Database, 2022, 1–10
DOI: https://doi.org/10.1093/database/baac038
Original article

Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders

T.M. Yates[1,2], A. Lain[3], J. Campbell[1,4], D.R. FitzPatrick[1,2,4] and T.I. Simpson[3,4,*]

[1]MRC Human Genetics Unit, Western General Hospital, Institute of Genetics and Cancer, The University of Edinburgh, Crewe Road South, Edinburgh EH4 2XU, UK
[2]Transforming Genetic Medicine Initiative, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3]Institute for Adaptive and Neural Computation, Informatics Forum, The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
[4]Simons Initiative for the Developing Brain, The University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XF, UK

[*]Corresponding author: Tel: +44 (0)131 6515637; Email: ian.simpson@ed.ac.uk

Citation details: Yates, T., Lain, A., Campbell, J. *et al.* Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders. *Database* (2022) Vol. 2022: article ID baac038; DOI: https://doi.org/10.1093/database/baac038

Abstract

There are >2500 different genetically determined developmental disorders (DD), which, as a group, show very high levels of both locus and allelic heterogeneity. This has led to the wide-spread use of evidence-based filtering of genome-wide sequence data as a diagnostic tool in DD. Determining whether the association of a filtered variant at a specific locus is a plausible explanation of the phenotype in the proband is crucial and commonly requires extensive manual literature review by both clinical scientists and clinicians. Access to a database of weighted clinical features extracted from rigorously curated literature would increase the efficiency of this process and facilitate the development of robust phenotypic similarity metrics. However, given the large and rapidly increasing volume of published information, conventional biocuration approaches are becoming impractical. Here, we present a scalable, automated method for the extraction of categorical phenotypic descriptors from the full-text literature. Papers identified through literature review were downloaded and parsed using the Cadmus custom retrieval package. Human Phenotype Ontology terms were extracted using MetaMap, with 76–84% precision and 65–73% recall. Mean terms per paper increased from 9 in title + abstract, to 68 using full text. We demonstrate that these literature-derived disease models plausibly reflect true disease expressivity more accurately than widely used manually curated models, through comparison with prospectively gathered data from the Deciphering Developmental Disorders study. The area under the curve for receiver operating characteristic (ROC) curves increased by 5–10% through the use of literature-derived models. This work shows that scalable automated literature curation increases performance and adds weight to the need for this strategy to be integrated into informatic variant analysis pipelines.

Database URL: https://doi.org/10.1093/database/baac038

Introduction

The use of genome-wide sequencing technologies combined with rational frequency, inheritance and consequence-based variant filtering strategies has transformed the diagnosis of genetically determined developmental disorders (GDD) (1–4). Although filtering efficiently reduces the number of genetic variants for diagnostic consideration, each of these have to be reviewed to determine if the clinical features (phenotype) of the individual being tested can be explained by one or more of these genotypes. This usually requires manual review of peer-reviewed case reports/series that describe relevant genotype/phenotype associations (5). The matching is based on expert opinion as few metrics exist to rank the associations with any statistical rigor.

Phenotypic data in the literature are not usually recorded in a standardized, computationally tractable format. Plain text descriptions of similar clinical features may be recorded in several different ways. For example, a technical term such as 'hypertelorism', may be recorded as its synonym 'widely spaced eyes'. In addition, case reports are found across a wide range of journals, with different structures and file formats for each publication.

The Human Phenotype Ontology (HPO) was developed to store phenotypic data in a computationally accessible format (6). Several initiatives have been developed to link diseases to phenotype data, in the form of HPO terms, for example, Online Mendelian Inheritance in Man (OMIM) (7) and Orphanet (8). However, these rely on manual expert curation and therefore are not inherently scalable and cannot be updated automatically.

Methods of extracting phenotype data from text at scale previously have relied on abstracts or open access

Received 2 November 2021; Revised 26 March 2022; Accepted 25 May 2022

papers (9, 10). At the time of writing, Europe PubMed Central (EPMC, https://europepmc.org/) contained approximately 39.5 million articles, of which only 3.8 million were open access. Therefore, there is likely a significant volume of phenotypic data that has not been used previously.

Our overall aim is to create systems that allow scalable, automated and clinically orientated literature curation to aid the robustness of diagnosis through genomic testing. Here, we present a method for creating disease models describing GDD, comprising lists of HPO terms, with weighting according to term frequency in the literature. Utilizing intellectual property law for research in the UK https://www.gov.uk/guidance/exceptions-to-copyright, we retrieve the full text of almost all relevant case reports and extract phenotypic data mapped to HPO terms for a set of GDD. We evaluate this against prospectively gathered patient phenotypes from the Deciphering Developmental Disorders (DDD) study (1) and compare to current widely used manually curated sources.

Materials and methods

Full-text retrieval

A test set of 99 GDD defined in the Developmental Disorders Genotype2Phenotype (DDG2P) database (4) were selected, to include conditions well-represented in the DDD study (1) and in OMIM (7). For each of these, a literature review was undertaken using PubMed searches. The initial search was by gene symbol in title—{gene symbol}[TI]. If this returned less than 300 results, an abstract review was undertaken to identify relevant case reports. If the initial search returned more than 300 results, modifier terms were added such as {gene symbol}[TI] AND {syndrome name} or {gene symbol}[TI] AND ‘intellectual disability’. Only papers that described case reports for variants in a single gene were included.

From this process, a list of case report PubMed IDs (PMIDs) was generated for each of the 99 diseases. These were inputted into the Cadmus full-text retrieval package (https://doi.org/10.5281/zenodo.5618052), which will be described in detail in a forthcoming publication. In brief, metadata obtained using each PMID was used to send requests for download for each paper to sources that authorize full-text retrieval for research purposes. Multiple sources were used to maximize the chances of download, including, but not limited to, Crossref, doi.org and EPMC. File formats generated through this include Hypertext Markup Language (HTML), eXtensible Markup Language (XML), Protable Document Format (PDF) and plain text. Where multiple formats were retrieved, a series of quality assessments were used to identify the best full-text version. The text was cleaned and converted to a string. The abstract and references were parsed out. This final document was used for all following steps and will be defined as the ‘full text’ in subsequent paragraphs.

