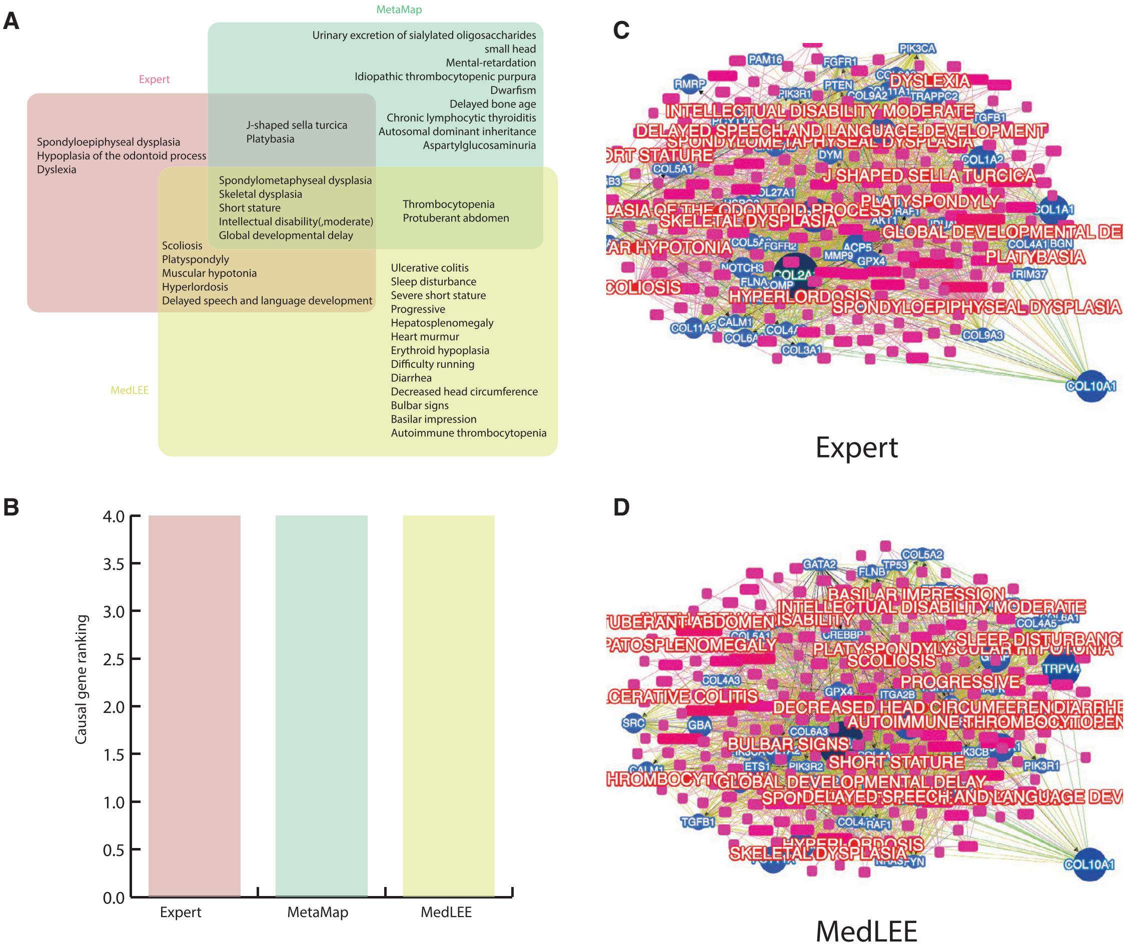


Figure 5. Phenolyzer Can Tolerate Inaccuracies in the Phenotype-Term Extraction of an Individual Affected by Schmid-type Metaphyseal Chondrodysplasia

(A) Only five phenotype terms were shared among three different phenotype-extraction methods.

(B) All three methods ranked the gene with a causal mutation as #4.

