THE UNIVERSITY *of* EDINBURGH

Systematising and scaling literature curation for genetically determined developmental disorders

Thabo Michael Yates

Doctor of Medicine

MRC Human Genetics Unit

Institute of Genetics and Cancer

The University of Edinburgh

2022

Declaration

I declare that I composed this thesis, and that the work contained within it is my own. This work has not been submitted for any other degree or professional qualification except as specified.

Thabo Michael Yates 4/3/22

Abstract

The widespread availability of genomic sequencing has transformed the diagnosis of genetically-determined developmental disorders (GDD). However, this type of test often generates a number of genetic variants, which have to be reviewed and related back to the clinical features (phenotype) of the individual being tested. This frequently entails a time-consuming review of the peer-reviewed literature to look for case reports describing variants in the gene(s) of interest. This is particularly true for newly described and/or very rare disorders not covered in phenotype databases. Therefore, there is a need for scalable, automated literature curation to increase the efficiency of this process. This should lead to improvements in the speed in which diagnosis is made, and an increase in the number of individuals who are diagnosed through genomic testing.

Phenotypic data in case reports/case series is not usually recorded in a standardised, 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. Several initiatives have been developed to link diseases to phenotype data, in the form of HPO terms. 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 developed to date have relied on abstracts or open access papers. 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 which has not been used previously at scale, due to difficulties accessing non-open access manuscripts.

In this thesis, I present a method for literature curation which can utilise all relevant published full text through a newly developed package which can download almost all manuscripts licenced by a university or other institution. This is scalable to the full spectrum of GDD. Using manuscripts identified through manual literature review, I use a full text download pipeline and NLP (natural language processing) based methods to generate disease models. These are comprised of HPO terms weighted according to their frequency in the literature. I demonstrate iterative refinement of these models, and use a custom annotated corpus of 50 papers to show the text mining process has high precision and recall. I demonstrate that these models clinically reflect true disease expressivity, as defined by manual comparison with expert literature reviews, for three well-characterised GDD.

I compare these disease models to those in the most commonly used genetic disease phenotype databases. I show that the automated disease models have increased depth of phenotyping, i.e. there are more terms than those which are manually-generated. I show that, in comparison to ‘real life’ prospectively gathered phenotypic data, automated disease models outperform existing phenotype databases in predicting diagnosis, as defined by increased area under the curve (by 0.05 and 0.08 using different similarity measures) on ROC curve plots.

I present a method for automated PubMed search at scale, to use as input for disease model generation. I annotated a corpus of 6500 abstracts. Using this corpus I show a high precision (up to 0.80) and recall (up to 1.00) for machine learning classifiers used to identify manuscripts relevant to GDD. These use hand-picked domain-specific features, for example utilising

specific MeSH terms. This method can be used to scale automated literature curation to the full spectrum of GDD. I also present an analysis of the phenotypic terms used in one year of GDD-relevant papers in a prominent journal. This shows that use of supplemental data and parsing clinical report sections from manuscripts is likely to result in more patient-specific phenotype extraction in future.

In summary, I present a method for automated curation of full text from the peer-reviewed literature in the context of GDD. I demonstrate that this method is robust, reflects clinical disease expressivity, outperforms existing manual literature curation, and is scalable. Applying this process to clinical testing in future should improve the efficiency and accuracy of diagnosis.

Lay Summary

Genetic testing is used to diagnose rare disorders which present before or soon after birth (developmental disorders or DD). Tests which look at all of an individual's genetic material (genome-wide testing) have become widely available and have increased the number of diagnoses that can be made for DD. However, a diagnosis is not made in a significant proportion of cases. One of the reasons for this may be that the large amount of data from a single genome-wide test is difficult to analyse. Locating a diagnostic genetic change (variant) in a genome-wide test file may be likened to finding the proverbial needle in a haystack. There are various methods used to narrow down the data which is analysed, for example excluding variants which are common in the general population. However, this still often leaves several candidate gene variants which may be the cause of an individual's condition. These then need to be analysed in relation to the person's phenotype. The phenotype is all physical characteristics which may be related to a genetic disorder, for example 'absent kidney', 'extra fingers'. To do this, a genetic specialist reviews case reports relevant to the candidate variants (i.e. describing individuals with variants in the same gene) from the scientific literature, to see if the phenotype in these relates to the phenotype in the individual being tested. This is a time-consuming process. To aid this, a number of gene-phenotype databases have been developed. However, they still rely on biocurators manually reviewing the scientific literature, which is impractical to scale up to the thousands of described DD. In this thesis, I present a method to automate and scale up literature curation to cover the full spectrum of DD, which should increase the rate at which diagnoses are made using genome-wide testing in future.

Using papers hand-selected from the literature, I show that disease models can be produced using text mining techniques. These models use standardised vocabulary so they can be used in computational applications, and they are weighted to show which phenotypic features are more common

for a condition. I test the processes used to generate these against manually annotated data to show that these are robust and high performance. I also test these models against those developed using manual curation. I show that models from the automated method are better at predicting a diagnosis using phenotypic data than those which are manually generated. I show that the structure of manuscripts describing DD may be utilised for separating out clinical information. I demonstrate that machine learning may be used to search the scientific literature to identify input papers for generating disease models.

In summary, I demonstrate a method for scalable automated literature curation. This should be useful for improving diagnostic rates for DD in future.

Acknowledgements

I would like to thank my supervisors David and Ian, without whom this work would not have been possible. In particular, I would like to thank David for his constant support and guidance, not only with this work but also for wider professional advice. David has opened up new and exciting opportunities for me in the field of bioinformatics and I am very grateful for this. I would like to thank Ian for his patience and valuable advice to help me progress in learning data science from scratch. I also very much enjoyed the board games nights with his group.

I would also like to thank Jamie and Antoine from Ian's group, who helped me with learning to code as well as developing the Cadmus package, without which this work would not be possible.

On a personal note, my wife Anne has been a source of endless support and motivation during the stressful times of writing up this thesis, and I would like to express my appreciation to her. Finally, my newly arrived daughter Frieda has been a wonderful distraction during the final stages of finishing this work.

Contents

Abbreviations

AJHG	American Journal of Human Genetics
API	Application Programming Interface
AUC	Area under the receiver operating characteristic curve
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
BN	Bayesian networks
CNN	Convolutional neural network
CNV	Copy number variant
CRF	Conditional Random Field
CUI	Concept Unique Identifier
DD	Developmental disorders
DDD	Deciphering Developmental Disorders study
DDG2P	Developmental Disorders Gene2Phenotype
DECIPHE R	DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources
DO	Disease Ontology
EHR	Electronic health records
EPMC	Europe PubMed Central
FN	False negative
FP	False positive
G2P	Gene2Phenotype
GDD	Genetically-determined developmental disorders
GenCC	Gene Curation Coalition
GWAS	Genome-wide association studies
HGNC	Human Genome Organisation Gene Nomenclature Committee
HOOM	Human Phenotype Ontology-Orphanet Rare Disease Ontology ontological module
HP ID	Human Phenotype (Ontology) unique identifier
HPO	Human Phenotype Ontology

HPO GS	HPO Gold Standard
HTML	Hypertext Markup Language
IC	Information Content
LGMDE	Locus-genotype-mechanism-disease-evidence
MeSH	Medical Subject Headings
MICA	Most informative common ancestor
MMI	MetaMap Indexing
MP	Mammalian Phenotype Ontology
NCBO	National Center for Biomedical Ontology
NDV	No derivational variants
NER	Named entity recognition
NPV	Negative Predictive Value
OBO	Open Biological and Biomedical Ontology
OMIM	Online Mendelian Inheritance in Man
P/R	Precision and recall
PDF	Portable Document Format
PhenIX	Phenotypic Interpretation of eXomes
PHIVE	Phenotypic Interpretation of Variants in Exomes
PMI	Pointwise mutual information
PMID	PubMed Identifier
POS	Part-of-speech
PPV	Positive Predictive Value
QDA	Quadratic Discriminant Analysis
RBF	Radial-Basis Function kernel
RBO	Rank Biased Overlap
ROC	Receiver Operating Characteristic
SNV	Single Nucleotide Variant
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TIAB	Title + Abstract

TN	True Negative
TP	True Positive
UMLS	Unified Medical Language System
VCF	Variant call format
WSD	Word sense disambiguation
XML	Extensible Markup Language

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

1.1.1 Traditional approach to diagnosis in genetic disease

Genetically-determined developmental disorders (GDD) are conditions which arise during embryogenesis or early development. These result from functionally deleterious genetic copy number variants (CNV) or single nucleotide variants (SNV -otherwise defined as mutations) primarily responsible for some or all of the clinical features under investigation.

Diagnosis of GDD has traditionally focused on the recognition of a constellation of phenotypic features in an individual which may be associated with a particular syndrome. This analysis would then direct genetic testing towards a given gene or set of genes. For example, an individual with widely spaced eyes (hypertelorism), bulbous nose, iris coloboma (gap in the structure of the iris) and frontal pachygyria (thickened area of cerebral cortex) may have Baraitser-Winter Cerebro-Fronto-Facial syndrome (1). Genetic testing will likely identify a mutation in the *ACTB* or *ACTG1* genes (1).

1.1.2 Application of genome-wide sequencing

Diagnostic innovation in this area of medicine has largely been driven by the development of genome-wide sequencing. This was initially mostly in the form of exome (all of an individual's protein-coding genes) sequencing (2,3), but whole genome sequencing is increasingly being used in clinical and research settings (4). This method has resulted in significant diagnostic yield, with the rate of diagnosis being 24-68% (mean 31%) according to one meta-analysis (5).

However, the genotype-driven approach has limitations. The volume of data generated necessitates informatic filtering approaches to narrow down the number of variants (usually SNV in the context of genome-wide sequencing) subject to detailed analysis. This may include eliminating common variants, and prioritising those which have occurred *de novo* (not present in either parent), as these are known to be enriched in GDD (6).

It is not usually possible to make a diagnosis based on the variants identified through this process alone. This is because variation with plausible disease-linked features is also present in the healthy population. For example, an individual will carry, on average, 200 rare SNV (allele frequency <0.1%) and around 27 SNV which have not been found in any other person (7). Additionally, the rate of *de novo* SNV is approximately 74 per generation (8).

1.1.3 Limitations of genotype-first approach

Given the limitations of bioinformatic filtering, candidate variants identified through diagnostic genomic sequencing often require detailed clinical review. GDD display an extreme degree of locus and allelic heterogeneity, with thousands of genes linked to these types of conditions (9). This means it is not usually possible to assess candidate variants without searching for, and evaluating, a number of case reports in the peer-reviewed literature.

This search and evaluation process is to determine if the phenotypic features of an individual might fit with those reported in association with variation in a particular gene (10). This may need to be repeated for many different genes, which is time-consuming and not always feasible in a busy clinical setting. Furthermore, technical limitations mean not all pathogenic variants are detected through genomic sequencing. In particular, there are a number of reports of SNV which have been missed through exome sequencing (11,12). Whilst whole genome sequencing may offer improved performance, coverage is still incomplete (4).

1.1.4 Phenotyping and genome-wide sequencing

The traditional diagnostic method, using subjective assessment of phenotypic similarity to compare an individual patient and those described in case reports/series, itself has limitations. For example, the increased diagnostic rate associated with genomic sequencing has resulted in phenotypic expansion. This may associate novel clinical features with well characterised disorders (13,14). Milder presentations of disease than previously recognized may be identified (15,16). Additionally, many disorders may be associated with apparently non-specific phenotypes such as intellectual disability (17).

1.1.5 Automated literature curation to improve diagnosis

The current molecular diagnostic rate of 24-68% (mean 31%) (5) of genome-wide sequencing means a diagnosis is not made in many cases of suspected GDD. A phenotype-driven approach to genomic diagnostics is likely to improve these figures. To address this, there is a need for a scalable, automated approach to extraction and analysis of phenotypic data from the peer-reviewed literature, which is the focus of this work. This should allow for:

1) Enhanced efficiency of diagnosis by increasing the speed at which candidate variants can be clinically reviewed.
2) Improvement in diagnostic rate through prioritisation of variants which fit the clinical phenotype. This may include integration with bioinformatic filtering pipelines and focused re-analysis of reduced coverage areas in genomic sequencing.
3) Deeper phenotyping of disorders than is possible through manual literature review, potentially allowing for subtle disease-specific/discriminant features to be recognised. This may enable delineation and diagnosis of GDD currently thought to present in a non-specific manner.

Several different components need to be combined to create a pipeline for automated literature curation, such as scalable identification and download of relevant manuscripts, and accurate phenotype extraction from text, using a standardised vocabulary. Phenotype models created using this pipeline need to be tested against the current widely used data standards, particularly to determine if they may be used to replace manual literature curation.

1.1.6 Principle hypothesis of this thesis

Following on from the above, the central hypothesis of this thesis is:

Development of automated literature curation may allow for accurate phenotypic models to be constructed for GDD, in a similar manner to manual curation.

An automated approach would be easily scalable, straightforward to update and provide computationally tractable data. It may also allow for a more comprehensive overview of the phenotypic spectrum of a given GDD than is feasible using manual curation.

In the subsequent sections, I review each of the components required to develop automated literature curation for GDD. This will inform the development of automated curation in subsequent chapters.

1.2 Human phenotype ontology

1.2.1 Standardisation of disease descriptors

Phenotypic data in the peer-reviewed literature is not usually recorded in a computationally-tractable format. Plain text descriptions of similar clinical features may be recorded in several different ways, depending on the preference of the recording clinician or researcher. For example, a technical

term such as ‘hypertelorism’, may be recorded as its synonym ‘widely spaced eyes’. Automated literature curation consequently requires the use of a standardised vocabulary to define phenotypic features.

Previous work in this area includes the London Dysmorphology Database, which used a tree-based structure to define term relationships (18). However, this did not define links between all terms, and could not incorporate multiple links to terms.

The Human Phenotype Ontology (HPO) is a standardised, machine-readable vocabulary covering a wide spectrum of human disease (19). It can define relationships between terms in detail using the ontology structure. The HPO has become the de facto standard for phenotyping in GDD. Therefore, it is used throughout this work, and will be reviewed in more detail here.

1.2.2 Ontology definition

The term ‘ontology’ has been widely used across different disciplines, including philosophy and many scientific domains (20). The definition of an ontology has correspondingly varied. For the purposes of this work, an ontology is defined as the explicit representation of knowledge within a particular domain (21). An ontology consists of concepts described by a standardised vocabulary, any attributes these may have, and the relationships between them (19).

1.2.3 Classes in human phenotype ontology

The HPO consists of concepts, or classes, which are text descriptors of clinical abnormalities. These are known as HPO terms. Each term is given an identifier (HP ID), which is unique and persistent across versions of the ontology. A term may have one or more synonyms, recording different words

or phrases used to describe the same abnormality, which are included under the same HP ID (22).

In the HPO, synonyms may include both biomedical technical descriptors and 'layperson' phrases in common usage. Cross-references to equivalent or closely related classes in other vocabularies or ontologies are often included, for example to Unified Medical Language System (UMLS) biomedical concepts (23). A term may also include a more detailed text definition, for example the term 'Hip dislocation' is defined as 'Displacement of the femur from its normal location in the hip joint' (22). An example of an HPO term is given in Figure 1-1.

Updates to the HPO may include defining deprecated terms as obsolete and removing them (although they remain in the ontology data for reference and may be mapped to other terms), or merging terms deemed to have similar meaning.

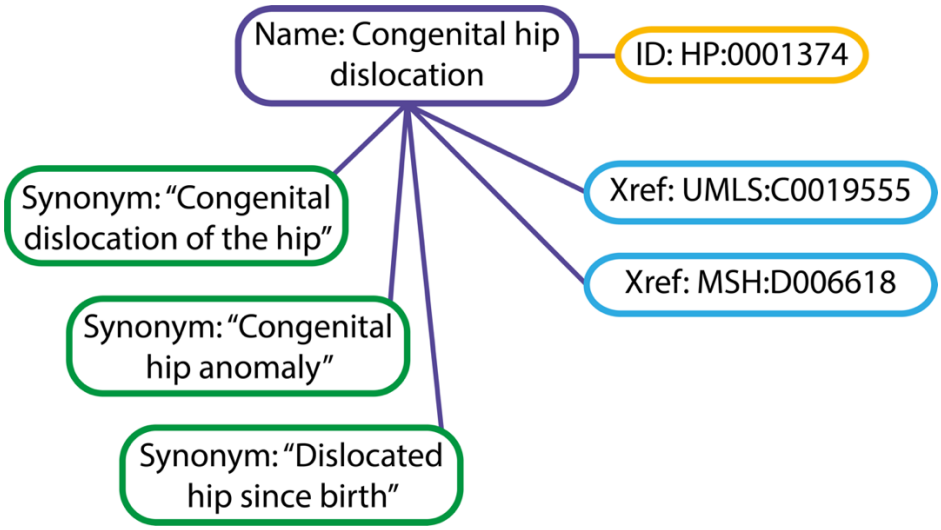


Figure 1-1. Example of class in Human Phenotype Ontology (22). Drawn from data in hp.obo file (version 1.2), downloaded on 08/10/2021. UMLS:C0019555 – Unified Medical Language System (UMLS) biomedical concept (23); MSH:D006618 – MeSH (Medical Subject Headings) descriptor (23).

1.2.4 Structure of the human phenotype ontology

The HPO is a directed acyclic graph. This means it contains directional relationships between terms in a tiered structure, all derived from a root node (22). This directionality of relationships means it is possible to travel 'up' or 'down' the ontology, and refer to 'parents'/'ancestors' or 'children'/'descendants' of terms.

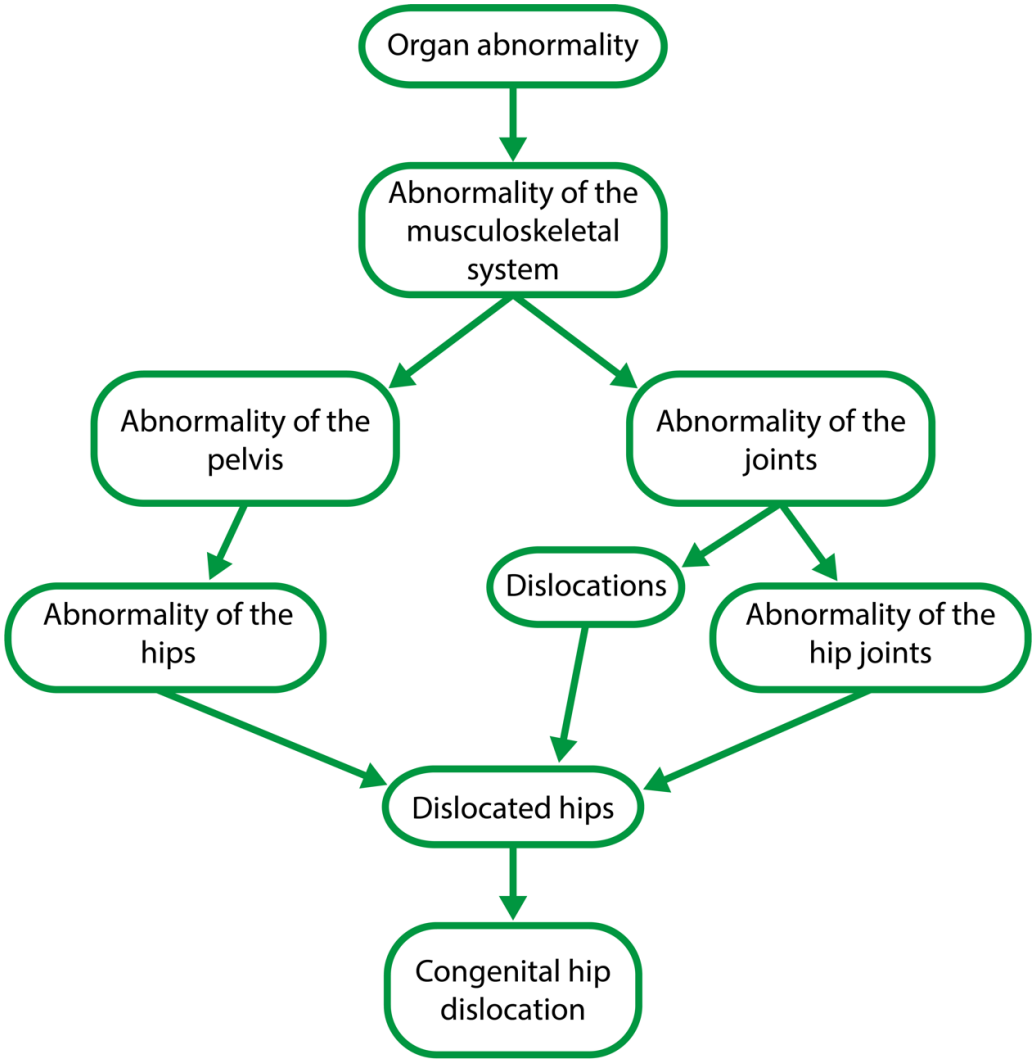


Figure 1-2. Example of all terms up to root for 'Congenital hip dislocation' in the HPO. Note all connections are directional. If a disease is annotated with a term, it will also be implicitly annotated with that term's ancestors. For example, a disease annotated with 'Congenital hip dislocation' will also be annotated with 'Abnormality of the hips' and 'Abnormality of the joints' and all other terms shown. Adapted from Robinson et al. (19).

Terms in the ontology may also be referred to by ‘level’ or ‘depth’, depending on how many ancestors a term has. Parent terms define broader concepts than more specific child terms. A term may have more than one parent. Sibling terms (those with the same parents) are not connected, meaning there are no cyclic relations between terms, hence the definition of HPO as acyclic.

Connections between terms in HPO are by ‘is-a’ (subclass-of) relationships e.g. ‘Broad eyebrow’ is-a ‘Abnormal eyebrow morphology’. Is-a connectors are transitive, which means annotated terms are inherited through all possible paths up to the root (22). An example of a term and its parents in the HPO is given in **Error! Reference source not found.**.

The root term of the HPO is ‘All’, however in practice the ontology is divided into five subontologies, each with their own root term. These are ‘Phenotypic abnormality’, ‘Mode of inheritance’, ‘Clinical modifier’, ‘Clinical course’ and ‘Frequency’. Terms with the root ‘Phenotypic abnormality’ form the bulk of the HPO, and are used in this work.

1.2.5 Limitations of the human phenotype ontology

The HPO has been successfully used to describe the phenotype of GDD in many studies (24–26), and has become the de facto standard for recording phenotypic data in this disease domain. However, there are limitations to its use, both in clinical diagnostic/research settings and in scalable disease analyses.

The same phenotypic feature for a given disease may be described in the free text of a manuscript using different words or phrases as mentioned in section 1.2.1. However, the same phenotypic feature may be recorded differently through manual input even when the HPO is used. For example, an individual may be described as having ‘Focal seizure with eyelid

myoclonia' or 'Focal myoclonic seizure'. Whilst the ontology structure may in theory mean it is simple to relate these terms – in fact one is a parent of the other – information which may be useful in disease discrimination or diagnosis is lost.

Furthermore, clinically similar terms may only be related in the ontology by a high level, non-specific common ancestor. For example, the terms 'Premature closure of fontanelles' and 'Craniosynostosis' have the common ancestor 'Abnormality of skeletal morphology'. Finally, the use of a standardised vocabulary inevitably means newly defined phenotypic features will not be included. It is straightforward to submit terms to HPO for consideration of inclusion (22) to address this. However, in practice the process is time consuming. If a phenotypic feature is not available in the HPO at the time of recording, this may mean phenotyping of an individual is incomplete, or uses non-specific terms.

1.3 Disease-phenotype databases in current usage

1.3.1 Overview of disease-phenotype databases

A number of initiatives have sought to address the challenge of collating and synthesising phenotypic data relating to genetic disease, from the peer-reviewed literature. These utilise HPO terms to describe phenotypic features and link them to a specified condition. The manner in which phenotypic data is collected and even how a genetic disease is defined differ between the databases.

An overview of the most prominent of these databases is given here, and their limitations with regard to phenotype data outlined. Automated literature curation in this work aims to address some of these shortcomings.

1.3.2 Online Mendelian Inheritance in Man

The Online Mendelian Inheritance in Man (OMIM) (27) database evolved from a text catalogue of Mendelian disorders. Each entry is defined by a stable identifier, known as a MIM number. There are entries for genes, and for disease phenotypes. Gene entries are linked to relevant disease phenotypes, and each disease is mapped to only one gene.

For each disease entry, the following information is included: disease name (and synonyms), MIM number, inheritance pattern, gene/locus (27). There is a focus on detailed free text descriptions in the database, including detailed descriptions of individual peer-reviewed manuscripts, with full references provided. Information from these papers is described under headings including ‘Clinical Features’ for phenotypic data and ‘Molecular Genetics’ for disease-associated mutations (27).

OMIM also records phenotypic descriptors in a more standardised manner as a ‘Clinical Synopsis’. This is a list of HPO terms stratified by body system (e.g. ‘head & neck’, ‘respiratory’), together with an inheritance pattern and molecular basis (‘caused by mutation in {gene name}’) (27). Diseases are mapped to other data sources e.g. Orphanet (28).

OMIM is maintained by manual biocuration at the McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, USA. The inclusion and curation criteria for OMIM are not publicly defined in detail. However, there is a statement on the OMIM website that ‘priority for inclusion is given to papers that provide significant insight into the gene-phenotype relationship, expand our understanding of human biology, or contribute to the characterization of a disorder’ (29).

OMIM is primarily accessed through a web interface; the database itself is not freely downloadable, although disease-phenotype annotations are

available through HPO (22). These consist of a disease with its MIM identifier, together with a list of HPO terms. The majority of the terms are not weighted, meaning there is no indication how commonly a given phenotypic feature is seen in association with a disorder.

1.3.3 Orphanet

Orphanet is an initiative which aims to collate information for rare diseases, to improve diagnosis and treatment (28). There are 41 countries in the network, coordinated at the French National Institute of Health and Medical Research in Paris, France (28). Orphanet offers a number of resources, of which the most pertinent to this work is the inventory of rare diseases. An entry in this inventory is for a clinical entity, which consists of a disease label (and synonyms) with a unique identifier – an ORPHAcode. Inheritance pattern, age of onset and mapping to other databases such as OMIM are included (27,28).

A free text description of a disease is included, written by a clinical expert, with categories including ‘Clinical description’, ‘Etiology’ and ‘Genetic counseling’. Orphanet defines a clinical entity as ‘a group of rare disorders, a rare disorder or a subtype of disorder’ which means one clinical entity can be mapped to multiple genes (28).

Phenotype annotations are provided as ‘Clinical signs and symptoms’, which consist of a list of HPO terms. These are uniformly weighted by HPO frequency terms, which include ‘Very frequent (present in 80-99% of cases)’, ‘Frequent (30-79%)’ and ‘Occasional (5-29%)’ (22). Per-individual frequencies are not available. Orphanet provides a website which allows for download of clinical entity-phenotype data at scale - http://www.orphadata.org.

1.3.4 DECIPHER

DECIPHER (DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources) is a database – which can be accessed interactively via a web interface – of genomic variants (including both SNV and CNV) associated with rare disease (30). These variants are uploaded by participating clinical genetics and genomics centres, and phenotype annotations can be added to the upload in the form of HPO terms.

A number of bioinformatic tools are provided to analyse patient variants, including associated phenotype data. For example, a search by gene symbol gives a list of patients with variants in the gene. These variants can be filtered by consequence, inheritance and pathogenicity. HPO terms present in multiple patients are given as a list with the number of patients per term shown. It is also possible to see all the HPO terms annotated for a given patient variant (30).

DECIPHER does not directly define disease entities, however there is gene/disease association information provided, linking to other databases such as OMIM (29) and Gene2Phenotype (31). DECIPHER does not allow for direct download of any files. However, it is possible to access genomic variant and phenotype data through a Data Access Agreement for research purposes.

1.3.5 Gene2Phenotype

Gene2Phenotype (G2P) was designed to facilitate diagnostic filtering of genome-wide data for SNV (31). The aim of this work – the development of automated literature curation – is ultimately designed to enable population of G2P entries with enhanced phenotype data. Diseases are defined in G2P through the linking of a gene with a disease phenotype, inheritance mechanism and mutation consequence. The confidence level of this

assertion is provided. Links to papers in the peer-reviewed literature are provided as evidence of the association.

Phenotype data is recorded as HPO terms, with no weighting. G2P is curated in panels relevant to different disease domains, including cancer, eye and skin G2P. The developmental disorders G2P database (DDG2P) is focused on GDD. DDG2P is curated by Clinical Geneticists and Clinical Genetic Laboratory Scientists. There are over 2000 entries in this database, and it has not been possible to comprehensively add phenotypic data to all of them. Automated literature curation aims to address this.

1.3.6 Disease definitions and mapping between datasets

Each of the databases described above contains a substantial body of knowledge which, in theory, describes the same genetic disease-phenotype domain. However, disease entities are not uniformly described across the datasets. OMIM defines a disease name based on a phenotype grouping, e.g. ‘Generalized epilepsy with febrile seizures plus’ or eponymous syndrome e.g. ‘Schuurs-Hoeijmakers Syndrome’ (29). The curators of OMIM explicitly do not wish gene symbols to be part of the name for a condition and prefer instead to add sequential numbering to define separate diseases e.g. ‘Intellectual developmental disorder, X-linked 29’ (32).

The phenotype MIM number associated with a disease is stable and avoids the issues associated with heterogenous disease name conventions. However, the molecular basis of a disease is not clearly defined in OMIM, with only gene and inheritance pattern used. For example, disease phenotypes associated with *SCN1A* and an autosomal dominant inheritance pattern in OMIM include ‘Developmental and epileptic encephalopathy 6B, non-Dravet’, ‘Dravet syndrome’, ‘Febrile seizures, familial, 3A’ and ‘Migraine familial hemiplegic, 3’. Detailed review of the free text for each disease entry may or may not reveal a different mutation spectrum and underlying

molecular mechanism in these cases. For a number of diseases in OMIM, it could be argued that multiple entries for the same gene need to be merged, as they represent part of the same phenotypic spectrum. For diseases that are truly allelic, OMIM does not offer a straightforward way of differentiating between them on a molecular basis.

Orphanet structures disease phenotype annotations into a 'HPO-ORDO ontological module' (HOOM), where ORDO is the Orphanet Rare Disease Ontology, which is designed to describe relationships between diseases and genes (28). This allows for associations to be made between clinical entities and phenotypic abnormality (HPO term), with frequency data and evidence for this assertion. However, this does not address the issue of Orphanet defining clinical entities as phenotype groupings which may be linked to multiple genes as discussed in section 1.3.3 (28). This means diseases linked to a unique ORPHAcode identifier do not map exactly to conditions associated with MIM numbers.

G2P defines genetic diseases on a molecular basis, with phenotypic information associated with inheritance mechanism and mutation consequence for a given gene. Therefore, each entry is uniquely mapped to a single gene and disease names are not essential, although these are provided (31). Disease names in G2P may follow historical convention or an OMIM name in many instances. However, newer entries follow the gene-phenotype dyad suggested by Biesecker et al. e.g. '*CFTR*-related cystic fibrosis' (33). There are no unique identifiers per G2P entry, although these may be added in future. G2P entries may differentiate between allelic disorders with different mutation consequence e.g. loss of function vs gain of function. Allelic diseases with the same mutation consequence are recorded as separate entries with a flag 'restricted repertoire of mutations', with SNV-specific information planned to be added.

The differences in disease definitions between these datasets, and the consequent difficulties mapping diseases between them, have been recognised. Several international initiatives have been devised to address this issue. The Mondo Disease Ontology defines diseases through semi-automated merging of multiple source ontologies (34). It gives 1:1 equivalence mappings across sources, e.g. OMIM and Orphanet (27,28). The Gene Curation Coalition (GenCC) was formed to harmonize gene-disease validity assertions (35). This provides a database of unified disease definitions from sources including OMIM, Orphanet and G2P (27,28,31), with the confidence assertions of gene-disease association from these sources also documented.

This work aims to develop automated literature curation with the aim of populating DDG2P entries with HPO terms weighted according to their frequency in relevant manuscripts. This should not only improve SNV filtering with G2P, but also feed into the knowledge base of unified disease-phenotype initiatives as described above.

1.4 Text mining on full text manuscripts

To develop automated literature curation, it is one of the hypotheses of this thesis that it is preferable to use the full text of a manuscript. Phenotype data extraction from text at scale previously has relied on abstracts or full text open access papers (36,37). Unsurprisingly perhaps, there is a significant loss of information in biomedical text mining when using just the abstract. For example, information including protein-protein interactions (38) and polymorphism-drug keywords (39) may be present in the full text only. Named entity recognition (NER) on a corpus of 15 million papers covering a range of scientific domains in full text significantly outperformed abstract-only analysis (40). Open access full text may be used to address this, however only a small proportion of peer-reviewed manuscripts are available through this route.

At the time of writing, Europe PubMed Central (EPMC) (41) contained approximately 39.5 million articles, of which only 3.8 million were open access. To the authors' knowledge, full text mining for phenotypes in the GDD domain has not been performed previously. Therefore, there is likely a significant volume of phenotypic data which has not been utilised at scale.

Legal download is possible for all papers accessible through university-wide access agreements (42). Scalable access to parseable, high-quality full text downloads is not straightforward. However, the Cadmus package has been developed for this purpose, with a high rate of successful downloads (43), and will be discussed in further detail as part of this work.

Another issue relevant to data extraction from the peer-reviewed literature is the heterogenous manner in which phenotypic information is presented. Manuscripts may have different underlying structures, which affects how easily they are parsed. For example, the American Journal of Human Genetics (AJHG) typically includes articles, which are subdivided into Introduction, Material and Methods, Results and Discussion; and reports which are single blocks of text (44). Therefore, parsing clinical information from articles is much more straightforward, as this can follow the structure of the manuscript. Extracting specific data from reports requires more advanced methods.

Phenotypic information is not just found in free text. Papers describing GDD frequently include tabular data, and often these may be the only individual-level patient descriptions in the manuscript. Additionally, supplemental files may contain detailed case reports, in free text or tabular form.

1.5 Named entity recognition for phenotype data

1.5.1 Overview of named entity recognition

NER is the process of identifying parts of text which correspond to defined entities, for example genes, proteins, diseases and phenotypic features (45). Given the scale of the biomedical literature, with millions of papers in EPMC alone (41), NER has become an increasingly important technique for extracting information at scale, for application in fields including pharmaceuticals (46) and genomics (47).

NER for biomedical text offers specific challenges (48). Entities are usually defined using domain-specific technical vocabularies. Compound or nested terms may be present e.g. ‘renal and pancreatic cancer’, ‘there was tortuosity of many of the arteries, although this was markedly more severe in the carotid artery’. Modifiers may be used as part of descriptions e.g. ‘subtle thickening of the macula with oedema’. Synonyms or non-standard metaphorical descriptors can be used e.g. ‘hitchhiker thumb’. Finally, ambiguous abbreviations are often present e.g. ‘ASD’ can refer to ‘atrial septal defect’ or ‘autism spectrum disorder’ (49).

1.5.2 Corpora generation

Biomedical text in NER needs to pass through several stages to allow accurate identification of defined entities. First, a corpus of domain-specific documents needs to be identified, for example through PubMed search. The challenge here is to create a corpus enriched for appropriate papers without including an excess of non-relevant work. It is possible to narrow down results using appropriate filters, however searches may still return thousands or millions of results. Manual assessment of all papers found using an appropriate strategy, as used for systematic review, remains the benchmark in this area, even when using highly filtered searches (50).

Corpora may be annotated with entities for the purposes of training/testing NER techniques. A gold-standard corpus is one which is annotated manually by experts according to well-defined guidelines (48). To aid development of biomedical NER, a number of annotated corpora have been made publicly available. Significant effort is required to generate these, meaning the volume of text included is often small, particularly for HPO/phenotype specific corpora. For example, Groza et al. released a corpus of 228 abstracts annotated with HPO terms, called the HPO Gold Standard (HPO GS) corpus (49). This example illustrates further some of the difficulties inherent in creating a gold standard corpus, with inconsistencies in manual annotation. It was annotated by three experts, who in fact created the HPO itself, using strict documented criteria. Despite this, Lobo et al. (51) identified 881 entities not annotated in the HPO GS, and added these to create a corpus designated HPO GS Plus.

Larger corpora are available, however these are annotated using multiple biomedical source vocabularies/ontologies e.g. the MedMentions corpus containing 4000 abstracts (52) and the CRAFT corpus which utilises 97 full text papers (53). To increase the size of these corpora, computerized methods may be used for annotation ‘silver standard’, although this necessarily means a decrease in accuracy compared to manual methods (54).

1.5.3 Sentence boundary detection, tokenization and part-of-speech tagging

NER requires text to be processed into a machine-readable form. This initially requires download of relevant abstracts/manuscripts and cleaning/parsing into a unified format. This will be discussed in further detail later in this work, with relevance to the Cadmus package (43).

Once this step is complete, the text needs be divided into readable units. First, sentence boundaries need to be determined. In theory, this can be

simply achieved by splitting text at full stops (“.”). However, there are many instances in biomedical text where these do not indicate sentence boundaries, for example abbreviations (“e.g.”, “Am. J. Hum. Genet.”), numbers (“2.4”) and copy number variants (“1q21.1 deletion”) (55). Next, the sentences are split into units of meaning, called tokens (tokenization). These may be words, characters or parts of words (n-grams). This could be achieved by splitting on whitespace, but more ambiguous dividers may be present, e.g. parentheses (“Ca(2+)”), and hyphens (“intravenous feed-dependent”) (55).

Given the potential complexity of sentence boundary detection and tokenization for biomedical text, a number of specialist parsers have been developed for these tasks, with generally high performance achieved. For example, Verspoor et al. evaluated the performance of three sentence boundary detectors with an F1 score of 0.8-0.98; and four tokenizers, with F1 of 0.88-0.95 (53). Part-of-speech (POS) tagging is then applied, assigning tokens to language classes, e.g. noun, verb adjective. This is important for the accuracy of NER, as it influences phrase construction. In one example, the phrase “history of significant left lower quadrant pain” was mistakenly parsed into “history of significant left” and “quadrant pain” due to erroneous tagging of “left” as a noun (56). Again, POS tagging benefits from training on domain-specific text for optimal performance (57).

1.5.4 Dictionary-, rule- and machine learning-based NER

NER methods in biomedicine can be broadly divided into three categories, dictionary-based, rule-based and machine learning-based. The above processing steps in general apply to both techniques. Dictionary-based NER maps text to a well-defined vocabulary. Rule-based systems use defined linguistic structural features to identify entities. Machine learning-based NER develops a representation of observed data using specified features by training on annotated text (48).

In practice, NER methods may use a hybrid of different approaches, particularly for dictionary + rule-based strategies. Examples of NER approaches used in the biomedical domain, particularly with relevance to phenotype data and GDD, will be reviewed here.

1.6 Named entity recognition tools focused on the HPO

1.6.1 Overview of NER tools

In this section, I have reviewed several tools which may be used for NER in biomedical text. I have particularly focused on those which have been tested using the HPO. A number of other tools are available in this domain (58,59). These may use the full UMLS vocabulary, which includes the HPO (19,60). These could therefore be adapted for use in phenotype-specific NER, but in the absence of literature specifically testing this, they are beyond the scope of this thesis.

1.6.2 NCBO annotator

The NCBO (National Center for Biomedical Ontology) annotator is a highly configurable dictionary-based system which runs primarily as a web service, but is also available to install locally in linux via a virtual machine (61). It uses text metadata to annotate input with biomedical ontology concepts from the UMLS Metathesaurus (23) and NCBO BioPortal (62). The NCBO annotator uses a dictionary constructed of strings corresponding to ontology concepts.