Phenotype extraction

Phenotypic features in text were identified and mapped to Unified Medical Language System (UMLS®) concept unique identifiers (CUIs) using the 2018 release of MetaMap (11). The source vocabulary for MetaMap was restricted to CUIs corresponding to the HPO (6), which is Category 0 under the UMLS Metathesaurus® licence. This version of MetaMap used the 2018_07_23 version of the HPO. MetaMap includes negation information for each phenotypic feature identified; the frequency of each non-negated feature in the text was used for term weighting. Negated terms, whilst potentially useful to identify phenotypic features that are not associated with a given disorder, were not included in this work as the comparison DDD dataset does not include these. CUIs were then mapped to HPO terms, using the mappings in the UMLS Metathesaurus® (Release 2020AA).

MetaMap performance evaluation

To test the performance of MetaMap (11), a set of 50 papers—randomly selected from the list of full-text downloads described above—were manually annotated. Non-negated phenotypic features in the text were annotated directly to HPO terms Against this standard, MetaMap was evaluated for precision (fraction of true positive terms in output) and recall (fraction of true positive terms compared to all true terms in full text), using a variety of usage options. The F1 score (harmonic mean of precision and recall) was also calculated.

Disease model creation

For each group of PMIDs corresponding to a given disease, extracted HPO terms and their frequency were aggregated to create a ranked, weighted disease model. Corresponding diseases in manually curated sets were mapped using the disease MIM identifier, and the HPO annotated file genes_to_phenotype file downloaded from http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt on 22 April 2021. This includes disease-specific HPO term lists (models) from OMIM (mim2gene) (7) and Orphanet (8). OMIM-derived terms mostly do not include frequency/weighting, whereas Orphanet terms are uniformly annotated. The frequency in both cases was recorded as an HPO frequency term, e.g. Very frequent (HP:0040281), present in 80–99% of the cases. OMIM models corresponding to the 99 G2P/DDD disease set were used for the majority of analyses. A subset of 43 Orphanet models were used for weighted analyses, as the Orphanet annotations were not available for the full set. HPO term models from the DDD study (1) were created using aggregated lists from probands with diagnoses corresponding to the 99 disease set. These were recorded prospectively by clinicians recruiting individuals to the study with a suspected, undiagnosed GDD.

Disease model evaluation

Evaluation of literature-derived models through comparison with DDD models was designed to assess their similarity to real life, clinical, prospectively gathered data. OMIM models were used as an example of widely used manual curation (7). Two similarity metrics were used: rank-biased overlap (RBO) (12) and semantic similarity using information content (IC) as defined by Resnik (13).

RBO is a method of comparing ranked lists that is top-weighted, can compare lists that contain differing members and is monotonic with increasing depth of list (12). RBO allows for the top-weightedness parameter p to be fine-tuned to weight the score more or less towards higher-ranked items in the list. For this analysis, p was equal to 0.98, weighting towards the top 50 terms in a list. RBO may be expressed as a min–max range; however, for this work, the extrapolated RBO_{EXT} point score was used for ease of comparison. RBO was calculated for all models in the literature-derived set vs the DDD set, OMIM vs DDD and literature vs OMIM (1, 7). Literature- and DDD-derived models were ranked according to term frequency. OMIM model terms are not generally annotated with a frequency; the frequency of each term across all OMIM models was used for ranking.

For the semantic similarity measure, the HPO terms in both comparison datasets, e.g. literature vs DDD, were used to calculate the IC for each term following the method used by Helbig *et al.* (14). If f is the number of diseases annotated with an HPO term g and n is the total number of diseases, the IC_g is defined as—$\log_2(f/n)$ (15). The most informative common ancestor (MICA) of two terms is the parent term in the ontology with the highest IC. For a disease-disease comparison, a matrix m is created with HPO terms of one disease (l terms) as the rows and the terms of the other (k terms) as the columns. Each position in the matrix (m_{ij}) is a comparison between pairs of HPO terms and is populated with the MICA for that pair. The similarity score between diseases is computed by summing the average of the rows and the columns, with a normalization measure (14).

$$sim(D_1,D_2) = \frac{1}{2}\Big(\frac{1}{l}\sum_{j=1}^{l}\max_{1\le i\le k} m_{ij} + \frac{1}{k}\sum_{j=1}^{k}\max_{1\le i\le l} m_{ji}\Big)$$

We calculated the semantic similarity between literature-dervied, DDD, and OMIM sets using unweighted models.

A comparison of weighted models using the Orphanet (8) subset was undertaken; however, there was no straightforward method of normalizing the frequency weightings between datasets. For the Orphanet models, which consisted of a flat list of phenotype terms annotated with HPO frequency terms, the percentage range in HPO frequency annotation was mapped to the mean of this range. For example, a term annotated as Very frequent with the range 80–99% was mapped to 89. Each term in a disease list was repeated according to its frequency mapping, thereby creating a weighted model. For literature-derived and DDD models (1), frequency annotations were binned into four bins using numpy.histogram, corresponding to HPO terms Very frequent, Frequent, Occasional and Very rare. Frequency weighting was then applied as per the Orphanet models. For models weighted in this manner, l row terms and k column terms in the disease comparison matrix m therefore may contain repeats, with the MICA sum average altering accordingly (16). Comparisons were calculated for the 43 diseases in the Orphanet set vs DDD models and for the corresponding 43 literature-derived models vs DDD.

Results

Full-text papers describing GDD were downloaded, phenotypic features extracted and weighted disease models constructed. These were evaluated against data from the DDD study (1) and manually curated models. An overview of this process is shown in Figure 1.

Full-text retrieval and phenotype extraction

For 99 GDD in the test set, 1018 relevant case reports were identified (Supp. Table S1). Cadmus (https://doi.org/10.5281/zenodo.5618052) successfully downloaded at least one format (HTML/XML/PDF/plain text) for 962/1018 papers (94.5%) (Supp. Table S2). There were significantly more HPO terms in full text than in title + abstract after phenotype extraction (Figure 2). There were also more terms in

