ARTICLE | Genetics inMedicine

Novel phenotype–disease matching tool for rare genetic diseases

Jing Chen, PhD[1], Huan Xu, MS[2], Anil Jegga, DVM, MRes[1], Kejian Zhang, MD, MBA[3,4], Pete S. White, PhD[1,3] and Ge Zhang, MD, PhD[3,4]

Purpose: To improve the accuracy of matching rare genetic diseases based on patient's phenotypes.

Methods: We introduce new methods to prioritize diagnosis of genetic diseases based on integrated semantic similarity (method 1) and ontological overlap (method 2) between the phenotypes expressed by a patient and phenotypes annotated to known diseases.

Results: We evaluated the performance of our methods by two sets of simulated data and one set of patient's data derived from electronic health records. We demonstrated that the two methods achieved significantly improved performance compared with previous methods in correctly prioritizing candidate diseases in all of the three sets. Our methods are freely available as a web application (https://gddp.research.cchmc.org/) to aid diagnosis of genetic diseases.

Conclusion: Our methods can capture the diagnostic information embedded in the phenotype ontology, consider all phenotypes exhibited by a patient, and are more robust than the existing methods when phenotypes are incorrectly or imprecisely specified. These methods can assist the diagnosis of rare genetic diseases and help the interpretation of the results of DNA tests.

Genetics in Medicine (2019) 21:339–346; https://doi.org/10.1038/s41436-018-0050-4

Keywords: Human Phenotype Ontology; Mendelian disease; Diagnosis

INTRODUCTION

Although genotype-based clinical diagnosis for genetic diseases has recently gained success with the advances of clinical-exome sequencing technology and corresponding analytical methods, diagnosis remains a substantial challenge for many genetic diseases.[1] Considerable effort has been made to develop computer-aided clinical diagnostic systems based on phenotypic information from the patients.[2–4] Recently certain of these methods have been shown to facilitate differential diagnosis[2] and prioritization of candidate disease-associated genes.[5,6] Despite varying details, the underlying computational approaches supporting phenotype-based clinical diagnostics are largely similar, typically involving two main components: (1) a disease knowledgebase annotated by standard vocabularies or ontologies used to describe the phenotypic traits of different diseases, and (2) a computational or statistical method that predicts diagnosis by searching the knowledgebase for diseases that best match the phenotypes manifested in the patient.[4]

The Human Phenotype Ontology (HPO)[7] is a hierarchically structured term set to describe phenotypic traits in human diseases. With different levels of specificity, HPO is especially effective in annotating phenotypes for genetic disorders. Many public disease knowledgebases, such as MedGen[8] and Orphanet,[9] have adopted HPO as the standard vocabulary to annotate phenotypes for diseases.

Computational methods utilizing HPO for clinical differential diagnostics can be generally grouped into two types: semantic similarity-based methods, such as Phenomizer[4] and Disease Phenotypes,[10] which evaluate and rank phenotypic similarity between queries and hereditary diseases annotated by HPO. Alternatively, nonsemantic similarity-based methods, such as Bayesian ontology query algorithm (BOQA),[11] integrate ontological analysis with methods to compensate for noise, imprecision in query terms, and consideration of attribute frequencies using a Bayesian network model. Central challenges underlying these approaches include how to maximize utilization of the diagnostic information embedded in the phenotype ontology, consider all phenotypes exhibited by a patient, and maintain robustness when phenotypes are incorrectly or imprecisely specified.

Here, we utilize HPO[7] as the standard phenotype vocabulary, and MedGen[8] as the disease phenotype knowledgebase. We develop two computational methods to evaluate and rank the similarity between a set of query HPO terms and HPO terms annotated to a disease. The first method, based on

[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA; [2]Division of Biostatistics and Bioinformatics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA; [3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA; [4]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. Correspondence: Jing Chen (jing.chen2@cchmc.org) or Ge Zhang (ge.zhang@cchmc.org)

Submitted 10 October 2017; accepted: 18 April 2018
Published online: 12 June 2018

semantic similarity, integrates semantic similarities from multiple HPO terms in a query to prioritize diseases. The second method prioritizes diseases by evaluating the significance of the overlap between the HPO terms in the query with the HPO terms in all diseases in the disease knowledgebase. Using simulated patients as well as patient phenotypic data derived from electronic health records, we show that these two methods are superior in ranking candidate diseases compared with current computational approaches. We have implemented our methods as a user-friendly web-based application that is available for general use at https://gddp.research.cchmc.org/.