(C and D) The network of prioritized genes and phenotype terms, where the phenotypes were extracted by an expert (C) or by MedLEE (D). *COL10A1* with a causal mutation is highlighted in the network. The size of each pie section is positively related to the Phenolyzer ranking.

Several computational tools, including Exomiser, allow integrated analysis of phenotype and genotype data. In contrast, EHR-Phenolyzer uses only phenotype data to rank genes with causal variants among all possible candidate genes without considering genotype information. This design principle of EHR-Phenolyzer allows flexible downstream variant analysis with various different computational tools; nevertheless, we will provide helper scripts that combine results from EHR-Phenolyzer with ANNOVAR-generated variant annotations (in “multianno” formats) to facilitate users who choose to use ANNOVAR for analysis of exome data. Similarly, we made EHR-Phenolyzer modular and flexible such that it can interoperate with many different NLP tools to process clinical notes. Indeed, besides MetaMap and MedLEE, we have now incorporated Annotator from the National Center for Biomedical Ontology[46] as another text-mining option.

A third limitation is that we used only a particular type of EHR narrative, and more studies are warranted for testing the portability of the NLP pipeline to other types of EHR notes. Our results also contained inaccurately extracted concepts, especially in the very dense sections of physical examination, where lack of punctuation rules and conjunctions caused negated concepts to be falsely recalled as a concept. Therefore, we advise that as is, the current implementations of these NLP methods still need additional improvements for high-quality phenotyping curation for phenotype databases. However, we believe that with appropriate optimization (such as HPO class filtering or UMLS “semantic types” filtering, as we demonstrated with MetaMap), our NLP methods can be utilized for such tasks in the future.

Clinical Significance of the Combined Genotype-Phenotype Analysis

As shown by the results from four independent cohorts, in more than half of the individuals, the genes with disease-causing mutations can be prioritized within the top 100

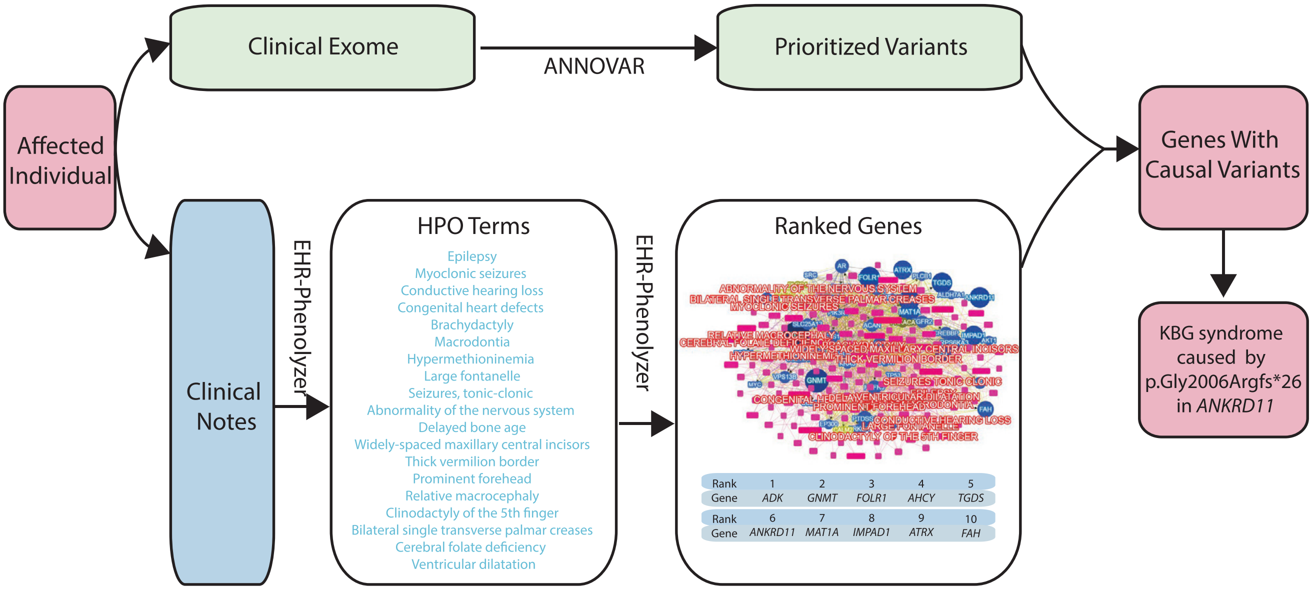


Figure 6. Molecular Diagnosis of KBG Syndrome in an Individual with a Frameshift Mutation in *ANKRD11* through Combined Genotype and Phenotype Analysis