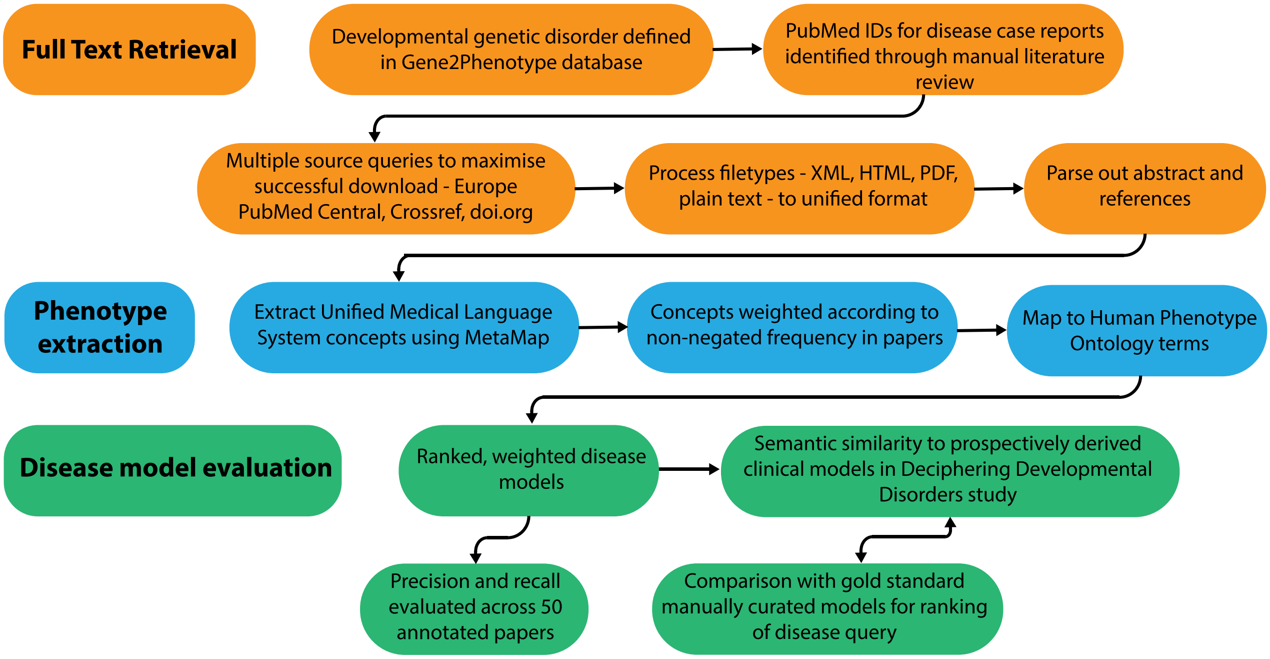
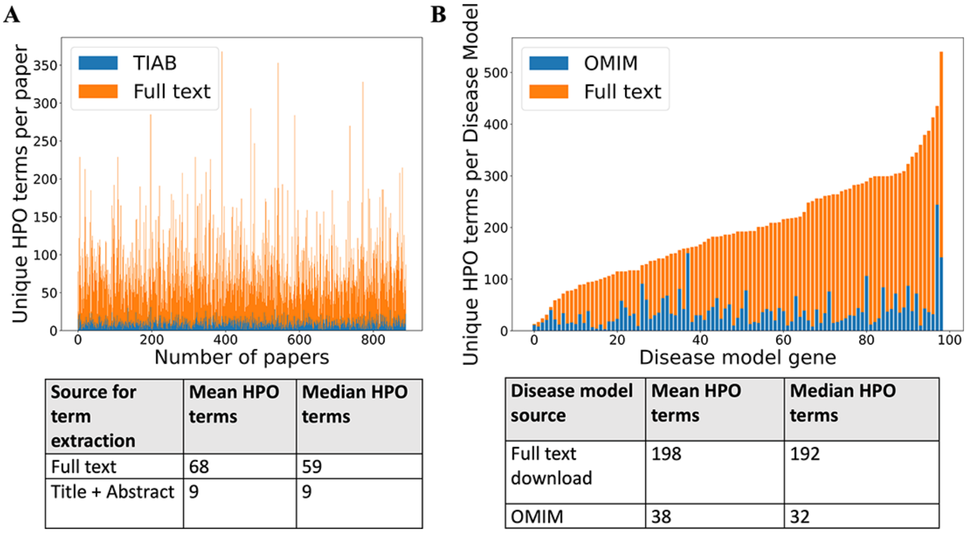

Full Text Retrieval
Developmental genetic disorder defined in Gene2Phenotype database
PubMed IDs for disease case reports identified through manual literature review
Multiple source queries to maximise successful download - Europe PubMed Central, Crossref, doi.org
Process filetypes - XML, HTML, PDF, plain text - to unified format
Parse out abstract and references
Phenotype extraction
Extract Unified Medical Language System concepts using MetaMap
Concepts weighted according to non-negated frequency in papers
Map to Human Phenotype Ontology terms
Disease model evaluation
Ranked, weighted disease models
Semantic similarity to prospectively derived clinical models in Deciphering Developmental Disorders study
Precision and recall evaluated across 50 annotated papers
Comparison with gold standard manually curated models for ranking of disease query

Figure 1. Overview of the disease model pipeline. Input was PMIDs for case reports describing developmental genetic disorders. Full-text downloads were performed using the Cadmus package. Output was disease models consisting of HPO terms weighted according to their frequency in full text. These were evaluated against 'real life' models from the Deciphering Developmental Disorders study and against gold standard manually curated models.



Source for term extraction	Mean HPO terms	Median HPO terms
Full text	68	59
Title + Abstract	9	9

Disease model source	Mean HPO terms	Median HPO terms
Full text download	198	192
OMIM	38	32

Figure 2. (A) Comparison of number of unique HPO terms extracted from full text vs title + abstract for sample of 962 papers, using full-text download pipeline. (B) Comparison of number of unique HPO terms in disease models generated by full-text download pipeline vs manually curated models in OMIM, for a sample of 99 diseases.