MATERIALS AND METHODS

Overview of the methods

Our methods take a patient's phenotypes coded by HPO terms (query terms, $Q = \{q_i, i \in \{1, \ldots, m\}\}$) as input and prioritize disease diagnosis based on HPO ontological similarity between the query terms and phenotype terms annotated to diseases. Multiple phenotypes may be annotated to a disease (D_k), denoted as $D_k = \{d_j, j \in \{1, \ldots, n_k\}\}$. For the current study, we utilized the phenotype annotations of 7036 OMIM diseases ($\mathcal{D}$) extracted from the National Center for Biotechnology Information (NCBI)'s MedGen resource[8] on 23 January 2017.

Two methods were developed and tested in this study. In method 1, ontological similarity between the query terms (Q) and phenotypes annotated to a disease (D_k) are calculated by integrating semantic similarities between HPO terms. In method 2, the similarity between the query terms (Q) and phenotypes annotated to a disease (D_k) is measured by the overlapping of HPO terms. The following sections explain the methods in detail.

Method 1: integrated semantic similarity

This method evaluates similarity between query terms Q and phenotypes annotated to a disease D_k using semantic similarities. This procedure involves (1) evaluating similarities between all pairs of phenotype terms, i.e., (q_i,d_j), and (2) calculating a similarity score to summarize the similarities between all the query terms (Q) and the HPO terms annotated to a target disease (D_k).

Semantic similarity between a query HPO term and a disease

Method 1a. Our first method evaluates semantic similarity between two HPO terms based on Resnik's method:[12]

$$\mathrm{sim}_a(t_1, t_2) = \mathrm{IC}(\mathrm{MICA}(t_1, t_2)), \qquad \text{where } \mathrm{IC}(t) = -\log(p(t))$$

MICA(t_1,t_2) is the most informative common ancestor of two HPO terms (t_1 and t_2) on the ontology. IC(t) = −(p(t)) is the information content of a phenotype term (t) in the MedGen database defined in the same way as in Resnik's method, which is the negative log frequency of the term.

Method 1b. The alternative method is based on the first method, but reduces the similarity between two terms to zero if the two terms are not on the same lineage in HPO ontology, to emphasize the difference between distinct lineages. Two terms are on the same lineage if one term is an ancestor of the other term. A similar method was used in GeneMANIA[13] to create negative gene list based on Gene Ontology functional annotations. Formally, we define the similarity as:

$$\mathrm{sim}_b(t_1, t_2) = \begin{cases} \mathrm{IC}(\mathrm{MICA}(\mathrm{t}_1, \mathrm{t}_2)), \text{if } \mathrm{t}_1 \text{ and } \mathrm{t}_2 \text{ are on the same lineage} \\ 0, \text{otherwise} \end{cases}$$

For both methods 1a and 1b, the "best match" between each query HPO term (q_i) and the HPO terms annotated to an OMIM disease in MedGen ($d_j \in D_k$) is selected to represent the "similarity score" between a query term q_i and the disease (D_k):

$$s_{ik} = \max_{d_j \in D_k} (\mathrm{sim}(q_i, d_j))$$

Integration of semantic similarities of multiple query HPO terms

We used a Fisher's method–based procedure, similar to the framework used in ToppGene,[14] to summarize the semantic similarities between a set of query terms (Q) and a disease (D_k). First, the semantic similarity score between a query HPO term and a disease (s_{ik}) is converted to a nominal p value according to its rank within all diseases ($\mathcal{D}$, $N = 7036$):

$$p_{ik} = \frac{N - \mathrm{rank}(s_{ik}) + 1}{N}$$

This p value can be interpreted as when comparing a query term q_i against all diseases ($\mathcal{D}$), the proportion of diseases with a higher semantic similarity score than the one observed between the query term (q_i) and the disease (D_k). The p value measures how specific an HPO term (q_i) is to a disease (D_k) when compared with all other diseases. The p values between multiple query HPO terms (Q) and a disease (D_k) are then combined using Fisher's method as the overall similarity score between the query terms and a disease.

$$S_k = -2 \sum_{s_{ik} > \gamma} \ln(p_{ik})$$

As query terms that are observed for multiple genetic diseases containing decreasing diagnostic information content for a disease (when the semantic similarity [s_{ik}] is low), only those p_{ik} whose corresponding s_{ik} is greater than or equal to a certain semantic similarity cutoff (γ) are combined together.