A concept identifier called Mgrep (63) is used to match input text to terms from the dictionary. These annotations may be expanded to related concepts e.g. using ontology parent-child relationships, MeSH (Medical Subject Headings) term siblings or cross-ontology mappings (61).

1.6.3 MetaMap

MetaMap is a highly configurable system designed to identify concepts from the UMLS Metathesaurus in biomedical text (64). It can be installed to run natively on linux, although a licence is required from the National Library of Medicine (23). MetaMap utilises both dictionary and linguistic rule based analysis. Input text firstly undergoes tokenization, sentence boundary determination and POS tagging as described above. Input words are then looked up in the SPECIALIST lexicon (65), which includes biomedical concepts sorted into syntactic categories.

Phrases are identified by the SPECIALIST parser (65). Tables containing variants of the words in these phrases are used to identify likely candidate terms. These are then evaluated to determine the best match to the text, with the option of weighting towards concepts which are semantically consistent with the adjacent text (64).

1.6.4 OBO annotator

The Open Biological and Biomedical Ontology (OBO) annotator matches biomedical concepts using indexed data structures (66). Lexical indexes are built by putting each concept in a reference ontology through stages including changing to lower case, tokenizing and using the stems of terms to generate lexical variants. Words in the input text are processed in a similar manner.

A window including x number of words is then run over the text to match word sequences to the lexical index. If no match is found for a given size of window, the word sequence is subdivided, e.g. “brain and cerebellar atrophy” is cut to “brain atrophy” and “cerebellar atrophy”. The longest word sequence is chosen where more than one concept is present e.g. where “seizures” is contained in “myoclonic seizures”.

1.6.5 Monarch annotator

The Monarch Initiative has developed a knowledge graph from multiple sources with input databases including information on, for example, animal models, gene expression, protein-protein interaction, as well as human gene-disease phenotypes (OMIM, Orphanet) (27,28,67). The graph contains over 33 million nodes and 160 million edges (67). The Monarch annotator is available through a web interface, which allows identification of ontology concepts in free text. However, the precise NER technique used has not been made publicly available (67).

1.6.6 BioLarK

Bio-LarK is a rule-based concept recognizer explicitly designed to identify HPO terms. It uses linguistic rules to normalise input terms and map lexical variants. These are mapped to indexed HPO terms. Conjunctive terms e.g. ‘straight and narrow eyebrows’ can be optionally processed using pattern matching based on manually derived rules applied over sentence structure (49).

1.6.7 Identifying Human Phenotypes

Identifying Human Phenotypes combines machine learning and rule-based approaches (51). Input text is processed using a Conditional Random Field (CRF) model (68) trained on the HPO annotated corpus provided by Groza et al. (49). A CRF may be used to derive the conditional probability of a label (in this case an HPO concept) given a sequence of tokens from the input text.

Entities recognized through this step then undergo a manual validation stage, which includes exact matching to concepts in an HPO dictionary, identification of sequences containing nested and compound terms.

Concepts containing punctuation errors, abnormal structure (according to position of POS tags) and stop words are removed (51).

1.6.8 Neural concept recognizer

Arbabi et al. developed a system called the Neural Concept Recognizer based on a convolutional neural network (CNN), trained on the HPO (69). Concepts in the ontology were represented as a matrix, with one concept per row. The model learned features of each concept which were novel compared to its ancestors. The embedding of each concept was ultimately the embedding of its parent (to represent its location in the embedding space) plus the raw concept embedding (to enable learning of the local location compared to the parent).

Input text words were initially represented as a bag of character n-grams. Each of these was represented as a vector and their sum was used as a representation of the word (70). The word vectors for a given phrase in text then went through a convolutional layer, i.e. the dot product of the vector and a filter array of weights, to generate an encoded representation of the phrase.

The dot product of this encoded representation and concept embeddings was used as a similarity score for mapping. Concept recognition in a sentence was performed by matching all one to seven word n-grams to concepts. For n-grams overlapping the same text, the shorter n-gram was retained if they match the same concept, and the longer n-gram if different concepts were mapped.

1.6.9 BioBERT and PhenoTagger

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a language representation model specific to the biomedical domain (71). It is based on BERT (72), which is a large

transformer-based language model, pre-trained on large general English corpora: Wikipedia and BooksCorpus (73). This uses bidirectional transformers to learn representations of words in context. Bidirectional in this case means sentences can be analysed both left-to-right and vice versa, which is more powerful for predicting meaning (74). This is enabled using masking, to predict randomly masked words in a sequence (72).

After pre-training, BERT can be fine-tuned on text mining tasks. BioBERT uses the same structure and initial weights as BERT, but with pre-training on PubMed abstracts and PubMed Central full-text articles for higher performance in the biomedical domain (71). Other BERT models are available with a biomedical focus (75,76). However, BioBERT in particular has been utilised in the context of GDD and the HPO to develop a method called PhenoTagger (54).

PhenoTagger is a method utilising a combination of BioBERT and dictionary-based methods to identify HPO concepts (54). This uses a training corpus developed from 150,052 open access PMC articles relating to ‘disease and mutation’. A dictionary based tagger annotated HPO concepts in the corpus, and this was used to fine-tune BioBERT. This model was then used to identify HPO concepts in biomedical text.

1.7 Performance of NER methods on biomedical text

Annotator	Source	Test Corpus	Precision	Recall
NCBO	Shah et al., 2009 (77)	PubMed	0.77	n/a
NCBO	Taboada et al., 2014 (66)	CX	0.97	0.49
NCBO	Groza et al., 2015 (49)	HPO GS	0.54	0.39
NCBO	Lobo et al., 2017 (51)	HPO GS	0.69	0.46
NCBO	Arbabi et al., 2019 (69)	HPO GS	0.80	0.49
NCBO	Arbabi et al., 2019 (69)	UD	0.37	0.20
NCBO	Oellrich et al., 2015 (78)	ShARe/CLEF	0.04	0.51
OBO	Taboada et al., 2014 (66)	CX	0.94	0.61
OBO	Groza et al., 2015 (49)	HPO GS	0.69	0.44
OBO	Lobo et al., 2017 (51)	HPO GS+	0.77	0.34
OBO	Arbabi et al., 2019	HPO GS	0.80	0.59
OBO	Arbabi et al., 2019 (69)	UD	0.29	0.20
OBO	Luo et al., 2021 (54)	HPO GS+	0.81	0.57
MetaMap	Shah et al., 2009 (77)	PubMed	0.76	n/a
MetaMap	Reátegui & Ratté, 2018 (79)	i2b2	0.78	0.91

Annotator	Source	Test Corpus	Precision	Recall
MetaMap	Oellrich et al., 2015 (78)	ShARe/CLEF	0.04	0.35
Bio-LarK	Groza et al., 2015 (49)	HPO GS	0.65	0.49
Bio-LarK	Lobo et al., 2017 (51)	HPO GS	0.65	0.49
Bio-LarK	Arbabi et al., 2019 (69)	HPO GS	0.77	0.66
Bio-LarK	Arbabi et al., 2019 (69)	UD	0.29	0.22
IHP	Lobo et al., 2017 (51)	HPO GS	0.56	0.79
IHP	Lobo et al., 2017 (51)	HPO GS+	0.87	0.85
NCR	Arbabi et al., 2019 (69)	HPO GS	0.81	0.68
NCR	Arbabi et al., 2019 (69)	UD	0.27	0.29
Doc2HPO	Luo et al., 2021 (54)	HPO GS+	0.77	0.62
Doc2HPO	Liu et al., 2019 (80)	CN	0.47	0.76
MI	Luo et al., 2021 (54)	HPO GS+	0.76	0.61
PhenoTagger	Luo et al., 2021 (54)	HPO GS+	0.77	0.74

Table 1-1. Comparison of precision and recall figures for example annotators used on biomedical text. Annotators used for HPO terms and/or GDD preferentially included where possible. Annotators: NCBO – NCBO Annotator (61), OBO – OBO Annotator (66), MetaMap (64) , Bio-LarK (49), IHP – Identifying Human Phenotypes (51), NCR - Neural Concept Recognizer (69), Doc2HPO – utilises MetaMap in concept recognition (80), MI – Monarch Initiative Annotator (67), Phenotagger (54). Test corpora: PubMed – 200 lines

chosen randomly from PubMed abstract downloads (77) CX - 50 PubMed abstracts for case reports describing 'Cerebrotendinous Xanthomatosis' (66), HPO GS – HPO Gold Standard (49); 228 abstracts cited by OMIM (27) describing GDD, annotated by three experts with HPO terms, UD - 39 clinical patient reports for undiagnosed diseases, ShARe/CLEF (81) – 99 annotated documents from corpus of clinical reports, HPO GS+ - HPO Gold Standard Plus (51); HPO GS with additional annotated entities, i2b2 – Informatics for Integrating Biology to the Bedside (82); 1237 clinical discharge summaries for individuals with obesity and diabetes, CN (80) – 18 clinical notes from New York-Presbyterian/Columbia University Irving Medical Center.

Annotator	Mean precision	Mean recall
NCBO	0.60	0.42
OBO	0.72	0.46
MetaMap	0.53	0.63
Bio-LarK	0.59	0.47
IHP	0.72	0.82
NCR	0.54	0.49
Doc2HPO	0.62	0.69

Table 1-2. Mean precision and recall for selected NER methods. Annotators included only where >=2 test examples available. Annotators: NCBO – NCBO Annotator (61), OBO – OBO Annotator (66), MetaMap (64) , Bio-LarK (49), IHP – Identifying Human Phenotypes (51), NCR - Neural Concept Recognizer (69), Doc2HPO – utilises MetaMap in concept recognition (80).

The performance of NER methods can be evaluated using precision and recall (P/R) metrics. Precision is the proportion of true matches in the retrieved instances. Recall is the proportion of all true matches returned. Table 1-1 gives P/R figures for the NER methods reviewed above. This illustrates some of the limitations with assessment of these techniques. Firstly, there is a lack of large gold standard corpora for testing, particularly in relation to HPO/phenotype annotations, as discussed previously. The HPO

GS corpus was used in many instances, comprising only 228 abstracts; no significant full text corpus was utilised.

Secondly, markedly different results may be obtained for the same annotator/NER method. The replicability of NER performance may be expected to vary depending on which test corpora are used, for example patient notes and PubMed abstracts will likely have very different structures and described entities. This particularly affects machine-learning based methods. Domain-specific training has been shown to improve results (83). However, given the small size of many training corpora in the rare disease domain, there is a danger of over-fitting, with resulting poorer performance when extrapolating results from one corpora to another (84).

In several cases, divergent results were obtained even for the same annotator apparently using equal corpora. For example, Groza et al. (49), Lobo et al. (51) and Arbabi et al. (69) evaluated the NCBO annotator (61) on the HPO GS (49) corpus with precision of 0.54, 0.69, 0.80 and recall of 0.39, 0.46, 0.49. However, on closer inspection, only Groza et al. utilised the full corpus for testing. The other two groups split the corpus for the purposes of training machine learning NER methods, and hence used only a proportion of annotated abstracts.

Additionally, many of the annotators shown here are highly configurable, and the settings used for testing are not given in the majority of examples above. Table 1-2 demonstrates the apparent significant differences in performance using mean P/R figures. There are instances where several annotators are tested on the same corpus by the same authors. For example, Arbabi et al. compared OBO, NCBO and BioLarK (69). However, there is no work directly comparing these to other annotators of interest, such as MetaMap, of which I am aware. Overall, it is clear that performance metrics for NER methods need to be compared using the same corpus and test criteria for meaningful, replicable analysis.

1.8 Phenotype models and similarity measures

1.8.1 Disease models and comparisons

Using the strategies described above, diseases, and particularly GDD, may be defined phenotypically using lists of standardised terms, in this case the HPO. These disease models may be useful in themselves, particularly if they synthesise different sources of information/publications. For example, the clinical synopses presented by OMIM may be used in variant interpretation.

The full potential of these disease models, however, is realised through computational analyses, particularly through comparison of models. For example, HPO terms recorded for an individual suspected with GDD could be compared with disease models to help identify the correct diagnosis. Simple exact matching between lists of terms in this case may not always reflect true disease similarity. Clinically similar terms may be recorded differently between models, depending on user preference, e.g. “Intellectual disability” and “Intellectual disability, moderate”, and hence not be matched.

Fuzzy matching, for close matches between the words/characters in a term, is also problematic as unrelated terms may share a significant number of characters, e.g. “Dental deformity” and “Sternal deformity” or “Hypoplastic palate” and “Hypoplastic pons”. However, given sufficient numbers of terms being compared, list matching may still be useful, particularly if weighting is applied to prioritise higher-ranked terms.

More complex approaches utilise semantics, i.e. the relationship between concepts, taking advantage of the ontology structure. Strategies for phenotype comparison will be further discussed in the following sections.

1.8.2 Item-based comparisons

Disease models may consist of groups of unweighted phenotypic descriptors, or items. In some cases, these may be weighted according to concept frequency – in the peer-reviewed literature (28) or in clinical studies (24) – to create ranked lists of phenotype descriptors. There are several properties of these type of itemisation or lists which need to be considered.

Conjointness describes whether lists contain the same items (85). In the GDD domain, comparison lists of this kind will inevitably be non-conjoint, therefore, list-based phenotype comparators used need to have the capacity to handle these. Additionally, it would be preferable to use a measure which is top-weighted (85), i.e. prioritises higher ranked terms, as these are more likely to be characteristic for a disease.

Most list comparators either require conjoint lists e.g. Kendall tau (86), but can be weighted (87) or are unweighted, but can use non-conjoint data (88). Webber et al. describe a method called rank-biased overlap (RBO) which can compare non-conjoint and weighted lists, by using the weighted mean overlap of terms at different evaluation depths (89), to address this.

Other methods for item-based comparison include those based on information. For example, pointwise mutual information (PMI) is a measure of how likely events are to co-occur, given their individual probabilities (90). Konagurthu et al. describe a method based on the compressibility of the item groupings, and uses the size of a lossless encoding of each to generate a similarity metric (91). However, neither of these address the issue of non-conjointness.

1.8.3 Semantic similarity

List-based similarity methods, however, do not take advantage of the relationships between terms defined in an ontology such as the HPO. These can be utilised through approaches which define similarity based on the meaning or semantic content of terms (semantic similarity).

Initial use of these methods in biomedical ontologies was primarily based on the Gene Ontology (92). These can be divided into node-based, utilising the terms themselves (93), or edge-based, using the links or relationships between terms (94). Semantic similarity methods using the HPO have mainly been based on node-based techniques, particularly utilising information content (IC).

1.8.4 Information content

Intuitively, terms in the HPO vary in their informativity, or specificity when used in semantic similarity methods. For example, a high-level term such as ‘Abnormality of the cardiovascular system’ (HP:0001626) is likely to be annotated to many diseases, with a low specificity. In contrast, a low-level term such as ‘Limbal dermoid’ (HP:0001140) is likely to be rare and hence more informative. The specificity of a term in an ontology can be defined by its IC, which is related to frequency (95). The frequency is the number of times a term, and all of its ancestors, appears per annotated disease. Annotation of a term with its ancestors is assumed by the structure of the HPO (96).

Use of these ancestral annotations means the true frequency of a term may be established. For example the high level term ‘Abnormality of the cardiovascular system’ (HP:0001626) mentioned above may well not be used for input annotation explicitly, however there will likely be many of its more specific children utilised, and therefore its true frequency is high (97). IC is

then calculated as $-\log_2$(frequency) for each term (96,98). This means that IC is potentially affected by the number of annotations and diseases in a dataset. If this number, N, is large enough, however, the IC will stabilise as N $\rightarrow \infty$ (99,100).

1.8.5 Most informative common ancestor (MICA)

For a given pair of terms, Resnik proposed a similarity metric based on their most informative common ancestor (MICA) in the ontology (100) (Figure 1-3). This is the method used most commonly in relation to the HPO, and hence will be discussed further below. However, it should be noted that modifications have been made to Resnik's method. For example, Lin adapted the measure to include an assessment of the closeness of comparison terms to the MICA (101). Schlicker et al. modified this further to include the specificity of the MICA (93). This addresses the issue that two high-level terms sharing a common ancestor may appear to have strong similarity using MICA, whereas intuitively two lower-level terms with a common ancestor should score higher, as they are more specific. These may prove useful when refining HPO-based disease model similarity comparisons in future.

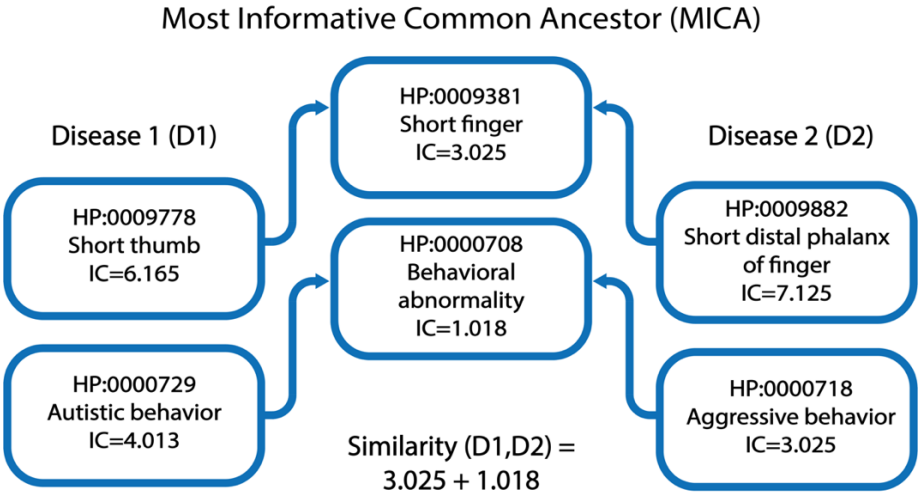


Figure 1-3. Most Informative Common Ancestor (MICA) semantic similarity metric. Adapted from Helbig et al. (98). MICA scores for sets of terms can be summed to create similarity scores for disease models. IC – information content.

1.8.6 Semantic similarity and HPO

The initiation of the HPO allowed semantic similarity metrics to be developed in the human genetic disease phenotype domain. These primarily utilise the MICA method. For example, Köhler et al. proposed the 'Ontological Similarity Search' (96). Here, each input term from a query is matched to the maximum MICA amongst disease annotated terms, and the mean of these values for all queries used to create a similarity score. The sum of maximum MICA scores may also be used to generate a similarity metric (98) (Figure 1-3).

The raw similarity score, however, is influenced by the number of annotations for a given disease. Larger numbers of annotations tend to result in a higher score (102). Therefore, Köhler et al. generate empirical p-values for each query-disease comparison, by comparing the similarity score with randomly generated models of the same size as the query (96). These are corrected for multiple testing using the Benjamini-Hochberg method for detecting the false discovery rate (103). This utilises the distribution of simulated p-values to reject those which lie below the line of α/N, where α is the threshold false discovery rate and N is the number of p-values generated.

The MICA-based semantic similarity method has been used in a number of applications. For example, similarity scores were used to predict diagnoses for simulated patient models using OMIM data (29,96). Phenotypically similar patients in a developmental and epileptic encephalopathy cohort were found to have the same de novo missense SNV in *AP2M1,* with subsequent identification of other individuals with the same variant (98). Distinct phenotypic signatures have also been identified for a subset of epileptic encephalopathy genes (97). Mutations in different parts of the glycosylphosphatidylinositol-anchor biosynthesis pathway have been shown to result in discrete phenotype differences (104).

The above MICA semantic similarity methods use unweighted models, that is each HPO term occurs only once per disease/patient entity. Köhler demonstrated an adaptation of Resnik's work to include weighting by repeating HPO terms according to their frequency per entity (which is defined according to the source used to generate these) (100,105).

Adaptations to IC based measures using the structure of HPO have also been proposed. For example, Xue et al. outline a measurement called DisPheno where the edges of the HPO graph were weighted according to the information content of genes/diseases annotated to each term (106). Similarity was calculated according to the probability of terms in query and disease overlapping using PMI. This was reported as performing better than the methods of Resnik, Lin and Schlicker for patient-disease matching on a simulated dataset (93,100,101,106).

1.8.7 Vector-based similarity

Lists of ontology terms can be represented in vector space. Each entry in the vector can be a binary representation of term annotation per model, or can be used for another property, for example the IC. Vector-based similarity methods can then be applied in analysis.

These include cosine similarity, where the cosine of the angle between vectors is used, and the Jaccard index, which is the ratio of the intersection and the size of the union of two sets (107). Mistry et al. showed that the IC method of Resnik and vector-based methods are correlated, using the Gene Ontology (92,100,108). Vector based similarity methods have not been used commonly with HPO, although Köhler showed similar performance using these to Resnik IC (105).

1.8.8 HPO diagnostic systems

Systems for diagnostic prediction using semantic similarity and HPO have been made publicly available. The basic MICA-based method by Köhler et al., described above, has been incorporated into a web application called the Phenomizer (96), where users can manually input HPO terms to generate a list of suggested diagnoses. Zemojtel et al. incorporate the Phenomizer into a system called Phenotypic Interpretation of eXomes (PhenIX), where the input is a variant call format (VCF) sequencing file, as well as patient HPO terms. Genes within which variants (SNV) are found in the VCF file are ranked according to the semantic similarity of the input patient terms to those annotated to diseases in OMIM (27,109).

James et al. describe a similar approach for the OMIM Explorer, where input phenotypes are used to prioritise SNV according to semantic similarity (110). This offers the additional functionality of a graphical interface where users can exclude clinically non-relevant genes, altering the underlying prioritisation scores.

It should be noted that other systems have been created to leverage phenotypic data for variant prioritisation, although these are beyond the scope of this work. For example, Phevor integrates data from multiple ontologies (111), including HPO, to combine with input phenotype terms for ranking variants. The eXtasy system utilises genes annotated with functional data from multiple sources, including HPO terms mapped to descriptions of diseases in HGMD using Phenomizer, as well as variant effect prediction, to prioritize variants (96,112,113). PHIVE (Phenotypic Interpretation of Variants in Exomes) uses similarity for both human and mouse phenotypes to prioritise variants from sequencing data (114).

1.8.9 Automated disease-phenotype mapping

The methods described above for text mining and matching of disease phenotypes largely rely on using manually identified manuscripts/abstracts, which is resource intensive and difficult to scale. Standardized vocabularies, such as the HPO, and text mining can be utilised to form comprehensive knowledge bases covering all diseases in a particular domain.

For example, Collier et al. developed a system called PhenoMiner (36). They searched Europe PMC (41) for all OMIM disease names and their synonyms (27). Each paper identified through this search was annotated to its canonical disease name. NCBO annotator (61) was used to extract HPO terms from these papers, regardless of source, i.e. any organism or mode of study could be included.

Association rule mining was used to identify rules describing the co-occurrence of HPO terms and diseases (115). This is a technique where association rules are inferred from frequently occurring items in a transaction set, often used in retail analysis. For example, if a customer buys milk, there is a 60% chance they will buy bread. The disease-phenotype mapping from this technique was assessed on a sample of 200 terms by a panel of expert reviewers as 38% accurate (36). The same group further developed this concept into PheneBank (116). This used the co-occurrence of disease-phenotype and gene-phenotype terms in Medline abstracts or excerpts from PubMed open access papers. Fisher Exact Testing was utilised to determine significant associations. The PheneBank database is available to search via a web interface (116).

Hoehndorf et al. also used the co-occurrence of disease and phenotype terms to create the ‘human diseasosome’ (117). From an indexed database containing all Medline (23) abstracts, they identified approximately 5 million documents containing at least one HPO or Mammalian Phenotype Ontology

(MP) term or its synonym (22,118). They used AberOwl, which uses semantic querying to identify ontology concepts (119), to look for HPO, MP, and disease ontology (DO) (120) terms in these abstracts. Disease models were constructed by using PMI (90) scores to select highly associated phenotype terms. They demonstrated an increased number of terms per model compared to those in OMIM (27), and a similar performance when predicting disease-associated genes using semantic similarity (117).

Xu et al. used prior knowledge from disease-phenotype associations in the UMLS (121) as a basis for learning linguistic patterns in Medline citations (122). An example pattern is {disease x} *is characterized by* {phenotype term}. The extraction corpus was 120 million sentences from Medline citations (23) indexed over a 40 year period. There was some correlation with OMIM disease genes for the extracted disease-phenotype pairs (27,122).

The methods for automated phenotype mining described above all emphasise creating a large corpus for data extraction, over selection of high-quality literature describing a relevant disease. Only the PhenoMiner system (36) describes filtering search results in biomedical databases, using OMIM disease names (27). However, it is likely these searches will still contain large numbers of non-relevant results. Indeed Collier et al. state that PhenoMiner includes phenotype concepts regardless of organism or type of study (36).

1.8.10 Automated literature search

It would seem reasonable to assume that the accuracy of disease-phenotype extraction from an automated extraction process would increase if the source data was enriched for relevant case reports/case series in humans. To my knowledge, this problem has not been addressed in the HPO/GDD domain. There are, however, studies looking at automated search from a medical

perspective, particularly in relation to identifying high-quality articles for systematic reviews.

Aphinyanaphongs et al. assessed the performance of various classifiers compared to PubMed Boolean queries to identify high-quality manuscripts for selected medical applications. Approximately 400 articles were pre-selected as high-quality in the ‘treatment’ or ‘etiology’ categories by the American College of Physicians journal club according to defined criteria (123). Features used included MeSH terms and publication type. From a (train/test split) corpus of 16000 Medline articles containing the high-quality articles, they found the support vector machine (SVM) classifier had the highest performance, and achieved higher recall than Boolean queries for both ‘treatment’ (0.80 vs 0.40) and ‘etiology’ (0.76 vs 0.28).

Kilicoglu et al. used an ensemble learning method, using the results of several different machine learning classifiers, including Naïve Bayes and polynomial SVM, to identify high-quality, methodologically rigorous papers (124). Features included words in title + abstract, citation metadata and UMLS Metathesaurus (121) concepts in title + abstract. A corpus of (train/test split) 12000 documents designated as ‘rigorous’ or ‘non-rigorous’ as part of the development of the PubMed search filter function (23) was used. They found that the ensemble method outperformed any individual method, with precision of 0.75 and recall of 0.86.

Kim et al. developed a system to identify high-quality articles for systematic reviews, using the hypothesis that articles commonly excluded from reviews covering multiple topics would be useful in training a classifier (125). They identified 7200 articles which had been included/excluded from 126 systematic reviews. Features used for learning were words in title + abstract and MeSH terms. They showed that the performance of an SVM, as assessed by AUC (area under the receiver operating characteristic curve), was significantly higher when using the commonly excluded articles for

training. This has implications for a GDD-focused classifier, as there may be commonly excluded articles in this domain which could be utilised.

Cohen et al. also explored the utility of using cross-topic inclusion/exclusion data for systematic reviews (126). A corpus of 50927 articles was used, describing multiple drug treatments, annotated as excluded or included according to individual systematic review criteria. For a given drug topic, an SVM model was trained on data from other systematic reviews, before being trained on topic-specific articles. They showed a performance enhancement, as assessed by AUC, for this method where topic-specific data was scarce. There was no impairment to performance with higher levels of topic-specific information (126). This may be relevant to a GDD classifier, given the often very small numbers of papers available for a given disorder.

A different approach was explored by Bian et al, who assessed whether features related to a paper's impact, i.e. whether it is likely to affect clinical practice, may be useful in developing a classifier (127). They chose time-agnostic features, meaning these could be established soon after publication, rather than waiting for e.g. citation metrics. These included journal impact factor, number of associated grants, number of authors, and number of references. A corpus of articles identified as high impact in systematic reviews was used for positive examples. The results of PubMed (23) searches related to the systematic review criteria were used as a negative corpus. They showed that a decision tree classifier performed similarly to the PubMed Best Match algorithm – which is based on term frequency and machine learning classifiers – and the method of Kilicoglu et al. (124,128). Performance overall was poor, with precision of 0.39 and recall of 0.09.

For well-characterised GDD, citation metrics may prove more useful than the immediate impact-based features described above. Bernstam et al. showed citation-based search was more effective than simple keyword search in PubMed for identifying high-quality articles (23,129,130). However, Bian et

al., using a similar impact-based classifier to that described above by the same group, found that adding or subtracting citation count from the feature list did not markedly affect performance (131).

The datasets used in the examples given above are almost entirely affected by imbalance, where the useful/positive examples are significantly outnumbered by those which are non-relevant/negative. Extreme imbalance may be present in literature search in the biomedical domain. For example, Bian et al. used a corpus with 541 positive citations and 45,012 negative citations (127). Imbalanced data has been shown to negatively affect the performance of classifiers (132). There are several methods which could be used to alleviate this issue, when considering automated search for GDD-relevant publications. PubMed keyword search + Boolean filters should help reduce imbalance in the data, prior to classification. It is likely that intelligent use of staged search criteria, for example adding “NOT Cancer” where a gene is highly studied by oncology focused groups, will be effective. Ensemble learning, where multiple machine learning classification methods are hybridized, has been shown to be increase performance when used on imbalanced data (133). Cost-sensitive learning may also be used, where costs are applied to different types of misclassification errors (134).

Finally, it is noted that the peer-reviewed literature is not the only source of disease-phenotype annotations. Electronic health records (EHR) represent an increasingly rich resource of phenotypic data, particularly with the trend towards recording clinical information in a computationally readable form, rather than using handwritten notes (135). However, there are significant obstacles to the use of EHR in a research setting. These include safeguarding of confidentiality, and heterogenous data recording and management systems across different clinical providers (136). Using the HPO, text mining of EHR has been shown to enrich phenotypic annotations in OMIM (27,137,138).

1.8.11 Aims and organisation of the thesis

The previous sections presented an overview of text mining in the context of phenotype data for GDD. Following on from this, I generated the following research questions, which this thesis aims to address:

1. Is it possible to generate weighted disease models describing GDD from full text peer-reviewed literature?
2. Do these models reflect true disease expressivity?
3. How do the models compare to manually curated databases?
4. Can these models be used for prediction of GDD diagnosis?
5. Is it possible to identify GDD – relevant manuscripts using automated search alone?

This thesis is divided into three major methodological chapters, in which the above questions are explored in detail. In the first of these – Chapter 2 – I outline the complexity of extracting phenotypic data from the biomedical literature. I show that weighted disease models can be created from the full-text literature using several different NER methods. I describe the process of evaluating the performance of these NER techniques. I discuss the limitations of the HPO and strategies to address these. I also clinically evaluate the created disease models.

In Chapter 3, I describe the utility of phenotypic disease models, and review diagnostic methods using these. I outline the process of creating a large set of disease models describing GDD. I use this set to analyse disease model structure, and explore the effects of modifications to the methods used to create them. I use similarity metrics to evaluate the performance of the full text-derived models to those from widely used manual curation in predicting GDD diagnoses. The process of disease model creation and analysis is summarised in Figure 1-4.

In Chapter 4, I present preliminary data describing methods to scale up disease model creation to the full spectrum of GDD. I analyse the data architecture of case reports/case series describing GDD, to inform document parsing. I evaluate different PubMed search strategies to create corpora pre-enriched for GDD. I describe the process of using these searches to create a large corpus of annotated manuscripts, with relevance to GDD. I use this to test numerous machine learning classifiers for biomedical text and evaluate their performance. I explore the next steps to build upon this work in future. I discuss improvements to manuscript classification, text mining and disease evaluation, with reference to methodological studies from the literature which could be applied to these problems.

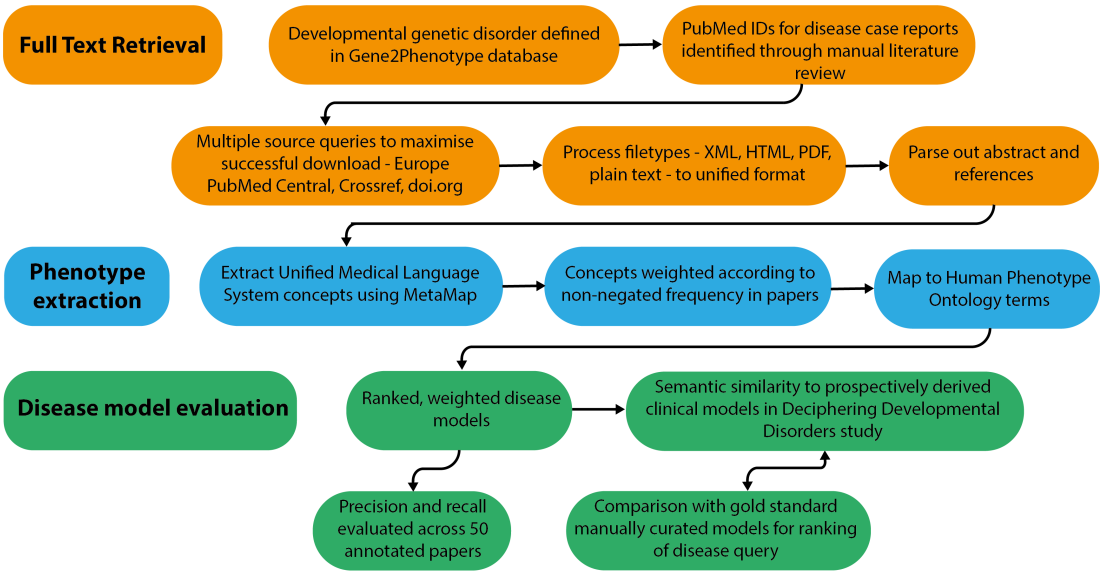


Figure 1-4. Overview of disease model creation and evaluation. Analysis of each of these steps forms the majority of this thesis.

Chapter 2 Creation of weighted phenotypic disease models

2.1 Introduction

In this chapter, I present an iterative process used to develop the disease model concept, where these comprise lists of weighted HPO terms (22), describing GDD. I first use manually-extracted phenotype terms to create disease models as a pilot study or proof-of-concept. This mimics the widely used manually curated databases OMIM and Orphanet, which store disease models generated by manual curation (27,28). The phenotypic descriptors stored in these databases have been directly extracted from the peer-reviewed literature by expert biocurators. These databases are widely used clinically, as it is likely this manual extraction method produces models which are reflective of true disease expressivity. OMIM and Orphanet are therefore useful sources of comparison for the disease models developed in this work.

Other reported methods, using an automated approach, extract phenotype descriptors from a wide range of manuscripts with little discrimination, for example including animal models (36). The technique used here involves creating disease models using automated phenotype extraction on specific case reports/case series. This combination of disease phenotypes highly specific to a disease, which is scalable to the full spectrum of GDD using text mining, should prove useful in variant interpretation in future.

Automation of disease model creation requires access to relevant text and accurate phenotype concept extraction. In this chapter, I demonstrate the use of the Cadmus package (43), developed at the University of Edinburgh for downloading full-text manuscripts at scale. I use data from this to create a test corpus for the purpose of testing NER methods. MetaMap (64), in particular, was chosen as it is straightforward to set up, highly configurable, regularly updated, and its reported performance was similar to other

techniques, as reviewed in section 1.7, and shown in Table 1-1. I evaluated MetaMap on my own corpus, and then used it to create disease models. These were analysed in detail to assess their relevance to reported phenotypes, or true disease expressivity, in the peer-reviewed literature.

2.2 Development of literature-derived disease models

2.2.1 Parsing the Human Phenotype Ontology

For all analyses involving the Human Phenotype Ontology (22) structure in this work, a copy of the hp.obo flat file format (version 1.2) was used, downloaded on 08/10/2021. This was parsed using the Python package Ontobio (139), for example to derive parent/ancestor terms and child/descendant terms. An HPO term is a text description of a clinical abnormality; an HP ID is the unique identifier for a given term.

Each HPO term entry may contain multiple synonyms, which may be clinical variants, description of a feature in common parlance or pejorative terms which are not preferentially used. Small changes in spelling may be included, for example to account for British vs American language. For example, the term ‘Cupped ear’ (HP:0000378) has the synonyms ‘Cup-shaped ears’, ‘Capuchin ears’ and ‘Cupped ears’ (22). For all instances where the HPO is used for mapping in this work, the HP ID was used, as this is stable and includes all synonyms associated with a given term.

Only terms which have the root ‘Phenotypic abnormality’ were used. This forms a sub-ontology of the HPO from the root term ‘All’ (HP:0000001). Other HPO terms derived from children of ‘All’ are modifiers and were not considered relevant to this work. There were 15985 terms and 14474 synonyms in this sub-ontology. The HPO contains obsolete terms which are deprecated (22). These are usually mapped to an up-to-date term within the

HPO. The current term was used to replace any which were obsolete during mapping from other sources.

2.2.2 Pilot study: comparison of literature-derived and DECIPHER phenotypes

To test the hypothesis that phenotype descriptors extracted from the literature reflect true clinical expressivity, a pilot study was undertaken. This compared phenotypic terms extracted using manual annotation, and prospectively gathered clinical models in DECIPHER (30).

2.2.2.1 Literature search for GDD

Disease models for three exemplar GDDs – Tatton-Brown Rahman syndrome (caused by SNV in the *DNMT3A* gene), Shuurs-Hoeijmakers syndrome (*PACS1*) and *GRIN2B*-associated Neurodevelopmental Disorder (*GRIN2B)* – were generated from the peer-reviewed literature and from DECIPHER (30). These syndromes were selected firstly because they were well-represented in the literature (which in this case means being reported in more than one or two case reports/case series) and reported for multiple individuals in DECIPHER. Secondly, these conditions were chosen because they had subjectively diverse phenotypic profiles, meaning each disease phenotype was clinically recognisable and could be distinguished from the others.

For the literature-derived models, a manual HGNC (HUGO Gene Nomenclature Committee) {gene symbol}[TI] search was performed on PubMed, as this strategy was thought to generate results enriched for relevant manuscripts, based on previous manual curation experience. Deprecated gene symbol or full gene name searches were not used, as these were thought to increase the number of search results without

significantly enriching for relevant case reports/case series, based on previous experience.

Every abstract from this search was reviewed to determine if it contained case series/case reports describing the search condition. Manuscripts which included descriptions of more than one disease were excluded. This process identified seven, six, and eight relevant manuscripts for *DNMT3A, PACS1* and *GRIN2B* respectively.

2.2.2.2 Manually created disease models

The full text for each identified paper was mapped to HPO terms using Doc2HPO (80). This is a web application, with limited capacity for large text files. The full text for each manuscript identified was copied and pasted into Doc2HPO, in batches where necessary. Doc2HPO offers several matching options, with increasing accuracy paired with increased computational power. The most basic option – string-based matching – was used as the other options did not work with larger text inputs.

Doc2HPO enables manual correction of mapped terms in text, via a graphical user interface, and this was used for each file to create an accurate download file of annotations per manuscript (80). A list of HPO terms per disorder was created by amalgamating the frequency per term from all papers describing a given disorder. Each term was therefore weighted by its frequency across all input manuscripts. This list can be described as a weighted disease model.

For the DECIPHER (30) models, a gene symbol search was performed on www.deciphergenomics.org. In the ‘Matching patient variants’ tab, the filters ‘De novo’ and ‘Pathogenic’/’Likely Pathogenic’ were applied, as these select SNV highly likely to be diagnostic. DECIPHER offers a ‘Phenotypes present in multiple matching patients’ tab, containing HPO terms together with their

frequency amalgamated across patients selected using the filters above. All of these HPO terms and frequencies were recorded to form a disease model. Comparison of the disease models generated revealed significantly more terms in the literature-derived models than recorded in DECIPHER (Table 2-1). Of note, this table does not show all the literature-derived terms for each disease as these numbered 50-100. All the DECIPHER-derived terms are shown.