Table 1. Number of unique HPO terms across 99 disease test set

Source	Full-text download	DDD	OMIM	Full text + DDD	DDD + OMIM	Full text + DDD + OMIM
Number of unique HPO terms in set	3234	1569	972	3741	1935	3864

Table 2. Performance of MetaMap usage options

MetaMap options	Precision	Recall	F1 score
Word sense disambiguation, no derivational variants, restrict to HPO	0.80	0.70	0.74
Word sense disambiguation, restrict to HPO	0.81	0.69	0.74
No derivational variants, restrict to HPO	0.79	0.70	0.74
Blanklines off, restrict to HPO	0.76	0.73	0.74
Restrict to HPO	0.77	0.71	0.74
Blanklines off, word sense disambiguation, no derivational variants, conjunction processing, restrict to HPO	0.79	0.69	0.74
Blanklines off	0.82	0.67	0.74
No options	0.84	0.65	0.73
Conjunction processing, restrict to HPO	0.77	0.70	0.73

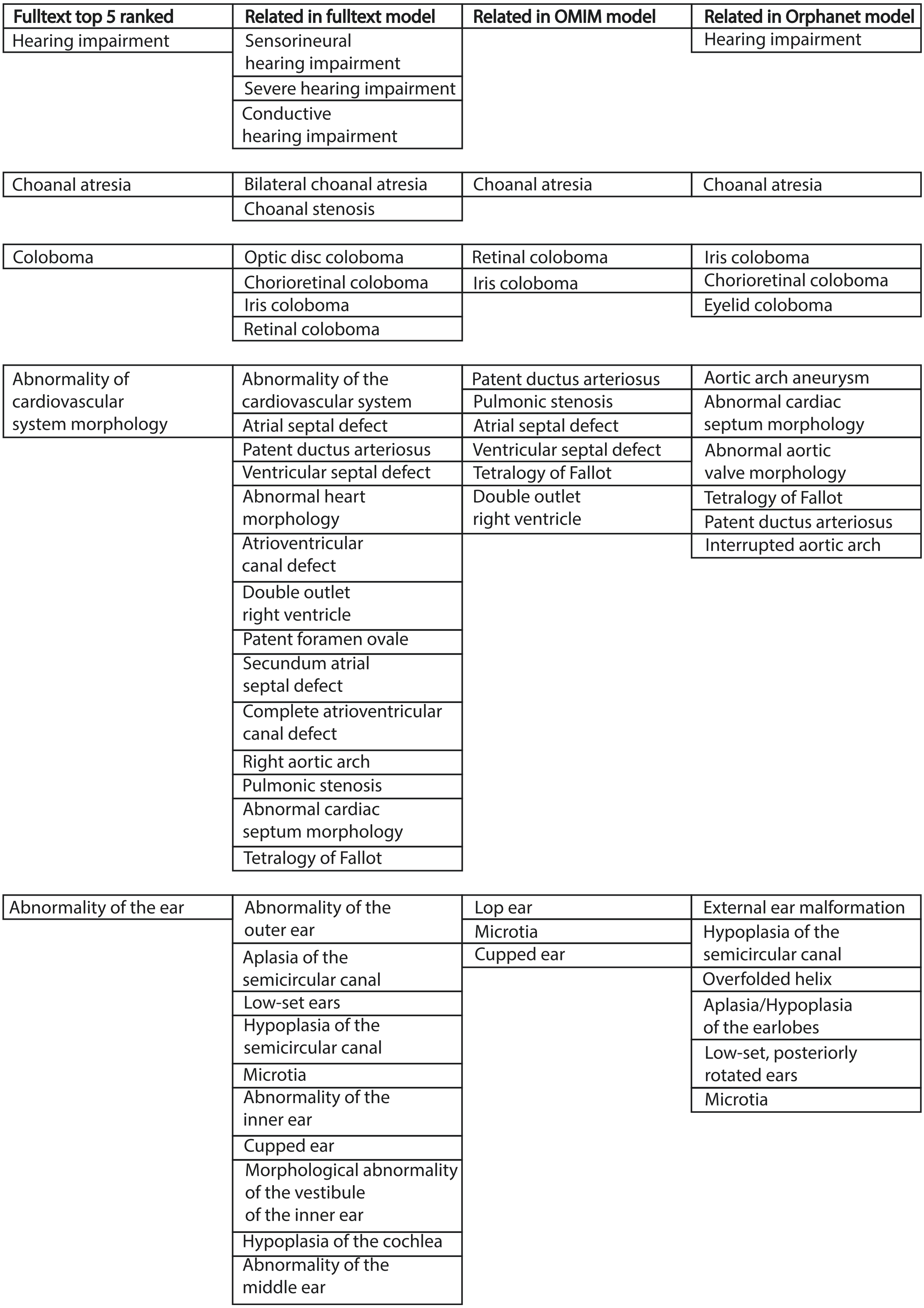
Evaluated against 50 manually annotated papers describing developmental disorders. Word sense disambiguation—disambiguate concepts with similar scores. Restrict to HPO—use only HPO for mapping concepts. No derivational variants—compute word variants without using derivational variants. Blanklines off—process text as a whole document. Conjunction processing—join conjunction-separated phrases.

the full-text-derived models than in OMIM (Figure 2). Table 1 shows the number of unique terms across the whole dataset by source.

The performance of phenotype concept extraction and HPO mapping in a 50 paper sample from the above set was evaluated using precision, recall and F1 score for a number of different MetaMap (11) output options (Table 2). The source was restricted to the HPO alone as performance was similar to other options but faster to process due to the smaller vocabulary size. This configuration was used for all the other analyses in this work.

Evaluation of disease models

The comparison of weighted disease models constructed from full text to models derived from other sources was not straightforward. The example model for *CHD7* in Figure 3 illustrates some of the issues. This describes the condition CHARGE syndrome, which is an acronym for Coloboma of the eye, Heart defects, Atresia of the choanae (choanal atresia), Restriction of growth and development and Ear abnormalities/deafness. The top-five ranked terms in the full-text-derived model are therefore highly relevant to this condition. However, a number of terms clinically relevant to these were also present in the same model, and this pattern is repeated across comparison datasets. Similar phenotypic features may be recorded in a heterogeneous manner by biomedical annotators using the same source documents. This inter- and intra-observer variability is a well-known phenomenon (17, 18). However, the issue of clinical phenotype heterogeneity across disease models is less well studied. For example, for CHARGE syndrome in Figure 3, a number of children of the term ‘Abnormality of the ear’ were included across the different datasets. These terms vary in their discriminant power for this condition—‘Hypoplasia of the semicircular canal’ is more specific for CHARGE than ‘Microtia’ (19). It is possible to use the ontology structure in HPO to relate similar terms, as per the semantic similarity MICA method. However, if two terms only share a high-level ancestor as in the example above, a significant loss of informativity will result, hindering meaningful comparison.