Method 2: weighted overlapping

In this method, the phenotypes of a patient (query terms, Q) and HPO terms annotated to diseases (D_k) are first "up-induced" based on HPO tree structure so that if an HPO term is annotated to a patient/disease, all of its ancestors are also annotated to the patient/disease. To compare the query terms (Q) with the terms annotated to a disease (D_k), we can construct a weighted 2×2 contingency table (Table 1) that

Table 1 Weighted 2 × 2 contingency table between query and a disease

	D	$\overline{D}$
Q	$a = \sum_{t\in(Q\cap D)} \mathrm{IC}(t)$	$b = \sum_{t\in(Q\cap\overline{D})} \mathrm{IC}(t)$
$\overline{Q}$	$c = \sum_{t\in(\overline{Q}\cap D)} \mathrm{IC}(t)$	$d = \sum_{t\in(\overline{Q}\cap\overline{D})} \mathrm{IC}(t)$

The values in this Table are the counts of HPO terms weighted by their information content (IC). a: the count of terms in *Q* and *D* (true positive); b: the count of terms in *Q* but not in *D* (false positive); c: the count of terms in *D* but not in *Q* (false negative); and d: the count of terms not in *Q* and not in *D* (true negative)

contains the weighted counts of HPO terms shared or not shared between the query terms and the terms annotated to a disease. A Fisher's exact test similar to that employed by Alexa and colleagues[15] is then applied to this 2 × 2 contingency table and the *p* value from the test can be used to rank the concordance/discordance between the query terms and the phenotypes of a disease.

Implementation

All analysis was implemented using the R platform.[16] Ontology-related manipulation and similarity measure was implemented based on Bioconductor packages dnet[17] and ontologyIndex.[18] Fisher's method to integrate multiple *p* values is available in R package metap.[19] The NOBLE coder[20] program was used for HPO Concept Recognition and integrated with R script by rJava.[21] An interactive web application that implemented our methods was developed using shiny.[22]

For comparison purposes, we also implemented the best-match average combination method (BM.ave) as used in Phenomizer[4] and the Bayesian ontology query algorithm (BOQA)[11] using R, according to the description by Bauer and colleagues,[11] which does not consider the phenotype frequency information for each disease.

Evaluation

We used simulated cases as well as real patient data to evaluate the performance of our methods. We also compared the performance of our methods with the current methods, BM.ave (Phenomizer) and BOQA.

Generation of simulated cases

Diseases and associated HPO annotations from Orphanet[9] were used to create simulated patients. Simulated patients were created based on the Orphanet data downloaded on 3 February 2017, which contains 2536 diseases. For each disease represented in Orphanet, associated phenotypes and the prevalence of each phenotype is provided by a frequency term (i.e., excluded, very rare, occasional, frequent, very frequent, and obligate). We converted these terms into numeric probability values (Supplemental Table 1).

A multistep procedure was applied to generate simulated patients with controlled noise level. In the first step, for each of these 1775 Orphanet diseases that can be mapped to at least one OMIM ID, 5 patients were created with HPO terms according to their occurrence probabilities provided by Orphanet. In the second step, HPO terms ("false negative") were randomly removed from each patient at a fixed probability β. In the third step, we randomly inserted HPO terms ("false positive") to each patient according to their relative frequencies in Orphanet diseases. The expected number of HPO terms to be added for each patient was a constant α. In the last step, if more than 6 HPO terms were present in a patient, a random subset of 6 HPO terms was selected. Patients with only one phenotype were ignored. This procedure is similar to those used in Phenomizer[4] and BOQA[11] to create simulated cases with noise.

Extraction of patient phenotypic data from electronic health records

De-identified patient data was obtained from the i2b2 database (Informatics for Integrating Biology and the Bedside, https://i2b2.cchmc.org/) at the Cincinnati Children's Hospital Medical Center (CCHMC). Patients whose records were assigned one or more ICD-10 codes representing an OMIM disease in their diagnosis (based on the International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes for OMIM diseases downloaded from Orphanet) were extracted from the database. Phenotype descriptions were originally coded either as ICD-10 codes or free text, and were converted to HPO terms by the NOBLE coder.[20]

Performance evaluation

For each simulated or real patient, the corresponding set of HPO terms was used as the query input for the computational models. The diagnosis was considered correct if the actual disease was ranked in the top 1, 2, 3, or up to 10 among all diseases, depending on different levels of specificity. To summarize the performance, we plotted receiver operating characteristic (ROC) curves. Specifically, sensitivity was defined as the proportion of "true diagnosis" that is ranked above a particular threshold (e.g., top 10), and specificity as the percentage of diseases ranked below the threshold. The area under the ROC curve (AUC) was calculated. This "full" ROC curve, however, is not very informative for the high specificity range, which is of particular interest in evaluating diagnostic performance.[23] For example, to make the prediction useful for clinical applications, we are more interested in the top 10 predicted diagnosis among the possible 7036 diseases in the reference database, which corresponds to a specificity close to 99.86%. Therefore, we also plotted the "partial" ROC with cutoff ranking up to 10, which corresponds to the specificity range [0.9986, 1]. The partial area under the ROC curve (pAUC)[24] of the same range was calculated to evaluate the performance of different methods.