This increased number of terms in the literature-derived models may reflect the difference in collection methods for the phenotype terms. Case reports in the literature are likely to contain as much detail as possible, to define the phenotypic spectrum of novel disorders. Clinicians inputting HPO terms for the purposes of interpretation of diagnostic sequence may only include the minimum number they deem sufficient for meaningful analysis. A rule-of-thumb of five terms per patient has been suggested for this purpose (140).

For the most frequent terms from the literature, there were a number of exact matches in the corresponding DECIPHER model (
Table 2-2). There were also a number of terms which were thought to be clinically similar but not exact matches, which is relevant to the construction and comparison of disease models.

The terms in the literature-derived models appeared to reflect those expected clinically. For example, the model for the overgrowth syndrome associated with *DNMT3A* (Tatton-Brown Rahman) included the terms ‘Macrocephaly’ and ‘Tall stature’. There were multiple seizure-related terms in the *GRIN2B* model, and this is a known epilepsy-associated gene.

These GDD were chosen because they are phenotypically diverse. The literature-derived terms reflected the diversity not only across the phenotypic spectra of individual disorders, but also across body systems. For example,

there were terms included describing growth and intellectual development, dysmorphic features, neurological disease and cardiac disease.
There were a number of terms in the DECIPHER data which were not found in the literature-based models. In general, individuals are recorded on DECIPHER as part of investigations into an undiagnosed genetic disorder. DECIPHER is not a curated database for phenotypic data.

It is possible that the terms on DECIPHER may reflect atypical examples of a condition, or some phenotypic features may be unrelated to the underlying diagnosis. This may also reflect an issue with the accuracy of the literature-derived models. However, the overall subjective impression was of a significant number of overlapping terms.

In conclusion, the results from the pilot study provided provisional evidence that phenotypic descriptors in literature-derived models were subjectively similar to clinically derived data. I am not aware of this direct literature-derived to clinically-derived term comparison being performed previously. This supported pursuing an automated approach to constructing these models.

Disease gene	**Number of relevant manuscripts**	**Number of unique literature-derived HPO terms**	**Number of unique DECIPHER HPO terms**
DNMT3A	7	118	9
PACS1	6	116	21
GRIN2B	8	56	23

Table 2-1. Summary of literature-derived and DECIPHER-derived disease descriptors found for three exemplar GDD.

***DNMT3A* – Tatton-Brown-Rahman syndrome**

DECIPHER-derived HPO terms	DECIPHER frequency	Literature-derived HPO terms	Literature frequency
Global developmental delay	5	Tall stature	7
Macrocephaly	4	Intellectual disability	7
Tall stature	3	Generalized hypotonia	6
Frontal upsweep of hair	2	Narrow palpebral fissure	6
Joint hypermobility	2	Horizontal eyebrow	6
Large for gestational age	2	Thick eyebrow	5
Moderate global developmental delay	2	Macrocephaly	4
Proportionate tall stature	2	Joint hypermobility	4
Synophrys	2	Thin upper lip vermilion	4

***PACS1* – Shuurs-Hoeijmakers syndrome**

DECIPHER-derived HPO terms	DECIPHER frequency	Literature-derived HPO terms	Literature frequency
Constipation	4	Downslanted palpebral fissures	5
Global developmental delay	4	Bulbous nose	5
Moderate global developmental delay	4	Hypertelorism	4
Cryptorchidism	3	Highly arched eyebrow	4
Delayed speech and language development	3	Iris coloboma	4
Epicanthus	3	Thin upper lip vermilion	3
Gastroesophageal reflux	3	Low-set ears	3
Hypertelorism	3	Gastroesophageal reflux	3
Microcephaly	3	Seizures	3
Thin upper lip vermilion	3	Downturned corners of mouth	3
Anteverted nares	2	Constipation	3
Depressed nasal bridge	2	Wide mouth	3
Downslanted palpebral fissures	2	Ventricular septal defect	3
Low-set ears	2	Single transverse palmar crease	2
Seizures	2	Pectus excavatum	2
Severe global developmental delay	2	Optic nerve coloboma	2

Short stature	2	Renal duplication	2
Sleep disturbance	2	Scoliosis	2
Telecanthus	2	Absent speech	2
Thin vermilion border	2	Aggressive behavior	2
Ventricular septal defect	2	Anteverted nares	2

***GRIN2B* – *GRIN2B*-associated Neurodevelopmental Disorder**

DECIPHER-derived HPO terms	DECIPHER frequency	Literature-derived HPO terms	Literature frequency
Global developmental delay	9	Global developmental delay	4
Severe global developmental delay	6	Epileptic spasms	3
Seizures	5	Intellectual disability, mild	3
Delayed speech and language development	4	Absent speech	3
Generalized hypotonia	4	Feeding difficulties	2
Intellectual disability, severe	4	Hypsarrhythmia	2
Cerebral visual impairment	3	Generalized tonic-clonic seizures	2
Abnormality of eye movement	2	Muscular hypotonia of the trunk	2
Abnormality of the nervous system	2	Muscular hypotonia	2
Absence seizures	2	Infantile spasms	2
Absent speech	2	Intellectual disability, moderate	2
Constipation	2	Microcephaly	1
Epicanthus	2	Status epilepticus	1
Failure to thrive	2	Intellectual disability, severe	1
Gastroesophageal reflux	2	Short attention span	1
Generalized-onset seizure	2	Seizures	1
Macrocephaly	2	Sleep disturbance	1
Moderate global developmental delay	2	Hyperactivity	1
Muscular hypotonia	2	Motor delay	1
Pectus excavatum	2	Myoclonus	1
Postnatal microcephaly	2	Poor eye contact	1
Stereotypy	2	Mood swings	1
Strabismus	2	Generalized tonic-clonic seizures with focal onset	1

Table 2-2. Comparison of HPO terms derived from clinically-derived patient data in DECIPHER and manual annotation of the peer-reviewed literature, describing three GDD. Terms highlighted in green are exact matches between datasets. Terms in orange are non-exact matches which are clinically similar. All terms for each DECIPHER model shown. The full list of literature-derived terms for *DNMT3A* (n=118), *PACS1* (n=116), and *GRIN2B* (n=56) not displayed for brevity.

2.2.3 Generation of 50 paper annotated test corpus

To test the performance of different text-mining methods, a corpus consisting of the full text of 50 papers containing case series or case reports relevant to a variety of GDD was generated. The manuscripts were randomly selected from those identified during curation of the DDG2P database (31).

Identifiers for each manuscript in the form of PubMed IDs (PMIDs) were used as input for the Cadmus full text retrieval package (43), which was developed at the University of Edinburgh as part of a wider project to optimise biomedical text mining and literature curation. The process used by Cadmus issummarised in the following paragraph. It should be noted that I used this package but was not involved in development of the source code. The summary here is for information regarding the method through which Cadmus obtains full text data.

For each PMID, PubMed metadata was obtained. The metadata includes information such as the title, abstract, authors, journal title, publication type and MeSH terms. It may also include a doi reference. This metadata was then used to send requests for download for each paper to sources which authorise full text retrieval for research purposes. Typically, a publisher or other database will offer an API (Application Programming Interface) to enable this. Multiple sources were used to maximise the chances of download, including, but not limited to, Crossref, doi.org and EPMC. File formats for downloads generated through this process varied; these included HTML (Hypertext Markup Language), XML (Extensible Markup Language),

PDF (Portable Document Format) 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, using a format-specific method, and converted to a string. The abstract and references were identified using keyword matching, and parsed out to create a final 'full text' document. This document is referred to as the 'full text' throughout this work.

Cadmus output was in the form of a Pandas dataframe which comprises metadata for each paper as well as full text, including: PMID, title, abstract, publication type and file formats downloaded (PDF/XML/HTML/plain text). A unique identifier for each paper was also included to allow for continuity in data processing. An outline of the process used is shown in Figure 2-1.

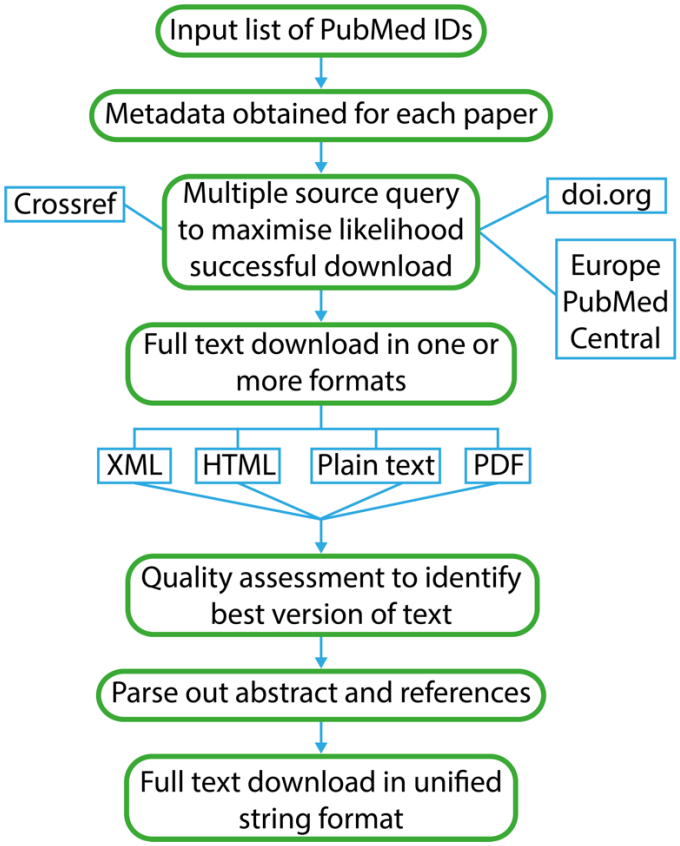


Figure 2-1. Overview of Cadmus full text download pipeline.

I then annotated the Cadmus-derived full text for each of the 50 papers. Text describing non-negated phenotypic features was highlighted using the Skim

PDF reader (141). Quantitative phenotypic features were excluded. This included quantitative statements which could be extrapolated to HPO terms, such as a small occipitofrontal head circumference measurement, which could be mapped to 'Microcephaly' (HP:0000252). Skim enables export of all highlighted text descriptors to a csv file (141). Exact string matching was used to map these text spans to HPO terms. Every instance where this was not successful was manually reviewed and assigned to an HPO term using the HPO website (22).

A summary of the phenotypic terms included in this corpus is shown in Table 2-3. Of note, there were 12 terms highlighted in the text which did not map to HPO, even on manual review. These included 'strephenopodia' (medial deviation of the forefoot) and 'triangular shaped conchae'. This illustrates the potential limitations of using a structured vocabulary, although it is straightforward to submit new terms to the HPO for consideration (22).

Test corpus summary	
Number of papers annotated	50
Total HPO terms	5450
Unique HPO terms	866
Mean HPO terms per paper	109
Median HPO terms per paper	83

Table 2-3. Summary statistics for annotated corpus of 50 papers for testing text named entity recognition methods.

2.2.4 Named Entity Recognition – FlashText, SpaCY and MetaMap

The performance of three NER methods was tested on the 50 paper test corpus, compared to manual annotation. These represent different levels of sophistication in terms of the matching algorithm used, and in specificity for

biomedical text. The use of these particular methods represents an evolution in my abilities to configure and use programmable text extraction methods. FlashText (142) was used as a type of basic exact string matching according to a defined list of terms, in this case the HPO. The algorithm prevents matching of partial strings within other words. SpaCy PhraseMatcher also matches strings from a defined list with more advanced tokenization and document processing (143). FlashText and SpaCy were more straightforward for me to use when starting from a non-computational background.

MetaMap is a highly configurable NER program developed by the U.S. National Library of Medicine for the purpose of identifying concepts in biomedical text (64). This is a more complex, and more powerful technique than the two listed above. This represented a significant step up for me in my ability to utilise natural language processing techniques. The most up-to-date techniques, such as BioBERT, (71) require pre-training and detailed optimisation before implementation, and it was thought that this process would be too lengthy for my purposes in relation to this work.

For FlashText and SpaCy, the Cadmus-derived full text for all 50 papers was concatenated into a single document. The FlashText algorithm version 2.7 (142), was used for exact string matching. This matches only whole words/phrases and not substrings. The FlashText keyword_processor module uses a dictionary of defined terms to match those in text. The HPO terms used for this dictionary were all descendants of ‘Phenotypic abnormality’ (HP:0000118) and their synonyms, as described in section 2.2.1.

SpaCy is a highly configurable natural language processing system (143). SpaCy version 2.1.8 was first used to split the input text into tokens, according to linguistic rules. This tokenized document was then used as input for the SpaCy PhraseMatcher module, which is used for matching text to terms in large terminology lists. Again, the HPO terms used for this

terminology list were all descendants of ‘Phenotypic abnormality’ (HP:0000118) and their synonyms, as described in section 2.2.1.

For MetaMap, the 2018 release was used to process each Cadmus-derived full text in the 50 paper test corpus via the command line in Linux, using the subprocess package to access this through Python. Running each text individually through MetaMap results in significant time constraints. Therefore, the multiprocessing Python package was used to assign input papers to different processors, with an output file generated for each paper. MetaMap output was in the form of UMLS concepts. These correspond to biomedical entities contained within the UMLS Metathesaurus (60). This includes multiple source vocabularies which standardize and codify terms related to, for example, diseases, genes and drugs. The HPO is one of the UMLS source vocabularies (22).

MetaMap by default uses all source vocabularies contained within the UMLS Metathesaurus, with each term defined by a Concept Unique Identifier (CUI) (60). In this case, the source vocabulary was restricted to CUIs corresponding to the HPO (22), which is Category 0 under the UMLS Metathesaurus licence, which means it can be used for research purposes.

MetaMap offers various output options. Fielded MetaMap Indexing (MMI) was used in this case as it is straightforward to parse, and contains detailed information about each matched concept. An example of MMI output is given in Figure 2-2. This comprises “|” separated output for each UMLS biomedical concept (60) identified in the text. Pertinent MMI output includes, for each concept: a unique identifier (CUI), a score for the relevance of the concept according to the MMI algorithm (based on word frequency and relevance of concept), the full text name of the concept, trigger information (the actual text mapped by MetaMap to a concept, with a negation flag; this may cover multiple spans of text), and positional information to identify triggers in the text.

The number of non-negated triggers for a CUI, as defined by MetaMap, was used as the frequency at which a concept is identified in the full text, for the purposes of term weighting. The UMLS Metathesaurus (Release 2020AA) provides direct mappings of CUI to HP ID. These were used to map all MetaMap output concepts to the HPO. In some cases, the CUI mapped to more than one HP ID. In this case, whichever term was higher-level in the ontology was mapped to the CUI. The "|" separated MMI output was parsed to give a list of tuples comprising (HP ID, non-negated frequency of concept) per CUI, for each full text document. This was merged with Cadmus-generated metadata, to give a weighted list of HP IDs per PMID.

```
USER|MMI|26.12|Anophthalmos|C0003119|[cgab]|["Anophthalmia"-tx
-1-"anophthalmia"-noun-0,"Anophthalmia"-tx-1-"anophthalmia"-no
un-1]|TX|[248/12],[327/12];[248/12],[327/12]|C11.250.080;C16.1
31.384.159\n
```

Figure 2-2. Example of fielded MetaMap Indexing (MMI) output. Adapted from National Library of Medicine (8). Blue highlight – UMLS Concept Preferred Name. Yellow highlight – UMLS Concept Unique Identifier (CUI). Green highlight – text from document mapped to UMLS concept. Orange highlight – negation flag: 1 if negated, 0 if not.

2.2.5 Named Entity Recognition performance vs gold standard corpus

HPO term + frequency lists generated by Flashtext, SpaCY and MetaMap after processing the 50 paper test corpus, as detailed in section 2.2.4, were compared to the list of terms from manual annotation. This was to assess the performance of each method. True positive (TP), true negative (TN) and false positive (FP) rates were calculated using the frequency per term and summed across all terms, for each method. Performance metrics were then generated using the following formulas:

$$Precision = \frac{TP}{TP/(TP + FP)}$$

$$Recall = \frac{TP}{TP/(TP + FN)}$$

$$F1\ score = \frac{TP}{TP/(TP + (0.5 \times (FP + FN))}$$

Precision is equivalent to positive predictive value, i.e. the proportion of terms returned which are true matches. Recall is equivalent to sensitivity, i.e. the proportion of true matches in the document which were returned. The F1 score is the harmonic mean of the precision and recall.

The results of this comparison are shown in Table 2-4. The performance of both PhraseMatcher and FlashText was similar, with a higher precision than recall, reflected by the same F1 score. The low recall for each of these methods means that between 44% and 47% of potential matches in the document were missed. MetaMap was uniformly superior to the other two methods, resulting in a 10 point higher F1 score. However, there were still 35% of potential matches in the document missed.

Text extraction method	Precision	Recall	F1
Flashtext	0.75	0.53	0.62
PhraseMatcher	0.69	0.56	0.62
MetaMap	0.84	0.65	0.73

Table 2-4. Comparison of NER methods, using annotated 50 paper test corpus.

For the unique phenotype terms in the test corpus, 180/866 (20.8%) were not recognised by any method (Figure 2-3). 91/866 (10.5%) exact matches were missed by MetaMap, but identified by PhraseMatcher. It was hypothesised that this discrepancy may be explained by MetaMap mapping phenotypic descriptors in text to different concepts, rather than to their exact matches in HPO.

To explore this, MetaMap (n=137) and annotated terms (n=180) which did not intersect were manually reviewed to determine if there were clinical correlates between the two sets. The full set of terms reviewed is in Supplementary Table 1. There were 35/137 (25.5%) terms in the MetaMap set which had clinical matches with manually annotated terms. This means that the true recall of MetaMap, from a clinical perspective, is likely higher than the figure generated previously, although it is difficult to apply this information in practice.

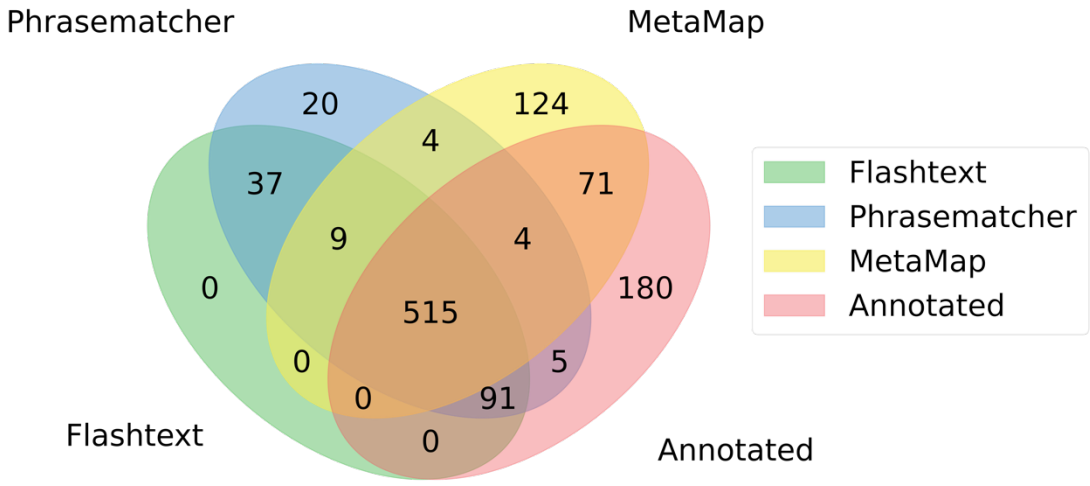


Figure 2-3. Overlap of all unique terms using NER methods on annotated 50 paper test corpus.

Given the superior performance of MetaMap against the other methods tested, it was used to develop literature-derived disease models, using the configurable processing options as discussed below.

2.2.6 Assessment of MetaMap processing options

The 50 paper test corpus was processed using MetaMap (64) as detailed in section 2.2.4, with additional processing options added. All options were

reviewed, and only those which were thought likely to increase performance in full text analysis were added. These were used singly and in combination.

The options were:

1. Word sense disambiguation (WSD) – returns a single mapping for terms with ambiguous meanings.
2. No derivational variants (NDV) – Stops derivational variants being used in concept mapping, for example hyperplastic and hyperplasia.
3. Restrict to sources (R-) – concept mapping restricted to specific vocabularies within the UMLS Metathesaurus (60), for example the HPO (R-HPO).
4. Blanklines off – prevents MetaMap from using any whitespace line as a separator for input. Normally this is the default behaviour. MetaMap recommends this option particularly for processing clinical text (64).
5. Conjunction processing – recombining phrases with a conjunction e.g. 'lung and breast cancer' would be processed as 'lung cancer', 'breast cancer' instead of the default 'lung' 'and' 'breast cancer'.

All MetaMap options had very similar F1 scores (

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

). However, there was variation in the balance between precision and recall.

Adding options appeared to increase recall at the cost of precision. There was also significant processing time differences between options, with word sense disambiguation and no derivational variants, in particular, markedly slowing performance. Restricting source vocabulary to HPO with no other options was chosen as it offered a good balance between precision and recall, without a large decrease in processing speed. This configuration was used for all the other analyses in this work.

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

Table 2-5. Performance of MetaMap usage options. Evaluated against 50 manually-annotated papers describing developmental disorders. Word sense disambiguation – disambiguate concepts with similar scores. Restrict to HPO – use only human phenotype ontology for mapping concepts. No derivational variants – compute word variants without using derivational variants. Blanklines off – process text as whole document. Conjunction processing – join conjunction-separated phrases.

2.2.7 Clinical assessment of MetaMap derived single gene disorder models

The clinical relevance of disease models created using MetaMap was reviewed in more detail. To enable this, a further set of models were created for *SOX2-, EFTUD2-* and *ASXL3*-related disorders. Relevant case series/case reports were identified through PubMed search and processed using Cadmus and MetaMap as detailed in sections 2.2.3 and 2.2.4. The (HP ID, frequency) lists for each paper were merged, to create final disease models. As a comparison, phenotypic descriptors from expert literature reviews were used. The top five phenotypes present in multiple matching patients with pathogenic variants in DECIPHER (30) were also compared. Note there were no patients available for comparison in DECIPHER for *SOX2*.

The literature-derived *SOX2* model is shown in Figure 2-4. Clinically, the main clinical features of *SOX2* disorder are anophthalmia and microphthalmia. Other associated phenotypic descriptors include coloboma, growth restriction, learning disability/global developmental delay, seizures, malformation of the hippocampus, pituitary hypoplasia, corpus callosum hypoplasia and genital abnormalities (144). All of these were included in the top 15 terms of the MetaMap *SOX2* model (n=380) (Figure 2-4) except growth restriction and malformation of the hippocampus. However, the terms

‘Short stature’ (HP:0004322) and ‘Hypoplastic hippocampus’ (HP:0025517) were included in the model at lower frequency.

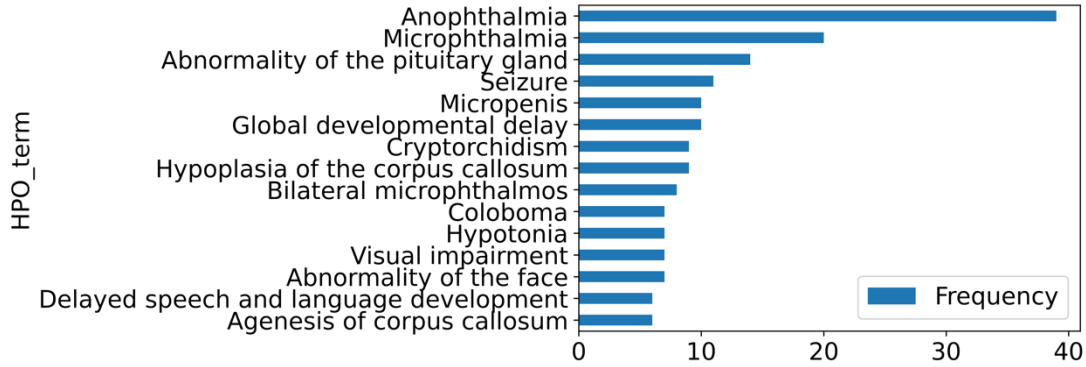


Figure 2-4. Example disease model generated using MetaMap for *SOX2*-related disorder. Top 15 most frequent terms shown (full model n=380 terms). Terms generated from the full text of 38 relevant case reports/case series.

SNV in *EFTUD2* result in the disorder Mandibulofacial dysostosis with microcephaly (145). The literature-derived model in Figure 2-6 indeed has ‘Mandibulofacial dysostosis’ (HP:0005321) and ‘Microcephaly’ (HP:0000252) as the highest weighted terms. The other main features of this disorder are intellectual disability, hearing loss and ear malformation, and these are included in the model (145). Of note ear malformation is represented by the general term ‘Abnormality of the pinna’ (HP:0000377) as well as more specific terms including microtia (HP:0008551), meaning small ears, ‘Preauricular skin tag’ (HP:0000384), and ‘Pretragal ectopia’ (HP:0030024).

The other characteristic features of this condition are cleft palate, choanal atresia, facial asymmetry, cardiac abnormalities, thumb abnormalities, oesophageal atresia, short stature, spinal abnormalities and seizures (145). All of these are present in the top 15 terms of the model, except cardiac, thumb and spinal abnormalities, and short stature. However, the terms ‘Abnormality of cardiovascular system morphology’ (HP:0030680), ‘Atrial septal defect’ (HP:0001631), ‘Ventricular septal defect’ (HP:0001629), ‘Abnormal thumb morphology’ (HP:0001172), ‘Proximal placement of thumb’

(HP:0009623), ‘Scoliosis’ (HP:0002650) and ‘Short stature’ (HP:0004322) were in the model at lower frequencies.

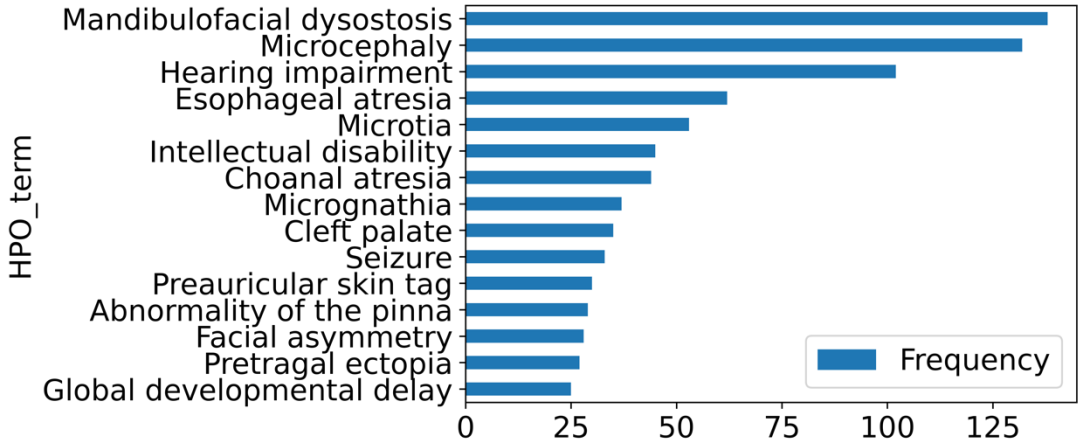


Figure 2-5. Example disease model generated using MetaMap for EFTUD2-related disorder. Top 15 most frequent terms shown (full model n = 278 terms). Terms generated from the full text of 12 relevant case reports/case series.

The top five terms in DECIPHER for *EFTUD2*-related disorder were ‘Microcephaly’, ‘Preauricular skin tag’, ‘Moderate global developmental delay’, ‘Delayed speech and language development’ and ‘Facial asymmetry’ (30). These were all present in the top 15 terms of the literature-derived model (Fig. 2-6), except that delayed development was represented by the parent term ‘Global developmental delay’ in the model.

The main features of *ASXL3*-related disorder are developmental delay/intellectual disability, speech delay, autism or autistic traits, feeding difficulties, hypotonia, failure to thrive, seizures, visual abnormalities such as strabismus and skeletal abnormalities, as well as dysmorphic features including prominent forehead, highly arched eyebrows and downslanted palpebral fissures (146). All of these are present in the top 15 model terms in Figure 2-5, except ‘Strabismus’ (HP:0000486) which is represented at a lower frequency.

Notably, there are several terms in the top 15 which were not described in a comprehensive review of this condition (146). These are ‘Microcephaly’ (HP:0000252), ‘Premature birth’ (HP:0001622), ‘Trigonocephaly’ (HP:0000243) and ‘Nevus’ (HP:0003764). Of these, microcephaly is present

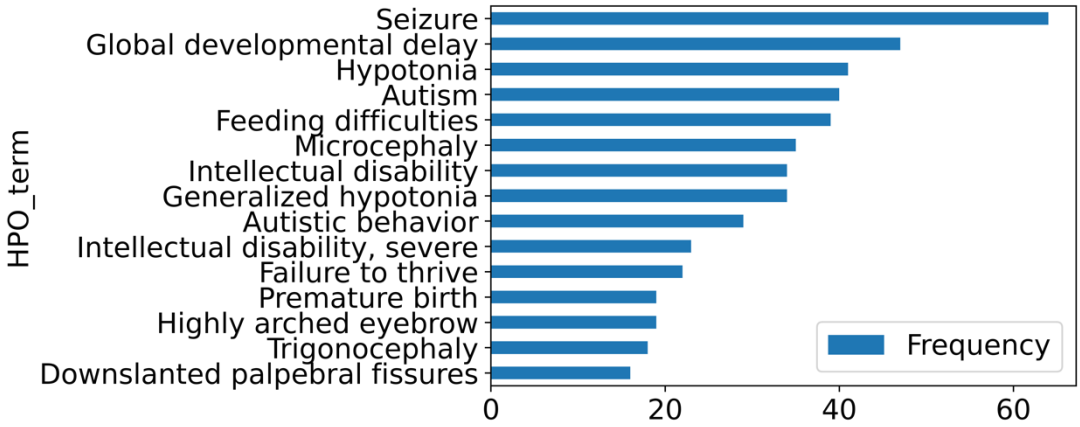


Figure 2-6. Example disease model generated using MetaMap for ASXL3-related disorder. Top 15 most frequent terms shown (full model n = 275 terms. Terms generated from the full text of 13 relevant case reports/case series.

in several of the reviewed papers (147,148), perhaps illustrating an error in manual literature review which is not missed using the automated method. The other features are not commonly associated with *ASXL3*-disorder, and may represent phenotypic expansion hitherto unnoticed without the benefit of deep computational phenotyping, or could be erroneous mappings by MetaMap (64,146).

In DECIPHER, the top five terms were ‘Global developmental delay’, ‘Feeding difficulties in infancy’, ‘Hypertelorism’, ‘Strabismus’, and ‘Generalized hypotonia’ (30). Of these, all were present in the literature-derived model except ‘Hypertelorism’. ‘Strabismus’ was present at a lower frequency, as above, with the other DECIPHER terms being in the top 15 of the model.

In conclusion, detailed clinical assessment of literature-derived disease models showed that these were reflective of true disease expressivity. This

applied to both the phenotypic terms used and to the weighting, which reflected the most common features of a given disorder.

2.2.8 Disease models for copy number variants

Disorders resulting from copy number variation (CNV) are often more complex to define than those caused by single nucleotide variants (SNV). This is because CNV involve the gain or loss of multiple genes, and the precise genes involved are variable between individuals. There are, however, recurrent CNV which result in recognizable syndromes, although the phenotypic spectrum may still be broad.

Literature-derived disease models were created, using methods as described in sections 2.2.3 and 2.2.4, for the well-characterised 16p11.2 deletion syndrome (Figure 2-7) and 22q11.2 deletion syndrome (Figure 2-8). This was to assess whether the literature-derived disease model concept may be applied to CNV syndromes as well as single gene disorders. As for SNV syndromes, expert literature reviews and the top five terms from DECIPHER were used as comparators.

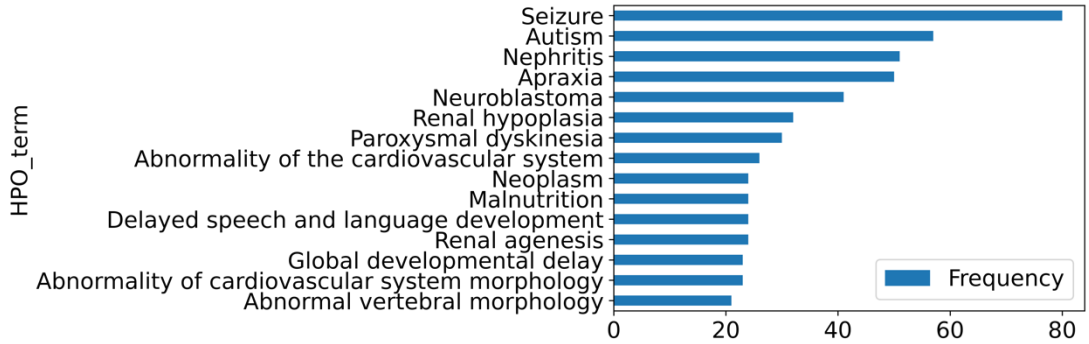


Figure 2-7. Example disease model generated using MetaMap for the 16p11.2 deletion syndrome. Top 15 most frequent terms shown (full model n = 206). Terms generated from the full text of 8 relevant case reports/case series.

The main features of the 16p11.2 deletion syndrome are motor and speech impairment, (particularly apraxia), intellectual disability/global developmental delay, autism/autistic behaviour, macrocephaly, Chiari I malformation, seizures, vertebral abnormalities and obesity (149). Of these, all are present in the model (Figure 2-7), including the lower frequency terms 'Macrocephaly' (HP:0000256), 'Arnold-Chiari type I malformation' (HP:0007099) and 'Obesity' (HP:0001513).

The top five DECIPHER terms 'Intellectual disability', 'Delayed speech and language development', 'Global developmental delay', 'Obesity', and 'Autism' were present in the model. However, there are a number of terms in the literature-derived model which are not usually associated with this condition, including 'Nephritis' (HP:0000123) and 'Neuroblastoma' (HP:0003006).

To explore this, the contribution of each input paper was assessed (Table 2-6). This showed that 'Nephritis' (HP:0000123) and 'Neuroblastoma' (HP:0003006), in particular, were heavily weighted in the model by single papers which were especially focused on these conditions. It is easy to exclude these papers manually before phenotype extraction, but this is not scalable in a fully automated system. In future, different methods of weighting phenotypic descriptors may help to ameliorate this issue.

Title	Highest ranked HPO term	Frequency
Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma.	Neuroblastoma	41
Human and mouse studies establish TBX6 in Mendelian CAKUT and as a potential driver of kidney defects associated with the 16p11.2 microdeletion syndrome.	Nephritis	50
Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion.	Speech apraxia	38
16p11.2 deletion syndrome.	Autistic behavior	10
Ocular Findings in the 16p11.2 Microdeletion Syndrome: A Case Report and Literature Review.	Hypertelorism	10

PRRT2-related phenotypes in patients with a 16p11.2 deletion.	Paroxysmal dyskinesia	28
Intrauterine phenotypic features associated with 16p11.2 recurrent microdeletions.	Abnormality of the cardiovascular system	26
Neurodevelopmental trajectory and modifiers of 16p11.2 microdeletion: A follow-up study of four Chinese children carriers.	Autistic behavior	9

Table 2-6. Manuscripts used to create 16p11.2 deletion syndrome disease model, with highest ranked HPO term per paper shown.

A similar process was undertaken for the 22q11.2 deletion syndrome (Figure 2-8). This is characterized by features including congenital cardiac disease (especially ventricular septal defect, tetralogy of Fallot, interrupted aortic arch, and truncus arteriosus), abnormalities of the palate, immunodeficiency, dysmorphism and intellectual disability. Psychiatric disease is more common, particularly schizophrenia (150).

These were all present in the model (Figure 2-8), with ‘Cleft palate’ (HP:0000175) and ‘Immune dysregulation’ (HP:0002958) at lower frequency. However, diabetes is not usually associated with this condition, and was the highest ranked term. There also seemed to be a predominance of cardiovascular terms.

The DECIPHER top terms were ‘Intellectual disability’, ‘Ventricular septal defect’, ‘Micrognathia’, ‘Hypocalcaemia’ and ‘Delayed speech and language development’. These were all present in the literature-derived model, except ‘Micrognathia’.

Analysis of the manuscript contributions to the 22q11.2 deletion model (Table 2-7) showed that papers focused on cardiovascular manifestations did contribute significantly to the frequency weighting. However cardiovascular disease is a well-known major feature of the condition (150), so this is

perhaps unsurprising. There was one paper included focusing on the risk of diabetes with 22q11.2 deletion, and this contributed the term 'Type II diabetes mellitus' (HP:0005978) at a much higher frequency than any other term across all manuscripts.

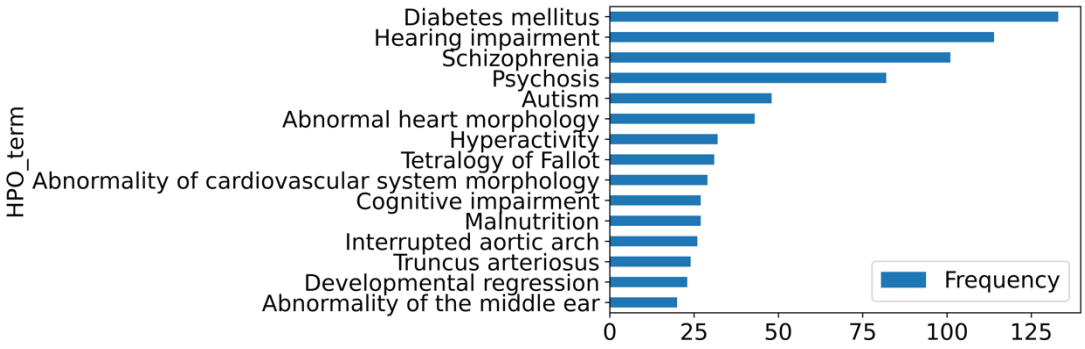


Figure 2-8. Example disease model generated using MetaMap for the 22q11.2 deletion syndrome. Top 15 most frequent terms shown (full model n = 181). Terms generated from the full text of 11 relevant case reports/case series.

Title	Highest ranked HPO term	Frequency
Complete Sequence of the 22q11.2 Allele in 1,053 Subjects with 22q11.2 Deletion Syndrome Reveals Modifiers of Conotruncal Heart Defects.	Tetralogy of Fallot	11
Clinical features of 22q11.2 deletion syndrome related to hearing and communication.	Hearing impairment	37
Movement Disorder Phenotypes in Children With 22q11.2 Deletion Syndrome.	Dystonia	8
Cognitive deficits in childhood, adolescence and adulthood in 22q11.2 deletion syndrome and association with psychopathology.	Autistic behavior	26
Assessing auditory processing endophenotypes associated with Schizophrenia in individuals with 22q11.2 deletion syndrome.	Schizophrenia	21

22q11.2 deletion syndrome and congenital heart disease.	Tetralogy of Fallot	17
Medical and dental characteristics of children with chromosome 22q11.2 deletion syndrome at the Royal Children's Hospital, Melbourne.	Abnormality of cardiovascular system morphology	14
Lymphoproliferative disorder with polyautoimmunity and hypogammaglobulinemia: An unusual presentation of 22q11.2 deletion syndrome.	Abnormal pulmonary interstitial morphology	5
Magnitude and heterogeneity of brain structural abnormalities in 22q11.2 deletion syndrome: a meta-analysis.	Psychosis	22
The 22q11.2 Microdeletion in Pediatric Patients with Cleft Lip, Palate, or Both and Congenital Heart Disease: A Systematic Review.	Cleft upper lip	1
22q11.2 microdeletion and increased risk for type 2 diabetes.	Type II diabetes mellitus	120

Table 2-7. Manuscripts used to create 22q11.2 deletion syndrome disease model, with highest ranked HPO term per paper shown.