Fulltext top 5 ranked	Related in fulltext model	Related in OMIM model	Related in Orphanet model
Hearing impairment	Sensorineural hearing impairment		Hearing impairment
	Severe hearing impairment		
	Conductive hearing impairment		
Choanal atresia	Bilateral choanal atresia	Choanal atresia	Choanal atresia
	Choanal stenosis		
Coloboma	Optic disc coloboma	Retinal coloboma	Iris coloboma
	Chorioretinal coloboma	Iris coloboma	Chorioretinal coloboma
	Iris coloboma		Eyelid coloboma
	Retinal coloboma		
Abnormality of cardiovascular system morphology	Abnormality of the cardiovascular system	Patent ductus arteriosus	Aortic arch aneurysm
	Atrial septal defect	Pulmonic stenosis	Abnormal cardiac septum morphology
	Patent ductus arteriosus	Atrial septal defect	Abnormal aortic valve morphology
	Ventricular septal defect	Ventricular septal defect	Tetralogy of Fallot
	Abnormal heart morphology	Tetralogy of Fallot	Patent ductus arteriosus
	Atrioventricular canal defect	Double outlet right ventricle	Interrupted aortic arch
	Double outlet right ventricle		
	Patent foramen ovale		
	Secundum atrial septal defect		
	Complete atrioventricular canal defect		
	Right aortic arch		
	Pulmonic stenosis		
	Abnormal cardiac septum morphology		
	Tetralogy of Fallot		
Abnormality of the ear	Abnormality of the outer ear	Lop ear	External ear malformation
	Aplasia of the semicircular canal	Microtia	Hypoplasia of the semicircular canal
	Low-set ears	Cupped ear	Overfolded helix
	Hypoplasia of the semicircular canal		Aplasia/Hypoplasia of the earlobes
	Microtia		Low-set, posteriorly rotated ears
	Abnormality of the inner ear		Microtia
	Cupped ear		
	Morphological abnormality of the vestibule of the inner ear		
	Hypoplasia of the cochlea		
	Abnormality of the middle ear		

Figure 3. Example of top-five ranked terms in the disease model for CHD7/CHARGE syndrome (left column). Clinically related terms in the remainder of the disease model ($n = 540$), OMIM model ($n = 71$) and Orphanet model ($n = 82$) shown.

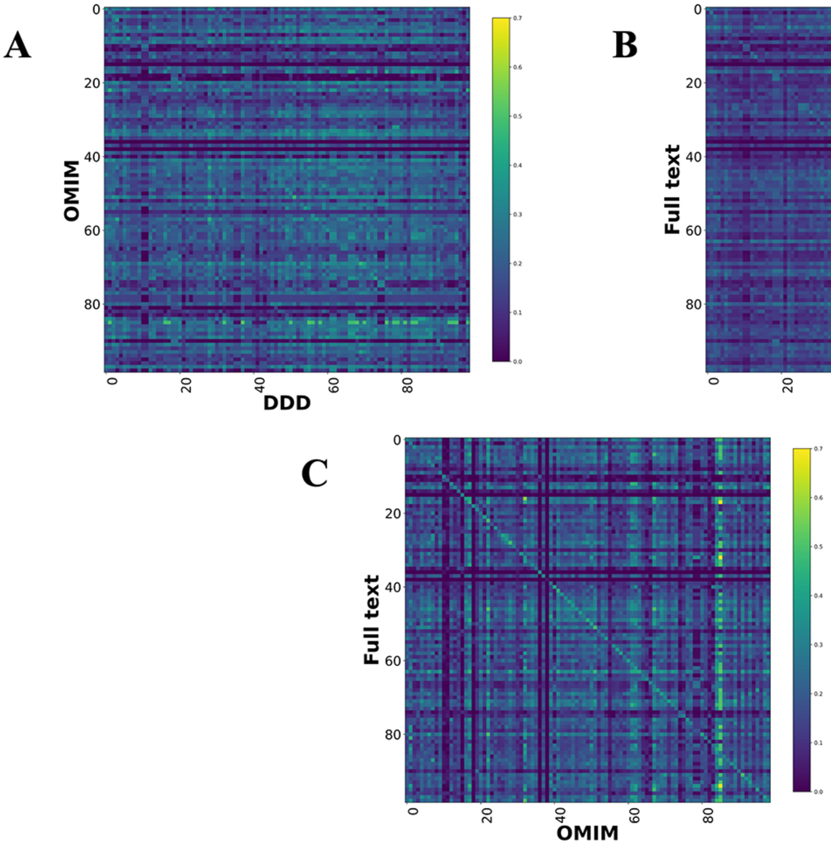


Figure 4. Disease model comparison heatmaps using RBO for literature-, OMIM- and DDD-derived models. Each model on the *y*-axis is compared with every model in the DDD set. Disease/DDD models describing the same disorder are on the rightward slanting diagonal. (A) compares OMIM and DDD models. (B) compares literature-derived and DDD models. (C) compares literature-derived and OMIM models. 99 diseases in comparison set.

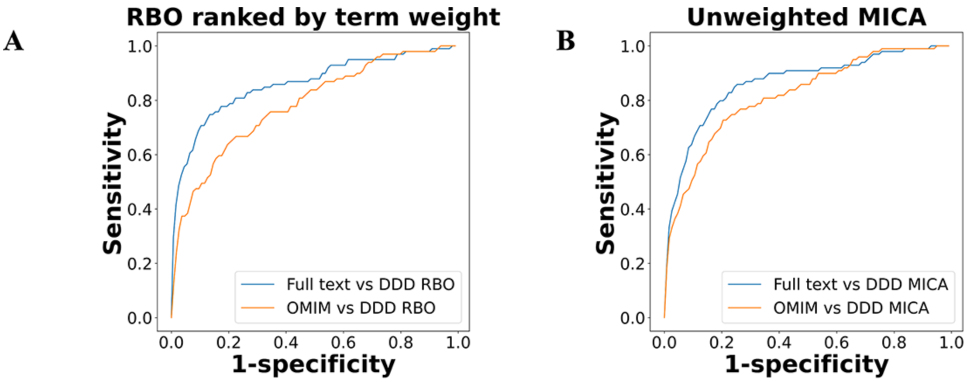
Comparison heatmaps using RBO for literature-derived, DDD and OMIM datasets (1, 7) (Figure 4) were generated, where every disease in one dataset is compared to every disease in the other. This was to determine if a corresponding pair, e.g. *CHD7*-full text vs *CHD7*-DDD, is more similar than any other in the comparison. These show a weak, but recognizable signal for the literature vs DDD and OMIM vs DDD. There is a clear signal for the literature vs OMIM models, showing that full-text-mined models are similar to manual curation.