and in some cases even within the top ten. In clinical practice, this information can greatly reduce the effort in manually searching for candidate genes when analyzing WES data. Furthermore, as illustrated in the combined analysis of genotype and phenotype for genetic diagnosis of two individuals, the genes with causal variants were ranked as the top gene, which showcased its practical significance in clinical diagnostic settings of joint analysis of phenotype and genomic data. The validation of our method in four independent cohorts from two different institutions also demonstrated the possibility of extending such approaches to other institutions with different informatics infrastructures. Meanwhile, we acknowledge that this study did not include large-scale genotypic WES data; additional studies on more affected individuals with paired WES data from other institutions are preferred for the evaluation of the generalization of EHR-Phenolyzer in the future.

Future Perspectives

In addition to addressing the aforementioned limitations, subsequent research efforts will focus on further improving the EHR-Phenolyzer framework. These include exploring concept recognition from structured EHR data in addition to unstructured clinical narratives, such as laboratory testing results and radiographic findings. Because HPO already integrates many of these concepts, such procedures can potentially further improve this process of automated EHR-phenotype-driven gene prioritization if these concepts are not recorded within the clinical notes. Mapping from other well-established standard terminologies, such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), to HPO has been shown to be feasible by Dhombres et al.[47] Given that many institutions annotate clinical notes and findings by using SNOMED CT or terminologies other than HPO, we will integrate other popular terminologies and their concept mappings to HPO into EHR-Phenolyzer in the future.

Another future area that we will explore is evaluating the transferability of the proposed methods to different healthcare systems that leverage different EHRs. In the current study, we examined and confirmed that the EHR-Phenolyzer method can be utilized in two different healthcare systems with a relatively small set of samples. We expect to significantly expand the number of sites to be analyzed by EHR-Phenolyzer in the future and examine how to adapt the method to different settings across institutions to enable the delivery of more benefits to the broader community.

We also plan to build an individual-facing Phenolyzer that allows people to enter self-reported phenotypes not captured in EHRs. With this feature, we will explore whether individual-provided information can further improve the accuracy for gene ranking when the genomic analysts have access to such information. In order to accommodate users who speak different languages from all over the world, we might also extend the EHR-Phenolyzer to accommodate phenotypes entered in non-English languages.

Finally, an effort to curate phenotype data in a systematic manner requires the recognition of the importance of phenotype information. As more high-quality genomic and phenotype information is collected with collaborative efforts such as the Monarch Initiative, PhenomeCentral, and HPO, we believe that approaches driven by phenotype data will become more robust and effective. With the continuing growth of HPO, the continued development of new techniques and optimization of pre-existing NLP techniques is likely to improve term normalization across the field of genomic medicine, making these efforts easier and more effective in the future.

Supplemental Data

Supplemental Data include five figures and three tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.05.010.

Acknowledgments

We thank the individuals and their family members who participated in this study to improve the diagnostic rates in clinical exome testing. We also thank the genetic counselors and clinical geneticists who provided the clinical notes that our algorithms used to infer automated phenotype terms. We thank colleagues at the Institute of Genomic Medicine, especially Dr. David Goldstein, Nick Stong, and Louise Bier, for suggesting phenotype-driven analysis and helpful discussion and comments on the study. We thank the developers of the Human Phenotype Ontology for continuous development of this ontology over the past few years, which greatly facilitated and standardized clinical diagnosis of individuals with suspected genetic disorders. This study was supported by grants from the JPB Foundation (to W.K.C), National Human Genome Research Institute (HG006465 to K.W. and HG008680 to A.G., G.H., and C.W.), National Institute of Mental Health (MH108728 to K.W.), National Library of Medicine (R01LM009886 to C.W.), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK105124, UG3DK114926, and RC2DK116690 to K.K.), and National Center for Advancing Translational Sciences (U01TR02062 to H.L., F.S., and L.W.).