2.3 Discussion

2.3.1 Disease model creation

In this chapter, I have demonstrated a method for creating disease models from full-text manuscripts in the peer-reviewed literature. The models were constructed using standardised vocabulary from the HPO, and terms weighted according to their frequency in the literature. I analysed the clinical similarity of these models to the phenotype of well-defined GDD.

2.3.2 Pilot study – disease model proof-of-concept

First, I conducted a pilot study to determine proof-of concept for the text-mining disease model concept. I showed that literature-derived phenotypic features reflect those seen in prospectively gathered clinical data, with a number of exact term matches between sets. However, this exercise also highlighted potential issues, where clinically similar phenotypic features were defined differently in each model. These were apparent on manual review, but pointed towards complexities in computational phenotype extraction and review, which were analysed further as part of the development of literature-derived phenotyping in this work.

2.3.3 Literature review for disease model creation

Full-text derived models were created for conditions relating to SNV in *SOX2*, *EFTUD2* and *ASXL3*, as well as from the CNV syndromes related to 16p11.2 and 22q11.2 microdeletions. PubMed search was used to identify relevant papers. This used only the current HGNC gene symbol. Deprecated symbol and/or full gene name searches could be added in future for a potentially more comprehensive overview of a given condition. However, for an automated search this could significantly increase the number of non-relevant search results. This could adversely affect the clinical accuracy of disease models generated.

For case reports/case series identified, careful exclusion criteria were applied to minimise phenotypic noise. For example, manuscripts containing descriptions of more than one disease were not included. This reduced the potentially confounding effect of other disease descriptors being included in the models created. However, this meant the exclusion of potentially valuable phenotypic data. However, given the whole-manuscript nature of phenotypic extraction described here, it is likely that descriptors of other diseases were included. For example, authors may discuss related conditions to those being reported, in the introduction and discussion sections of a paper. These factors should be considered in the development of disease models derived

from automated literature search, as this is likely to result in increased phenotypic noise.

In future work, all phenotypic descriptors extracted from a manuscript should ideally be associated with single individuals in case reports/case series. This would not only refine the disease model, but would also allow for exact frequencies per term to be used as weightings. One method of doing this may be to use tabular data; this is more complex to parse but often includes per-individual phenotypes.

2.3.4 Cadmus for full-text download

I used the Cadmus package (43) to download full-text manuscripts for the creation of disease models, leveraging university-wide permissions to access almost all of the biomedical literature. Most work in phenotype text mining has used only title + abstract (151), or in some cases open access full text (36), which represents a minority of the published literature (41). Therefore, the use of Cadmus in this work should allow for more comprehensive breadth of coverage of the relevant literature than previously possible. The use of full text should also enable greater depth of phenotyping compared to title + abstract (40). This ought to translate into more clinically accurate disease models, reflecting the full phenotypic spectrum of a given disorder. Cadmus also greatly simplifies the process of accessing full text, providing a corpus of parsed documents from a list of PMIDs, ready for phenotype extraction.

2.3.5 NER method evaluation

After full-text download, the next step in disease model construction is phenotypic feature extraction. Three methods were tested for this purpose – FlashText, SpaCY and MetaMap (64,142,143). These were chosen for pragmatic reasons, being straightforward to set up and use, compared to more complex NER techniques such as BioBert (71). Unsurprisingly, the

simpler string-matching methods FlashText and SpaCY were outperformed by MetaMap. The precision and recall figures obtained for MetaMap of 0.77 and 0.71 also compare favourably to those reported in the literature, as reviewed in section 1.7. This is the case for similar rule-based methods such as the NCBO annotator, OBO annotator and Bio-LarK (49,61,66).

However, similar performance metrics were also obtained for the most advanced HPO annotator I am aware of, PhenoTagger, which utilises BioBert (54,71). This had precision of 0.77 and recall of 0.74 when tested on a corpus of HPO-annotated abstracts (51,54). It is clear from the review in section 1.7. that very variable performance can be obtained from the same annotator using different corpora. Therefore, the results obtained here are not sufficient to directly compare to reported NER methods. Nevertheless, MetaMap appears to be at least comparable to other techniques, and was thought to be sufficiently performant to use for disease model creation.

2.3.6 Test corpus – advantages and limitations

The corpus created here for the purposes of testing NER methods comprises 50 full text manuscripts. Phenotypic features in the text were directly annotated to HPO terms. As far as I am aware, this is one of the largest test corpora available which specifically uses full-text HPO terms. This is illustrated by the fact that most HPO-specific NER methods in the literature, reviewed in 1.7 utilise the ‘HPO Gold Standard’ (HPO GS) corpus, as created by Groza et al., and modified by Lobo et al. (49,51). This comprises 228 abstracts.

Given the advantages of using full text outlined above, it is likely that the 50 paper full-text corpora here is enriched for HPO terms, compared to the ‘Gold Standard’, and offers a more in-depth performance test for NER methods.

The HPO GS did (in theory) utilise more rigorous annotation methods in its creation (49). Three expert annotators (the creators of the HPO) were used, two for the initial annotation and a third for a consistency/completeness check, paired with one of the others (49). Annotation was done according to a set of defined guidelines. However, one of these was that non-canonical phenotypes should not be annotated, e.g. include 'hypoplastic nails' but not 'nails were hypoplastic' (49). This may be one of the reasons Lobo et al. were able to add 881 entities to the HPO GS (51). This case illustrates that manual annotation is essentially an imperfect process. The 50 paper test corpus demonstrated here is likely to be a useful resource not only for testing NER methods, but may also be used for training machine-learning based techniques in future.

2.3.7 Clinical expressivity in disease models

Clinical review of exemplar disease models created for *SOX2*, *EFTUD2*, *ASXL3*, 16p11.2 deletion and 22q11.2 deletion showed that they reflect true disease expressivity, as defined by comprehensive reviews of these conditions (144–146,149,150). In particular, weighting of terms according to their frequency in input manuscript appeared to prioritise the major features of each condition. Manual review of these models has limitations. Most of them contain hundreds of terms, more than the scope of phenotypic features covered in review papers. It is not possible to definitively assess whether a model of this size matches the true clinical phenotype of a disorder without detailed annotation of all disease-relevant manuscripts. Clinical review of the generated disease models was promising, in that they appeared to reflect true disease expressivity, particularly in relation to the top-ranked terms. Nonetheless, comparison of the full scope of these models is only possible using computational methods, which will be covered in Chapter 3.

2.3.8 Weighting bias from single manuscripts

Whilst the exemplar disease models shown here appeared to reflect disease expressivity overall, there were individual high-ranked terms present which were clearly not part of the defined phenotypic spectrum. This was particularly true for the CNV models. One of the highest-weighted terms in the 16p11.2 deletion syndrome was ‘Neuroblastoma’ (HP:0003006). This was only present in the model because one of the input papers was ‘Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma’. Similarly, for the 22q11.2 deletion, the top term was ‘Type II diabetes mellitus’ (HP:0005978), included due to the paper ‘22q11.2 microdeletion and increased risk for type 2 diabetes’.

This effect may be particularly pronounced for these CNVs, as they have been well-characterised for a relatively long period, therefore recently published reports are more likely to focus on a single phenotypic feature. The same is likely to be true for some single gene disorders, particularly those for which the phenotype was well defined prior to the molecular basis being discovered. In future work, it would be useful to refine term weighting to prevent single paper phenotypes skewing the model. Ultimately, extraction of phenotypic features on an individual basis, as discussed in section 2.3.3, should help ameliorate this issue.

2.3.9 Conclusion

In conclusion, I have presented here a method for generating weighted disease models for GDD from the full text of manuscripts in the peer-reviewed literature. I tested several phenotype extraction methods, and showed that MetaMap (64) likely performs as well as other reported techniques. I assessed these models from a clinical perspective, and showed that they reflect true disease expressivity, with some caveats regarding the

accuracy of text mining. In the next chapter, I scale up the disease model concept, and test these against gold standard manual curation.

Chapter 3 Disease matching and evaluation

3.1 Introduction

Several literature-derived models discussed in section 2.2.7 appeared to clinically reflect true disease expressivity. Given these encouraging results, this chapter focuses on testing a set of these models at a larger scale to determine whether they maintain clinical utility.

Phenotype models for GDD using HPO terms have previously been created using two methods: manual curation and automated disease-phenotype linking. Manual curation, for example OMIM and Orphanet (28,29), utilises disease-phenotype annotations extracted directly from relevant case reports/case series by expert reviewers. This method may be scaled up – in the version of OMIM used in this work, there were 5574 diseases annotated (29). 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 represents a significant challenge given the volume of new publications describing GDD on a monthly basis (152).

The disease models in these databases have limitations. OMIM does not generally directly link phenotype terms to the manuscript from which they were extracted, i.e. only amalgamated models are presented. Curation for these databases links phenotypic descriptors to disease entities. It is also useful to quantify the expressivity of these phenotypic features, i.e. how commonly a given phenotype is seen in association with a particular disease. HPO terms in OMIM do not usually have a frequency weighting to describe the most common features of a disease, although these are typically present in Orphanet, as discussed in sections 1.3.2 and 1.3.3.

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. (153) used PMI (Pointwise Mutual Information) (90) to rank genotype-phenotype associations for both HPO and the Mouse Ontology (22,118) 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 (154). 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 (23,116). This study used the Mondo ontology (67) to define disease names, including GDD.

The database created by Pilehvar et al. (116), PheneBank, is the most comprehensive automated GDD-phenotype curation work of which I am aware. However, this study and those of Kafkas et al. and Li et al. (153,154) have the significant limitation that the abstracts/manuscripts used were not filtered for human case reports/case series, meaning 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 means the underlying molecular mechanism is 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 (116), but there was no method of differentiating somatic from germline variation, therefore these were likely to include, for example, cancer-related manuscripts. In reality, these methods as demonstrated are not sufficiently clinically accurate for diagnostic use. A degree of manual curation is still needed, for example through selecting input manuscripts.

The literature-derived disease models described in section 2.2.7 were generated using manually-selected manuscripts to ensure these contained single gene case reports/case series relevant to a particular GDD. It was hypothesised that this method could be scaled to a larger set of disorders whilst maintaining clinical utility. This increased number of models would also allow testing of this hypothesis using comparative similarity metrics.

In this chapter, the process of generating a larger set of disease models is detailed. Several modifications to the process of generating disease models were also tested, to determine whether they increased performance. This was assessed through similarity to prospectively gathered clinical data from the DDD study (3).

The larger disease model test set also allowed testing of the hypothesis that automated literature curation could produce data similar to the highly disease-specific curation used in OMIM and Orphanet (28,29), but in a less resource intensive manner. Literature-derived disease models in the test set are compared to those generated from manual curation in this chapter using several similarity metrics. The performance of models from each of these three datasets (literature-derived, OMIM, Orphanet) at predicting diagnoses from DDD is also assessed using ROC curves (3,28,29).

3.2 Generation of larger scale disease model test set

3.2.1 Generation of literature-derived disease test set

The first step in testing the clinical utility of a larger set of disease models was their generation from literature sources. To enable this, a set of 99 GDD defined in the DDG2P database (31) were selected from the top 150 genes most frequently associated with diagnoses in the DDD study (6). Diseases were selected to include a range of different allelic requirements (monoallelic, biallelic, hemizygous and X-linked dominant), and for conditions which

already had phenotypic annotations in OMIM (29). Genes with more than three associated diseases were excluded. Each GDD was defined by a locus-genotype-mechanism-disease thread in DDG2P, derived from a DD gene-disease pairs and attributes file downloaded from www.ebi.ac.uk/gene2phenotype/downloads on 29/4/21.

For each of these diseases, a literature review was undertaken using manual PubMed searches. The initial search was by gene symbol in title – {gene symbol}[TI] – as this strategy was found to be enriched for relevant GDD during curation of DDG2P. Of note, the 'case report'[publication type] filter was not used as it is my experience that many case reports/case series in GDD are not tagged as such in PubMed. Case reports in rare disease often include a review of the previous literature, and this can lead to an article being tagged as 'review'.

If the search strategy returned less than 300 citations, every abstract in the results was reviewed to identify relevant case reports. If the initial search returned more than 300 citations, disease-specific modifier terms were added such as {gene symbol}[TI] AND {syndrome name} or {gene symbol}[TI] AND 'intellectual disability'. Only papers which described case reports/case series relating to a single gene were included. Reviews (without any novel data) and reports relating to CNV were not included. The PMID for each manuscript was added to the DDG2P data to form a locus-genotype-mechanism-disease-evidence (LGMDE) thread for each disease.

This PubMed search and review process identified 1018 PMID for the 99 disease test set (Supplementary Table 2). These were used as input for Cadmus (43). There was a successful download in at least one format (HTML/XML/PDF/plain text) for 962/1018 papers (94.5%). There were no diseases with zero successful downloads. In several instances, there was a download in more than one format (Figure 3-1). Cadmus includes a quality assessment step to identify the best format to use in these cases.

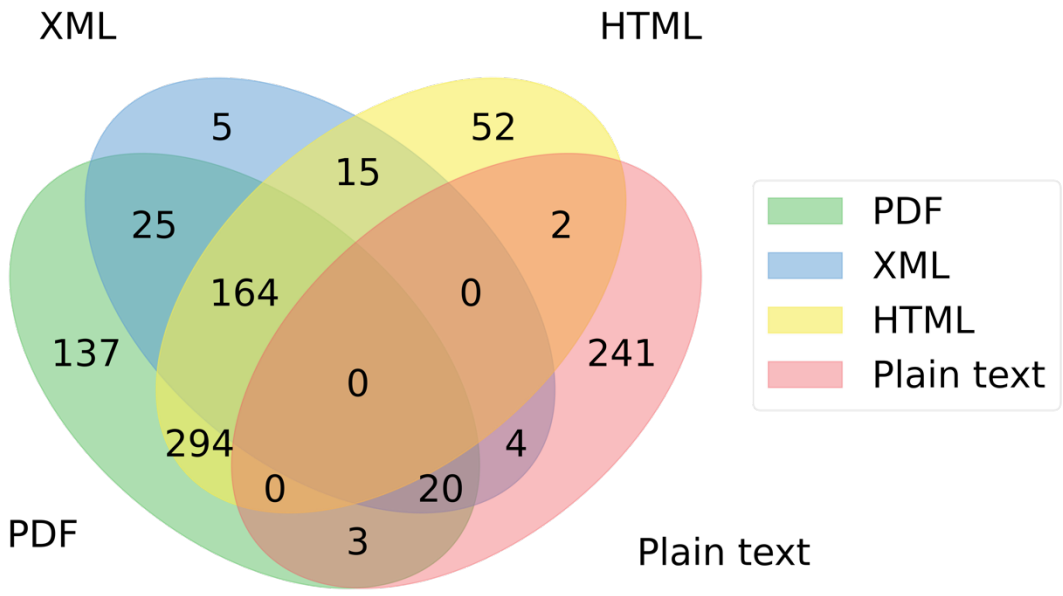


Figure 3-1. Venn diagram of formats downloaded successfully by Cadmus per PMID. PMIDs identified for 99 disease test set used as input, with 962/1018 total successful downloads.

The Cadmus downloads were then used to construct a disease model per DDG2P entry using MetaMap (64), as described in section 2.2.4. Of note, Cadmus by default includes the abstract as 'full text' output if a download was unsuccessful. The abstract in this case was not used for phenotype extraction in this work. MetaMap was configured to restrict the UMLS source vocabulary to the HPO alone (22,60). No other output options were used, based on the analysis in section 2.2.6. The title and abstract for each PMID were derived from PubMed metadata contained in Cadmus output. These were also run through MetaMap to obtain weighted HPO term lists.

3.2.2 Parsing OMIM and Orphanet

To allow comparison of literature-derived disease models to manual curation, the HPO annotated file genes_to_phenotype was downloaded from http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt on the 22nd of April 2021. This contained disease-specific HPO term models derived from OMIM (mim2gene) (29) and Orphanet (28). Each GDD in the 99 disease set from DDG2P was manually annotated with a MIM number (one-to-one mapping), which is a unique identifier used in the OMIM database (29). This was used to match diseases in genes_to_phenotype and map phenotype annotations from these to the DDG2P set.

Orphanet-defined disease entities may map to multiple genes. For example, the ORPHAcode 98938 (‘Colobomatous microphthalmia’) was associated with 14 genes at the time of writing (28). Only those within the set where there was a one-to-one gene-ORPHAcode correspondence were included. This meant there were only 43 entries in the Orphanet subset.

Frequency annotations for each term in the OMIM and Orphanet sets were recorded as part of the model, where present. 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), which is defined in the HPO as ‘present in 80 to 99% of the cases’ (22).

3.2.3 Disease models from DDD study

To enable comparison of literature-derived models to clinical data, a set of HPO terms per gene in the 99 disease set from the DDD study (6) was obtained, with permission. This study used exome sequencing in individuals with a suspected, undiagnosed GDD. HPO terms (non-weighted) were recorded for each proband, as part of the study, by recruiting clinicians at the

time of enrolment. For each gene in the 99 disease set, DDD probands with likely disease-related SNV were identified. The HPO terms per individual identified with one of these variants were amalgamated. This generated a set of DDD disease models where terms were weighted by frequency per proband.

3.2.4 Unified disease model test set

From the above work, a unified 99 disease dataset was created. Each entry in this consisted of a G2P-defined GDD with a corresponding literature-derived (from both title+abstract and full text), OMIM and DDD model. (3,22,29,31). A model consisted of a list of HPO terms, with a frequency weighting if present. Orphanet models were included where available (28). This dataset was used for all further analyses in this chapter.

3.3 Structure and vocabulary of full text models in disease test set

An initial examination of the structure and vocabulary of models in the 99 disease set was undertaken. This was to identify patterns which might inform similarity analyses in subsequent sections.

3.3.1 HPO terms across disease test set

An assessment of the vocabulary used across datasets was carried out. The HPO terms derived from the full text were compared to those from DDD and OMIM (Figure 3-2). The full text set contained more unique HPO terms overall, with 2869, compared to 1569 from DDD and 972 from OMIM. There was significant overlap with DDD and OMIM (Figure 3-2). This meant there were only 630 extra unique terms in the combined DDD + OMIM set compared to those derived from full text alone.

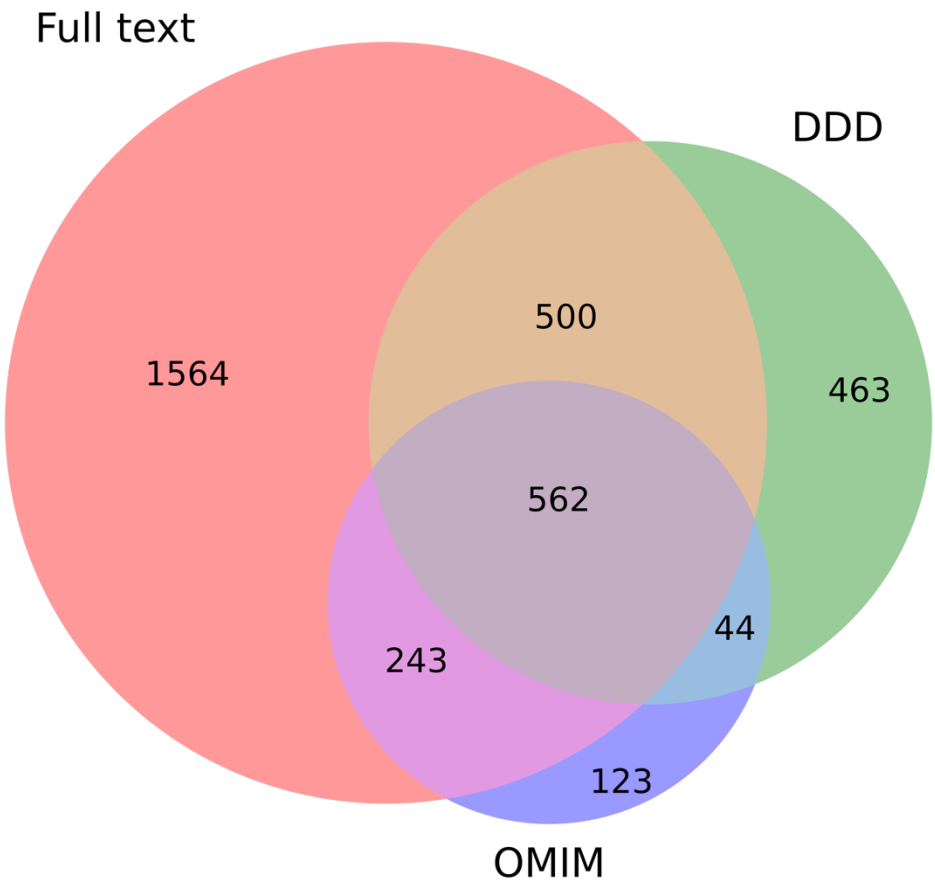
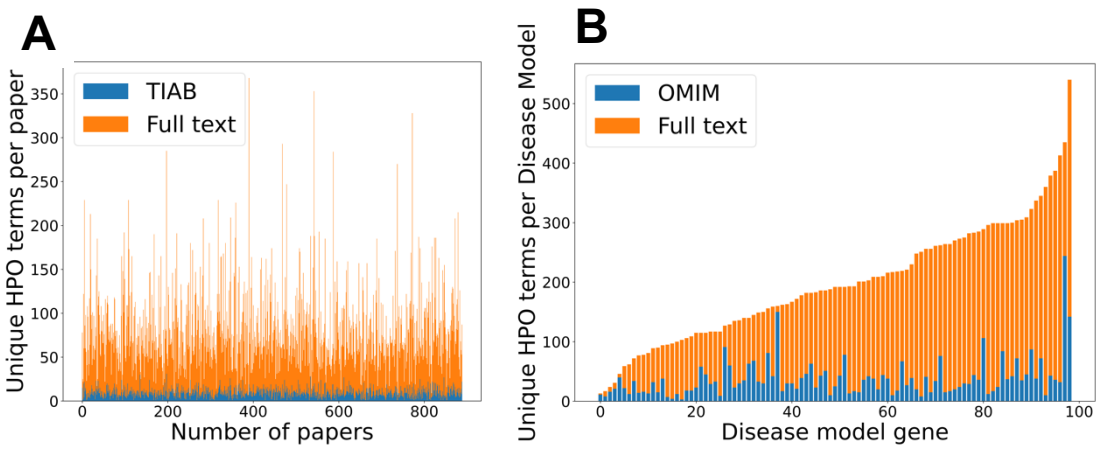


Figure 3-2. Venn diagram of unique terms in full text-derived, DDD, and OMIM vocabularies. Terms taken from all models in 99 disease test set. DDD – Deciphering Developmental Disorders study, OMIM – Online Mendelian Inheritance in Man.

3.3.2 Disease model size using different data sources

The disease model size was defined as the number of unique HPO terms contained within each model. It was hypothesised that model size correlates with phenotyping depth, where depth equates to a more comprehensive representation of the true phenotypic spectrum of a disorder.



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 3-3. (A) Comparison of number of unique HPO terms extracted per paper, comparing title+abstract to the full text. For a sample of 962 papers, HPO terms were extracted using the full text download pipeline. For each manuscript, the number of unique terms extracted from the title+abstract were compared to the equivalent number from the full text. (B) Comparison of number of unique HPO terms in disease models. The disease set, comprised 99 genetically-determined developmental disorders. For each of these, disease models were generated using the full text download pipeline. The number of unique HPO terms per literature-derived model was compared to the number of terms in the corresponding OMIM model. The OMIM models were generated through manual curation.

Unsurprisingly perhaps, there were significantly more HPO terms per paper in the full text than in the title + abstract after phenotype extraction (Figure 3-3A). This may indicate that there is increased depth of phenotyping using full text extraction as demonstrated in this work, compared to title+abstract alone, which has been more commonly used previously (49). Comparison to

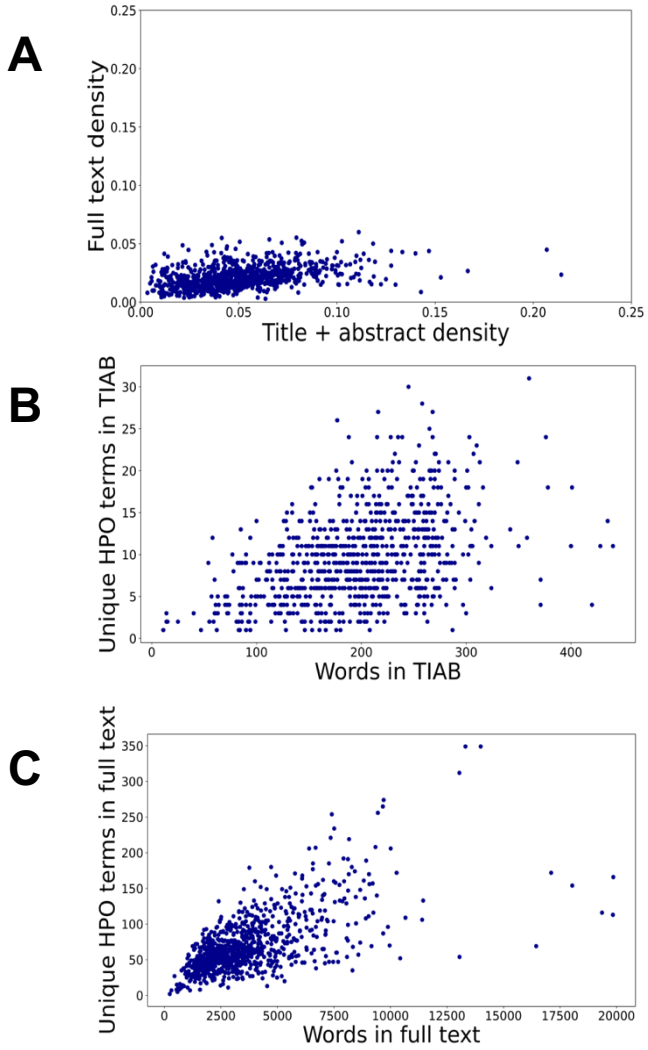


Figure 3-4. Relationship between words in title+abstract/full text and number of unique HPO terms. Density is the number of unique HPO terms/number of words in text. Number of words determined by whitespace separated tokens. TIAB -title + abstract; HPO – Human Phenotype Ontology.

clinically-derived data is made later in this thesis (section 3.5) to determine if increased model size might mean better phenotypic coverage.

This figure also demonstrates considerable variability in the number of terms per paper. It was hypothesised that the number of HPO terms extracted could be related to the length of the text used. Figure 3-4A shows that the density (number of HPO terms/number of words in text) for title + abstract was in fact higher than that of the full text. There was no obvious relationship

between number of HPO terms and length of text for title + abstract (Figure 3-4B). For the full text, there was more of a correlation between number of terms extracted and length of text (Figure 3-4C). However, there was still considerable variability, meaning the increased number of terms from full text was not simply a function of the size of the paper.

There were also more terms per full text-derived disease model than in OMIM (Figure 3-3B). This may indicate that there is increased depth of phenotyping using automated phenotype extraction from full text, compared to manual curation. It is likely a proportion of these terms are false positives from the NER method. It should also be noted that NER misses a proportion of terms present in clinical data (Figure 3-2).

There may be less terms in OMIM because manual annotators were not adding the full spectrum of terms present in the manuscripts analysed, or that less papers were used per disease to create models in the OMIM set compared to the literature searches used here. The number of references per OMIM model were not checked as OMIM do not publish a clear methodology defining how they extract phenotypic terms. It is possible a subset of papers are used, or even that manuscripts which are not referenced by OMIM are utilised, although this seems less likely.

There did not seem to be a correlation between the size of disease models between datasets. It might have been expected there would be a relationship between, for example, disorders defined years ago with many published reports and the size of the corresponding disease model for both the literature-derived and OMIM sets. This may reflect generally less papers being used for curation in OMIM, as entries in this database generally include references sufficient to define a disorder rather than a completely comprehensive overview of all case reports/case series. The source papers used for phenotype curation were not available. Therefore, it was not

possible to test whether the differences in model sizes between these datasets was due to length of input text.

In conclusion, phenotypic extraction from full text results in increased numbers of terms per disease compared to title+abstract mining and manual curation. This may reflect increased depth of phenotyping and/or increased phenotypic noise with an automated approach. Comparison is made to prospectively derived clinical data later in this work (section 3.5) to help clarify this.

3.3.3 Frequency weightings in full text models

Each full text disease model contains HPO terms weighted by their summed frequency across relevant case series/case reports. Normalization of these frequency weightings would be useful for inter- and intra-model comparison between the data sources in the 99 disease set (full text literature, G2P, OMIM, Orphanet, DDD) (3,28,29,31). In this case, normalization means adjusting frequency weightings to lie on the same scale. Without this, it is not possible to compare weighted terms between models. Therefore, an analysis of the frequency weightings for each model was undertaken.

The distribution of frequencies was highly skewed, as shown in Figure 3-5. There were also a large number of terms per model with a frequency of one. Furthermore, there was no denominator for each frequency as the phenotype extraction was on a per-paper rather than per-individual basis. Therefore, it was not possible to determine if a high frequency value in itself meant a phenotypic feature was highly expressed in a particular disease. There were other potential sources of highly weighted terms, for example if an input paper had a particular focus on a given phenotype. Overall, these issues offered significant obstacles to normalization of frequency data across the set.

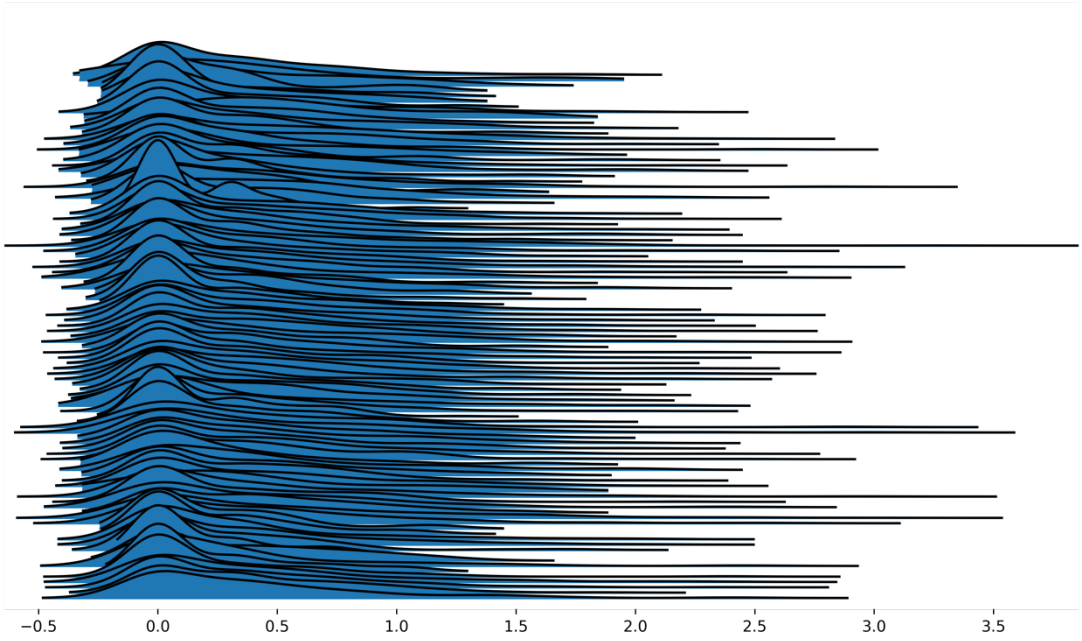


Figure 3-5. Ridgeplot of term weightings per model in 99 disease test set. Log_{10} transform applied to frequency values. Each model is represented as a density subplot. The subplot is generated using Gaussian kernel density estimation. Note probability density estimations of values near 0 will result in plots showing negative x-axis values; all input weightings are positive.

Analysis was undertaken of the most frequent terms in the literature-derived models by weighting across the test disease model set (without collapsing terms) (Table 3-2). The top five terms appeared in 78-95% of all models. The terms themselves were expected common features in the GDD domain, including seizures, intellectual disability and autism.

The terms ‘Intellectual disability’ (HP:0001249) and ‘Global developmental delay’ (HP:0001263) are not directly related in the HPO. However, they are clinically similar as they both refer to global impairment of learning. The same child may have global developmental delay as an infant, and develop intellectual disability when they are older. These terms could therefore be collapsed together, e.g. to ‘Intellectual disability’, for the purposes of disease model comparison. The same is true for ‘Autistic behavior’ (HP:0000729) and ‘Autism’ (HP:0000717). This again illustrates the issue of equivalent features being recorded with different HPO terms.

HPO Term		Frequency weighting across all literature-derived models	% models containing term
Seizure		9016	87.88
Intellectual disability		5520	94.95
Global developmental delay		2173	93.94
Autistic behavior		1858	79.80
Autism	1439	77.78	

Table 3-2. Most frequent terms in literature-derived models by weighting across 99 disease test set.

Gene	Disease	HPO term	Frequency in literature-derived model
SCN8A	Epileptic encephalopathy, early infantile, 13	Seizure	1635
KCNQ2	Epileptic encephalopathy early infantile type 7	Seizure	974
STXBP1	Epileptic encephalopathy early infantile type 4	Seizure	882
SCN2A	Infantile epileptic encephalopathy	Seizure	841
KCNQ2	Benign neonatal epilepsy type 1	Seizure	724
KMT2B	Complex early-onset dystonia	Dystonia	613

Table 3-1. Terms with weighting of >500 in individual literature-derived disease models, from 99 disease test set.

Examination of the top terms per model revealed a significant representation of epilepsy syndromes (Table 3-1). Predictably, ‘Seizure’ (HP:0001250) was the most frequent term for each of these, and contributed to this being the top-ranked term across all models (Table 3-1). ‘Dystonia’ (HP:0001332) also appeared frequently for the *KMT2B*-related dystonia syndrome.

3.3.4 Heterogeneous recording of similar phenotypic features

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 3-6 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 (‘Hearing impairment’ (HP:0000365), ‘Choanal atresia’ (HP:000453) ‘Coloboma’ (HP:0000589), ‘Abnormality of cardiovascular system morphology’ (HP:0030680), and ‘Abnormality of the ear’ (HP:0000598) are therefore very pertinent for this condition, although some of them are higher-level and less specific. However, a number of terms clinically relevant to these were also present in the same model, and this pattern is repeated across comparison datasets. This again shows that similar phenotypic features may be recorded in a heterogenous manner, and complicates comparison between models. This may be addressed to some extent using semantic similarity with the HPO structure at the cost of losing specificity with higher-level terms.

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	
	Ventricular septal defect	Ventricular septal defect	Abnormal aortic valve morphology
	Abnormal heart morphology	Tetralogy of Fallot	Tetralogy of Fallot
	Atrioventricular canal defect	Double outlet right ventricle	Patent ductus arteriosus
	Double outlet right ventricle		Interrupted aortic arch
	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 3-6. Example top five ranked terms in disease model for CHD7/ CHARGE syndrome (left column). Clinically related terms in remainder of disease model (n=540), OMIM model (n=71) and Orphanet model (n=82) shown.

3.4 Effects of modification to disease model generation

Iterative modification of the process used to create disease models was undertaken, to identify features which may be used to refine automated literature curation. The performance of these changes was evaluated through comparison with DDD models (3). Increasing similarity of literature-derived models to DDD was thought to represent an improvement to the disease model creation process.

DDD data here was used as a proxy for the 'real life' clinical expressivity of a given disorder. Therefore, DDD represents the ground truth for the phenotype of a condition here. This is not perfect, for example individuals were recruited to DDD because they were undiagnosed using conventional clinical evaluation and testing. This meant some individuals had atypical examples of well-characterised disorders. However, DDD represents a rich source of phenotypic data verified with molecular genetic testing. An intra-model comparison was also undertaken, to assess internal consistency and potential bias.

3.4.1 Rank biased overlap to assess disease model similarity

Similarity between models was assessed using rank biased overlap (RBO). This is a method of comparing ranked lists which is top-weighted, can compare lists which contain differing members, and is monotonic with increasing depth of list (89). This method therefore does not take into account the HPO ontology structure.

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, iterative similarity testing of different values of p was undertaken. From this process, the greatest discriminant power was found when p was

equal to 0.98, weighting towards the top 50 terms in a list. This value of p was used for all analyses in this work. RBO output is usually expressed as a min-max range, however for this work the extrapolated RBO_{EXT} point score was used for ease of comparison (89).

A Python implementation of RBO (rbo version 0.1) was used for all calculations in this work (155). For each model in one comparison set, RBO was calculated compared to all models in the other data source. DDD-derived frequency was the amalgamated count per proband for a disease gene.

Literature- and DDD-derived models were ranked according to term frequency. Literature-derived frequency was the number of times a term occurred across all input manuscripts for a given model. For example, if a term was mentioned 100 times in one manuscript, and once in another, the aggregate frequency would be 101 for the disease model.

OMIM model terms are not generally annotated with a frequency, and where this is present, it is in a frequency range such as ‘Very frequent’ (99-80%) (22). To weight the terms in OMIM models for the purpose of list ranking used by RBO, the frequency of each term across the whole OMIM dataset was used.

Comparison heatmaps using RBO for literature-derived and DDD comparison datasets (3) were used to visualise the effects of alterations to full text model generation in the following sections (Figure 3-7, Figure 3-10, Figure 3-12). Here, every disease in one dataset was 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. Similarity between corresponding models was indicated by a diagonal line seen from top left to bottom right.

3.4.2 Stratification of models by MetaMap score

MetaMap provides a score for each biomedical concept identified, to quantify the confidence level for each mapping (156). This is based on the frequency of a concept in a document and its relevance to the text in context. Full text-derived disease models for the 99 disease set were modified to assess the effect of stratification of phenotype terms by MetaMap score.