The performance of full-text-derived models vs OMIM models in ranking the correct corresponding model in the DDD set was evaluated using receiver operating characteristic (ROC) curves (Figure 5). The full-text-derived models outperformed OMIM in both similarity metrics—ranked lists using RBO (12) and unweighted semantic similarity (14)—as defined by an increase in the area under the curve (AUC). A similar semantic similarity analysis using a subset of models from Orphanet showed comparable performance to those derived from full text when unweighted. However, weighted models did not show similar performance for either Orphanet or the full-text-derived set (Supp. Figure 1).

The results using a relatively simple list-based comparator—RBO (12)—and more advanced ontology-derived semantic similarity were perhaps more similar than may have been expected. This generated the hypothesis that the signal from RBO-based comparisons was dependent more on set overlap than term rankings. This could partially explain the similar results seen with the unweighted MICA method (14). To test this, the percentage of exact term matches across comparison datasets was plotted against the RBO score for each disease model, including the OMIM data (Figure 6). RBO scores did not clearly correlate to increasing set overlap. The RBO scores were in a broadly similar range regardless of high or low set overlap. This indicated that ranking of terms is an important determinant of similarity for this metric, not just set overlap. Of note, there was a significant overlap of exact match terms between the literature-derived and OMIM models. This indicates these were highly similar.

Discussion

Here, we demonstrate a method for automated literature curation that enables the creation of disease models based on ranked and weighted lists of clinical features from any of the most widely used structured vocabularies. This method can be compared to previous work linking genotype–phenotype data using manual and automated approaches. Databases such as OMIM and Orphanet (7, 8) are widely used in clinical and research settings because the disease–phenotype relationships contained therein are of high quality. This means that phenotype descriptors have been manually reviewed by expert curators directly from the source literature. However, this is a highly resource-intensive approach. Curation time needs to be spent not only documenting newly described disorders but also regularly updating existing entries. This



C

Comparator	Similarity metric	AUC
Full text vs DDD	RBO	0.850
OMIM vs DDD	RBO	0.774
Full text vs DDD	MICA	0.853
OMIM vs DDD	MICA	0.808

Figure 5. ROC curves using threshold ranking for literature-derived/OMIM disease models compared to real-life terms in the DDD study, across sample of 99 diseases, with a disease model and DDD model for each. Each disease model is compared to every model in the DDD set. (A) uses RBO to compare ranked lists of terms. Literature-derived and DDD models were ranked according to the model term frequency. OMIM models were ranked according to the frequency of terms across all OMIM models. (B) uses mean MICA to compare models, with IC calculated according to Resnik. Unweighted models were used for comparison, meaning each term in a model appeared only once, and term frequencies were not utilized. (C) shows the AUC for each model comparison.

represents a significant challenge given the volume of new publications describing GDD on a monthly basis (20). It is therefore likely that these manually curated databases do not include a truly comprehensive overview of the peer-reviewed literature for each GDD described.

Given the significant resources required to create and update manually curated disease–phenotype databases, studies have been undertaken to extract this information in an automated manner. For example, Kafkas *et al.* (21) used pointwise mutual information (22) to rank genotype–phenotype associations for both HPO and the Mouse Ontology (6, 23) in sentences extracted from a corpus of PubMed open access papers. Li *et al.* used a similar approach to generate autism-related gene–phenotype associations, although the corpus in this case was filtered using relevant search terms (24). Pilehvar *et al.* mapped phenotypic features to diseases by using Fisher exact testing to determine significant co-occurrence of disease-phenotype terms in Medline abstracts or paragraphs from PubMed open access articles (25, 26). This study used the Mondo ontology (27) to define disease names, including GDD. The database created by Pilehvar *et al.* (26), PheneBank, is the most comprehensive automated GDD-phenotype curation work of which we are aware. However, this study and those of Kafkas *et al.* and Li *et al.* (21, 24) had the significant limitation that the abstracts/manuscripts used were not filtered for human case reports/case series. This means phenotype associations were likely made from other sources, for example, animal models. The use of names only, i.e. text strings, to define disease entities also meant the underlying molecular mechanism was not defined. This means, for example, that there is no way to differentiate between gene-specific phenotypes where the disease name relates to multiple genes. Pilehvar *et al.* did also extract gene–phenotype mappings (26), but there was no method of differentiating somatic from germline variation; therefore, these were likely to include, for example, cancer-related manuscripts. This would also not differentiate between disorders caused by, for example, gain-of-function or loss-of-function variants in the same gene.

In this work, we utilized case series/case reports highly specific to individual GDD for automated text mining. This was to test the hypothesis that this approach would allow for the generation of disease–phenotype relationships, which closely replicated the true expressivity of a condition, in a similar manner to manual curation, with the benefits of a less resource-intensive automated system. Therefore, given the limitations of current automated disease models as discussed above, manually curated databases were used as a comparison in this work.

We tested this hypothesis using a subset of GDD to generate models using the HPO. We showed that full-text download can be achieved using standard licence agreements with both the journals and the online search engines at scale for a corpus of over 1000 papers, with close to 95% retrieval rate. Unsurprisingly, our annotation using full-text returned more phenotypic descriptors per paper when compared to title + abstract and per disease cf. manually curated models.

We showed that full-text models had a high degree of set overlap with those created from manual curation in OMIM (Figure 6), although the full-text models were generally larger (Figure 2). The full-text models outperformed manual curation when identifying corresponding diagnoses from the DDD study, using two similarity metrics. The first of these, RBO, compared ranked lists of terms and therefore did not utilize the ontology structure. It was perhaps surprising to show similarity between disease models without the advantages of an ontology-based metric. This was not only due to set overlap (Figure 6) and indicates that term weighting may