Table 2 Results of different methods for simulated sets 1 and 2

	Simulated set 1					*Simulated set 2*				
Rank cutoff	*Method 1a*	*Method 1b*	*Method 2*	*BOQA*	*BM.ave*	*Method 1a*	*Method 1b*	*Method 2*	*BOQA*	*BM.ave*
1	38.7%	38.9%	45.8%	39.4%	2.6%	31.9%	32.6%	38.0%	31.2%	2.2%
3	51.9%	52.5%	58.0%	51.0%	8.3%	46.0%	47.0%	49.7%	42.2%	6.1%
10	65.6%	65.7%	69.4%	62.6%	26.2%	59.4%	59.8%	62.4%	54.3%	20.9%
pAUC	7.10e–4	7.16e–4	7.79e–4	6.93e–4	1.74e–4	6.32e–4	6.42e–4	6.82e–4	5.81e–4	1.35e–4

The numbers in this table represent the correct diagnosis rates with different ranks as cutoffs. For example, for simulated set 1, at rank 10, the correct diagnosis rate for method 2 is 69.4% for the 8798 simulated patients. For both methods 1a and 1b, the results are based on semantic similarity cutoff 1.0
BOQA Bayesian ontology query algorithm, *pAUC* partial area under the ROC curve, *BM.ave* best-match average combination method

RESULTS

Performance based on simulated patients

Simulated patients

Using Orphanet, we simulated 5 patients for each of the 1775 diseases represented in the database. Two sets of patients with different noise levels were created (see Methods for details). The first set was created with a probability of 0.1 for removing any HPO term and on average inserting 2 random HPO terms into each patient record (i.e., $\alpha = 2$, $\beta = 0.1$). The second set was created with a higher noise level (i.e., $\alpha = 3$, $\beta = 0.2$). The predicted diagnosis was considered correct if the actual disease of the patient was ranked within or equal to the cutoff (from top 1 to 10 in our evaluation).

The effect of semantic similarity cutoff γ on method 1

Our methods 1a and 1b only consider query terms that have semantic similarity score with a disease larger than a cutoff (γ) (see Methods for details). Therefore, we first studied the impact of different cutoff γ on the performance. In both simulated sets, the best correct diagnosis rates were obtained when $\gamma = 1.0$ (Supplemental Table 2). The correct diagnosis rates and pAUC scores were lower in simulated set 2 as expected, because this set contains higher noise. To test the robustness of similarity cutoff $\gamma = 1.0$, we included a third simulation set with higher noise ($\alpha = 4$ and $\beta = 0.3$) and the result (Supplemental Table 2) suggested that the same $\gamma = 1.0$ gave the best performance.

Comparing performance between methods 1a and 1b

Next, we tested method 1b, which disregards the semantic similarity between two HPO terms to zero if they do not arise from the same lineage. As can be seen from Table 2, method 1b performed slightly better in both sets of simulated patients. For simplicity, this table only shows the results for $\gamma = 1.0$.

Comparing performance between methods 1b and 2 with existing methods

The correct diagnosis rates and the pAUC scores for the two simulated sets are summarized in Table 2. The corresponding partial ROC curves are displayed in Fig. 1a,b. The full ROC curves and their AUC are plotted in Supplemental Fig. 1. Although our method 1b and method 2 are quite different, their performance were comparable in both simulated data sets, and both methods resulted in improved performance compared with the existing methods BM.ave and BOQA. In the less noisy simulation set 1, method 2 had an improved diagnostic rate of 45.8 vs. 39.4% for BOQA and 2.6% for BM.ave at order rank 1, and 69.4 vs. 62.6% (BOQA) and 55.3% (BM.ave) at rank ≤10 (Table 2). The p values of the improvement at rank 10 were very significant ($<1.0 \times 10^{-10}$) by McNemar test (Supplemental Table 3). Similar improvements in performance were observed for the noisier simulation set 2. As simulated set 2 represents a set of patient phenotypes with higher noise, this suggested our methods are more robust for noisy queries.

Evaluation using phenotypic data from electronic health records

To evaluate performance using real patient data, we selected 10 ICD-10 codes representing 10 OMIM diseases to query our institutional electronic health records. A list of the 10 OMIM diseases is shown in Table 3. The number of patients for each disease ranged from 4 for toxic epidermal necrolysis to 232 for double outlet right ventricle yielding 462 patients in total. The numbers of HPO terms for each patient ranged from 1 to 65, with a median value of 12. This data set was extracted directly from the clinical records research data warehouse without manual inspection or curation.