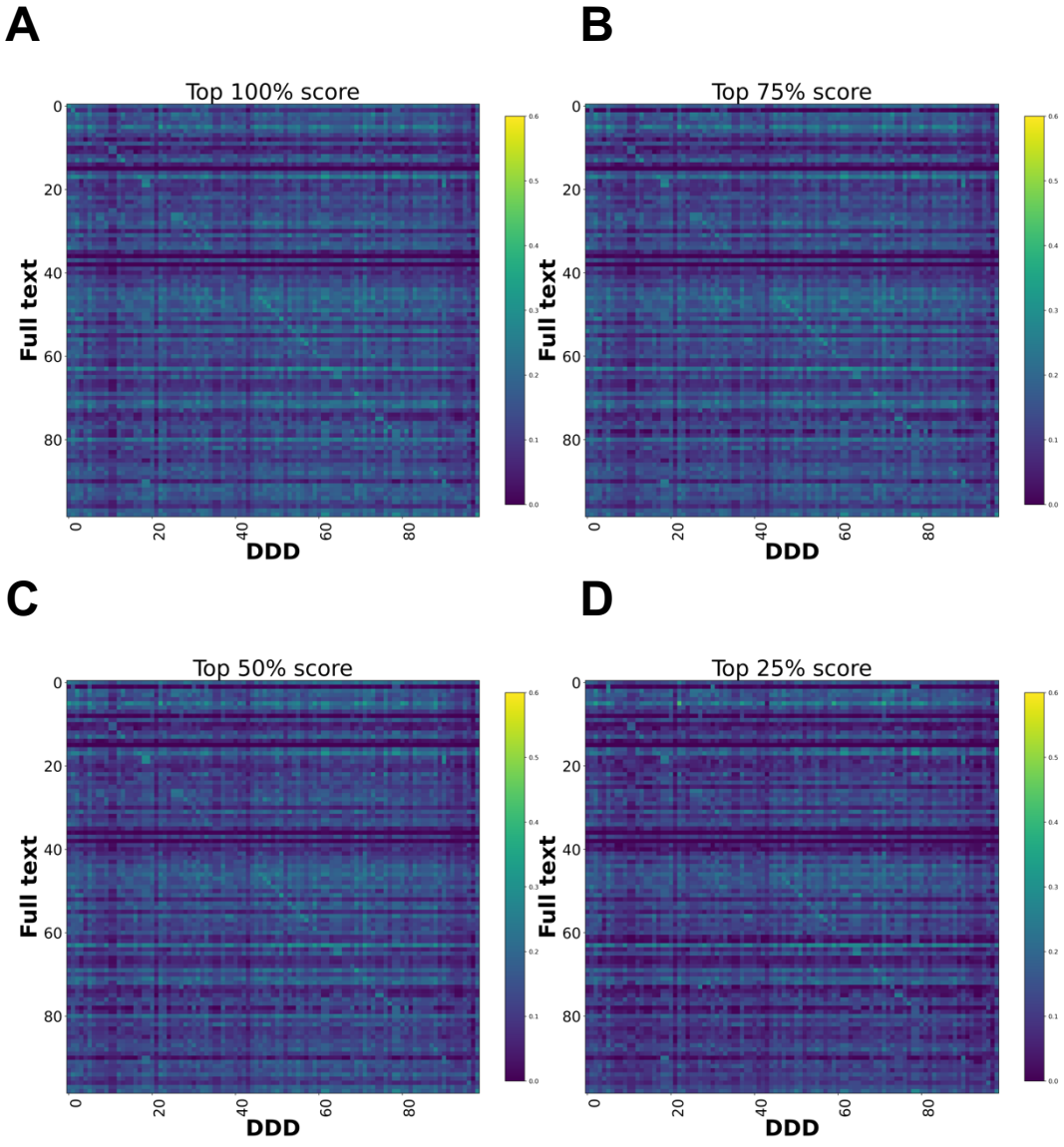


Figure 3-7. Full text-derived models in 99 disease test set compared to DDD, using rank biased overlap. HPO terms in full text models subset according to MetaMap score.

The full text for each PMID was processed with MetaMap as described in section 2.2.4. For each PMID, the mapped HPO terms were divided into quartiles according to their individual MetaMap scores. Disease models were constructed as previously described, except these were divided into the top 100%, 75%, 50% and 25% terms per PMID, by MetaMap score. These were compared to DDD models.

There was no improvement in full text-DDD similarity with selection for higher MetaMap score (Figure 3-7). Rather, there was a tendency towards less similarity overall with selection for the top 25% score, although this may also be associated with a reduction in the number of terms per model.

3.4.3 Collapsing clinically similar HPO terms

Given the possibility of recording the same or similar phenotypic feature using different HPO terms (or CUIs), it was hypothesised that a method of collapsing together clinically relevant terms could increase the performance of literature-derived disease models, as measured by similarity to DDD and OMIM data.

Collapsing terms could result in loss of informativity, but may also increase the power of HPO term-based models in disease similarity matching. For example, two individuals with the same disorder may have ‘Unilateral microphthalmos’ (HP:0011480) and ‘Bilateral microphthalmos’ (HP:0007633) respectively. It may be important to discriminate between bilateral and unilateral features, but in practice it could be more relevant to record ‘Microphthalmia’ (HP:0000568) as a discriminant clinical feature of the condition.

In theory, the structure of the HPO should make this straightforward, with similar terms being collapsed to their common ancestor. In the example given, ‘Microphthalmia’ is the parent term of ‘Unilateral microphthalmos’ and

‘Bilateral microphthalmos’. However, there are instances of clinically similar (using subjective medical judgement) terms which do not share an informative common ancestor in the HPO. Figure 3-8A shows clinically discriminant terms which share the high level, less informative common ancestor ‘Abnormality of skeletal morphology’ (HP:0011842). Therefore, collapsing terms solely using the HPO structure was not pursued.

However, there was a subset of HPO terms identified which exactly included their parent, and which could be collapsed to this parent without losing significant discriminant power. These largely comprise include terms containing modifiers, for example relating to severity – ‘Intellectual disability, severe’ (HP:0010864) collapsing to ‘Intellectual disability’ (HP:0001249) (Figure 3-8B) – and to clinical descriptors – ‘Atypical absence seizure’ (HP:0007270) collapsing to ‘Absence seizure’ (HP:0002121). Iterative collapsing was also possible: ‘Midline facial capillary hemangioma’ (HP:0007601) to ‘Facial capillary hemangioma’ (HP:0000996) to ‘Capillary hemangioma’ (HP:0005306).

This method, collapsing to parent HPO terms if the parent was exactly contained as a substring within the child, was scripted and applied to the whole HPO in an automated manner. This enabled testing of the hypothesis that this collapse process would increase the performance of literature-derived disease models. The total list of collapsed terms generated was manually reviewed to ensure this process did not significantly reduce clinical informativity. A small list of terms were additionally identified for manual collapse from this, including ‘Autism’ and ‘Autistic behavior’.

It should be noted, however, that this collapsing technique was not successful in many instances. The examples given above – ‘Unilateral microphthalmos’ (HP:0011480) and its parent ‘Microphthalmia’ (HP:0000568) – do not collapse. There are numerous shared characters in these, and a fuzzy matching strategy was trialled to address this, to collapse where a term

closely matched its parent (rather than an exact match as above). However, the results of fuzzy matching were not thought to be accurate enough for further use.

The exact matching collapse method was applied to the literature-derived models from the 99 disease set. The total number of unique HPO terms after collapse was 2360, compared to 2869 without applying this. Table 3-3 shows the top five terms across all 99 disease models before and after collapse. The top terms were largely unchanged across both sets overall. The frequency weighting of ‘Seizure’ and ‘Intellectual disability were increased. ‘Autism’ changed from the fifth most common to the third most frequent term. This was due to ‘Autistic behavior’ being collapsed into ‘Autism’. Arguably, ‘Global developmental delay’ and ‘Intellectual disability’ are clinically very similar, and these could be manually collapsed, which would further increase the frequency of ‘Intellectual disability’.

RBO heatmaps were constructed to visualise the difference in similarity scores between non-collapsed and collapsed full text models. The method for generating heatmaps as described in section 3.4.1 was used, except the similarity score for each collapsed model comparison was subtracted from the equivalent non-collapsed score. The heatmaps for full text vs DDD and full text vs OMIM did not show any appreciable difference using this method (Figure 3-9). Overall, the exact match collapse method did not appear to significantly alter the performance of full text-derived disease models when assessed using the RBO similarity metric. Therefore, this was not used for the disease models further described in this work.

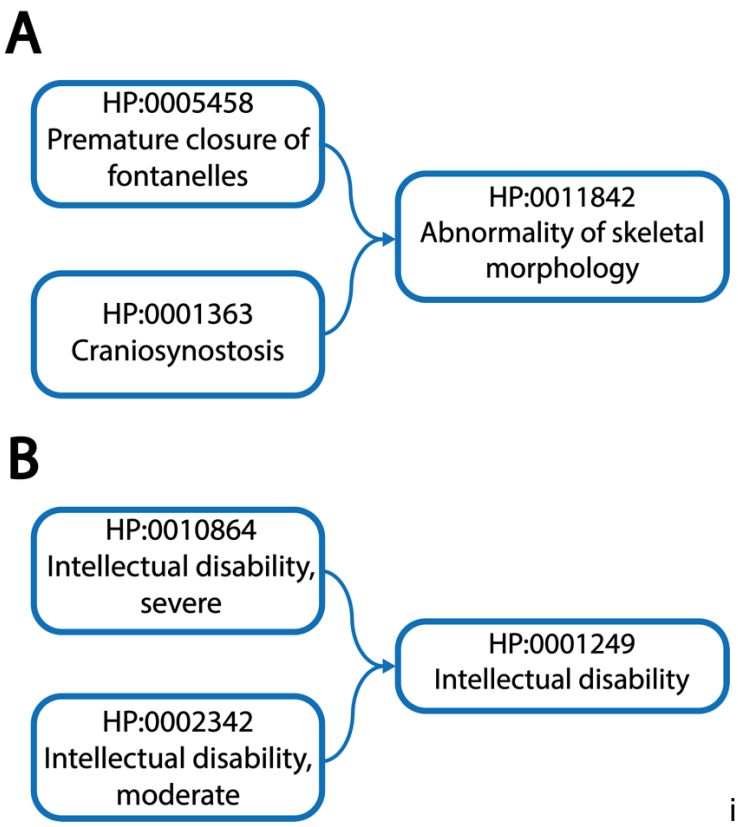


Figure 3-8. Clinical relatedness and collapsing HPO terms using ontology structure. (A) shows two terms which are clinically related but which have an uninformative common ancestor. (B) shows terms which contain their parent. These can be collapsed to the parent without significant loss of information.

Non-collapsed	Count	Collapsed	Count
Seizure	9016	Seizure	9773
Intellectual disability	5520	Intellectual disability	6870
Global developmental delay	2173	Autism	3297
Autistic behavior	1858	Global developmental delay	2307
Autism	1439	Hypotonia	2158

Table 3-3. Top five most frequent HPO terms in literature-derived models without and with iterative collapse method. Models from 99 disease test set corpus used. HPO – Human Phenotype Ontology.

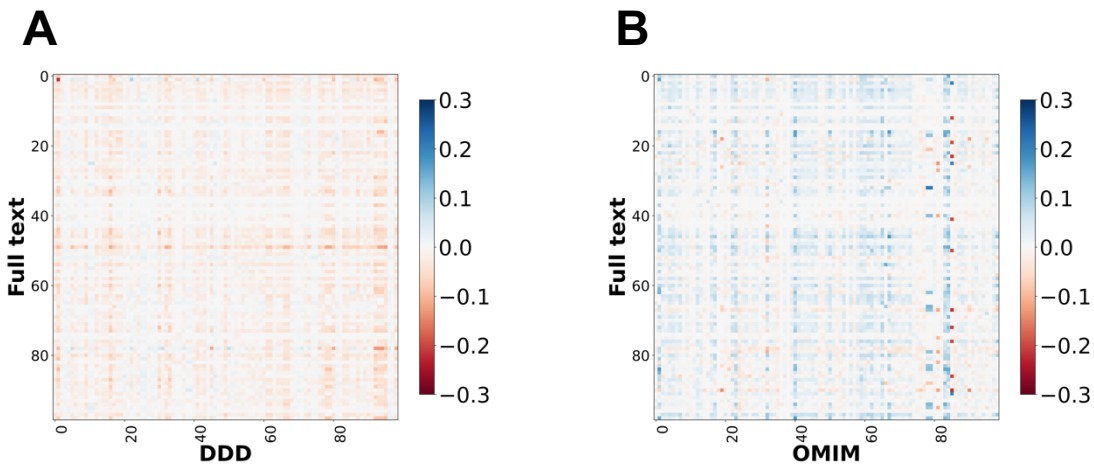


Figure 3-9. RBO difference heatmap comparing all full text, to full text with HPO terms collapsed to parent, if parent exactly contained in child. A) Difference for full text/collapsed full text similarity scores compared to DDD set. B) Difference for full text/collapsed full text similarity scores compared to OMIM set. RBO scores for full text models containing all terms subtracted from score for collapsed models to assess difference. RBO scores calculated compared to DDD models. RBO – Rank Biased Overlap, HPO – Human Phenotype Ontology, DDD – Deciphering Developmental Disorders Study, OMIM – Online Mendelian Inheritance in Man.

3.4.4 Intra-model comparison by splitting PubMed IDs

A comparison within disease models was undertaken by splitting each set of PMIDs per DDG2P entry in two (or n,n+1 where the list was of unequal length). This was to assess for internal consistency/bias. For example, if one paper used to construct a model described a significantly different phenotype, this would be seen in this analysis. Disease models were otherwise constructed using Cadmus and MetaMap as described in sections 2.2.3, 2.2.4 and 3.2.1.

The heatmap generated from this process showed that corresponding models were highly similar (Figure 3-10). This indicated that models in the disease test set were not unduly influenced by skewed input papers, for example those focused on a single phenotypic feature.

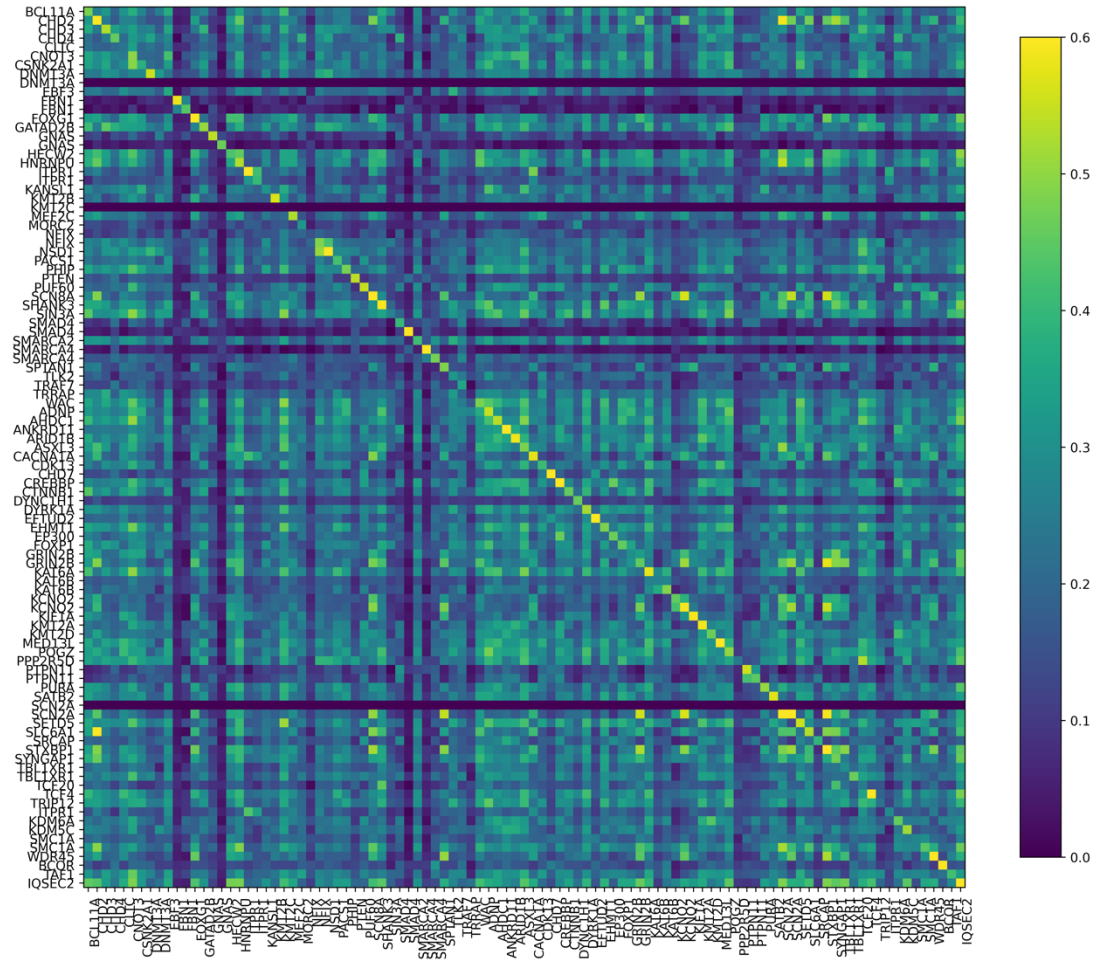


Figure 3-10. Intra-model comparison of full-text derived disease models, with each list of PubMed IDs used to derive the model divided into two. Therefore, each comparison involves two sets of different papers.

3.4.5 Effect of removing single occurrence HPO terms

Each full text-derived model contained a number of HPO terms which appeared only once (frequency weighting =1) (Figure 3-12). It was hypothesised that removing this long tail of low frequency terms would increase disease model similarity to DDD data, potentially removing phenotypic noise.

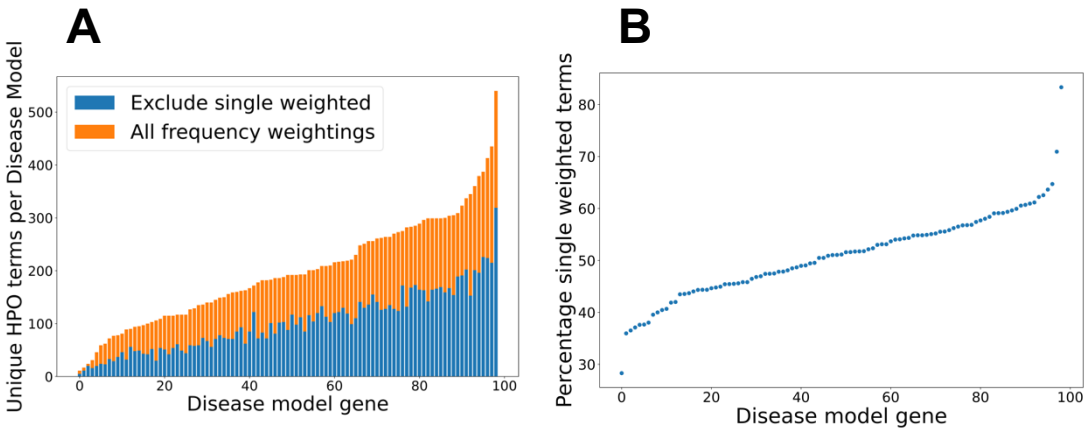


Figure 3-11. Terms with frequency weighting of one across 99 disease set. A) Unique terms per model including and excluding single weighted terms. B) Percentage of single weighted terms per model.

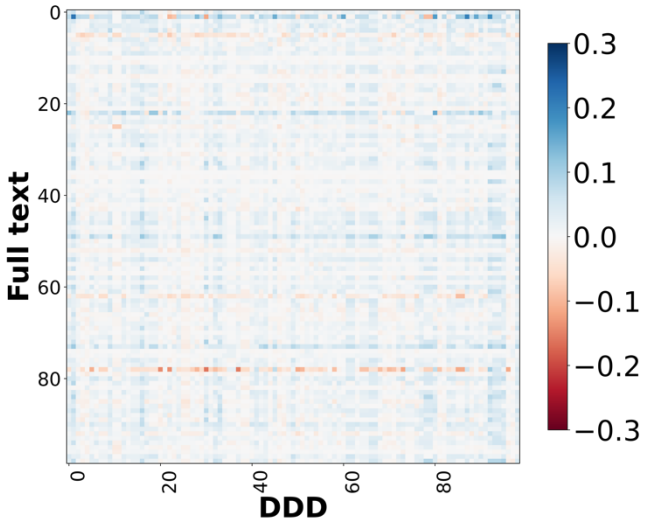


Figure 3-12. RBO difference heatmap comparing all full text, to full text with HPO terms occurring once removed. RBO scores for full text models containing all terms subtracted from score for model with single terms removed to assess difference. RBO scores calculated compared to DDD models. RBO – Rank Biased Overlap, HPO – Human Phenotype Ontology, DDD – Deciphering Developmental Disorders Study

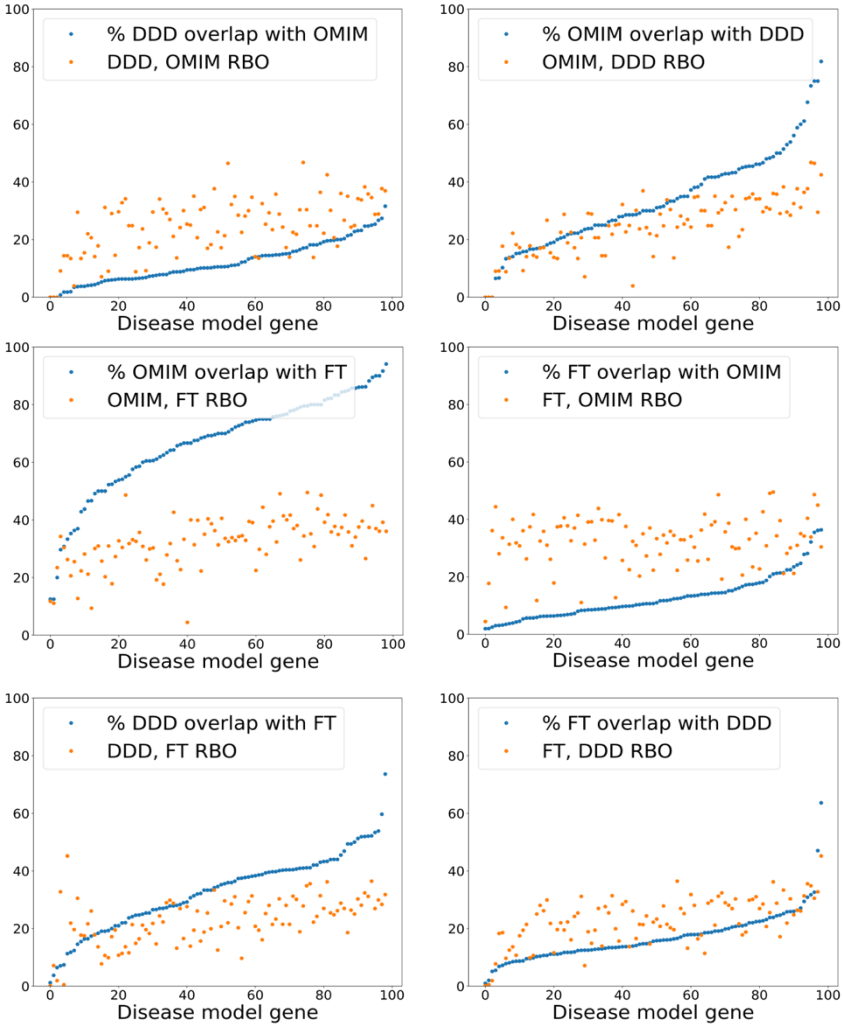


Figure 3-13. Exact model term overlap against RBO. For each pair of comparison models – A & B – the percentage of exact match terms (A in B / A) is shown. The corresponding RBO similarity score for A & B is also plotted. RBO scores multiplied by 100 to normalise to percentage range. FT – full text-derived, OMIM – Online Mendelian Inheritance in Man, DDD – Deciphering Developmental Disorders Study, RBO – Rank Biased Overlap.

A difference RBO heatmap was constructed to compare the similarity scores for full text-derived models to full text-derived models with single occurrence terms removed, according to the method in section 3.4.1. Removing terms with a single occurrence did not make an appreciable difference to the signal on this heatmap, when compared to the non-adjusted full models (Figure 3-11).

The ROC curve for RBO full text vs DDD was not significantly affected either, with AUC of 0.84 with single frequency terms removed, compared to 0.85 with all terms. It may be useful to simplify disease models by removing single occurrence terms in future, given they do not appear to significantly affect disease model expressivity.

This generated the hypothesis that the signal from RBO based comparisons was dependent more on set overlap than term rankings, which could partially explain the similar results seen with the unweighted MICA method (98). 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 3-13). 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.

In conclusion, the disease model modifications reviewed in this section did not significantly improve performance, as assessed by similarity to DDD. Therefore, these were not applied to the test dataset for analyses in the rest of this work.

3.5 Comparison of automated and manually curated disease models using similarity metrics

3.5.1 Similarity metrics for disease test set

Following the analyses above, further comparison was made between the full text, DDD (6) and OMIM (29) datasets. This was to determine:

1) The similarity of full text models to prospectively gathered clinical data, to assess if they reflect ‘real life’ phenotypic presentations.

2) Whether OMIM models, as an example of widely used manual curation, are more or less similar to DDD clinical data than those derived from full text.
3) The similarity of full text and OMIM models, to directly compare the method presented here with a widely used manually curated set.

Two similarity metrics were used for this comparison. One of these was RBO, as discussed in section 3.4.1 (89). The other was MICA-based semantic similarity using IC following the method of Resnik (100), as discussed in section 1.8.4. Heatmaps were generated as described in section 3.4.1. The comparisons computed were the literature-derived set vs the DDD set, OMIM vs DDD and literature vs OMIM (6,22).

3.5.2 MICA-based semantic similarity

For the MICA-based semantic similarity measure, the frequency of HPO terms combined across both comparison datasets, for example literature-derived and DDD, were used to calculate the IC for each term following the method used by Helbig et al. (98). 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 $-\log_2(f/n)$ (96).

The MICA of two terms is the shared ancestral term in the ontology with the highest IC. The higher-level/less specific the ancestor, the lower the IC is likely to be. 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 normalisation measure (98).

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

Semantic similarity scores between literature-derived, DDD and OMIM sets were performed using unweighted disease models, i.e. every term appeared uniquely per model.

3.5.3 Similarity comparison between annotated models

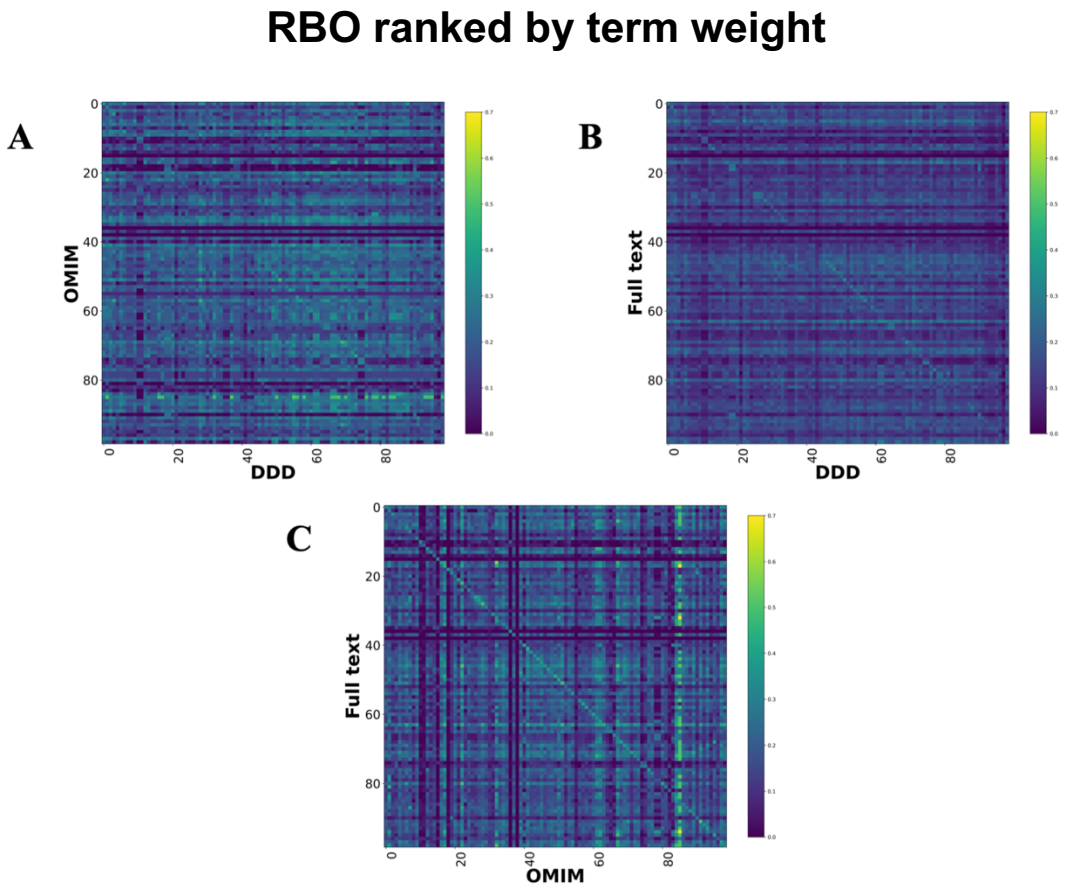


Figure 3-14. Disease model comparison heatmaps using rank biased overlap for literature-, OMIM- and DDD- derived models. Each model on the y axis is compared with every model in the DDD set. Disease/DDD models describing the same disorder are on the rightward downslanting diagonal. (A) compares OMIM and DDD models. (B) compares literature-derived and DDD models. (C) compares literature-derived and OMIM models. 99 diseases in comparison set.

RBO heatmaps showed a weak, but recognizable signal for literature-derived models vs DDD and OMIM vs DDD (Figure 3-14), showing that both curated sets appeared to besimilar to clinically-derived data. There was a much clearer signal for literature-derived models vs OMIM, showing that automated curation may create phenotypic data similar to that generated by expert manual curators.

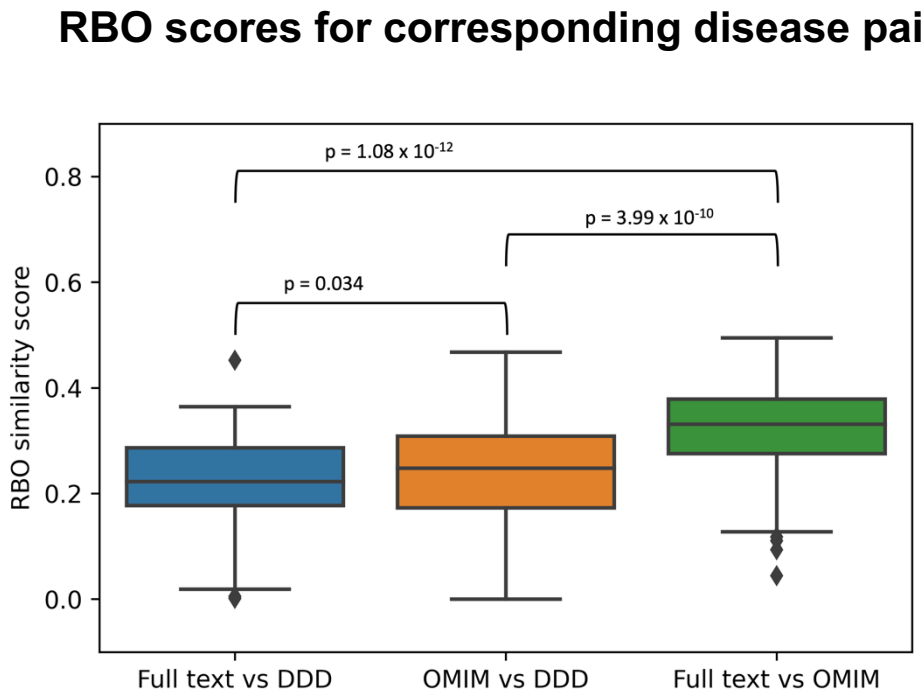


Figure 3-15. Boxplots of RBO similarity scores for corresponding disease pairs across different data sources for 99 disease set. This is the same data as seen in the diagonal on the heatmaps. P-values calculated using Wilcoxon signed-rank test.

A boxplot of these results (Figure 3-15) demonstrated slightly higher scores for OMIM compared to DDD than when using the full text models. This was significant, but the absolute difference in scores was small. There were significantly higher scores for full text vs OMIM, which reflects the strong signal seen on the heatmap.

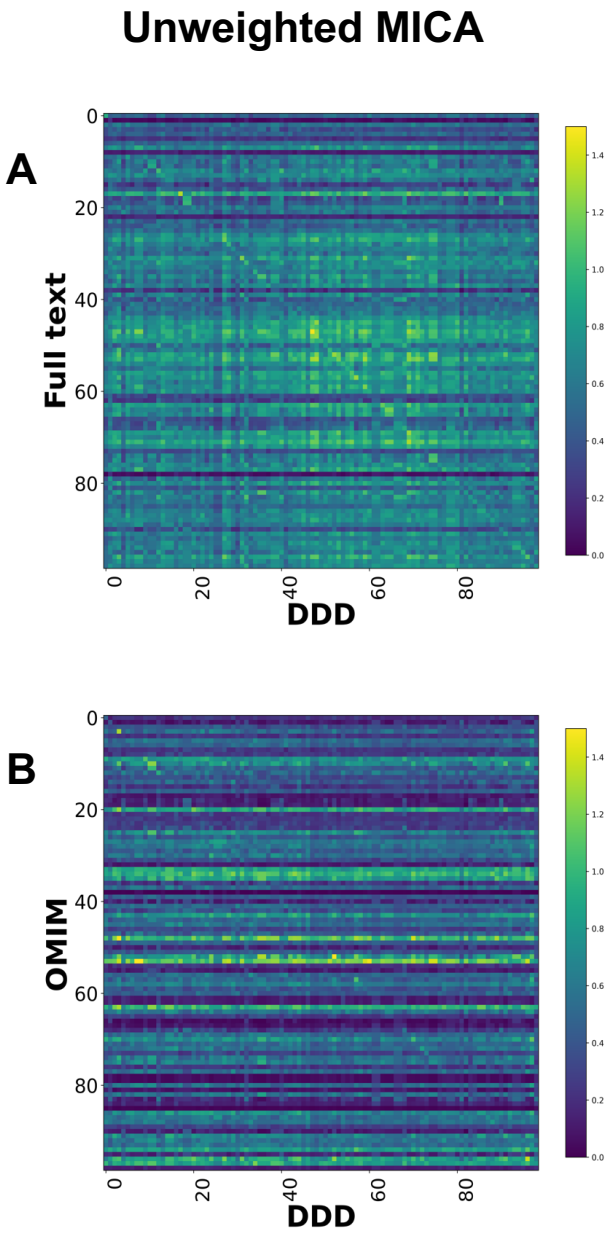


Figure 3-16. Disease model comparison heatmaps using unweighted semantic similarity (MICA) for literature-, OMIM- and DDD- derived models. Each model on the y axis is compared with every model in the DDD set. Disease/DDD models describing the same disorder are on the rightward slanting diagonal.

Comparison heatmaps using the unweighted semantic similarity/MICA method showed an identifiable signal between the full text and DDD models (Figure 3-16A). A faint signal was seen for OMIM against DDD, although

similarity scores were more homogenous across all models for this set (Figure 3-16B).

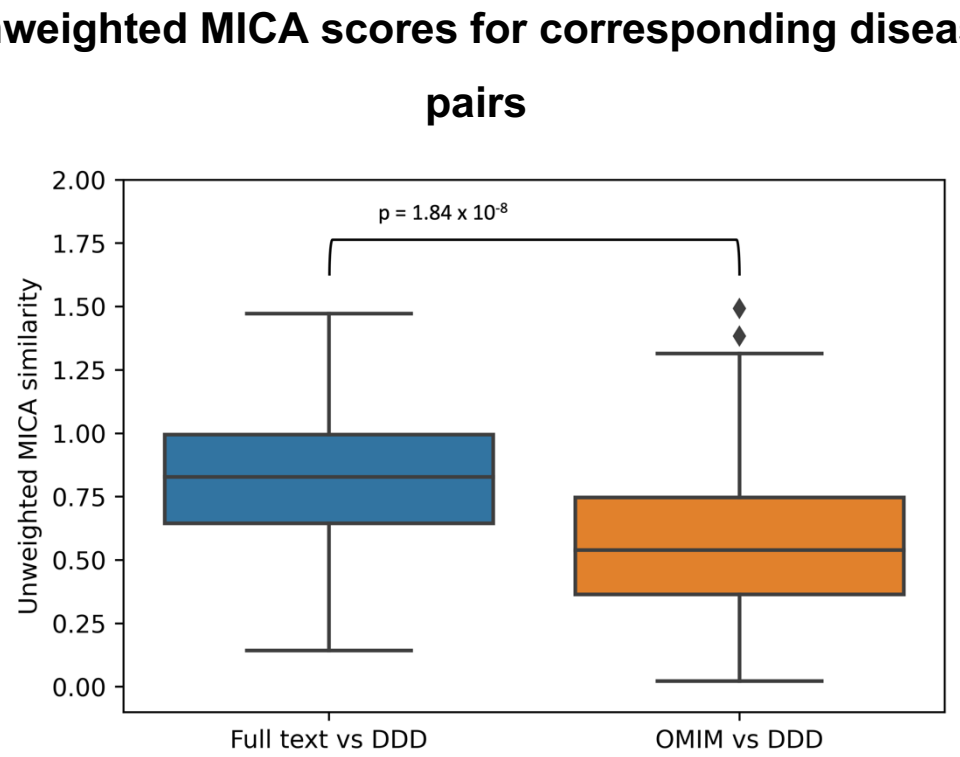


Figure 3-17. Boxplots of unweighted MICA similarity scores for corresponding disease pairs across different data sources for 99 disease set. This is the same data as seen in the diagonal on the heatmaps. P-value calculated using Wilcoxon signed-rank test.

A boxplot (Figure 3-17) showed that full-text models generated significantly higher similarity scores compared to the 'real life' DDD clinical data than those from OMIM, using the unweighted MICA method.

In conclusion, it appeared that full text-derived models were highly similar to manually curated OMIM data, using the RBO metric. This may be because of the significant degree of exact match overlap between these sets (Figure 3-13). There were slightly discordant results between the two similarity methods when comparing to DDD data. Similarity scores were slightly higher for OMIM than the full text models ($p = 0.034$). However, with the unweighted MICA method, similarity scores were significantly higher for the full text

models ($p = 1.84 \times 10^{-8}$). It is possible that this reflected a greater discriminant power of the MICA-based method to determine clinical difference, given the utilisation of the ontology structure.

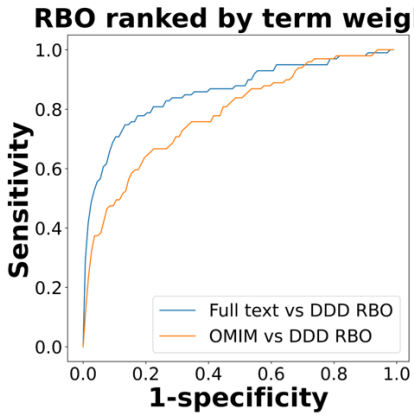
3.5.4 Receiver operating characteristic curves for full text, DDD and OMIM comparisons

To further test the full text models, it was hypothesised that these would be at least as performant as those derived from manual curation at predicting correct diagnoses in DDD. This was designed as a proxy for using curated disease models to predict a correct diagnostic gene using phenotype data alone, in a clinical setting.

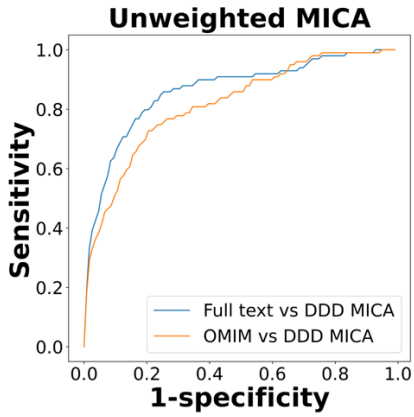
To assess this hypothesis, ROC (receiver operating characteristic) curves were generated from a matrix containing similarity scores (RBO or semantic similarity scores) for every disease in one dataset e.g. full text, compared to every disease in the other e.g. DDD. For a given disease, for a threshold x, where x is an integer between 1 and the number of diseases in the set (99), true positive, true negative, false positive and false negative figures were derived by determining if a corresponding disease pair similarity score, e.g. *CHD7*-full text vs *CHD7*-DDD was present in the top x scores across the 99 diseases. This was repeated for each disease. The AUC was calculated using the numpy trapz module (157).

The full text-derived models outperformed OMIM in both similarity metrics – ranked lists using RBO (89) and MICA-based semantic similarity (98) – as defined by an increase in the area under the curve (Figure 3-18). Of note, the MICA analysis did not use term weighting in this case. A similar semantic similarity evaluation using a subset of models from Orphanet showed comparable performance to those derived from full text when unweighted (Figure 3-19).

A



B



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 3-18. ROC curves using threshold ranking for literature-derived/OMIM disease models compared to real life terms in DDD study (3), 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 ranked biased overlap (RBO) (89) to compare ranked lists of terms. Literature-derived and DDD models were ranked according to model term frequency. OMIM models were ranked according to frequency of terms across all OMIM models. (B) uses mean most informative common ancestor (MICA) to compare models (96,98), with information content calculated according to Resnik (100). Unweighted models were used for comparison, meaning each term in a model appeared only once, and term frequencies were not utilised. (C) shows area under curve (AUC) for each model comparison.