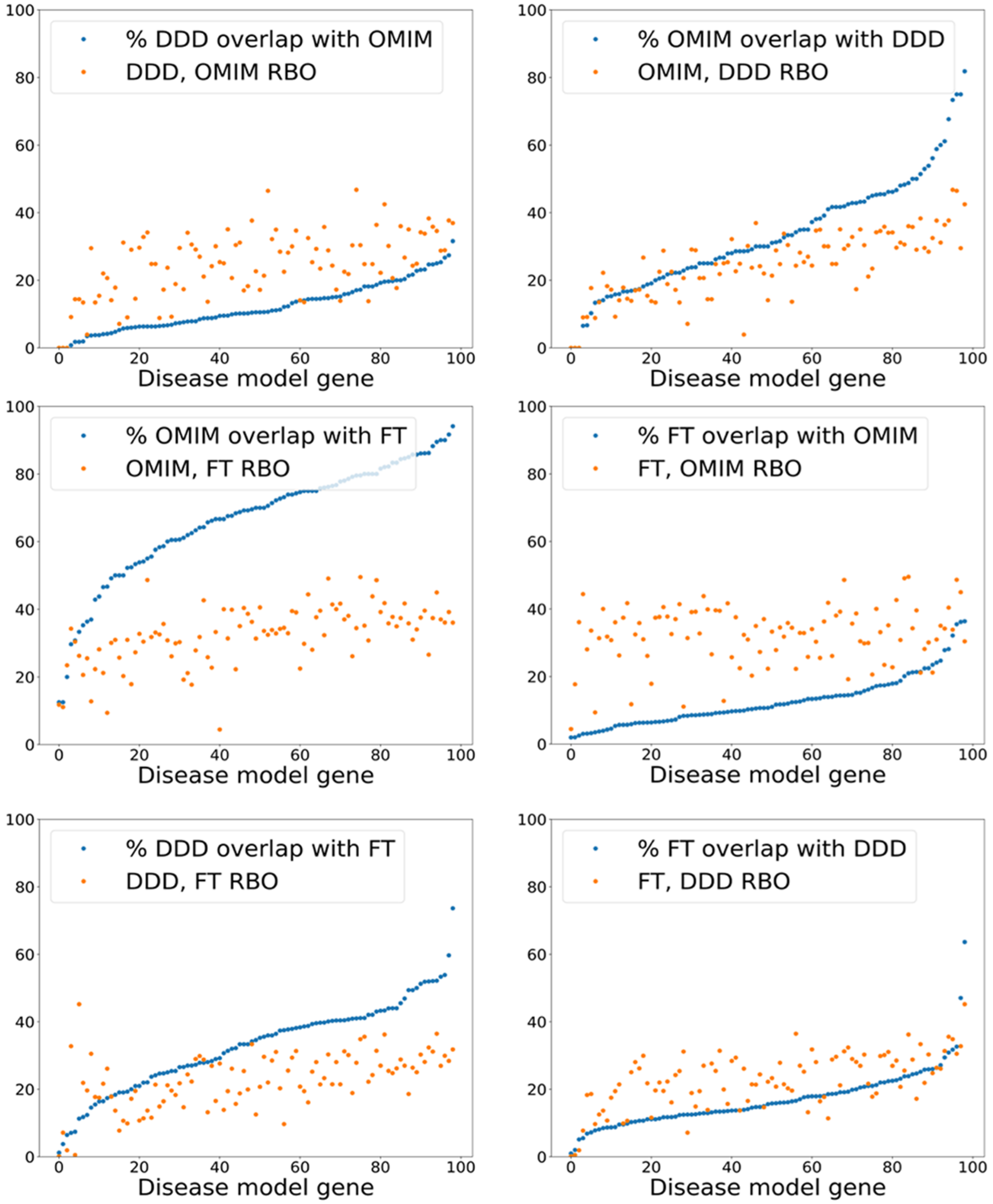


Figure 6. Exact model term overlap against RBO. For each pair of comparison models—A & B—the percentage of exact match terms (A in B/A) is shown. The corresponding RBO similarity score for A & B is also plotted. RBO scores multiplied by 100 to normalize to percentage range. FT—full-text-derived.

be important in developing phenotype similarity metrics in future. The performance improvement using full-text-derived models may reflect a higher number of disease-relevant terms per disease compared to OMIM, given the disparity in model sizes between them.

The second MICA-based similarity metric used the ontology structure of the HPO to compare terms. This measure should provide a more robust method of comparison than RBO as the ontology allows for a direct measure of the relatedness of a term pair. The performance of full-text-derived models was better than OMIM using this metric, even though term weighting was not utilized. However, the ROC curve derived from this was remarkably similar to the RBO-based measure. It is likely that developing condensed full-text-derived disease models, where clinically similar or discriminant terms are collapsed together, will demonstrate an improvement in predictive power using MICA-based similarity scoring. This task is not straightforward, as shown in Figure 3. We, and others, are currently working on methods to collapse clinically similar HPO terms without losing information.

Adding weighting, in the form of term frequency, to the MICA-based similarity comparison did not improve performance (Supp. Figure 1). This may be partly due to the challenge of the normalizing term weighting between full-text-derived, DDD, and manually curated datasets. The distribution of terms differed significantly between models, with generally much higher mean and median terms in the full-text-derived set. Improving phenotype extraction to identify terms on a per individual, rather than per paper, basis would allow for straightforward normalizing of term weighting, enable better comparison with manually curated models and potentially improve predictive power. Saklatvala *et al.* (28) used text mining on OMIM data to generate disease-associated phenotype terms weighted by frequency in a similar manner to the method in this work. They showed that weighting improved prediction of disease-associated genes for individual DDD probands, although sensitivity was low (23%). It would be interesting to repeat this analysis using the full-text models from this work in future.

Whilst the results we present here are encouraging with regard to the performance of automated literature curation, further improvements to disease models are needed before these can replace or minimize expert clinical interpretation of the peer-reviewed literature. This includes parsing out clinical descriptive text from papers to remove superfluous phenotypic descriptors in the introduction and discussion. Upweighting of disease-discriminant terms may also be helpful as well as collapsing clinically similar terms as mentioned above. Including negated terms, i.e. phenotypic descriptors that are explicitly never present in a disorder may yield further improvements.

The performance of MetaMap in named entity recognition (NER) demonstrated here is comparable to newer deep learning models such as PhenoTagger (29). This may be because MetaMap has been specifically designed—and regularly updated—to perform in the biomedical concept extraction domain. However, there could be improvements in precision and recall available through the use of a deep learning NER model, particularly if domain-specific training data are used. Additionally, representation of phenotypic features as word embeddings could allow for novel relationships between phenotypes and disease models to be elucidated. The results presented here showing that a non-ontology-based measure (RBO) is a useful similarity metric in the phenotype-disease space support this.

Here we have demonstrated and tested a method for generating HPO-based disease models for a small subset of the >2500 different GDD. It is possible that this technique could be applied to any set of diseases for which there is a reasonable level of aetiological homogeneity within case reports and/or case series, although this would require careful disease domain-specific evaluation. We consider this approach to be easily scalable, with the main bottleneck being in the efficient identification of the relevant clinical papers using online searches. We and others are actively developing systems to improve the discriminative power of the both the search strategies and the subsequent classification of the title/abstract/full text to minimize the requirement for manual review prior to full model creation. It is likely that leveraging annotated full-text downloads as demonstrated here will be useful in the classification of relevant papers for phenotype extraction.