Declaration of interests

A provisional patent application titled "Automated tool for diagnostic genomic predictions based on electronic health record data" has been filed by Columbia University. C.F. is a consultant for Health Fidelity, which has licensed MedLEE from Columbia.

Received: March 22, 2018
Accepted: May 24, 2018
Published: June 28, 2018

Web Resources

EHR-Phenolyzer, https://github.com/WGLab/EHR-Phenolyzer
HPO browser, http://compbio.charite.de/hpoweb/
OMIM, https://www.omim.org/
Phenolyzer, https://github.com/WGLab/phenolyzer
Phenomizer, http://compbio.charite.de/phenomizer/
R Project for Statistical Computing, https://www.r-project.org/

References

1. van Nimwegen, K.J., Schieving, J.H., Willemsen, M.A., Veltman, J.A., van der Burg, S., van der Wilt, G.J., and Grutters, J.P. (2015). The diagnostic pathway in complex paediatric neurology: A cost analysis. Eur. J. Paediatr. Neurol. *19*, 233–239.
2. Vissers, L.E.L.M., van Nimwegen, K.J.M., Schieving, J.H., Kamsteeg, E.J., Kleefstra, T., Yntema, H.G., Pfundt, R., van der Wilt, G.J., Krabbenborg, L., Brunner, H.G., et al. (2017). A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. Genet. Med. *19*, 1055–1063.
3. Graungaard, A.H., and Skov, L. (2007). Why do we need a diagnosis? A qualitative study of parents' experiences, coping and needs, when the newborn child is severely disabled. Child Care Health Dev. *33*, 296–307.
4. 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.
5. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.
6. Valencia, C.A., Husami, A., Holle, J., Johnson, J.A., Qian, Y., Mathur, A., Wei, C., Indugula, S.R., Zou, F., Meng, H., et al. (2015). Clinical impact and cost-effectiveness of whole exome sequencing as a diagnostic tool: A pediatric center's experience. Front Pediatr. *3*, 67.
7. 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.
8. Eldomery, M.K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J.A., Gambin, T., Stray-Pedersen, A., Küry, S., Mercier, S., Lessel, D., Denecke, J., et al. (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. *9*, 26.
9. 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.
10. Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K.G., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. Genet. Med. *18*, 696–704.
11. Shashi, V., McConkie-Rosell, A., Rosell, B., Schoch, K., Vellore, K., McDonald, M., Jiang, Y.-H., Xie, P., Need, A., and Goldstein, D.B. (2014). The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genet. Med. *16*, 176–182.
12. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., et al. (2006). Gene prioritization through genomic data fusion. Nat. Biotechnol. *24*, 537–544.
13. Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am. J. Hum. Genet. *94*, 599–610.
14. Javed, A., Agrawal, S., and Ng, P.C. (2014). Phen-Gen: Combining phenotype and genotype to analyze rare disorders. Nat. Methods *11*, 935–937.

15. Sifrim, A., Popovic, D., Tranchevent, L.-C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J.R., Aerts, J., De Moor, B., and Moreau, Y. (2013). eXtasy: Variant prioritization by genomic data fusion. Nat. Methods *10*, 1083–1084.

16. Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci. Transl. Med. *6*, 252ra123.

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

18. 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.

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

20. Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. *20*, 490–497.

21. Robinson, P.N., Krawitz, P., and Mundlos, S. (2011). Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin. Genet. *80*, 127–132.

22. Smedley, D., and Robinson, P.N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Med. *7*, 81.

23. 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.

24. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. *83*, 610–615.

25. 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.

26. Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2017). The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. *45* (D1), D712–D722.

27. 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.

28. Deans, A.R., Lewis, S.E., Huala, E., Anzaldo, S.S., Ashburner, M., Balhoff, J.P., Blackburn, D.C., Blake, J.A., Burleigh, J.G., Chanet, B., et al. (2015). Finding our way through phenotypes. PLoS Biol. *13*, e1002033.

29. Smedley, D., Köhler, S., Czeschik, J.C., Amberger, J., Bocchini, C., Hamosh, A., Veldboer, J., Zemojtel, T., and Robinson, P.N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics *30*, 3215–3222.

30. Jagadeesh, K.A., Wu, D.J., Birgmeier, J.A., Boneh, D., and Bejerano, G. (2017). Deriving genomic diagnoses without revealing patient genomes. Science *357*, 692–695.

31. Lata, S., Marasa, M., Li, Y., Fasel, D.A., Groopman, E., Jobanputra, V., Rasouly, H., Mitrotti, A., Westland, R., Verbitsky, M., et al. (2018). Whole-exome sequencing in adults with chronic kidney disease: A pilot study. Ann. Intern. Med. *168*, 100–109.

32. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., and Johnson, S.B. (1994). A general natural-language text processor for clinical radiology. J. Am. Med. Inform. Assoc. *1*, 161–174.

33. Friedman, C., Johnson, S.B., Forman, B., and Starren, J. (1995). Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc. Annu. Symp. Comput. Appl. Med. Care, 347–351.

34. Aronson, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc. AMIA Symp. *2001*, 17–21.

35. 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.

36. R Core Development Team (2017). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). https://www.R-project.org/.

37. Devuyst, O., Knoers, N.V., Remuzzi, G., Schaefer, F.; and Board of the Working Group for Inherited Kidney Diseases of the European Renal Association and European Dialysis and Transplant Association (2014). Rare inherited kidney diseases: Challenges, opportunities, and perspectives. Lancet *383*, 1844–1859.

38. Stokman, M.F., Renkema, K.Y., Giles, R.H., Schaefer, F., Knoers, N.V., and van Eerde, A.M. (2016). The expanding phenotypic spectra of kidney diseases: Insights from genetic studies. Nat. Rev. Nephrol. *12*, 472–483.

39. Lachman, R.S., Rimoin, D.L., and Spranger, J. (1988). Metaphyseal chondrodysplasia, Schmid type. Clinical and radiographic delineation with a review of the literature. Pediatr. Radiol. *18*, 93–102.

40. Kleyner, R., Malcolmson, J., Tegay, D., Ward, K., Maughan, A., Maughan, G., Nelson, L., Wang, K., Robison, R., and Lyon, G.J. (2016). KBG syndrome involving a single-nucleotide duplication in *ANKRD11*. Cold Spring Harb. Mol. Case Stud. *2*, a001131.

41. Shi, L., Li, B., Huang, Y., Ling, X., Liu, T., Lyon, G.J., Xu, A., and Wang, K. (2014). "Genotype-first" approaches on a curious case of idiopathic progressive cognitive decline. BMC Med. Genomics *7*, 66.

42. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

43. Stessman, H.A., Xiong, B., Coe, B.P., Wang, T., Hoekzema, K., Fenckova, M., Kvarnung, M., Gerdts, J., Trinh, S., Cosemans, N., et al. (2017). Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. Nat. Genet. *49*, 515–526.

44. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H., Gorham, J., et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science *350*, 1262–1266.

45. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat. Genet. *49*, 1593–1601.

46. Jonquet, C., Shah, N.H., and Musen, M.A. (2009). The open biomedical annotator. Summit On Translat. Bioinforma. *2009*, 56–60.

47. Dhombres, F., and Bodenreider, O. (2016). Interoperability between phenotypes in research and healthcare terminologies–Investigating partial mappings between HPO and SNOMED CT. J. Biomed. Semantics *7*, 3.