3.5.5 Weighting MICA comparison using Orphanet

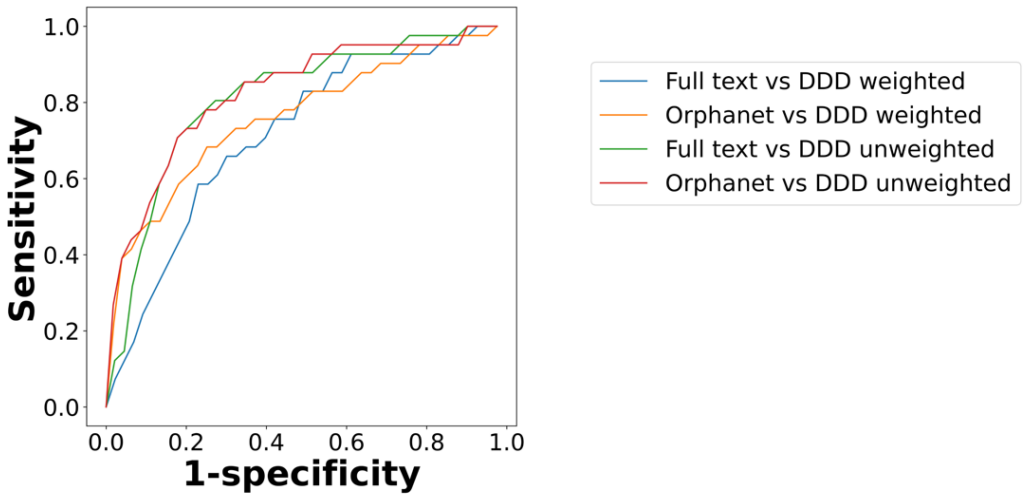
Unweighted models were used for the MICA comparisons above. These appeared to show that full text models outperformed manual curation in predicting diagnosis. Term weighting in disease models appeared to prioritise the main clinical features of a condition, as discussed in section 2.2.7 it was hypothesised that adding term weightings would increase the performance of full text-derived disease models, as defined by AUC.

To test this hypothesis, a comparison of weighted models using the Orphanet (28) subset was undertaken. This was because Orphanet models consistently had term weighting applied whereas those from OMIM did not (28,29). However, there was no straightforward method found for normalizing the frequency models between datasets. This was due to the different methods of recording frequency between datasets for literature and DDD vs OMIM and Orphanet; and the skewed distribution of frequency annotations in the literature set.

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 to 99% was mapped to 89. Each term in a disease list was repeated according to the number of its frequency mapping value, thereby creating a model with weighting suitable for use in the MICA equation.

For literature-derived and DDD models (6), frequency annotations were split into four bins using the numpy.histogram module (157), corresponding to the HPO terms ‘Very frequent’ (HP:0040281), ‘Frequent’ (HP:0040282), ‘Occasional’ (HP:0040283) and ‘Very rare’ (HP:0040284) . Numerical frequency weighting according to these categories was then applied as per the Orphanet models. For models weighted in this manner, *l* row terms and *k* column terms in disease comparison matrix *m* (as discussed in section 3.5.2) therefore may contain repeats. This accordingly alters the MICA sum average (105). Comparisons using this weighted data were calculated for the

43 diseases in the Orphanet set vs equivalent DDD models, and for the corresponding 43 literature-derived models vs DDD.



Comparator	AUC
Full text vs DDD weighted	0.694
OMIM vs DDD weighted	0.739
Full text vs DDD unweighted	0.788
OMIM vs DDD unweighted	0.802

Figure 3-19. Precision curve using threshold ranking for full text-derived/Orphanet disease models compared to real life terms in DDD study, across sample of 41 diseases. Orphanet models are weighted by HPO term frequency e.g. HP:0040281 (Very Frequent 99-80%). The full text and DDD model weightings were binned according to the mean of each HPO term frequency range.

ROC curves for full text/Orphanet vs DDD when unweighted were very similar, indicating there was similarity between the two literature-based sets (Figure 3-19). Adding weighting actually decreased the area under the ROC curve (Figure 3-19). This may be because adding weighting to HPO terms does not increase the clinical disease similarity of a model. However, given the results in section 2.2.7, where top ranked terms reflected the most important features of the conditions analysed, this seems less likely.

Of note, the binning normalization method used here likely resulted in less accurate weighting for the full text-derived models. It is possible that a different method of normalization, if developed in future, would result in an increased AUC for the comparisons shown here. This would particularly be true if the granularity of full text model weightings could be retained.

3.6 Comparison of diseases in same biological pathway

It was hypothesised that literature-derived disease models could also be used to identify diseases relating to genes in the same biological pathway. An exploratory analysis was carried out to test this. Three groups of GDD were selected, each of which was associated with similar phenotypes in multiple genes due to perturbation of the same underlying biological pathway. For each group, three genes were selected and a literature review performed using {gene symbol}[TI] PubMed search. Full text-derived disease models and an RBO heatmap were constructed as described in sections 2.2.3, 2.2.4 and 3.4.1.

The groups (and genes) chosen were ciliopathies (*BBS1, BBS10, AHI1*), RASopathies (*HRAS*, *BRAF*, *SOS1*), and Cornelia de Lange syndrome (*SMC1A*, *HDAC8*, *NIPBL*). An RBO heatmap (Figure 3-20) showed increased intra-group similarity, particularly for Cornelia de Lange, with no significant inter-group relationship. If proved at scale, this phenotype clustering using disease models could be directly applied to diagnostic bioinformatic variant filtering in the context of GDD. If phenotype matching could narrow down analysis to a group of genes in the same pathway, clinical resources could be focused earlier and in more detail on variants in these. This would improve speed and accuracy of diagnosis when filtering genomic data.

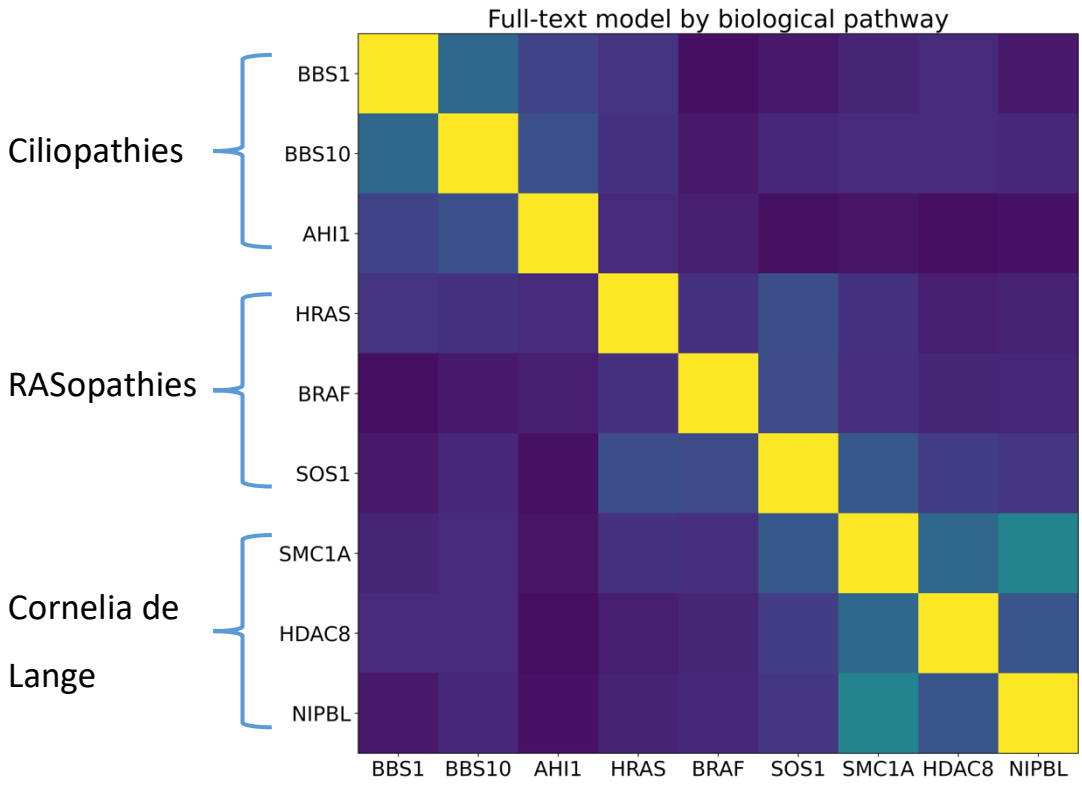


Figure 3-20. Rank biased overlap heatmap for full text-derived models describing genes from three well-defined groups of GDD. Mutations in genes within each group affect similar biological pathways, resulting in similar disease phenotypes.

3.7 Discussion

This chapter focused on the creation and analysis of an example set of 99 literature-derived disease models. These were mapped to phenotypes in prospectively derived clinical data from the DDD study (3) as well as to manually curated diseases in OMIM and Orphanet (28,29). Comparative analysis was carried out using the RBO and MICA-based similarity metrics (89,98,100).

3.7.1 Scaling disease model creation using full text downloads

The set of 99 diseases was created with aim of showing proof-of-concept for the full-text mining phenotype disease model method presented here, over a larger sample size. This could then provide evidence for the utility of scaling up these models to cover the full spectrum of GDD.

Cadmus allows access to downloadable full text of almost all the biomedical literature (43). This proved to have high performance on the set of 1018 papers identified, with a 94.5% successful download rate. This package should prove useful not only to scaling up disease models for GDD, but to any researchers working on the applications of text mining. As discussed previously, the majority of the biomedical peer-reviewed literature is not open access (41). The full text manuscripts made accessible by Cadmus represents a huge hitherto untapped resource.

The potential of using full text was demonstrated by the mean number of HPO terms per paper being 68 using the whole manuscript, versus nine when using the title + abstract alone. This increased depth of data extraction may result in better coverage of the full phenotypic spectrum of a disorder. These results support the hypothesis that using full text in the context of GDD allows access to knowledge which is not present in the abstract alone. This echoes the findings of Westergaard et al. (40), where full text mining outperformed abstract only across multiple scientific domains.

The increased number of terms in literature-derived models (mean 198) compared to OMIM (mean 38) may also describe greater coverage of disease-specific phenotypic spectra. This may represent an increased proportion of terms per paper being extracted using text mining on the full manuscript. It may also be related to the generally larger number of manuscripts used to construct the literature-derived models compared to

OMIM. In the version of OMIM annotations used in this work, there were 5574 diseases with a mean term annotation of 17 and median of 13 (29). Scaling up the literature-derived disease model concept could result in a similar broad coverage of disease to OMIM, but with greater depth of phenotyping.

3.7.2 Structured vocabulary in the GDD domain

The disease test set created here allowed for an analysis of the vocabulary used in the GDD domain. For these 99 conditions, a total of 3499 terms were used across all three comparison datasets. This compares with 15985 terms in the full HPO (version 1.2) (22). This disease set only covers a fraction of the thousands of GDD in DDG2P (31). However, these results indicate there may be a GDD domain-specific vocabulary which uses only part of the HPO. This is supported by the use of only 4158 unique HPO terms used for all DDD probands (unpublished data), compared to 15985 terms in the version of HPO used in this work. This result may inform the development of a GDD-specific sub-ontology in future. This could improve the performance of, for example, MICA-based semantic similarity.

3.7.3 Modification of disease model construction

It was hypothesised that systematic modifications such as removing terms with a weighting of one, and selecting terms with a high MetaMap score, would improve performance as assessed by disease matching. However, this proved not to be the case, as shown in section 3.4. Therefore, the full models were used in this work.

Of note the performance measure used, where increased similarity to DDD equalled improved disease models, could select against novel phenotype identification from full text. However, in the absence of a definitive truth set, it was thought likely that comparison to real clinical data should be a helpful

measure of whether disease models reflect true disease expressivity. It remains possible that adjustments to disease model construction could increase performance in future. This could include, for example, upweighting of terms which are highly discriminant between disorders.

3.7.4 Utility of HPO structure

The disease comparisons shown here not only act as a functional assessment of literature-derived models, but also allow for examination of the differences between list-based and ontology-based similarity metrics. The ROC curves generated using RBO (89) and MICA (98,100) were remarkably similar, as demonstrated in Figure 3-18, with the caveat that RBO used term weightings for ranking, whereas the MICA models were unweighted. This could indicate that the HPO-defined term relationships are not providing useful information for the purposes of disease similarity matching. For example, an interesting analysis in future would be to see how many of the MICA terms in this study were high-level, and therefore less specific, in the ontology.

Use of a modified IC measure, such as that of Schlicker et al. (93), which accounts for the specificity of the MICA, may demonstrate improved performance using semantic similarity than RBO. Additionally, the IC measure (defined by Resnik (100)) is influenced by the corpus size used to generate it. The IC will stabilise as corpus size increases (99). Therefore, MICA-based semantic similarity may improve in disease matching if it is used in future on a scaled up set of thousands of GDD, such as the full DDG2P set (31).

3.7.5 Comparative disease model similarity

It was hypothesised that the literature-derived models described here could be as, or more, descriptive of true disease expressivity than those generated

by manual biocurators. As it was not straightforward to assess this through manual review of hundreds of terms per model, comparative similarity metrics were used. There was a noticeable similarity between corresponding literature-derived and OMIM models on the RBO heatmap in Figure 3-14B. There was significant overlap of exact match terms between these models (Figure 3-13). These results were encouraging, as they indicated that a computational, scalable method of literature mining is capable of extracting phenotype data which is at least comparable to that generated through resource-intensive manual curation.

However, this did not provide evidence for whether the literature-derived models were similar to disease expressivity in a clinical setting, and how they would compare to manual curation in this context. Similarity analysis was therefore carried out in comparison to DDD data. This showed a recognizable RBO heatmap signal for literature vs DDD models, which was similar to that seen for OMIM vs DDD (Figure 3-14). The signal for OMIM vs DDD appeared weaker than literature vs DDD using the semantic similarity metric (Figure 3-16). Overall, this provided evidence that literature-derived models were similar to prospectively gathered clinical data.

It is possible that the similarity seen between sets was related to the larger number of terms in the literature-derived models, as shown in Figure 3-3B. There is evidence that highly-annotated entities result in increased semantic similarity score (102). However, as disease gene prediction was based on similarity scores across the 99 GDD set, this effect should have applied across all models. There is evidence that using empirically generated p-values for similarity scores may increase their predictive power and correct for annotation number (96,98). Therefore, p-values could be beneficial if applied to these disease models in future. Nevertheless, it is also possible the weaker signal seen for OMIM vs DDD using semantic similarity was because the ontology-based metric was more effective at showing that these

models were less reflective of clinical expressivity than those derived from text mining.

3.7.6 Disease model prediction by threshold ranking

Following on from the results above, it was hypothesised that literature-derived models may be used for prioritising disease genes from the DDD clinical set when using ranking thresholds. ROC curves were constructed to assess this. The AUC was 0.85 for both the RBO and MICA-based comparisons, indicating that this approach was effective (Figure 3-18). Furthermore, the AUC was greater than for the OMIM (29) models, implying that the literature-based method outperformed manual curation. The (unweighted) AUC was similar when compared to Orphanet (28) models (Figure 3-19), however it is difficult to extrapolate whether these are comparable to OMIM, as the analysis was on a subset of 41 models.

These results should be directly applicable when considering integration of phenotype disease models into bioinformatic pipelines for genome-wide sequencing. All probands for this whom this type of test is requested should have a list of HPO terms recorded by the referring clinician. This list could be compared with a set of disease models, with a list of top-ranked models given as likely diagnoses according to similarity score. The ROC curve could be used to assess the sensitivity and specificity thresholds for this test. This process would require two steps for validation: scaling up disease models to cover all of DDG2P, and testing against individual level DDD data – aggregated phenotypes were used here.

3.7.7 Term weighting and normalization

It was hypothesised that adding term weighting to the MICA-based semantic similarity comparison should improve performance, given that this appeared to reflect disease expressivity, as shown in section 2.2.7. Normalization of

term weightings in different formats across datasets, to enable comparison, was not straightforward, as shown in section 3.3.3. Indeed, a method where weightings were put into bins corresponding to HPO frequency terms resulted in reduced AUC for both literature-derived and Orphanet-derived models (Figure 3-19). It was not clear whether this result was due to the normalization method used, or if term weighting truly does decrease disease model matching efficacy.

Nevertheless, the results here demonstrate a need for improved term weighting in future work. This may be aided by relatively simple measures such as a per-paper weighting, where a term is only counted once per manuscript. On the other hand, this problem would be solved if term extraction was linked to individuals described in case reports/case series. This would mean weighting could be easily normalized, expressed as a percentage for conversion to HPO frequency terms, and used for meaningful comparison between models.

3.7.8 Conclusions

In this chapter, I have demonstrated the construction of literature-derived disease models on a larger scale. This enabled testing of these models against clinically-derived and manual biocuration-derived datasets. Using several similarity metrics, I showed that the literature-derived models reflect disease phenotypes as defined by manual curation in OMIM and Orphanet (28,29). I also demonstrated that the full text models outperform OMIM when predicting diagnostic genes in DDD data.

One of the strengths of this work is that the disease models are based on high-quality manuscripts, i.e. those which have been assessed as containing detailed case reports describing molecularly confirmed GDD. This should mean that the models better reflect true disease expressivity compared to automated manuscript-disease linking. For example, Pilehvar et al. used the

co-occurrence of a disease and phenotype in abstracts to create disease models (116). This likely includes phenotypes which are not clearly linked to a defined molecular disease mechanism.

However, the careful selection of case reports/case series also acts as a significant bottleneck to scaling up disease model creation, as it is not feasible to perform a manual literature review for thousands of genes. In the next chapter, I present an exploratory analysis of methods to automate literature searching, including the development of a machine learning classifier for this purpose. I also evaluate the data architecture of published manuscripts, to inform parsing of clinical data in future.

Chapter 4 Parsing and automatic search of the peer-reviewed literature

4.1 Introduction

In the previous two chapters, I demonstrated the construction and testing of disease models using the full text biomedical literature. Here, I present preliminary analyses which will inform development of a text mining pipeline in future. This is designed to eventually feed into the development of an automated curation system for DDG2P, for use in clinical variant interpretation. First, I evaluate the data architecture of the peer-reviewed literature, to inform improved parsing of clinical information from text. Second, I demonstrate the development of a machine learning abstract classifier, for computational GDD-relevant manuscript identification. I then discuss the next steps to develop automated literature curation, following on from this work.

4.2 Structure of manuscripts describing GDD in the peer-reviewed literature

4.2.1 Background

Phenotypic features extracted from full text manuscripts in the peer-reviewed literature may not always be relevant to the disorder/gene being reported. In this context, non-relevant features may include those generally describing a disorder, which may also be a different condition to that forming the focus of a case report. Relevant features are those tightly linked to an individual patient, and ideally to a specific SNV or CNV, in a case report/case series.

Therefore, it was hypothesised that parsing out descriptions of only the individuals being reported in a manuscript would allow for more precise phenotypic association to specific patients, and thereby increase the clinical accuracy of literature-disease models.

To inform this, an analysis of the underlying structure of published case series/case reports describing GDD was undertaken. This was aimed at identifying features which could enable more accurate full text parsing in future. This strategy should represent an improvement on the disease models described previously in this work. These models utilised the full text as parsed using Cadmus (43). This removes the abstract and references only.

4.2.2 Annotation of single journal output

Manuscripts from a single journal covering a full calendar year were annotated. The AJHG (44) was chosen, as newly-described GDD are well represented in the journal. All AJHG abstracts from 2017-2018 were reviewed. Of these, papers describing childhood or earlier onset genetically determined disease were selected. Full-text manuscripts were then annotated using the published version online, using Hypothesis (158), a free, open source annotation tool. All supplemental files were reviewed, and those which contained free text case reports were annotated instead of the full text manuscript.

Annotation was of text spans corresponding to phenotypic features. Compound terms e.g. ‘there was patellar and radial hypoplasia’ were separated into individual annotations. Hypothesis allows for tags to be added to each annotated text span. Each phenotypic feature was tagged as relevant or not relevant to the individuals newly described in the paper. The type of manuscript (‘Article’ or ‘Manuscript’) was also recorded per document to determine how this might affect parsing out clinical information. ‘Articles’

have clearly defined results sections which are relatively straightforward to parse out; ‘Reports’ are a single block of text and it is less simple to parse these.

4.2.3 Structure and relevance of phenotypic data in the literature

Annotation of the AJHG corpus (Table 4-1) showed the minority of phenotypic data was in Articles, with most contained in Reports or Supplemental files. Articles and Reports contained a significant proportion of non-relevant phenotypic data (38% and 32% respectively). Non-relevant phenotypic descriptors included, for example, those relating to previously reported individuals in the literature with the same condition, or describing diseases caused by variants in related genes.

Manuscript type	Manuscript count	Relevant/ total number of phenotype annotations	Relevant annotations per document (median)
Article	25	1277/2052 (62%)	67
Report	43	1796/2623 (68%)	57
Supp.	33	4726/5027 (94%)	124
Total	101	7799/9702 (80%)	67

Table 4-1. Distribution of Human Phenotype Ontology terms in AJHG manuscripts from year 2017-2018, describing childhood-onset genetic disorders. Relevant terms are those directly describing an individual or group of individuals with a genetically-determined condition forming part or whole of the research focus of the manuscript, as determined by my manual annotation. Non-relevant terms may include references to previously described individuals with the same or similar conditions. Supp – supplemental. AJHG – American Journal of Human Genetics.

Supplemental files had a much higher proportion of relevant data (94%), largely because these contained detailed individual level case reports. This demonstrates the value of parsing out relevant clinical reports/results from a manuscript, or using supplemental case reports where available, as the phenotypic data obtained is likely to be more relevant to the paper being analysed.

4.2.4 Strategies for manuscript parsing using data architecture

The results from this analysis of the data architecture of GDD-relevant manuscripts can be used for improved parsing of the full text in future. It should be straightforward to parse out clinical data from Articles. This is especially the case if they are in a structured document created using a markup language, such as HTML or XML. Here, the relevant sections would be contained within sections headed as, for example, ‘Results’ or ‘Case reports’. Regex patterns for string matching section titles could be used if the download is in a PDF format.

Articles represented the minority (25%) of manuscripts here, however. If this pattern is replicated across the wider literature, it will necessitate a strategy for parsing reports. This would also apply to the commonly used format ‘Letter to the Editor’. Reports are generally a single block of text, with no obvious handles for parsing. Extracting clinical information only from these will require the development of more advanced techniques. If it becomes possible to extract phenotype data per proband, rather than per manuscript, this could be applied to reports.

Supplemental files, where present, should be utilised, given their enrichment for phenotypic data as shown in Table 4-1. The text for these should be relatively simple to parse, as they are often in the form of structured case series. However, the challenge here is identifying which files contain

phenotypic data. Most manuscripts have supplementary data included in a variety of formats, for example PDFs and spreadsheets. Many of these files do not contain phenotypic information. Therefore, a method needs to be developed to automatically identify the correct files for processing.

4.2.5 Using data architecture to inform NER

During the annotation process described above, each phenotypic feature was recorded as a text span only. This means these were not manually mapped to HPO terms (22). Accurate biomedical NER is challenging, due to the often complex manner in which phenotypic features are recorded in text. This is discussed in section 1.5. Exact matches to HPO terms in text represent only a proportion of available phenotype data, as demonstrated in section 2.2.5. To analyse this issue further, the fraction of annotated text spans which could be mapped to HPO in the AJHG corpus, using simple string matching, was assessed. This aimed to give a more comprehensive overview of the need for more advanced NER methods in the GDD domain. This should inform the development of NER in this field in future.

4.2.6 Mapping HPO to AJHG corpus

To determine the proportion of annotations which could be mapped using simple methods, exact string matching and fuzzy matching using cosine similarity were applied to the list of phenotypic text spans. Exact string matching was using FlashText (142) as described in section 2.2.4.

Cosine similarity was applied following a customised method utilising Scikit-learn version 1.0.1 (159,160). In brief, for each paper, the annotated text spans were converted to three-grams, which consist of all possible three character combinations. All text spans were then converted to a sparse TF-IDF (term frequency-inverse document frequency) matrix using Scikit-learn

TfidfVectorizer (160). In this case, TF-IDF multiplies the frequency of an n-gram in a text span by its inverse frequency per document in all text spans per paper. This therefore upweights less common n-grams per text span. The same process was applied to all HPO terms and synonyms (22). For each text span, the TF-IDF vector was compared with all HPO vectors using cosine similarity. This is a measure of the cosine of the angle between two vectors. The top cosine similarity value for each comparison was returned for each text span. After manual review, a value of >= 0.8 was chosen as the cosine similarity threshold for matching. Matched HPO terms below this did not accurately match the annotated text.

4.2.7 Exact/fuzzy string matching in AJHG corpus

The proportion of terms annotated as phenotypic features which a) mapped exactly to HPO b) mapped closely to HPO using a fuzzy matching score and c) did not map easily, indicating a need for more sophisticated NER methods such as MetaMap (64), were assessed.

Manuscript type	Total annotations with exact match to HPO	Total annotations with match to HPO above cosine similarity threshold 0.8
Article	680/2052 (33%)	1216/2052 (59%)
Report	854/2623 (33%)	1659/2623 (63%)
Supplemental	3172/5027 (63%)	3214/5027 (64%)
Total	4706/9702 (49%)	6089/9702 (63%)

Table 4-2. Proportion of text spans in annotated AJHG corpus which map to HPO terms, using exact and fuzzy matching. HPO – Human Phenotype Ontology, AJHG – American Journal of Human Genetics.

Annotations which exactly matched HPO were present in 49% of terms overall (Table 4-2). Close HPO matches to annotated text, as defined by a cosine similarity score of >=0.8 were found in 63% of cases (Table 4-2). This shows that a significant proportion of phenotypic features in a manuscript can be identified using relatively simple techniques. However, this also confirms the necessity for using more advanced text extraction methods such as MetaMap (64) for a significant performance advantage, as described in section 2.2.5.

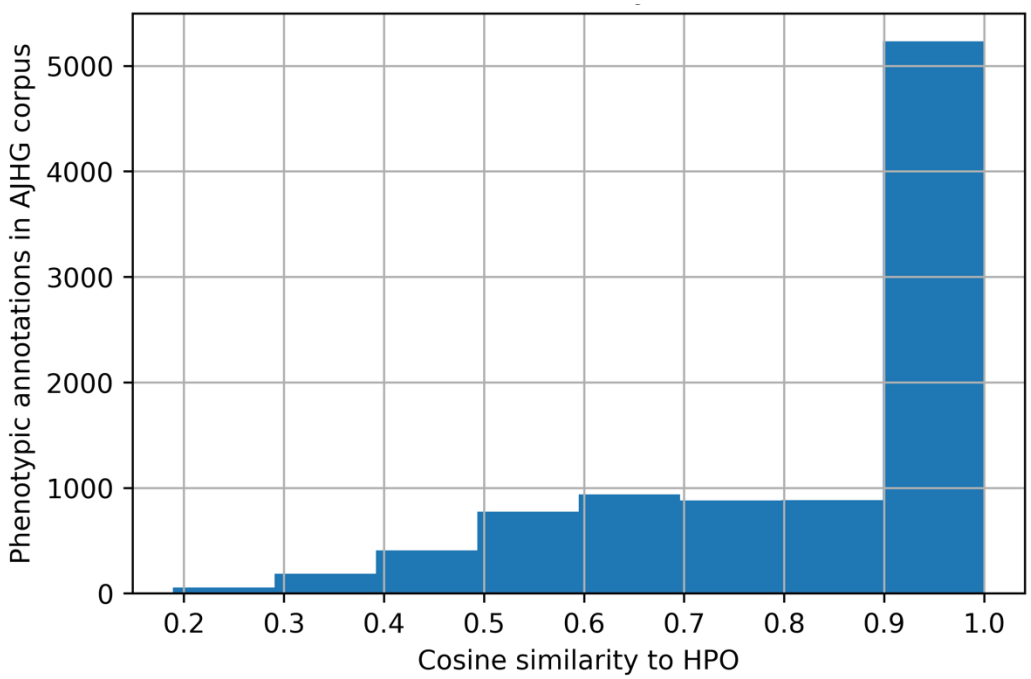


Figure 4-1. Distribution of fuzzy matched annotated phenotypic descriptors across AJHG corpus. Annotations were matched to HPO terms using cosine similarity. A higher cosine similarity indicates a closer match to an HPO term. A cosine similarity of 1.0 is an exact match. AJHG – American Journal of Human Genetics, HPO – Human Phenotype Ontology.

4.2.8 Linguistic analysis of phenotypic descriptors

Cosine similarity matching in the AJHG corpus allowed for assessment of the linguistic structure of annotated text. Table 4-3 demonstrates examples of

high-scoring cosine matches to HPO. Word order changes compared to canonical terms prevent identification of the corresponding term ('language delayed', 'delayed language'). However, these are relatively simple to capture using fuzzy matching. Compound phrases such as 'Small, low set ears' match to only one HPO term 'Low set ears' (HP:0000369) using this technique. More advanced NER may be expected to pick up the other: 'Microtia' (HP:0008551) (synonym is 'Small ears').

HPO term	Matched phenotype annotation	Cosine similarity
Wide nasal bridge	Wide nasal bridge and ridge	0.854010
Language delayed	Delayed language	0.853888
Neonatal asphyxia	Perinatal asphyxia	0.853581
Low set ears	Small, low set ears	0.853545

Table 4-3. Examples of text in AJHG corpus annotated as phenotypic features, mapped to HPO terms using cosine similarity. AJHG – American Journal of Human Genetics.

HPO term	Matched phenotype annotation	Cosine similarity
Purpura	Non-purposeful use of hands	0.300348
Thin skin	Skin is remarkably thin	0.321925
Myopathic face	EMG had some features of a myopathic process	0.350894
Cat-like cry	Rett-like hand automatisms	0.362302
Small cerebellum	Cerebellum was of slightly reduced size	0.368094
Ischemic stroke	Perinatal hypoxic-ischemic event	0.371900

Table 4-4. Selected terms in AJHG corpus with cosine similarity <0.4. AJHG – American Journal of Human Genetics.

Perhaps more relevant to the development of NER are the difficult-to-match examples (Table 4-4). These include the use of subjective modifiers in between phenotype descriptors ('remarkably', 'some features of', 'slightly reduced'). 'Non-purposeful use of hands' requires the linking of an object ('hands') with a dysfunction ('non-purposeful'). This may be straightforward for the reader, but is not easy to assess computationally. There is also inferred meaning in the phrase 'Rett-like hand automatisms'. This carries the implicit assumption that the reader is familiar with the phenotypic features of Rett syndrome. These may include hand stereotypies such as hand wringing, clapping, and washing automatisms (161). Additionally, there is no HPO term for 'Hand stereotypies'. Therefore, there is a significant challenge for an NER method. First, it would need to recognise that 'Rett-like hand automatisms' maps to the phenotypic features of a particular GDD. Then it needs to identify a matching HPO term from this. Possible maps include 'Stereotypy' (HP:0000733) or 'Stereotypical hand wringing' (HP:0012171).

4.2.9 Conclusions

The analysis of one year of AJHG manuscripts here represents a comprehensive overview of GDD-relevant literature. The results demonstrate a number of features which should be useful in developing automated literature curation in future. I show that parsing out clinical data from full text should improve the accuracy of phenotype extraction. This is because there are a significant proportion of phenotypic descriptors which are not relevant to newly-described probands in each manuscript. 'Articles' are more straightforward to parse for this purpose than 'Reports'.

The use of a corpus solely from one journal may have introduced some biases. It is possible that other journals may record phenotypic information differently. The manuscripts analysed from AJHG largely pertained to rare, newly described disorders. These papers may include detailed literature review and comparison to other previously published reports. Manuscripts

describing rare complications of well-known disorders in other journals may not include phenotypic descriptors from previous work. It is also possible that manuscripts from other journals may be structured differently, although the single paragraph ‘Report’ or ‘Letter to the Editor’ format is common across many publishers. The same is true for manuscripts with structured headings such as ‘Introduction’, ‘Results’, ‘Discussion’.

It is clear that supplementary case reports should be used, where available, in place of the full text manuscript. The supplementary data is enriched for relevant phenotype descriptors. Additionally, I further demonstrate the need for advanced NER methods as simple string matching misses a significant proportion of terms in this corpus. I gave some examples of edge cases, or phenotypic annotations which do not easily map to HPO. It would be interesting to test the performance of the most up-to-date NER techniques on this corpus, to see if these are correctly analysed.

4.3 Scaling up manuscript identification

4.3.1 Background

The disease models described in earlier chapters were generated using hand-selected manuscripts containing case series/case reports describing GDD. This was to test phenotypic extraction without the complicating factor of non-disease-relevant input. The importance of this was demonstrated in section 2.2.8. This showed that individual input papers could significantly alter the clinical relevance of weighted disease models. However, to scale up the disease model method to cover all 2000+ GDD, accurate automated identification of case series/case reports would be required. Here, I present preliminary work towards this goal.

I assessed different PubMed search strategies. From these, a custom annotated corpus was created, labelling manuscripts (rather than entities within them as previous) as relevant or not to GDD. Features derived from this were used to assess supervised learning abstract classifiers.

4.3.2 Searching PubMed at scale

Automating literature curation requires an accurate, scalable method of searching for peer-reviewed manuscripts. PubMed contains over 33 million citations (23), only a small proportion of which will describe GDD. Therefore, a robust search strategy and classification method is required to correctly identify relevant papers.

In my experience manually curating DDG2P, a gene symbol search in PubMed is often effective in identifying case series and case reports describing GDD. Permutations of this search with different filters, using all 2164 genes in DDG2P were performed. These used modified Python scripts kindly provided by Jamie Campbell (162). In brief, Biopython (version 1.79) (163) Entrez search (23) returned a list of PMID and metadata for a given string search term.

The input used was every gene symbol in the DDG2P DD gene-disease pairs and attributes file downloaded from www.ebi.ac.uk/gene2phenotype/downloads on 29/4/21. Modifier tags were added to the {gene symbol} search to determine how these affected the number of PMID returned per gene. These were: search in title only ({gene symbol}[TI]), search in title + abstract only ({gene symbol}[TIAB]), and search in title + abstract only, where ‘Gene’ appears in the MeSH headings (23) (Gene[MESH] {gene symbol}[TIAB]).

4.3.3 Identification of effective PubMed search strategy

The results of these searches are shown in
Table 4-5. Using {gene symbol} alone to search resulted in almost 14 million results, which is a significant proportion of PubMed overall, and is unlikely to be enriched for GDD-relevant papers. The number of results was highly influenced by the small number of genes which returned a large number of citations (**Error! Reference source not found.**). Adding filters reduced the number of results, at the cost of more genes having no citations returned. Using {gene symbol}[TI] (symbol in the title only) achieved the best balance between generating a parsable, manageable number of results – approximately 400,000 for 2164 genes in this case– with searches which do not identify any papers at all for a gene – 30/2164 for the title search. These missed genes would require a different search strategy, for example using the associated disease name could be considered.

Search term	**Total results**	**Genes with zero results**
{gene symbol}	13,872,488	8
{gene symbol}[TI]	411,783	30
{gene symbol}[TIAB]	2,342,061	9
Gene[MESH] {gene symbol}[TIAB]	136,370	100

Table 4-5. PubMed results for all (2164) genes in DDG2P database, using gene symbol + differing filters. DDG2P – Developmental Disorders Gene2Phenotype, TI – search in title, TIAB – search in title + abstract, Gene[MeSH] – include only results with Gene in MeSH metadata.

4.3.4 Filtering gene searches with high results

The distribution of results per gene for the {gene symbol}[TI] search showed significant right skewing (**Error! Reference source not found.**).Therefore, only a minority of genes returned individually large numbers of citations. This should mean parsing out relevant papers for the majority of genes with lower numbers of results is more straightforward.

Table 4-6 shows the top five genes by number of results. These include gene symbols which are also words (*PIGS, SET*), chemical symbols (*CA2* for calcium, usually $Ca2^{+}$), highly studied molecules (*MTOR,* mTOR pathway) or acronyms (*NHS* or National Health Service). This helps explain why these genes return high numbers of results.

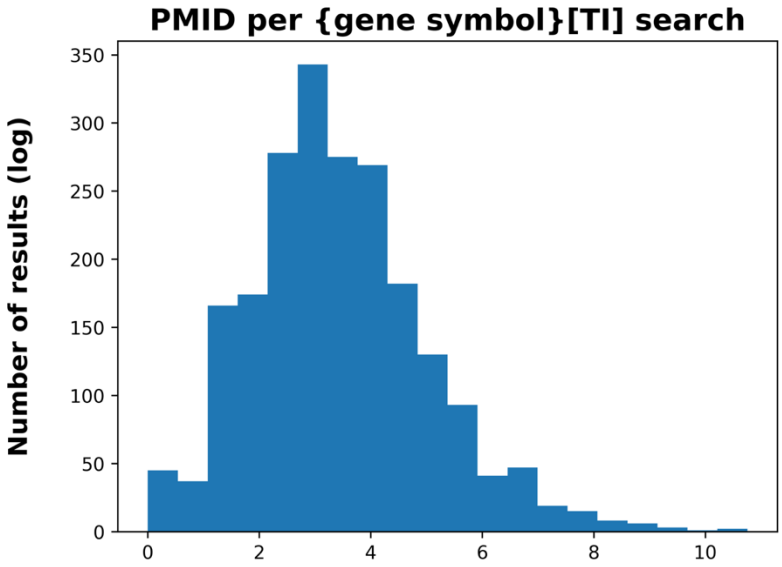


Figure 4-2. Distribution of results per {gene symbol}[TI] search for 2164 genes in DDG2P. PMID – PubMed ID, DDG2P – Developmental Disorders Gene2Phenotype. Log transform used for number of results per gene.

Further specific search strategies may be useful in these instances. For example, for the *NHS* gene, a search for *NHS*[TI] AND “gene” cut the number of results from 10270 to 41, a number of which were case reports/case series describing the GDD Nance-Horan syndrome. Using the disease name could also be effective – a search for *NHS*[TI] AND “Nance-Horan syndrome” returned 24 results, of which all but three were relevant case reports/case

series. Similarly, the mTOR pathway is highly studied, particularly in the context of cancer. A search for *MTOR*[TI] NOT "Cancer" reduced the results to 6867 compared to 11678 using *MTOR*[TI] alone. This particular technique has the issue that it may exclude GDD reports which do mention cancer as a feature of a given disorder. Overall these examples illustrate that an extra step to focus searching for genes with high numbers of results may be helpful when applied to a classification method.

Gene symbol	Results from {gene symbol}[TI] search
PIGS	46901
CA2	40345
SET	21054
MTOR	11678
NHS	10270

Table 4-6. Top five genes by number of results returned from {gene symbol}[TI] search in 2164 gene DDG2P set. DDG2P – Developmental Disorders Gene2Phenotype.

4.3.5 Annotation of papers to test classification strategies

Following the identification of {gene symbol}[TI] as an effective PubMed search strategy, this was used to create a corpus of annotated citations. This was to enable testing of automated classification strategies in identifying papers for text extraction. First, the top diseases in DDG2P by number of diagnoses in DDD were reviewed, and entries selected to represent more common GDD, as well as to include a variety of allelic requirements (monoallelic, biallelic, hemizygous and X-linked dominant). Genes with more than two associated diseases were excluded.