We then applied the four computational methods (method 1b with $\gamma = 1.0$, method 2, BOQA, and BM.ave) on all 462 patients. The correct diagnosis rates at rank 10 for different methods are summarized in Table 3, and the corresponding partial ROC curves are displayed in Fig. 1c. The full ROC curves and their AUC are plotted in Supplemental Fig. 1. Overall, method 1 outperformed method 2 (32.5 vs. 28.1%). Both methods 1 and 2 outperformed either BOQA (19.7%) or BM.ave (4.1%). The performance improvement at rank 10 was very significant ($<1.0 \times 10^{-10}$) (Supplemental Table 3). For all four methods, the correct diagnosis rates were much lower than for the simulated data sets, suggesting much higher noise levels in the electronic health records. This complexity could be caused by multiple comorbidities present in the patients, adverse events from treatment, or inaccurate mapping from ICD-10 codes to HPO terms. Therefore we evaluated the effect of number of phenotypes in the patient on the performance of different methods (Supplemental Fig. 2).

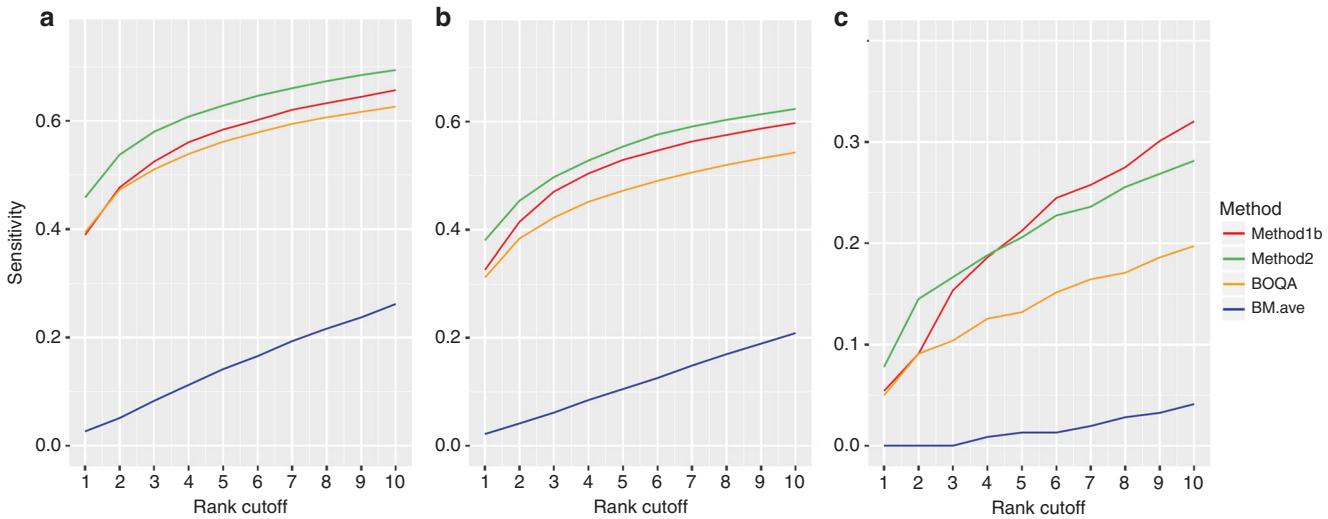

Fig. 1 The partial receiver operating characteristic (ROC) curves of all methods for simulated patients and patient data from electronic health records. **a**, **b**, and **c** correspond to the partial ROC curves for simulated patient set 1, simulated patient set 2, and real patient data derived from electronic health records respectively. The *x*-axis of the plot is the rank cutoff for correct diagnosis ranging from 1 to 10. The *y*-axis of the plot is the proportion of patients with correct diagnosis at the rank cutoff

Based on the result, the performance of methods 1 and 2 remained stable for patients with many phenotypes, while the performance of BM.ave and BOQA peaked for patients with 5 to 15 phenotypes and deteriorated fast when the number of phenotypes increased.

Implementation

We implemented both of our methods (methods 1 and 2) in a web-based application called GDDP (Genetic Disease Diagnosis based on Phenotypes), freely available at https://gddp.research.cchmc.org. This application takes a set of HPO terms or free text describing a patient's clinical phenotypes as input, and ranks disease diagnosis using either method 1 or 2. The output of the application is a list of diseases, sorted by the similarity between patient's phenotypes and phenotypes annotated to diseases (Fig. 2a). The application also generates interactive plots to demonstrate the detailed similarity map between the query HPO terms and the HPO terms annotated to a candidate disorder (Fig. 2b). Such plots can provide valuable information to guide the further differential diagnosis. In the example in Fig. 2, the diagnosis was supported by the partial matching (light blue line) between "cerebellar atropy" (query term) and "cerebellar cortical atrophy" and perfect matching (red lines) of several other HPO terms. More specific clinical examination for symptoms like "cerebellar cortical atrophy," "limb ataxia," etc. will further confirm or revoke the diagnosis.