Scaling up the method shown here will enable automatic addition of new disorders as well as updates to the phenotypic spectrum of known conditions. This means curation of databases such as DDG2P should become significantly less time- and resource-intensive. Ultimately, the expansion of automated curation to all GDD should enable testing to determine if phenotypic models as presented here improve diagnostic rates through computational comparison of individual patient phenotypes to those generated from the literature. This could be could be, for example, through the addition of automated full-text disease models to diagnostic systems developed for this purpose, such as LIRICAL and PhenIX (30, 31), which currently rely on manual curation such as OMIM (7). Alternatively, other similarity metrics could be used, for example, using vectorization of phenotype terms (16, 28). In time, this should allow for the incorporation of in-depth phenotypic data generated from automated literature curation into genomic bioinformatic analysis pipelines.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank the DDD study for access to the aggregated phenotype information on specific genetic disorders. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003). This study makes use of DECIPHER (http://www.deciphergenomics.org), which is funded by Wellcome. See Nature PMID: 25533962 or www.ddduk.org/access.html for full acknowledgement. MetaMap is provided courtesy of the U.S. National Library of Medicine.

Funding

Wellcome Strategic Award 200990/Z/16/Z as part of the Transforming Genetic Medicine Initiative and the Simons Initiative for the Developing Brain (SFARI—529085); MRC Human Genetics Unit University Unit Award to the University of Edinburgh to D.R.F.

Conflict of interest

None declared.

References

1. Deciphering Developmental Disorders Study. (2017) Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, **542**, 433–438.
2. Short,P.J., McRae,J.F., Gallone,G. *et al.* (2018) De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, **555**, 611–616.
3. Kaplanis,J., Samocha,K.E., Wiel,L. *et al.* (2020) Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, **586**, 757–762.
4. Thormann,A., Halachev,M., McLaren,W. *et al.* (2019) Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.*, **10**, 2373.
5. Richards,S., Aziz,N., Bale,S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–423.
6. Köhler,S., Gargano,M., Matentzoglu,N. *et al.* (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
7. OMIM®. McKusick-*Nathans Institute of Genetic Medicine Johns Hopkins University (Baltimore, MD) Online Mendelian*

Inheritance in Man. Online Mendelian Inheritance in Man https://omim.org/ (22 April 2021, date last accessed).

8. Orphanet©. *INSERM Orphanet: an online rare disease and orphan drug data base. Orphanet: an online rare disease and orphan drug data base* http://www.orpha.net (22 April 2021, date last accessed).
9. Collier,N., Groza,T., Smedley,D. *et al.* (2015) PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. *Database*, **2015**, bav104.
10. Wei,C.-H., Allot,A., Leaman,R. *et al.* (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.
11. 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.
12. Webber,W., Moffat,A. and Zobel,J. (2010) A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, **28**, 1–38.
13. Resnik,P. (1995) *Using Information Content to Evaluate Semantic Similarity in a Taxonomy. arXiv:cmp-lg/9511007.*
14. Helbig,I., Lopez-Hernandez,T., Shor,O. *et al.* (2019) A recurrent missense variant in AP2M1 impairs clathrin-mediated endocytosis and causes developmental and epileptic encephalopathy. *Am. J. Hum. Genet.*, **104**, 1060–1072.
15. Köhler,S., Schulz,M.H., Krawitz,P. *et al.* (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, **85**, 457–464.
16. Köhler,S. (2018) Improved ontology-based similarity calculations using a study-wise annotation model. *Database*, **2018**, bay026.
17. Miñarro-Giménez,J.A., Cornet,R., Jaulent,M.C. *et al.* (2019) Quantitative analysis of manual annotation of clinical text samples. *Int. J. Med. Inf.*, **123**, 37–48.
18. Martínez-demiguel,C., Segura-Bedmar,I., Chacón-Solano,E. *et al.* (2022) The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms. *J. Biomed. Inform.*, **125**, 103961.
19. van Ravenswaaij-arts,C.M., Hefner,M., Blake,K. *et al.* (1993) CHD7 disorder. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, Mirzaa GM, Amemiya A (eds). *GeneReviews*®. University of Washington, Seattle.
20. Bamshad,M.J., Nickerson,D.A. and Chong,J.X. (2019) Mendelian gene discovery: fast and furious with no end in sight. *Am. J. Hum. Genet.*, **105**, 448–455.
21. KafkasŞ. and Hoehndorf,R. (2019) Ontology based text mining of gene–phenotype associations: application to candidate gene prediction. *Database*, **2019**, baz019.
22. Church,K.W. and Hanks,P. (1990) Word association norms, mutual information, and lexicography. *Comput. Linguist.*, **16**, 22–29.
23. Eppig,J.T., Blake,J.A., Bult,C.J. *et al.*, The Mouse Genome Database Group. (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.*, **43**, D726–D736.
24. Li,S., Guo,Z., Ioffe,J.B. *et al.* (2021) Text mining of gene–phenotype associations reveals new phenotypic profiles of autism-associated genes. *Sci. Rep.*, **11**, 15269.
25. Sayers,E.W., Bolton,E.E., Brister,J.R. *et al.* (2021) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **49**, D10–D17.
26. Pilehvar,M.T., Bernard,A., Smedley,D. *et al.* (2021) PheneBank: a literature-based database of phenotypes. *Bioinformatics*, **38**, 1179–1180.
27. Shefchek,K.A., Harris,N.L., Gargano,M. *et al.* (2020) The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **48**, D704–D715.
28. Saklatvala,J.R., Dand,N. and Simpson,M.A. (2018) Text-mined phenotype annotation and vector-based similarity to improve identification of similar phenotypes and causative genes in monogenic disease patients. *Hum. Mutat.*, **39**, 643–652.
29. Luo,L., Yan,S., Lai,P.-T. *et al.* (2021) PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinforma. Oxf. Engl.*, **37**, 1884–1890.
30. Zemojtel,T., Köhler,S., Mackenroth,L. *et al.* (2014) Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.*, **6**, 252ra123.
31. Robinson,P.N., Ravanmehr,V., Jacobsen,J.O.B. *et al.* (2020) Interpretable clinical genomics with a likelihood ratio paradigm. *Am. J. Hum. Genet.*, **107**, 403–417.