From this list of diseases, the PMIDs from a PubMed {gene symbol}[TI] search, as described in section 4.3.2, were used to generate a corpus of papers for annotation. If a {gene symbol}[TI] search returned more than 100 results, only the first 100 were included for annotation, otherwise all results were used. Each PMID was mapped to a corresponding DDG2P entry according to gene symbol.

A Python script, kindly provided by Jamie Campbell, was modified and used to analyse the metadata for each PMID in the annotation set (164). The title and abstract were reviewed and assigned to one of three groups: 1 – Relevant paper containing case reports/case series describing the correct DDG2P entry, without including any other disease phenotypes. 2 – Relevant to the corresponding DDG2P entry, but not suitable for text extraction, e.g. included more than one gene/disease, included allelic disorders, included phenotypic data for CNV, or reviewed the condition rather than providing novel case reports. 0 – Not relevant to the corresponding DDG2P entry. Where an abstract was not available, the full text was reviewed by performing a manual PubMed search.

4.3.6 Annotated PubMed citations corpus

125 well-defined diseases in DDG2P were identified, including 109 genes. These formed the basis of a {gene symbol}[TI] search in PubMed (Table 4-7). 38 diseases returned more than 100 results. The search returned 6578 papers overall. The results were reviewed and annotated to create a gold standard corpus of relevant (to GDD) and non-relevant papers.

Diseases in set	Genes in set	Allelic requirement			
		Monoallelic	Biallelic	X-linked dominant	Hemizygous
125	109	87	26	7	5

Table 4-7. Summary of DDG2P entries used to test classification and PubMed search strategy. DDG2P – Developmental Disorders Gene2Phenotype.

Total annotated papers	Relevant papers	Genetic disease manuscripts not suitable for text extraction	Non-relevant papers
6578	1536/6578 (23%)	472/6578 (7%)	4570/6578 (69%)

Table 4-8. Annotations for papers returned from {gene symbol}[TI] PubMed search for genes in 125 disease test set. Papers marked as relevant if contained case report/case series describing corresponding DDG2P entry only. Non-relevant papers were subset into those which described a human genetic disease which did not correspond to a given DDG2P entry or contained reports of more than one disease; and those which did not contain any human disease phenotype data. DDG2P – Developmental Disorders Gene2Phenotype.

In the annotated set, 23% were deemed relevant, and 69% non-relevant (Table 4-8). A further 7% of papers were not suitable for text extraction, but did contain phenotypic data relevant to the DDG2P disease entry, for example those including case reports for more than one disorder. Examples of each of the three classification categories are given in Table 4-9.

Gene	DDG2P disease Name	PMID	Title	Classification
KIF1A	NESCAV Syndrome	25253658	*KIF1A* mutation in a patient with progressive neurodegeneration	1
KIF1A	NESCAV Syndrome	28332297	Hereditary spastic paraplegia caused by compound heterozygous mutations outside the motor domain of the *KIF1A* gene	2
KIF1A	NESCAV Syndrome	27484852	*KIF1A* mediates axonal transport of *BACE1* and identification of independently moving cargoes in living SCG neurons	0

Table 4-9. Example manual classifications of citations returned for *KIF1A*[TI] PubMed search. Results classified after review of title + abstract (not shown). Classification categories: 1 – Relevant paper containing case reports/case series describing the correct DDG2P entry, without including any other disease phenotypes. 2 – Relevant to the corresponding DDG2P entry, but not suitable for text extraction, e.g. included more than one gene/disease, included allelic disorders, included phenotypic data for CNV, or reviewed the condition rather than providing novel case reports. 0 – Not relevant to the corresponding DDG2P entry.

4.3.7 Feature selection in GDD manuscripts

The process of annotating relevant manuscripts described above allowed identification of groups of papers sharing a common theme/subject, to inform the development of features for machine learning classification. In this context, features are computationally tractable parts of the title, abstract or metadata thought to be descriptive of the manuscript's relevance to GDD.

Themes identified included : corrections/erratum, generation of induced pluripotent stem cells, GWAS (genome-wide association studies), cancer studies and mouse models. Therefore, these represent papers which are not relevant to GDD text extraction. Additionally, MeSH terms thought to be applicable to GDD were identified, for example 'Phenotype', 'Intellectual disability', 'X-linked', 'Exome'. The PubMed metadata was also utilised, for example where 'Case report' was included in 'Publication type'. Finally, a number of features thought to be specific to GDD case series/case reports and accurate text mining were identified. For example, where only one gene was mentioned in the title, and when the disease name (exact or fuzzy match) was present in the title.

33 classification features were designed to capture these, using regex or exact string matching for key phrases in the title, abstract or MeSH terms. These phrases were developed both from analysis of the annotated corpus and from background clinical knowledge. Examples of these features included:

1. 'Mouse' OR 'Mice' NOT IN title
2. 'Pluripotent stem cell line' OR 'iPSC' NOT IN title
3. 'Leukaemia' OR 'leukemia' OR 'carcinoma' OR 'lymphoma' NOT IN abstract
4. {Disease name} IN title
5. {Fuzzy match to disease name} IN title
6. 'Exome' IN MeSH terms

7. ‘Genes’ AND ‘Dominant’ OR ‘Heterozygote in MeSH terms

As above, the features which related to manuscripts which should be excluded, e.g. mouse models, were coded as NOT IN. Features relevant to GDD papers, e.g. including ‘Exome’ were coded as IN. These features were applied computationally to the whole annotated corpus and coded as 1 (present) or 0 (not present).

The performance of these features individually was assessed using positive and negative predictive values (PPV/NPV). Where a manuscript was manually annotated as relevant, and a feature was positive (coded as 1), this corresponded to a true positive (TP) result, or a false negative (FN) where the feature was coded as 0. True negatives (TN) corresponded to manually annotated non-relevant manuscripts with negative (coded as 0) features, and false positives (FP) where the feature was positive (coded as 1). PPV and NPV were then calculated according to the following equations:

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Table 4-10 demonstrates the performance of the top 15 classifiers, sorted by NPV. Most of the top NPV classifiers excluded only small numbers of manuscripts. The most successful features in excluding larger numbers of papers (>1000) with a high NPV were for mouse-related terms in the abstract, and animal study-related MeSH terms. A feature describing papers with more than one gene in the title (which cannot be used for text extraction currently using the methodology previously described) also excluded >1000 papers.

In contrast, the top classifiers by PPV (
Table 4-11) mostly included clinically-relevant MeSH terms. The PPV values were generally lower than NPV. The full list of features and their performance against the annotated corpus is shown in Supplementary Table 3.

Classification feature	TP	FP	TN	FN	PPV	NPV
Erratum not in publication type	1535	4989	54	0	0.24	1.00
>1 gene in PubMed metadata	1535	5039	4	0	0.23	1.00
Corrected not in PubMed metadata	1535	5042	1	0	0.23	1.00
IPSC not in abstract	1535	4972	71	0	0.24	1.00
IPSC not in title	1535	4998	45	0	0.23	1.00
GWAS not in title	1535	4989	27	0	0.24	1.00
Mouse not in title	1535	5016	311	4	0.23	0.99
Review not in abstract	1531	4732	66	1	0.24	0.99
Correction not in title	1534	4977	54	1	0.25	0.98
Cancer not in title	1534	4989	492	12	0.24	0.98
>1 gene in title	1523	4551	1053	27	0.25	0.98
GWAS not in abstract	1508	3990	229	7	0.27	0.97
Mouse not in abstract	1528	4814	1124	49	0.24	0.96
Cancer not in abstract	1486	3919	770	37	0.27	0.95
Animals not in MeSH	1498	4273	1756	90	0.26	0.95

Table 4-10. Performance of string-based classifiers for selecting non-relevant papers in annotated n=6578 set. Top 15 results sorted by NPV shown. TP – true positive, FP – false positive, TN – true negative, FN – false negative, PPV – positive predictive value, NPV – negative predictive value.

Classification feature	TP	FP	TN	FN	PPV	NPV
Disease name in title	290	115	4928	1245	0.72	0.80
Exome in MeSH	188	83	4960	1347	0.69	0.79
Fuzzy match to disease name in title	601	279	4764	934	0.68	0.84
Intellectual disability in MeSH	329	168	4875	1206	0.66	0.80
Infant in MeSH	437	254	4789	1098	0.63	0.81
Case report in publication type	763	486	4557	772	0.61	0.86
Dominant in MeSH	150	105	4938	1385	0.59	0.78
Phenotype in MeSH	513	372	4671	1022	0.58	0.82
Recessive in MeSH	49	37	5006	1486	0.57	0.77
X-Linked in MeSH	30	23	5020	1505	0.57	0.77
Mutation in MeSH	1028	1103	3940	507	0.48	0.89
Family in MeSH	57	65	4978	1478	0.47	0.77
Animals not in MeSH	1445	3287	1756	90	0.31	0.95
Humans in MeSH	1259	3048	1995	276	0.29	0.88
Mouse not in abstract	1486	3919	1124	49	0.27	0.96

Table 4-11. Performance of string-based classifiers for selecting relevant papers in annotated n=6578 set. Top 15 results sorted by PPV shown. TP – true positive, FP – false positive, TN – true negative, FN – false negative, PPV – positive predictive value, NPV – negative predictive value.

4.3.8 Supervised learning abstract classifiers

It was hypothesised that automation of abstract identification would be effective using a supervised machine learning classifier. This learns a predictive function from labelled training data, in this case PMIDs with the features and relevancy annotations described above. The function can then

be used to predict the likely relevancy of new input examples. The performance of a number of different supervised learning algorithms was assessed.

First, a binary matrix was constructed on Cadmus (43) output using Boolean filtering for the features described above. Each entry corresponded to a PMID, classified as relevant (1) or non-relevant (0) (to include papers annotated as 2 – related to the corresponding DDG2P entry, but not suitable for text extraction). Each feature was coded as present (1) or absent (0). Additionally, for each PMID, the corresponding DDG2P allelic requirement was coded as a feature i.e. monoallelic (1/0), biallelic (1/0) and X-linked (1/0).

Using Scikit-learn version 1.0.1 (160), the annotated papers were split into a 75% training set and 25% test set. The following Scikit-learn classifiers were used on these data: K Nearest Neighbours, Linear SVM, RBF (radial-basis function kernel) SVM, Gaussian Process, Decision Tree, Random Forest, Neural Net (multi-layer perceptron), AdaBoost, Naive Bayes, QDA (quadratic discriminant analysis). Given the skew towards non-relevant papers in the dataset, the balanced_accuracy_score metric as well as precision, recall and F1 score were used to evaluate the results. True positive/negative and false positive/negative classifications were also assessed.

4.3.9 Classification of disease-relevant manuscripts

The results for these classifiers are shown in Table 4-12. The most performant by F1 score was Gaussian Process (0.78), although Random Forest was very close to this (0.77). Interestingly, there were two classifiers – Naïve Bayes and QDA – which had almost perfect recall. These had negligible false negative predictions, which could be a useful property in a multi-step classification system.

Classifier	TP	FP	TN	FN	Accuracy	Precision	Recall	F1 score
Nearest Neighbours	260	115	1170	100	0.82	0.69	0.72	0.71
Linear SVM	237	71	1214	123	0.80	0.77	0.66	0.71
RBF SVM	222	84	1201	138	0.78	0.73	0.62	0.67
Gaussian Process	285	89	1196	75	0.86	0.76	0.79	0.78
Decision Tree	266	97	1188	94	0.83	0.73	0.74	0.74
Random Forest	286	100	1185	74	0.86	0.74	0.79	0.77
Neural Net	228	62	1223	132	0.79	0.79	0.63	0.70
AdaBoost	230	57	1228	130	0.80	0.80	0.64	0.71
Naive Bayes	356	974	311	4	0.62	0.27	0.99	0.42
QDA	360	924	361	0	0.64	0.28	1.00	0.44

Table 4-12. Performance of supervised learning classifiers for selecting relevant papers in annotated n=6578 set. Results from test set of 1645 papers (75%/25% train/test split) SVM – support vector machine, RBF – radial-basis function kernel, QDA – quadratic discriminant analysis. TP – true positives, FP – false positives, TN – true negatives, FN – false negatives.

4.3.10 Analysis of feature importance

The Random Forest algorithm was one of the highest performing classifiers tested (Table 4-12). This classifier uses averaged signals from different randomized decision trees. An advantage of this method is that it allows for assessment of how much particular features are contributing to performance. One method of doing this is the mean decrease in impurity. The decision trees in Random Forest are derived from a measure of how well a feature separates examples of different classes within nodes of a tree. This can be

used to quantify the discriminative value of each feature, based on how often they were used to split nodes (165,166).

Scikit-learn version 1.0.1 (160) was used to calculate this metric – the mean decrease in impurity – for the annotated set (Figure 4-3). This showed the most important features, in order, were: ‘case report’ in publication type, fuzzy match to disease name in title, ‘mutation’ in MeSH terms, ‘infant’ in MeSH terms, ‘phenotype’ in MeSH terms, and ‘intellectual disability’ in MeSH terms. That the highest performing feature is ‘case report’ acts as a sense check that the classifier is working as expected. These features can be used to inform optimisation of this classifier in future.

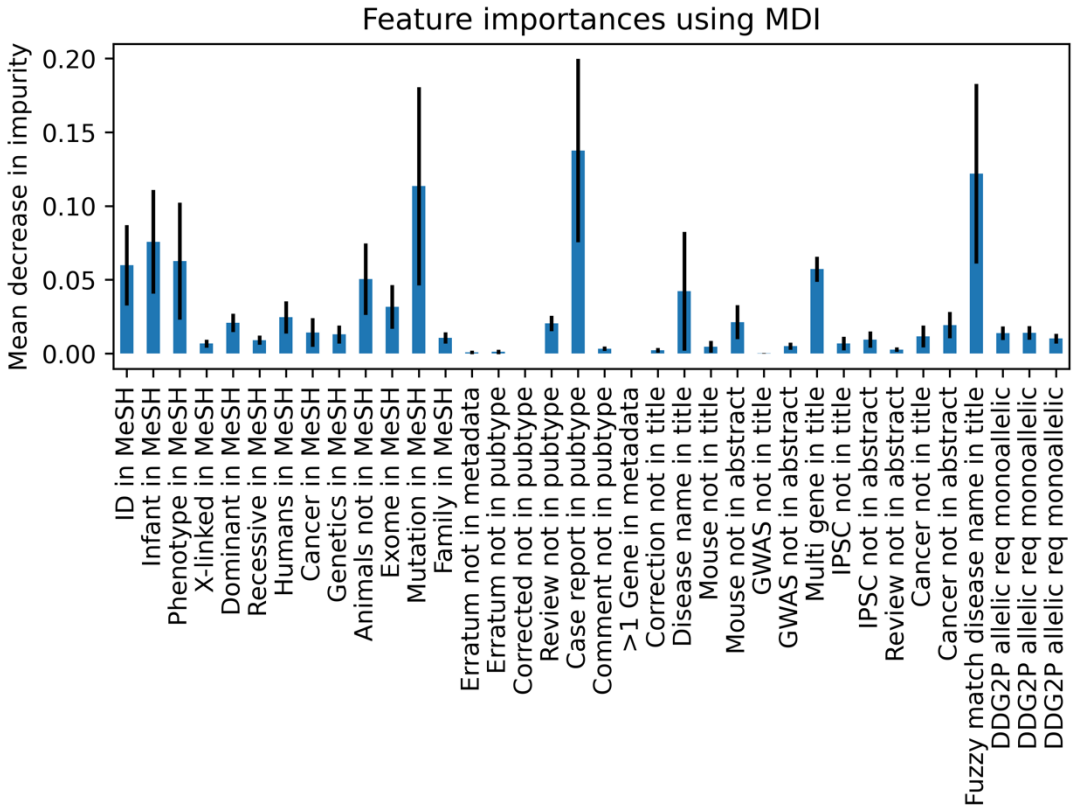


Figure 4-3. Random forest classifier feature importance. Calculated using mean decrease in impurity across all trees. ID – intellectual disability.

4.3.11 Discussion

This section demonstrates the development and testing of a supervised learning classifier for GDD-relevant abstracts. The data shown were preliminary. Nonetheless, the performance figures obtained were encouraging, and compare favourably to similar approaches in the literature. For example, Pham et al. developed an abstract screening system using Random Forest assisted by manual review, with an F1 score of 79% (167).

The maximum F1 obtained here (Table 4-12) was 78%. This is likely explained in part because the features used were domain-specific and carefully selected after review of thousands of abstracts. Using prior knowledge to select features in this manner has proven successful in other domains, for example predicting sensitivity to medications (168).

Nevertheless, there is considerable scope for improvement in the performance of the abstract classifier. The machine learning algorithms used were not modified from their default settings. Fine tuning of their hyperparameters should result in performance improvement. There is also scope to optimise the features used. Only 13 MeSH terms were used, of which four were highly discriminant (Figure 4-3). It is likely that further MeSH terms relating to GDD could be identified to use as features. Another possibility is using the entire list of MeSH terms per paper as a feature, although this may reduce domain specificity.

Another improvement would be optimising the disease name feature (Figure 4-3). Here the disease name used was only that defined by DDG2P. This database includes a number of synonyms for the disease name, which could be used to refine this feature. Furthermore, disease names from other sources, e.g. OMIM (27) could be added for further improvements. In this case, it may be beneficial to use terms from a unified disease ontology such as MONDO (67), to ensure consistent mapping of diseases across datasets.

Topic modelling could also be used to identify patterns in relevant and non-relevant papers, which could be utilised in feature selection (169).

An additional source of features could be the full text of manuscripts as downloaded by Cadmus (43). For example, the number of HPO terms extracted could be higher in papers describing GDD. However, this is much more computationally intensive than using the title and abstract. Therefore, this strategy may be best reserved for a filtered, enriched corpus in a multi-stage system. This will be further discussed below. The exceptions to this are manuscripts types such as 'Letter to the Editor' which do not have an abstract. Full text features may be particularly useful in these cases.

The ratio of relevant to non-relevant papers in the annotated corpus was approximately 1:4 (Table 4-8). This therefore represents skewed data, although this degree of skewing is relatively small. However, this masks a high degree of skew on a per disease basis. The relevant:non-relevant ratio for these was between 5:1 and 1:100. There are various methods for optimising classification on skewed data which could be applied, for example using ensemble learning (170). This combines the results of several different classifiers.

The results presented here could inform the structure of a multi-stage classification system. First, there were a small number of manuscripts which could be easily excluded at first pass. These included erratum/corrections, papers describing induced pluripotent stem cells, and GWAS studies. String matching of terms describing these groups had an NPV of 1 (Table 4-10). Second, a 'exclusion' step could cut down non-relevant results, particularly for those genes which return large numbers of citations. This might utilise the Naïve Bayes and QDA classifiers, which excluded around 20% of papers with almost zero error (Table 4-12). A classifier with a high precision such as Gaussian Process or Random Forest (Table 4-12) could then be used on an

enriched dataset. However, these would need to be trained and optimised on this new data.

It would also be beneficial to improve the PubMed searches used to derive data for classification. The {gene symbol}[TI] search appeared to generate a corpus enriched for GDD-relevant papers. However, 30 genes did not return any results at all (Table 4-5). An expanded search could be used in these cases, for example looking at the abstract as well as the title. Additionally, the {gene symbol}[TI] search will inevitably miss relevant papers. For example, a paper entitled ‘Genetic heterogeneity in Noonan syndrome: evidence for an autosomal recessive form’ (171) would not be included. These could be searched for through other strategies, such as using the disease name alone. However, this will inevitably include incorrect citations. In this case, Noonan Syndrome is linked to multiple different genes.

The {gene symbol}[TI] search uses part of the structure of G2P to identify diseases. A disease is defined in G2P using a locus-genotype-mechanism-disease-evidence thread (31). PubMed search for GDD may be improved by incorporating more of this disease definition, for example {gene symbol} AND {monoallelic}.

In conclusion, I demonstrate here the development and testing of abstract classifiers in the GDD domain. There is scope for optimisation of this system. However, the results demonstrate that an accurate automated literature search is feasible. This is an essential step towards scaling up the disease model concept to cover all GDD. Automated search would also make updating phenotypic data straightforward. In the next section, I outline possible future directions of enquiry to take the work here forward.

4.4 Future directions

4.4.1 Introduction

In this thesis, I have presented a method designed to automate literature curation. I demonstrated that disease models can be generated from the full text literature. These models reflected clinical expressivity. They were similar to manually curated data, as well as prospectively derived clinical phenotyping. However, further study is needed to refine the techniques used, and instigate a robust pipeline covering the full spectrum of GDD. Broadly speaking, this future work covers methodological improvements which could be made to disease model creation, and to phenotype matching/similarity. These will be discussed in the following sections.

4.4.2 Alternative data sources in full-text manuscripts and parsing

Literature-derived disease models in this work were derived from full text manuscripts downloaded using Cadmus (43). Supplemental case reports, where available, are enriched for phenotypic data (Table 4-1). Use of supplementary files should increase the depth of phenotyping and clinical accuracy of the disease model method. Tabular data is also likely to be a rich source of phenotypic information.

Accessing tables could be a significant challenge, given the heterogeneity with which information is presented (172). PDFs are likely to present particular difficulties, given the lack of parseable data structure in these. However, this problem is an active area of research. For example, Khusro et al. and Anand et al. proposed deep learning methods for table extraction in PDFs (173,174). Supplementary data may also contain tables, often in a more parseable format, such as spreadsheets.

Another source of information in full text manuscripts which was not utilised in this work is quantitative data. This is likely to be difficult to extract and parse at scale. For example, a text span could read ‘the head circumference was 36 cm at 3 months’. Here, the number 36 needs to be extracted, related to ‘cm’ to define it as a measurement, and to ‘head circumference’ to define the entity being measured. This then needs to be related to the age at measurement – ‘3 months’. This could be applied to a reference growth chart to allow mapping of the implied term ‘Microcephaly’. Overcoming the challenge of extracting quantitative data is likely to have clinical benefits. For example, quantitative information from the DDD study was used to identify likely diagnoses caused by SNV in a gene previously not associated with GDD (175).

4.4.3 Improved NER using BioBERT

MetaMap (64) was used for named entity recognition/phenotype extraction in this work. Performance as assessed by precision and recall (0.77, 0.71) appeared to be similar to more advanced methods such as PhenoTagger (0.77, 0.74), which uses BioBERT (54,71), although these were tested on different corpora. Although PhenoTagger utilises more sophisticated NER than MetaMap, it is possible that its performance was impaired due to the lack of domain-specific training data. PhenoTagger was trained on a set of 150 052 open access articles generated using the search string ‘disease and mutation’ in EPMC (41,54). A ‘silver standard’ annotated corpus was created using dictionary-based matching in this set. However, a ‘disease and mutation’ search is likely to identify a significant proportion of articles which do not describe GDD.

The abstract classifier discussed in 4.3.11, together with Cadmus, could be utilised to create a GDD-enriched training corpus. MetaMap could be used to annotate this corpus to create a ‘silver standard’ set. This could be used to

train a more domain-specific BioBERT model, which is likely to improve performance (71,83).

4.4.4 Weighting of phenotypic features

HPO terms used to create disease models in this work were weighted by their frequency across all input papers. These appeared to reflect true disease expressivity. However, there was evidence that individual papers could significantly bias weighting (section 2.2.8). Additionally, this measure was difficult to normalize (section 3.3.3). In future, different methods of weighting could be explored. For example, terms could be counted only once per PMID. This would not account for diseases which have highly-studied phenotypic features. For example, at the time of writing a PubMed search for “22q11.2 deletion” AND “schizophrenia” returned 400 results. Another approach could be to correct for terms appearing more frequently in longer papers through adjusting weighting by the number of tokens in a manuscript.

An interesting experiment could be to see how disease models change with date of publication. Reports on GDD typically follow a pattern of first molecular confirmation, larger generalized case series, then focus on individual specialty-specific phenotypic features as with 22q11.2 deletion above. It may be useful to create disease models using a defined time window, for example the first few years after molecular characterization. This could help to define the core phenotypic features of a particular disorder.

4.4.5 Enhancement of semantic similarity

MICA-based semantic similarity methods for comparing phenotype models would seem to have the inherent advantage of using an ontology structure to assess relationships between terms. However, these performed similarly to a list-based method in section 3.5.4. It is possible that adjustments to the

ontology structure may increase the performance of MICA-based similarity methods. For example, Ross et al. used automated literature curation to add new post-translationally modified protein forms to the Protein Ontology (176). It is possible that new GDD ontology classes could be learned using machine learning-based text mining approaches.

Improvements could also be made by adjusting the manner in which information content is calculated for MICA models. For example, using the methods of Lin or Schlicker et al. to account for the closeness of terms to the MICA or the specificity of the MICA respectively (93,101). Another important step is likely to be selecting disease-disease matches by empirical p-values, generated using randomly-sampled data (98). The use of p-values helps correct for semantic similarity metrics tending to generate a higher score for highly-annotated entities (102,177). Adding p-values has been shown to improve the performance of semantic similarity measures using the HPO (22,96).

4.4.6 Bayesian phenotype networks

An interesting application of literature-derived disease models would be to use these with Bayesian networks (BN). These are probabilistic graphical models where nodes represent variables in a given domain, with edges representing the conditional probability for these variables (178). Intuitively, this probability-based inference is similar to the process used in clinical decision making.

BN have been shown to be effective in an number of medical applications, including diagnosis (178). Bauer et al. integrated the HPO into a BN to compare disease models from Orphanet to individual patient phenotypes (28,179). They showed this outperformed MICA-based semantic similarity metrics when predicting GDD diagnosis. This appeared to be partly due to the resilience of BN against phenotypic noise, where a patient is annotated

with HPO terms not relevant to their condition. Using a BN with literature-derived disease models would potentially be a useful way to leverage phenotypic data in diagnosis.

4.4.7 Individual-level phenotype matching

As discussed above, one of the ultimate goals of this work is to use phenotypic data to identify a diagnosis in individual patients with suspected GDD. This would require generation of literature-derived disease models for all GDD in DDG2P. It is likely this would require a version of the abstract classifier discussed in section 4.3.11.

A large-scale test of this system could then be performed using DDD data (24). For thousands of individuals in this study, a list of candidate disease-associated variants was identified using bioinformatic filtering. These were then sent to the recruiting centre, and in many cases one SNV was deemed diagnostic after clinical review. Therefore, a test of the power of literature-derived disease models would be to see if they can be used to correctly identify the diagnostic variant from the filtered list. If this method proves effective, it could be directly applied to diagnosis in suspected GDD.

In most cases, however, the number of HPO terms used to describe an individual referred for genome-wide sequencing is low. This is because the terms need to be manually added by the referring clinician. To address this, text mining in electronic health records could be used. This is a rapidly expanding field of research in the GDD domain (180,181). For example, Hully et al. used a large EHR database to identify individuals with a phenotype matching *KCNA2*-related disease, confirmed on sequencing (182). The data from EHR could also be added to literature-derived models, potentially increasing their clinical relevance and predictive power. Slater et al. demonstrated that adding data mined from hospital notes to literature-derived

models improved differential diagnosis of common disease in the critical care setting (183).

4.4.8 Redefining phenotypic space

Another interesting avenue for future research would be to use methods other than the HPO to define phenotypic space in the GDD domain. Representing the HPO as a network could allow for analysis of disease models using graph-based measures. For example, Menche et al. showed that the overlap of common diseases modelled in a network correlated with biological similarity (184). A network model of GDD could also demonstrate phenotypic clustering, for example for diseases caused by SNV in the same biological pathway. This could be used for variant interpretation. For example, if a patient presented with a phenotype fitting with a particular cluster, genes in that group could be analysed for diagnostic variants. This could even lead to disease discovery, if a gene was identified through this method previously not associated with GDD.

4.4.9 Application to Clinical Genetics

When automated curation as described here is in a fully operational form, the impact on Clinical Genetics practice is potentially significant. First, from a computational perspective, these models could be used in bioinformatics filtering pipelines, for example following the work of Aitken et al. (185). This would allow phenotypic data to be directly used in identifying candidate disease-causing variants from genome-wide sequencing data. There is potential for a novel genotype-phenotype comparison system to be constructed for this purpose. This would build on the disease model and similarity metric methods mentioned in this thesis.

There are several systems which already use phenotype information for filtering genomic sequencing data. These include, for example, the Exomiser with PhenIX (109,186) and the OMIM explorer (110). Manual biocuration is used as the source data for these, such as OMIM (27). It would be straightforward to substitute literature-derived models into these systems. This would potentially enhance performance given the increase in phenotyping depth with full text mining compared to manual curation (section 3.3.2).

Secondly, these disease models may directly affect the clinical assessment of genetically-determined disease. Variant interpretation ultimately relies on a judgement of whether the phenotype of an individual being tested reflects that seen with similar variants in the peer-reviewed literature. Literature-derived disease models should make this process more efficient and accurate, through a comprehensive synthesis of known phenotypic data for a given disease.

There is also a potential direct impact on clinical decision-making. For example, knowledge of the phenotypic spectrum of a specific disorder may influence parental decision making when GDD are diagnosed *in utero*. In particular, the likelihood of severe complications manifesting is often an important factor. Accurate literature-derived disease models describing the frequency of phenotypic features should help inform decision-making in this situation.

Appendix

Publication of work from this thesis

Part of the work from this thesis has been published, as a preprint on medRxiv (https://doi.org/10.1101/2021.11.04.21265878) and in the journal Database:

Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders, Database (Oxford). 2022 Jun 7;2022:baac038. doi: 10.1093/database/baac038. PMID: 35670729.

The supplemental data includes the PMIDs used to create disease models.

Implementation and source code

Source code for the creation of disease models and semantic similarity analysis is available at https://github.com/tmyates/literature_to_pheno. Further source code and datafiles, which would allow for replication of the entire study, are maintained in private storage.

MetaMap not in annotated	Annotated not in metamap
decreased head circumference	dysgenesis of the hippocampus
abnormal nasal morphology	cerebral white matter atrophy
short 4th toe	flared nostrils
oral-pharyngeal dysphagia	abnormal myelination
hyperplasia of the maxilla	kinetic tremor
knee flexion contracture	melanoma
ablepharon	bradykinesia
clubbing of toes	reduced visual acuity
generalized osteoporosis	supraventricular tachycardia
primum atrial septal defect	abnormality of the hand
malignant gastrointestinal tract tumors	vocal cord dysfunction
steppage gait	abnormality of hindbrain morphology
fatigue	femur fracture
acute megakaryocytic leukemia	cerebral hemorrhage
chorioretinal coloboma	hypogonadotropic hypogonadism
hashimoto thyroiditis	increased reactive oxygen species production
patellar hypoplasia	iron accumulation in globus pallidus
abnormal lung morphology	focal aware tonic seizure
recurrent urinary tract infections	neonatal seizure
giant hypertrophic gastritis	deviation of the 2nd toe
abnormality of the cardiovascular system	recurrent ear infections
morphological abnormality of the vestibule of the inner ear	dermoid cyst
chronic kidney disease	scrotal hypoplasia
paralysis	abnormality of the seventh cranial nerve
t-cell acute lymphoblastic leukemias	memory impairment
hypoplasia of the cochlea	moderate global developmental delay
abnormality of the face	cough
scaling skin	intestinal polyp
postaxial hand polydactyly	internal carotid artery hypoplasia
congenital pseudoarthrosis of the clavicle	gastrointestinal hemorrhage
severe muscular hypotonia	genu recurvatum
sparse and thin eyebrow	abnormality of the vestibulocochlear nerve
head-banging	sudden unexpected death in epilepsy
pulmonary insufficiency	psychotic mentation
cortical cataract	short phalanx of the 4th toe
vascular dilatation	hyperalaninemia
cleft upper lip	abnormality of thyroid physiology
small intestinal polyp	deep cerebral white matter hyperdensities

episodic tachypnea	talipes valgus
aneurysmal bone cyst	microphthalmia
equinovarus deformity	hypoplasia of the olfactory bulb
alveolar rhabdomyosarcoma	pruritus
morphological abnormality of the middle ear	joint stiffness
intention tremor	2-3 finger syndactyly
optic nerve coloboma	hypoplasia of the vestibular nerve
keratoconjunctivitis sicca	ovarian carcinoma
tongue thrusting	linear earlobe crease
infertility	11 pairs of ribs
testicular dysgenesis	emphysema
transudative pleural effusion	low hanging columella
abnormality of metabolism/homeostasis	cleft lip
abnormality of the skin	prominent supraorbital ridges
generalized myoclonic seizure	dysmetria
thyroid dysgenesis	abnormality of subcutaneous fat tissue
toe syndactyly	spinal canal stenosis
hyperkinetic movements	ebstein anomaly of the tricuspid valve
morphological central nervous system abnormality	hypermelanotic macule
paroxysmal supraventricular tachycardia	glioma
cutaneous melanoma	abnormality of the nail
hypoplastic nipples	aortic aneurysm
chronic lung disease	2-3 toe syndactyly
transient myeloproliferative syndrome	fundic gland polyposis
mesiodens	mask-like facies
joint contracture of the hand	simple ear
abnormality of the thyroid gland	constrictive median neuropathy
toxemia of pregnancy	diminished mental health
gonadotropin deficiency	malnutrition
severe generalized osteoporosis	convulsive status epilepticus
small scrotum	pyloric stenosis
hypokinesia	hallux valgus
neural tube defect	migraine
poor fine motor coordination	neoplasm of the gastrointestinal tract
cavernous hemangioma	prominent digit pad
skin tags	abnormality of the vertebral column
facial paralysis	psychomotor retardation
abnormality of the pituitary gland	mitochondrial respiratory chain defects
truncal obesity	cerebral cortical atrophy
rigidity	constriction of peripheral visual field

hip contracture	abnormal incisor morphology
cerebral edema	hypoplasia of the vestibule of the inner ear
right ventricular failure	avascular necrosis
abnormality of the lower limb	pontocerebellar atrophy
abnormality of the palmar creases	leg dystonia
skin rash	hemihypertrophy of lower limb
melanocytic nevus	small cerebral cortex
breast hypertrophy	prominent nose
increased body weight	transient ischemic attack
missing ribs	spastic paraparetic gait
trigonocephaly	abnormal pons morphology
chronic active hepatitis	abnormal number of hair whorls
abnormal renal morphology	downturned corners of mouth
burkitt lymphoma	otitis media
fibular deviation of toes	renal agenesis
neonatal hypoglycemia	prominent ear helix
palate telangiectasia	abnormal cochlea morphology
aplasia/hypoplasia of the mandible	abnormality of the vestibular window
megalencephaly	severe infection
secundum atrial septal defect	lacticaciduria
choanal stenosis	abnormal autonomic nervous system physiology
piebaldism	flexion contracture of finger
abnormal skull morphology	axial muscle weakness
acute monocytic leukemia	abnormal emotion/affect behavior
fair hair	hip dysplasia
recurrent infections of the middle ear	abnormality of the clivus
bilateral cleft lip	immunodeficiency
focal dystonia	hydrops fetalis
disinhibition	abnormality of tibia morphology
nasal speech	overweight
hematological neoplasm	incomplete partition of the cochlea
malignant mesothelioma	renal dysplasia
acute otitis media	acute promyelocytic leukemia
facial tics	asthenia
abnormality of eye movement	hand muscle weakness
agenesis of cerebellar vermis	aplasia/hypoplasia of the cochlea
delayed eruption of teeth	abnormal palate morphology
excessive daytime somnolence	weak cry
corneal stromal edema	iron accumulation in substantia nigra
dyssynergia	gastrostomy tube feeding in infancy

prune belly	urethral valve
abnormal electroretinogram	muscle fibrillation
thoracic hypertrichosis	obstructive sleep apnea
abnormal peripheral nervous system morphology	tube feeding
excessive salivation	delayed skeletal maturation
impacted tooth	tonic seizure
apraxia	speech apraxia
alzheimer disease	delayed fine motor development
retinopathy	bilateral postaxial polydactyly
arrhinencephaly	abdominal aortic aneurysm
difficulty walking	flushing
renal tubular acidosis	primary microcephaly
crohn's disease	talipes equinovarus
epidermoid cyst	neoplasm of the lung
abnormality of vision	motor regression
lip telangiectasia	hepatosplenomegaly
clubbing	large intestinal polyposis
muscle spasm	coxa valga
hair-pulling	abnormal subcutaneous fat tissue distribution
	achilles tendon contracture
	appendicular hypotonia
	amenorrhea
	protruding tongue
	abnormal gallbladder morphology
	long toe
	migrating focal seizure
	venous malformation
	optic disc coloboma
	preeclampsia
	neurofibromas
	babinski sign
	abnormal cerebral cortex morphology
	nephropathy
	hypoxemia
	widely-spaced incisors
	absent earlobe
	esophageal carcinoma
	retinal degeneration
	palilalia
	biparietal narrowing

	cerebral hypoplasia
	decreased circulating gonadotropin concentration
	multifocal epileptiform discharges
	pica
	myotonia
	focal emotional seizure
	tented upper lip vermilion
	osteoporosis
	bilateral cleft lip and palate
	supernumerary tooth
	fractured humerus
	transposition of the great arteries
	hepatic failure
	aplasia/hypoplasia of the cerebral white matter
	fibroma
	laryngeal cleft
	prominent subcalcaneal fat pad
	abnormal basal ganglia mri signal intensity
	forceps delivery
	thick cerebral cortex
	drowsiness
	pes valgus
	abnormality of the kidney
	broad finger
	delayed ability to walk
	hypoplasia of the frontal lobes
	mild global developmental delay
	myoclonic seizure
	cholangiocarcinoma
	abnormality of the hairline
	mixed hearing impairment
	lower limb hypertonia
	chronic gastritis
	stuttering
	pulmonic stenosis
	long thumb
	tracheoesophageal fistula
	coronary artery aneurysm
	fetal distress
	hemiparesis

	esodeviation
	clubbing of fingers
	hemiatrophy
	abnormality of the autonomic nervous system
	abnormality of dental eruption
	hypoglycemia
	myotonia of the upper limb
	abnormality of the vestibular nerve
	oral cleft
	hair follicle neoplasm
	recurrent skin infections
	fused lumbar vertebrae
	cortical dysplasia
	thin corpus callosum
	increased csf alanine concentration
	punctate periventricular t2 hyperintense foci
	receptive language delay
	facial cleft
	iridodonesis
	abnormal respiratory system physiology
	abnormality of the nose
	aplasia of the olfactory bulb
	multiple cafe-au-lait spots
	continuous spike and waves during slow sleep
	abnormal aortic arch morphology
	retrognathia
	weak grip
	abnormal finger flexion creases
	wide nose
	enuresis
	simplified gyral pattern
	hypoplastic hippocampus
	communicating hydrocephalus
	neoplasm of the large intestine
	sparse eyebrow
	myoclonic absence seizure
	fusion of middle ear ossicles
	athetosis
	thin skin
	cerebral white matter hypoplasia

abdominal obesity
clonus
aplasia/hypoplasia of the patella
motor seizure
prominent nipples
specific learning disability
shock
paroxysmal bursts of laughter
pseudoarthrosis
abnormality of the ear
aplasia of the semicircular canal
dysgraphia
arteriovenous fistula
diminished ability to concentrate
recurrent otitis media
flexion contracture of digit
hypoplasia of the semicircular canal
thin eyebrow
neonatal respiratory distress
systemic lupus erythematosus
pyelonephritis
abnormality of the dentition
interictal epileptiform activity
abnormal ear morphology
dyspnea
clonic seizure
duodenal polyposis
anophthalmia
abdominal mass
abnormality of the gingiva
decreased circulating iga level
everted lower lip vermilion
prominent nasal tip
abnormal cerebral ventricle morphology
cerebral ischemia
bronchomalacia
nevus
microretrognathia

Supplementary table 1. Comparison of Metamap-extracted and annotated terms from test corpus, where there was no exact match between sets. Clinically similar terms highlighted in green.