DISCUSSION

Diagnosis of human disease is challenging because patients often manifest many phenotypic symptoms of varying specificity, and the cooccurrence of these symptoms may not always be recognizable in known syndromes. There has been considerable effort to develop more accurate and comprehensive methods for predicting disease diagnosis from patient phenotypes. As an example, Monarch Initiative[25] leverages large-scale integration of multiple phenotype data sources across many model organisms to collectively achieve better inference. In this study, we limit ourselves to disease and HPO annotations from the MedGen database and focus on computational methods to prioritize disease diagnosis. Using simulated and patient-based phenotypic data, we demonstrate that our methods outperform two current methods, the best matching average (BM.ave, the algorithm used by Phenomizer) and the Bayesian ontology query algorithm (BOQA).

Using simulated data, both of our methods achieved a correct diagnosis rate of 60% (at rank 10) and were more accurate than either the BM.ave or BOQA algorithms. To assess performance for clinical cases, diagnoses and associated phenotypes derived from electronic health records for 10 OMIM diseases were used. For these real cases, the correct diagnosis rates of all methods dropped substantially, likely due to increased levels of noise. Nevertheless, both of our methods (~30% correct at rank 10) performed substantially higher than the two current methods, each of which performed at rates below 20%.

Our method 1 employs a framework to integrate semantic similarities of multiple HPO terms to prioritize disease diagnosis. By converting similarity scores to a *p* value based on ranking among all diseases, our approach has three advantages over the current averaging method: (1) it provides a more straightforward way to interpret a similarity score as "specificity" of a phenotype pertaining to a disorder, (2) it enables combination of "specificity" for multiple query terms based on Fisher's method, and (3) by converting the original similarity scores into rank-based *p* values, the method is more robust to extreme values. In

Table 3 Results of different methods for patients from electronic health records

ICD10	*DX_DESCRIPTION*	*MIM ID*	*# patient*	*Method 1b (%)*	*Method 2 (%)*	*BOQA (%)*	*BM.ave (%)*
L51.2	Toxic epidermal necrolysis	608579	4	50.0	50.0	100.0	0.0
Q20.6	Left atrial isomerism	208530	8	25.0	87.5	25.0	0.0
E76.1	Hunter disease	309900	14	35.7	42.9	35.7	0.0
G12.0	Werdnig–Hoffmann disease	253300	36	16.7	22.2	22.2	0.0
G90.1	Dysautonomia, familial	223900	11	9.1	27.3	18.2	0.0
Q79.4	Eagle–Barrett syndrome	100100	38	57.9	68.4	63.2	0.0
I82.0	Budd–Chiari syndrome	600880	31	22.6	12.9	12.9	3.2
D59.3	Hemolytic–uremic syndrome	235400	49	12.2	12.2	20.4	0.0
Q76.0	Spina bifida occulta	600145	39	23.1	28.2	28.2	2.6
Q20.1	Double outlet right ventricle	217095	232	38.8	24.6	9.1	7.3
	Overall			32.5	28.1	19.7	4.1

The numbers in this table represent the correct diagnosis rates at rank 10. For method 1b, the results are based on semantic similarity cutoff 1.0
BOQA Bayesian ontology query algorithm, *BM.ave* best-match average combination method

addition to Resnik's method[12] (method 1a), we propose an alternative approach to evaluated semantic similarity between HPO terms. This method (method 1b) disregards similarity between two terms if they are not on the same lineage to account for reduced relatedness between different phenotypic lineages. Within the parameters of our evaluation, our results indicate that this semantic similarity measure is superior to the conventional Resnik method for predicting diagnosis. We also show that excluding terms of low information content (i.e., using certain semantic similarity cutoff, γ) improves diagnostic accuracy.

Our method 2 utilizes a weighted Fisher's exact test to evaluate the concordance/discordance between the query terms and the phenotypes annotated to a disease. This method captures the similarity between a query and a disease by overlapping "up-induced" HPO terms weighted by their information content. In contrast to our method 1 and other semantic similarity-based methods, this approach also considers information of "dissimilarity" in diagnosis.

Our methods are more robust than BM.ave and BOQA when the number of query phenotypes is large (Supplemental Fig. 2). Our method 1 is similar to BM.ave, but instead of using a symmetric similarity scheme (equation 2 of ref. [4]), our method 1 only considers similarities based on query to disease matches (equation 1 of ref. [4]). The inclusion of disease to query matches will generate substantial noise when the number of query phenotypes is large. We also applied a similarity cutoff γ to reduce noise due to noninformative matches of disease nonspecific phenotypes. On the other hand, BOQA requires a predefined constant false positive rate (α) and false negative rate (β) grid uniform prior. When the number of phenotypes is large, it is likely this prior is inappropriate and the performance decreases. Our method 2 evaluates the concordance as well as the discordance between the query terms and the phenotypes annotated to a disease, and therefore is robust to the disease nonspecific phenotypes because the noise introduced by the concordant matches of nonspecific phenotypes can be canceled out by the discordant matches of the nonspecific phenotypes.

Although both of our methods were effective in our evaluation, they have certain limitations. Each model relies on statistical tests that assume independence among features (i.e., query terms), which is an assumption that is not strictly true for phenotypes. Therefore the significance measures estimated by the methods are not quantitatively accurate. To improve significance estimation, a strategy similar to BOQA,[11] which incorporates frequency of phenotypes (for better modeling of incomplete penetrance of phenotypes) in the diagnostic model, can be used. However, this would require quantitatively accurate annotation of disease phenotypes (i.e., disease prevalence, variable expressivity of the same pathogenic variant), which is still sparse in most disease knowledgebases.

Our proposed methods (and other similar ones) are trying to match a patient's phenotypes to a reference knowledgebase that annotates the phenotypes of different diseases. The accuracy of these methods is therefore primarily dependent upon the quality of patient phenotyping as well as the accuracy and comprehensiveness of phenotypic annotations of disorders in the reference databases (e.g., MedGen and Orphanet). The phenotyping of the patient should be accurate (using the right terms) and precise (using specific terms with high information content whenever possible). The comprehensiveness of both patient phenotyping and the complete coverage of phenotypic abnormalities in the reference databases is also important as these tools usually integrate diagnostic information from all phenotypic features. Recent efforts to also include lab tests and cellular phenotype terms in HPO should substantially increase the power of these tools in clinical diagnosis. In addition, as patients often have incomplete penetrance and variable expressivity of different phenotypes, it is also important to include this information in the diagnosis (as described in the "Frequency" and "Clinical modifier" branches of HPO). For patients with family history information or genetic data of candidate pathogenic variants, the consideration of "Mode of Inheritance" will also help.

Computational analysis of phenotype data remains challenging because patient disease phenotypes are usually

a

Rank	OMIM	OMIM_name	Score	Plot	Gene
1	248800	MARINESCO-SJOGREN SYNDROME	102.7	View	SIL1
2	607426	COENZYME Q10 DEFICIENCY	63.77	View	COQ2
3	615356	MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE 2S	54.94	View	TRAPPC11
4	604168	CONGENITAL CATARACTS, FACIAL DYSMORPHISM, AND NEUROPATHY	54.54	View	CTDP1

b

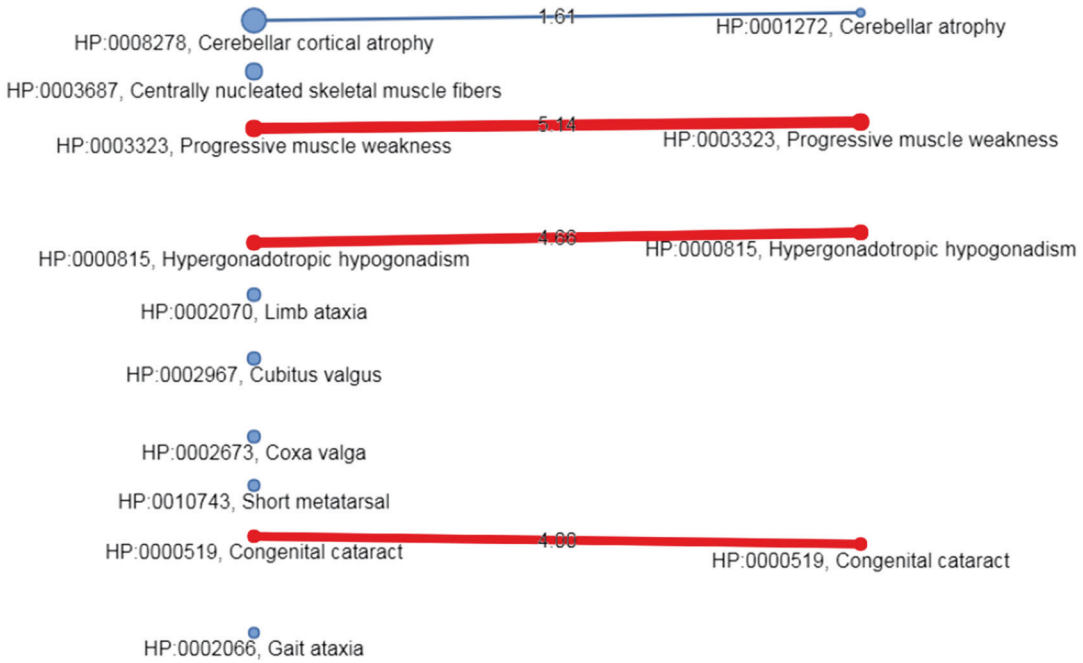

Fig. 2 Screen shots of the diagnostic reports generated by GDDP (computational Genetic Disease Diagnosis based on Phenotypes). a A list of candidate diagnoses ranked by similarity between a patient's phenotypes and phenotypes annotated to diseases. **b** Matching between the query Human Phenotype Ontology (HPO) terms (right side) and the HPO terms annotated to a candidate disorder (left side). The HPO terms annotated to a candidate disorder are sorted by their information content (also shown by the diameter of the dots). Perfect matches between a query and an HPO term annotated to the candidate disorder are highlighted in red. Partial matches are shown in blue. The numbers on the matching lines are the "similarity scores" as defined in the Methods section

incomplete and noisy. While HPO provides a structured vocabulary to relate all phenotypic terms, it does not explicitly link these terms to the genetic cause of disorders. In this study, we introduced new computational methods and demonstrated that these methods generally outperformed prior approaches. These initial findings await a more systematic exploration of how and why our methods relate to current approaches.

It's foreseeable that further improvements may be achievable by integrating multiple complementary information, such as mode of inheritance, genetic variants detected through diagnostic testing,[2,26] or associated phenotypic annotations derived from animal models.[27] While our results show promising improvement as decision aids, much additional experimentation is necessary to achieve prioritization algorithms that are sufficiently accurate to be considered as a first-line reasoning approach in a diagnostic setting.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0050-4) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Alka Chandel, Parth Divekar, and Diana Epperson for helping to query and organize clinical data from the i2b2 database. This study is partially funded by the Center for Pediatric Genomics, Cincinnati Children's Hospital Medical Center, and National Institutes of Health (NIH) grant U01 HG008666.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*. 2013;369:1502–11.
2. Zemojtel T, Kohler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*. 2014;6:252ra123.
3. Alves R, Pinol M, Vilaplana J, et al. Computer-assisted initial diagnosis of rare diseases. *PeerJ*. 2016;4:e2211.
4. Kohler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009;85:457–64.
5. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med*. 2015;7:81.
6. Masino AJ, Dechene ET, Dulik MC, et al. Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the Human Phenotype Ontology. *BMC Bioinformatics*. 2014;15:248.
7. Kohler S, Vasilevsky NA, Engelstad M, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res*. 2017;45(D1):D865–76.
8. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2017;45(D1):D12–7.
9. Orphanet: an online database of rare diseases and orphan drugs. 1997; http://www.orpha.net. Accessed 10 June 2018.
10. Hoehndorf R, Schofield PN, Gkoutos GV. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Sci Rep*. 2015;5:10888.
11. Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics*. 2012;28:2502–8.
12. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. *Int Joint Conf Artif*. 1995:448-53. Proceedings of the 14th International Joint Conference on Artificial Intelligence (Morgan Kaufmann, San Francisco), Vol 1, pp 448–453.
13. Mostafavi S, Ray D, Warde-Farley D, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008;9(suppl 1):S4.
14. Chen J, Xu H, Aronow BJ, et al. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinform*. 2007;8:392.
15. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006;22:1600–7.
16. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2016).
17. Fang H, Gough J. The 'dnet' approach promotes emerging research on cancer patient survival. *Genome Med*. 2014;6:64.
18. Greene D, Richardson S, Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*. 2017;33: 1104–1106.
19. Michael D. metap: meta-analysis of significance values. Rpackage version 0.8. 2017.
20. Tseytlin E, Mitchell K, Legowski E, et al. NOBLE—flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinform*. 2016;17:32.
21. Simon U. rJava: Low-Level R to Java Interface. R packageversion 0.9-9, 2017. https://CRAN.R-project.org/package=rJava.
22. Winston C, Joe C, JJ Allaire, et al. shiny: Web Application Framework for R. R packageversion 1.0.5., 2017. https://CRAN.R-project.org/package=shiny.
23. Ma H, Bandos AI, Rockette HE, et al. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat Med*. 2013;32:3449–58.
24. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making*. 1989;9:190–5.
25. Mungall CJ, McMurry JA, Kohler S, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2017;45(D1):D712–22.
26. Trakadis YJ, Buote C, Therriault JF, et al. PhenoVar: a phenotype-driven approach in clinical genomics for the diagnosis of polymalformative syndromes. *BMC Med Genomics*. 2014;7:22.
27. Robinson PN, Kohler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014;24:340–8.