Gene symbol	Gene MIM	Disease name	Disease Mim	Allelic requirement	Mutation consequence	PubMed IDs
CHD2	602119	EPILEPTIC ENCEPHALOPATHY	615369	monoallelic	loss of function	31993582 31677157 29740950 29529558 28960266 28910737 26754451 26262932 25672921 24614520
SCN2A	182390	NONSPECIFIC SEVERE ID	613721	monoallelic	loss of function	30062040
DNMT3A	602769	Microcephalic primordial dwarfism	618724	monoallelic	gain of function	30478443
KMT2C	606833	INTELLECTUAL DISABILITY	617768	monoallelic	loss of function	29069077
CNOT3	604910	CNOT3 syndrome	618672	monoallelic	loss of function	32720325 31201375
PPP2R5D	601646	INTELLECTUAL DISABILITY	616355	monoallelic	dominant negative	26576547 25972378
ITPR1	147265	Gillespie Syndrome	206700	biallelic	loss of function	29663667 29169895
CHD4	603277	Syndromic INTELLECTUAL DISABILITY with or without congenital heart disease	617159	monoallelic	loss of function	27616479 31388190
GRIN2B	138252	EPILEPTIC ENCEPHALOPATHY	616139	monoallelic	all missense/in frame	24272827 30151416 27605359 23934111 28377535 31085877
GNAS	139320	MCCUNE-ALBRIGHT SYNDROME	174800	mosaic	activating	1594625 15126527 1944469 31852070 29104223 26188235 17878646 16264125
GRIN2B	138252	MENTAL RETARDATION, AUTOSOMAL DOMINANT 6	613970	monoallelic	loss of function	20890276 23718928
CSNK2A1	115440	CSNK2A1 syndrome	617062	monoallelic	activating	29383814 28725024 27048600
ITPR1	147265	SPINOCEREBELLAR ATAXIA 29, CONGENITAL NONPROGRESSIVE	117360	monoallelic	all missense/in frame	27062503 22986007 31632679 29196976 28826917 28659154
TLK2	608439	TLK2 syndrome	618050	monoallelic	loss of function	29861108 33323470
SLC6A1	137165	EPILEPSY WITH MYOCLONIC-ATONIC SEIZURES	616421	monoallelic	loss of function	29315614 25865495 30132828 27600546 31176687 29961511 31516630
TBL1XR1	608628	Intellectual disability with autism spectrum disorder	616944	monoallelic	loss of function	25425123 29777588 25102098

ITPR1	147265	Gillespie Syndrome, monoallelic	206700	monoallelic	dominant negative	27108798 31340402 30249237 28698159
SMARCA4	603254	RHABDOID TUMOR PREDISPOSITION SYNDROME 2	613325	monoallelic	loss of function	20137775 33836796 31190001 29204511 25060813
CLTC	118955	Epilepsy and intellectual disability	617854	monoallelic	loss of function	26822784 31776469
BCL11A	606557	INTELLECTUAL DISABILITY	617101	monoallelic	loss of function	27453576 28960836 32903878
MORC2	616661	MORC2 - axonal neuropathy and neurodevelopmental disorder	619090	monoallelic	all missense/in frame	26497905 28771897 32693025 30624633 33844363
NFIX	164005	MARSHALL-SMITH SYNDROME	602535	monoallelic	dominant negative	32701632 24924640 28442439
KDM5C	314690	MENTAL RETARDATION SYNDROMIC X-LINKED JARID1C-RELATED	300534	hemizygous	loss of function	16538222 21575681 18697827 19826449 15586325 18203167
KCNQ2	602235	BENIGN NEONATAL EPILEPSY TYPE 1	121200	monoallelic	loss of function	10323247 17872363 9430594 11572947 9425895 11175290 28503627 28038823 24375629 23290024 22884718 20119593 19818940 18640800 18353052 18246739 15596769 15178210 12847176 10774989
HECW2	617245	Neurodevelopmental disorder with hypotonia, seizures, and absent language	617268	monoallelic	all missense/in frame	27334371 27389779 33205896 32814609 29807643 29395664
TRAF7	606692	Developmental Delay Congenital Anomalies and Dysmorphic Features	618164	monoallelic	all missense/in frame	29961569 32376980
SIN3A	607776	SYNDROMIC INTELLECTUAL DISABILITY	613406	monoallelic	loss of function	27399968 33437032 30267900
TBL1XR1	608628	Pierpont syndrome	602342	monoallelic	activating	28687524 30365874 26769062
TRIP12	604506	TRIP12-related intellectual disability with/without autism spectrum disorder	617752	monoallelic	loss of function	27848077 28251352 31814248
CHD3	602120	Macrocephaly and impaired speech and language	618205	monoallelic	all missense/in frame	30397230 33571694 33358638 32483341
FBN1	134797	WEILL-MARCHESANI SYNDROME AUTOSOMAL DOMINANT	608328	monoallelic	all missense/in frame	12525539 28696036 25142510 23897642
TRRAP	603015	Autism and Syndromic Intellectual Disability	618454	monoallelic	all missense/in frame	30827496 30424743

EBF3	607407	Intellectual Disability, Ataxia, and Facial Dysmorphism	617330	monoallelic	loss of function	28017370 28017373 28017372 28487885 29162653
PHIP	612870	Developmental delay, ID, obesity and dysmorphic features	617991	monoallelic	loss of function	29209020 27900362 31167805
GNAS	139320	ALBRIGHT HEREDITARY OSTEODYSTROPHY	103580	monoallelic	loss of function	1505964 2122458 8072545 8702665 11095461 9328353 10487696 11073544 17299070 30349702 25502941 20015054 18806481
KANSL1	612452	CHROMOSOME 17Q21.31 MICRODELETION SYNDROME	610443	monoallelic	loss of function	22544363 22544367 33361104 28211987
PACS1	607492	INTELLECTUAL DISABILITY	615009	monoallelic	activating	23159249 30113927 29550517 28975623 28111752 26842493 25522177
PTPN11	176876	NOONAN SYNDROME 1	163950	monoallelic	activating	12161469 19449407 12529711 12325025 15240615 11992261 15384080 11704759
SPTAN1	182810	EPILEPTIC ENCEPHALOPATHY EARLY INFANTILE TYPE 5	613477	monoallelic	dominant negative	22258530 20493457 33578420 33206935 32811770 31515523 30548380 29986434 29050398 22656320 22429196
SMARCA4	603254	COFFIN SIRIS	614609	monoallelic	all missense/in frame	32686290 31160358 30973214 28608987 24700502
CDK13	603309	Syndromic INTELLECTUAL DISABILITY with or without congenital heart disease	617360	monoallelic	all missense/in frame	29222009 29393965 27479907 28807008 29021403
MEF2C	600662	MENTAL RETARDATION-STEREOTYPIC MOVEMENTS-EPILEPSY AND/OR CEREBRAL MALFORMATIONS	613443	monoallelic	loss of function	20513142 23001426 30922778 30376817 29468350 29104469 28456137 28794905 27255693 22449245
SMC1A	300040	CORNELIA DE LANGE SYNDROME TYPE 2	300590	x-linked dominant	all missense/in frame	24124034 28102598 32532882 22106055 26354354 20635401
PTEN	601728	PTEN Hamartoma Tumor Syndrome	158350	monoallelic	loss of function	11238682 9140396 9425889 12844284 15805158 9832031 9241266 10353779 9467011

						9832032 10777358 17286265 9259288 10051160
SMC1A	300040	SMC1A-related Epileptic Encephalopathy	301044	x-linked dominant	loss of function	26358754 26752331 28166369 31185419 26386245 31098032 28677859
KIF1A	601255	NESCAV SYNDROME	614255	monoallelic	all missense/in frame	25265257 26125038 26486474 30385166 25253658 21376300 32096284 26354034
KAT6B	605880	GENITOPATELLAR SYNDROME	606170	monoallelic	dominant negative	22265014 30900427 28696035 22265017 31871732
KDM6A	300128	KABUKI SYNDROME 2	300867	x-linked dominant	loss of function	24527667 30509212 24664873 27028180 23076834 33674768
KCNQ2	602235	EPILEPTIC ENCEPHALOPATHY EARLY INFANTILE TYPE 7	613720	monoallelic	loss of function	23774309 24107868 31199083 24371303 22169383 28687180 12742592 31418850 28728838 27602407 28631195 29687029 30107960 23621294 23692823 28832002 22275249 27861786 25880994 22926866 25566516 30530441 31152295 25092550 31951342
EHMT1	607001	Kleefstra syndrome	610253	monoallelic	loss of function	28361100 19264732 23232695 28498556 27123477 16826528
TAF1	313650	Dysmorphic Features, Intellectual Disability, and Neurological Manifestations	300966	hemizygous	loss of function	32396742 31646703 31341187 30805980 32714589 26637982
GATAD2B	614998	NONSPECIFIC SEVERE ID	615074	monoallelic	loss of function	23644463 31949314 32688057 31205050 30482549 30346093 28077840
SYNGAP1	603384	MENTAL RETARDATION AUTOSOMAL DOMINANT TYPE 5	612621	monoallelic	loss of function	21237447 28721930 30685520 28576131 30800045 26110312 26079862 29381230 23141534 31395010

						30572772 23161826 26989088 30556619 19196676 23708187
KAT6B	605880	SAY-BARBER-BIESECKER-YOUNG-SIMPSON SYNDROME	603736	monoallelic	loss of function	23436491 24458743 30353918 22077973 27696664 28232779 28758091 26334766
DYNC1H1	600112	SPINAL MUSCULAR ATROPHY, LOWER EXTREMITY-PREDOMINANT, AD	158600	monoallelic	all missense/in frame	28554554 22459677 25484024 29306600 27066557 30122514 25609763 27331017 22368300 28193117 24307404
SMARCA2	600014	NICOLAIDES-BARAITSER SYNDROME	601358	monoallelic	all missense/in frame	32694869 22366787 32657847 31288860 28948053 27665729 22822383
PTPN11	176876	LEOPARD SYNDROME TYPE 1	151100	monoallelic	all missense/in frame	23799168 21747628 27484170 17875892 24790373 25884655 19659470 15520399 19054014 26377839 22822385 19768645 24820750 25917897 21677813 21365175 16733669 21910226 17927788 19864201
SRCAP	611421	FLOATING-HARBOR SYNDROME	136140	monoallelic	dominant negative	22965468 30425916 22265015 30304910 23621943 25433523 23165645 26788936 24375913 23763483
WDR45	300526	WDR45-RELATED NEURODEGENERATION WITH BRAIN IRON ACCUMULATION	300894	x-linked dominant	loss of function	23176820 31466010 27030146 29600274 30612247 28551038 26790960 30539914 29981852 26022463 28932395 29171013 27957548 28361255 29082105 27681470 29681108 30713886 26609730 28371320 26240209

TCF20	603107	TCF20 syndrome	618430	monoallelic	loss of function	25228304 28135719 30739909 27436265 30819258
WAC	615049	Desanto-Shinawi syndrome	616708	monoallelic	loss of function	26757981 26264232 33387902 32214004
IQSEC2	300522	MENTAL RETARDATION X-LINKED TYPE 1	309530	x-linked dominant	loss of function	31415821 23674175 24306141 26793055 27665735 20473311 31829726 29026562 28295038 28815955 30666632 30206421 26733290 31490346
KMT2B	606834	Complex early-onset dystonia	617284	monoallelic	loss of function	33816656 33300088 33150406 32877735 32546208 32634684 32241076 31338059 31216378 31165786 31061210 30935829 30253925 29653907 29396090 28921672 28520167 27992417 27839873
FBN1	134797	MARFAN SYNDROME	154700	monoallelic	loss of function	8428751 8504310 1631074 7611299 7762551 18412115 1301946 15287423 15032979 8040326 8101042 9101298 8406497 11175294 17366579 11702223 9837823 1569206 8136837 9241263 7633409 20082464 10441597 7911051 20979188 8281141 1852208
NFIX	164005	SOTOS SYNDROME 2	614753	monoallelic	loss of function	22301465 29897170 32193017 31369202 28584646 26193383 25118028 22982744
SETD5	615743	MENTAL RETARDATION, AUTOSOMAL DOMINANT 23	615761	monoallelic	loss of function	25138099 28905509 28549204 27375234 28881385 24680889 31656537
PURA	600473	INTELLECTUAL DISABILITY	616158	monoallelic	loss of function	32089526 25439098 29150892 29097605 25342064 27148565 29307761 31911028

SCN2A	182390	INFANTILE EPILEPTIC ENCEPHALOPATHY	613721	monoallelic	all missense/in frame	28489313 29635106 25457084 19783390 25772804 23935176 23988467 24579881 28379373 25459969 28254201 22591750 24659627 28709814 24814476 29625812 15028761 31966371 26311622 23827426 24710820 23550958 30203812 19786696 31439038 31204721 26291284
TCF4	602272	PITT-HOPKINS SYNDROME	610954	monoallelic	loss of function	22670824 20184619 29695756 27132474 18728071 19235238 17436255 23528641 22045651 19938247 30848346 29604340 20205897 17436254
CACNA1A	601011	EPILEPTIC ENCEPHALOPATHY	617106	monoallelic	all missense/in frame	28927557 28742085 29366381 33557884 33445191 33425808 33349592 32692472 32170034 31468518 26739101 25735478 20097664
FOXP1	605515	MENTAL RETARDATION WITH LANGUAGE IMPAIRMENT AND AUTISTIC FEATURES	613670	monoallelic	loss of function	29090079 29330474 24214399 20950788 28884888 28735298 25853299 30092897
SHANK3	606230	PHELAN-MCDERMID SYNDROME	606232	monoallelic	loss of function	22892527 17173049 32202324 30537371 29939863 29719671 29423971 28963116 28754298 27554343 25931020 26045941 23612248 21378602 20385823 18615476
DNMT3A	602769	Tatton-Brown Rahman syndrome (OVERGROWTH SYNDROME WITH INTELLECTUAL DISABILITY)	615879	monoallelic	loss of function	24614070 29900417 32435502 31905446 31685998 28941052

						28449304 28432085 27701732 27991732
CTNNB1	116806	MENTAL RETARDATION, AUTOSOMAL DOMINANT 19	615075	monoallelic	loss of function	25326669 28514307 26968164 24614104 30929091 27915094
AHDC1	615790	XIA-GIBBS SYNDROME	615829	monoallelic	loss of function	29696776 29230160 31182893 32256298 30622101 30858058 31812316 30152016 27148574 30729726 24791903
PUF60	604819	PUF60 syndrome	615583	monoallelic	loss of function	27804958 28327570 32851780 30569551 30352594 28471317 28074499
ASXL3	615115	BAINBRIDGE-ROPERS SYNDROME	615485	monoallelic	loss of function	29305346 23383720 29445472 29367179 31638014 29316359 24044690 28100473 27075689 27901041 28955728 31180560 32240826
EP300	602700	RUBINSTEIN-TAYBI SYNDROME TYPE 2	613684	monoallelic	loss of function	17299436 20014264 19353645 33442921 33043588 30076641 29506490 29133209 28027063 27581590 27964710 27465822 27648933 27386132 26374735 26279656 25712426 24476420 24352918
POGZ	614787	INTELLECTUAL DISABILITY	616364	monoallelic	loss of function	31136090 26763879 27148570 26942287 31347273 28480548 31782611 32103003 30879264 27103995 25694107 26739615
DYRK1A	600855	MENTAL RETARDATION AUTOSOMAL DOMINANT TYPE 7	614104	monoallelic	loss of function	21294719 25641759 28053047 31263215 29034068 31803247 25707398 26922654

						23099646 25944381 25920557
KMT2D	602113	KABUKI SYNDROME	147920	monoallelic	loss of function	27530205 28404210 29283410 21607748 29482518 25142838 20711175 21671394 25972376 31935506 25944076 24739679 30569626 29914387 27573763 27302555 28295206
HNRNPU	602869	EPILEPTIC ENCEPHALOPATHY	617391	monoallelic	loss of function	33914968 32319732 29858110 28944577 28815871 28393272
KAT6A	601408	MENTAL RETARDATION, AUTOSOMAL DOMINANT 32	616268	monoallelic	loss of function	25728775 27133397 30245513 25728777 31754438 29899504 30775047 32041641
SATB2	608148	GLASS SYNDROME	612313	monoallelic	loss of function	29436146 31021519 30648748 31420882 31392730 17377962 24301056 29739092 28139846 25885067 30848049 28787087 28211976 28151491 27774744 26596517 31333717 24363063 30575289
SMAD4	600993	Juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome	175050	monoallelic	loss of function	33370972 32944796 32556653 30196345 27375208 26159157 25931195 25705527 24312718 26181832 23239472 22617360 22331366 21572342 22056587 21465659 20685751 20101697 18355998 15990641 15754356 15031030 31394615 12116240 11920286 10455879 10398437 9811934 9582123
STXBP1	602926	EPILEPTIC ENCEPHALOPATHY EARLY INFANTILE TYPE 4	612164	monoallelic	loss of function	21062273 26212315 26384463 32105008

						29929108 28944233 18469812 20887364 26865513 21762454 22596016 29544889 27069701 24315539 19557857 29718889 29264391 31387522 27184330 21770924 21364700 26514728 30654231 24170257 25714420 24189369 23763664 20876469 31344879 23531706 24623842 25631041 25418441 24095819 23533165 23409955 29896790 21204804
ADNP	611386	MENTAL RETARDATION, AUTOSOMAL DOMINANT, 28	615873	monoallelic	loss of function	27031564 32275126 30107084 25169753 28407407 31127536 29724491 24531329 28221363 28475273 29475819
BCOR	300485	MICROPHTHALMIA SYNDROMIC TYPE 2	300166	x-linked dominant	loss of function	29974297 32748437 15957158 15004558 28317252 29974297 15770227 19367324 31048080
EFTUD2	603892	MANDIBULOFACIAL DYSOSTOSIS WITH MICROCEPHALY	610536	monoallelic	loss of function	25387991 26507355 24470203 25735261 27670155 23188108 23879989 23239648 31413053 22305528 30343593 28643921
FOXG1	164874	CONGENITAL VARIANT OF RETT SYNDROME	613454	monoallelic	loss of function	18571142 21441262 19564653 19578037 32757993 31316448 30533527 29396177 28851325 28781028 28661489 27029630 27001178 26364767 26344814 25266269 24836831 24388699

						22129046 22091895 21953941
ARID1B	614556	MENTAL RETARDATION, AUTOSOMAL DOMINANT 12	135900	monoallelic	loss of function	30349098 22426309 22405089 33936271 33714239 32618029 32339967 31981384 32161024 30933046 31628733 31421289 31105273 29504208 28323383 27474218 27570168 27511161 27672547 26395437 26376624 24569609
KMT2A	159555	WIEDEMANN-STEINER SYNDROME	605130	monoallelic	loss of function	27777327 32311999 24818805 22795537 24886118 27320412 28815892 25929198 31044088 27759909 29203834 31250358 31168168 30549396 30841869 25810209
SMAD4	600993	MYHRE SYNDROME	139210	monoallelic	activating	33428109 31837202 31654632 27302097 24715504 22711472 22585601 22243968 22158539
SCN8A	600702	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 13	614558	monoallelic	dominant negative	22365152 16236810 33827760 33915942 33007625 32920374 32853054 32846312 32040247 32509551 31672125 31675620 31402610 31335965 31174070 31026061 31010614 30968951 30851583 30685519 30171078 30078772 2972606 29677576 29432985 29263050 29128679 29121005 28923014 28676440 28702509 28084268 27875746 27900360 27210545 26677014 26553437

						26252990 26235738 25725044 25785782 25799905 25568300
NSD1	606681	SOTOS SYNDROME	117550	monoallelic	loss of function	12525543 16222665 11896389 33029244 30719864 30461603 30332768 29164086 28457852 27834868 25887879 24795065 23341071 21834047 21677402 21677396 20420030 19876911 19545651 17565729 16547423 16247291 16232326 15742365 15362962 14627693
MED13L	608771	INTELLECTUAL DISABILITY	616789	monoallelic	loss of function	25758992 28645799 25712080 29511999 25137640 29159987 24781760 29959045 23403903 28371282
ANKRD11	611192	KBG SYNDROME	148050	monoallelic	loss of function	30088855 27667800 23494856 28250421 28449295 25464108 27900361 25652421 23184435 21782149 25838844 30877071 29224748
CREBBP	600140	RUBINSTEIN-TAYBI SYNDROME TYPE 1	180849	monoallelic	loss of function	7630403 27311832 20684013 11331617 30737887 12566391 33747050 32839936 31637876 31566936 30892814 30770747 30633342 30614040 29745126 27342041 27165009 26956253 26603346 26275701 25768348 25388907 20949605 20583168 20689175 19852432 12114483
CHD7	608892	CHARGE SYNDROME	214800	monoallelic	loss of function	16400610 17334995 18978652 17661815 15300250 17937444

						18074359 33844462 32699053 31929333 32625235 32509017 32126561 31037873 31315586 29945602 29531775 29434620 28609304 26901670 27081570 26921530 26741373 26929907 26590800 26551301 26187070 25606431 24578717 24550764 23495722 23333604 23024289 22517486 22302456 21041284 20943277 20624498 19159393 19021638 18484313 18505430 18445044 18073582

Supplementary Table 2. DDG2P disease entries with PubMed IDs used to create 99 disease test set.

Feature	TP	FP	TN	FN	Sens	Spec	PPV	NPV
Erratum not in metadata	1535	4989	54	0	1	0.01	0.24	1
Intellectual disability in MeSH	329	168	4875	1206	0.21	0.97	0.66	0.8
Infant in MeSH	437	254	4789	1098	0.28	0.95	0.63	0.81
Phenotype in MeSH	513	372	4671	1022	0.33	0.93	0.58	0.82
X-linked in MeSH	30	23	5020	1505	0.02	1	0.57	0.77
Dominant in MeSH	150	105	4938	1385	0.1	0.98	0.59	0.78
Recessive in MeSH	49	37	5006	1486	0.03	0.99	0.57	0.77
Humans in MeSH	1259	3048	1995	276	0.82	0.4	0.29	0.88
Cancer in MeSH	33	568	4475	1502	0.02	0.89	0.05	0.75
Genetics in MeSH	1256	3315	1728	279	0.82	0.34	0.27	0.86
Animals not in in MeSH	1445	3287	1756	90	0.94	0.35	0.31	0.95
Exome in MeSH	188	83	4960	1347	0.12	0.98	0.69	0.79
Mutation in MeSH	1028	1103	3940	507	0.67	0.78	0.48	0.89
Family in MeSH	57	65	4978	1478	0.04	0.99	0.47	0.77
Erratum not in publication type	1535	4989	54	0	1	0.01	0.24	1
Corrected not in publication type	1535	5042	1	0	1	0	0.23	1
Review not in publication type	1460	4807	236	75	0.95	0.05	0.23	0.76
Case report in publication type	763	486	4557	772	0.5	0.9	0.61	0.86
Comment not in publication type	1530	4990	53	5	1	0.01	0.23	0.91
Multiple genes not in metadata	1535	5039	4	0	1	0	0.23	1
Correction not in title	1534	4989	54	1	1	0.01	0.24	0.98
Disease name in title	290	115	4928	1245	0.19	0.98	0.72	0.8
Mouse not in title	1531	4732	311	4	1	0.06	0.24	0.99
Mouse not in abstract	1486	3919	1124	49	0.97	0.22	0.27	0.96
GWAS not in title	1535	5016	27	0	1	0.01	0.23	1
GWAS not in abstract	1528	4814	229	7	1	0.05	0.24	0.97
Multiple genes not in title	1508	3990	1053	27	0.98	0.21	0.27	0.98
IPSC not in title	1535	4998	45	0	1	0.01	0.23	1
IPSC not in abstract	1535	4972	71	0	1	0.01	0.24	1
Review not in abstract	1534	4977	66	1	1	0.01	0.24	0.99
Cancer not in title	1523	4551	492	12	0.99	0.1	0.25	0.98
Cancer not in abstract	1498	4273	770	37	0.98	0.15	0.26	0.95
Fuzzy match to disease name in title	601	279	4764	934	0.39	0.94	0.68	0.84

Supplementary Table 3. Performance of individual features against annotated corpus of GDD-relevant manuscripts. GWAS – genome-wide association study. IPSC – induced pluripotent stem cells. TP – true positive. FP – false positive. TN – true negative. FN – false negative. Sens – sensitivity. Spec – specificity. PPV – positive predictive value. NPV – negative predictive value.

Bibliography

1. Verloes A, Di Donato N, Masliah-Planchon J, Jongmans M, Abdul-Raman OA, Albrecht B, et al. Baraitser-Winter cerebrofrontofacial syndrome: delineation of the spectrum in 42 cases. Eur J Hum Genet. 2015 Mar;23(3):292–301.
2. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10;461(7261):272–6.
3. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015 Mar 12;519(7542):223–8.
4. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868–80.
5. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med. 2018;3:16.
6. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017 Feb 23;542(7642):433–8.
7. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: lessons from gnomAD. arXiv:210711458 [q-bio] [Internet]. 2021 Nov 4 [cited 2022 Jan 2]; Available from: http://arxiv.org/abs/2107.11458
8. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011 Jun 12;43(7):712–4.
9. European Bioinformatics Institute. gene2phenotype [Internet]. [cited 2022 Jan 2]. Available from: https://www.ebi.ac.uk/gene2phenotype/
10. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. 2015 May;17(5):405–23.
11. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med. 2018 Apr;20(4):435–43.
12. Carss KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, et al. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am J Hum Genet. 2017 Jan 5;100(1):75–90.

13. Wang X, Posey JE, Rosenfeld JA, Bacino CA, Scaglia F, Immken L, et al. Phenotypic expansion in DDX3X - a common cause of intellectual disability in females. Ann Clin Transl Neurol. 2018 Oct;5(10):1277–85.
14. Rech ME, McCarthy JM, Chen CA, Edmond JC, Shah VS, Bosch DGM, et al. Phenotypic expansion of Bosch-Boonstra-Schaaf optic atrophy syndrome and further evidence for genotype-phenotype correlations. Am J Med Genet A. 2020 Jun;182(6):1426–37.
15. Sullivan JA, Stong N, Baugh EH, McDonald MT, Takeuchi A, Shashi V. A pathogenic variant in the SETBP1 hotspot results in a forme-fruste Schinzel-Giedion syndrome. Am J Med Genet A. 2020 Aug;182(8):1947–51.
16. Di Donato N, Kuechler A, Vergano S, Heinritz W, Bodurtha J, Merchant SR, et al. Update on the ACTG1-associated Baraitser-Winter cerebrofrontofacial syndrome. Am J Med Genet A. 2016 Oct;170(10):2644–51.
17. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012 Nov 10;380(9854):1674–82.
18. Guest SS, Evans CD, Winter RM. The Online London Dysmorphology Database. Genet Med. 1999 Aug;1(5):207–12.
19. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008 Nov;83(5):610–5.
20. Guarino N. Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. 1998;3–15.
21. Fung KW, Bodenreider O. Knowledge Representation and Ontologies. In: Richesson RL, Andrews JE, editors. Clinical Research Informatics [Internet]. London: Springer London; 2012 [cited 2022 Jan 3]. p. 255–75. (Health Informatics). Available from: http://link.springer.com/10.1007/978-1-84882-448-5_14
22. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Research. 2021 Jan 8;49(D1):D1207–17.
23. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2021 Dec 1;gkab1112.
24. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015 Apr 4;385(9975):1305–14.
25. Urreizti R, Lopez-Martin E, Martinez-Monseny A, Pujadas M, Castilla-Vallmanya L, Pérez-Jurado LA, et al. Five new cases of syndromic intellectual disability due to KAT6A mutations: widening the molecular and clinical spectrum. Orphanet J Rare Dis. 2020 Feb 10;15(1):44.
26. Kumar R, Palmer E, Gardner AE, Carroll R, Banka S, Abdelhadi O, et al. Expanding Clinical Presentations Due to Variations in THOC2 mRNA Nuclear Export Factor. Front Mol Neurosci. 2020;13:12.

27. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Research. 2019 Jan 8;47(D1):D1038–43.
28. Orphanet © INSERM. Orphanet: an online rare disease and orphan drug data base [Internet]. Available from: http://www.orpha.net
29. OMIM® McKusick-Nathans Institute of Genetic Medicine Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man [Internet]. Available from: https://omim.org/
30. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics. 2009 Apr;84(4):524–33.
31. Thormann A, Halachev M, McLaren W, Moore DJ, Svinti V, Campbell A, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. Nat Commun. 2019 Dec;10(1):2373.
32. Rasmussen SA, Hamosh A, Amberger J, Arnold C, Bocchini C, O'Neill MJF, et al. What's in a name? Issues to consider when naming Mendelian disorders. Genetics in Medicine. 2020 Oct;22(10):1573–5.
33. Biesecker LG, Adam MP, Alkuraya FS, Amemiya AR, Bamshad MJ, Beck AE, et al. A dyadic approach to the delineation of diagnostic entities in clinical genomics. Am J Hum Genet. 2021 Jan 7;108(1):8–15.
34. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2017 Jan 4;45(D1):D712–22.
35. DiStefano MT, Goehringer S, Babb L, Alkuraya FS, Amberger J, Amin M, et al. The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources [Internet]. Genetic and Genomic Medicine; 2022 Jan [cited 2022 Jan 5]. Available from: http://medrxiv.org/lookup/doi/10.1101/2022.01.03.21268593
36. Collier N, Groza T, Smedley D, Robinson PN, Oellrich A, Rebholz-Schuhmann D. PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. Database. 2015;2015:bav104.
37. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019 Jul 2;47(W1):W587–93.
38. Samuel J, Yuan X, Yuan X, Walton B. Mining online full-text literature for novel protein interaction discovery. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW) [Internet]. HongKong, China: IEEE; 2010 [cited 2022 Jan 5]. p. 277–82. Available from: http://ieeexplore.ieee.org/document/5703812/
39. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 2009 Feb 5;10 Suppl 2:S6.
40. Westergaard D, Stærfeldt HH, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018 Feb;14(2):e1005962.

41. European Bioinformatics Institute. Europe PubMed Central [Internet]. [cited 2022 Jan 6]. Available from: https://europepmc.org/
42. Intellectual Property Office UK. Exceptions to copyright [Internet]. [cited 2021 Oct 22]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf
43. Campbell, Jamie, Lain, Antoine, Simpson, Ian. biomedicalinformaticsgroup/cadmus: First Release of Cadmus (v1.0.0) [Internet]. Zenodo. Available from: https://doi.org/10.5281/zenodo.5618052
44. Cell Press. American Journal of Human Genetics [Internet]. [cited 2022 Jan 6]. Available from: https://www.cell.com/AJHG/
45. Perera N, Dehmer M, Emmert-Streib F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Front Cell Dev Biol. 2020;8:673.
46. Korkontzelos I, Piliouras D, Dowsey AW, Ananiadou S. Boosting drug named entity recognition using an aggregate classifier. Artif Intell Med. 2015 Oct;65(2):145–53.
47. Lee K, Wei CH, Lu Z. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. Brief Bioinform. 2021 May 20;22(3):bbaa142.
48. Campos D, Matos S, Luis J. Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. In: Sakurai S, editor. Theory and Applications for Advanced Text Mining [Internet]. InTech; 2012 [cited 2022 Jan 6]. Available from: http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining/biomedical-named-entity-recognition-a-survey-of-machine-learning-tools
49. Groza T, Kohler S, Doelken S, Collier N, Oellrich A, Smedley D, et al. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. Database. 2015 Feb 27;2015(0):bav005–bav005.
50. Agoritsas T, Merglen A, Courvoisier DS, Combescure C, Garin N, Perrier A, et al. Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews. J Med Internet Res. 2012 Jun 12;14(3):e85.
51. Lobo M, Lamurias A, Couto FM. Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules. BioMed Research International. 2017;2017:1–8.
52. Mohan S, Li D. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. arXiv:190209476 [cs] [Internet]. 2019 Feb 25 [cited 2022 Jan 8]; Available from: http://arxiv.org/abs/1902.09476
53. Verspoor K, Cohen KB, Lanfranchi A, Warner C, Johnson HL, Roeder C, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics. 2012 Aug 17;13:207.
54. Luo L, Yan S, Lai PT, Veltri D, Oler A, Xirasagar S, et al. PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology. Bioinformatics. 2021 Jan 20;btab019.

55. Tomanek K, Wermter J, Hahn U. A reappraisal of sentence and token splitting for life sciences documents. Stud Health Technol Inform. 2007;129(Pt 1):524–8.
56. Fan JW, Friedman C. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. J Biomed Inform. 2011 Oct;44(5):805–14.
57. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis P, Houstis EN, editors. Advances in Informatics [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005 [cited 2022 Jan 7]. p. 382–92. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Lecture Notes in Computer Science; vol. 3746). Available from: http://link.springer.com/10.1007/11573036_36
58. Gorrell G, Song X, Roberts A. Bio-YODIE: A Named Entity Linking System for Biomedical Text. 2018 [cited 2022 Oct 18]; Available from: https://arxiv.org/abs/1811.04860
59. Fraser KC, Nejadgholi I, De Bruijn B, Li M, LaPlante A, Abidine KZE. Extracting UMLS Concepts from Medical Text Using General and Domain-Specific Deep Learning Models. 2019 [cited 2022 Oct 18]; Available from: https://arxiv.org/abs/1910.01274
60. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-270.
61. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit Transl Bioinform. 2009 Mar 1;2009:56–60.
62. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-545.
63. Dai M, Shah NH, Xuan W, Musen MA, Watson SJ, Athey BD, et al. An efficient solution for mapping free text to ontology terms. AMIA summit on translational bioinformatics. 2008;21.
64. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May;17(3):229–36.
65. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc. 1993 Apr;81(2):184–94.
66. Taboada M, Rodríguez H, Martínez D, Pardo M, Sobrido MJ. Automated semantic annotation of rare disease cases: a case study. Database (Oxford). 2014;2014:bau045.
67. Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2020 Jan 8;48(D1):D704–15.
68. Okazaki N. CRFsuite: a fast implementation of Conditional Random Fields [Internet]. [cited 2022 Jan 8]. Available from: http://www.chokkan.org/software/crfsuite/

69. Arbabi A, Adams DR, Fidler S, Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. JMIR Med Inform. 2019 May 10;7(2):e12596.
70. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. TACL. 2017 Dec;5:135–46.
71. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, editor. Bioinformatics. 2019 Sep 10;btz682.
72. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] [Internet]. 2019 May 24 [cited 2022 Jan 11]; Available from: http://arxiv.org/abs/1810.04805
73. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. arXiv:150606724 [cs] [Internet]. 2015 Jun 22 [cited 2022 Jan 11]; Available from: http://arxiv.org/abs/1506.06724
74. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
75. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-Alignment Pretraining for Biomedical Entity Representations. 2020 [cited 2022 Oct 18]; Available from: https://arxiv.org/abs/2010.11784
76. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019 [cited 2022 Oct 18]. p. 72–8. Available from: http://aclweb.org/anthology/W19-1909
77. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics. 2009 Sep 17;10 Suppl 9:S14.
78. Oellrich A, Collier N, Smedley D, Groza T. Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. PLoS One. 2015;10(1):e0116040.
79. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018 Sep 14;18(Suppl 3):74.
80. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Research. 2019 Jul 2;47(W1):W566–70.
81. ShARe/CLEF. ShARe/CLEF corpus [Internet]. Available from: https://sites.google.com/site/shareclefehealth/

82. Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009 Aug;16(4):561–70.
83. Wilcox A, Hripcsak G, Friedman C. Using Knowledge Sources to Improve Classification of Medical Text Reports. KDD-2000 Workshop on Text Mining. 2000;2.
84. Collier N, Tran M vu, Paster F. The impact of near domain transfer on biomedical named entity recognition. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) [Internet]. Gothenburg, Sweden: Association for Computational Linguistics; 2014 [cited 2022 Jan 19]. p. 11–20. Available from: http://aclweb.org/anthology/W14-1103
85. Fitchett S, Cockburn A. AccessRank: predicting what users will do next. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [Internet]. Austin Texas USA: ACM; 2012 [cited 2022 Jan 14]. p. 2239–42. Available from: https://dl.acm.org/doi/10.1145/2207676.2208380
86. Kendall MG. Rank Correlation Methods. Journal of the Institute of Actuaries. 1949;75(1):140–1.
87. Shieh GS. A weighted Kendall's tau statistic. Statistics & Probability Letters. 1998 Jul;39(1):17–24.
88. Fagin R, Kumar R, Sivakumar D. Comparing Top k Lists. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms. USA: Society for Industrial and Applied Mathematics; 2003. p. 28–36. (SODA '03).
89. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010 Nov;28(4):1–38.
90. Church KW, Hanks P. Word Association Norms, Mutual Information, and Lexicography. Comput Linguist. 1990 Mar;16(1):22–9.
91. Konagurthu A, Collier J. An information measure for comparing top k lists. arXiv:13100110 [cs, math] [Internet]. 2013 Sep 30 [cited 2022 Jan 14]; Available from: http://arxiv.org/abs/1310.0110
92. The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research. 2021 Jan 8;49(D1):D325–34.
93. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 2006 Dec;7(1):302.
94. Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, et al. A Knowledge-Based Clustering Algorithm Driven by Gene Ontology. Journal of Biopharmaceutical Statistics. 2004 Dec 29;14(3):687–700.
95. Cover TM. Elements of information theory. John Wiley & Sons; 1999.
96. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. The American Journal of Human Genetics. 2009 Oct;85(4):457–64.
97. Galer PD, Ganesan S, Lewis-Smith D, McKeown SE, Pendziwiat M, Helbig KL, et al. Semantic Similarity Analysis Reveals Robust Gene-Disease Relationships in Developmental and Epileptic

Encephalopathies. The American Journal of Human Genetics. 2020 Oct;107(4):683–97.
98. Helbig I, Lopez-Hernandez T, Shor O, Galer P, Ganesan S, Pendziwiat M, et al. A Recurrent Missense Variant in AP2M1 Impairs Clathrin-Mediated Endocytosis and Causes Developmental and Epileptic Encephalopathy. The American Journal of Human Genetics. 2019 Jun;104(6):1060–72.
99. Zhou Z, Wang Y, Gu J. A New Model of Information Content for Semantic Similarity in WordNet. In: 2008 Second International Conference on Future Generation Communication and Networking Symposia [Internet]. Hinan, China: IEEE; 2008 [cited 2022 Jan 14]. p. 85–9. Available from: http://ieeexplore.ieee.org/document/4813554/
100. Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. arXiv:cmp-lg/9511007 [Internet]. 1995 Nov 29 [cited 2021 Nov 1]; Available from: http://arxiv.org/abs/cmp-lg/9511007
101. Lin D. An Information-Theoretic Definition of Similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. p. 296–304. (ICML '98).
102. Kulmanov M, Hoehndorf R. Evaluating the effect of annotation size on measures of semantic similarity. J Biomed Semant. 2017 Dec;8(1):7.
103. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995 Jan;57(1):289–300.
104. Carmody LC, Blau H, Danis D, Zhang XA, Gourdine JP, Vasilevsky N, et al. Significantly different clinical phenotypes associated with mutations in synthesis and transamidase+remodeling glycosylphosphatidylinositol (GPI)-anchor biosynthesis genes. Orphanet J Rare Dis. 2020 Dec;15(1):40.
105. Köhler S. Improved ontology-based similarity calculations using a study-wise annotation model. Database [Internet]. 2018 Jan 1 [cited 2021 Nov 1];2018. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bay026/4953405
106. Xue H, Peng J, Shang X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. BMC Syst Biol. 2019 Apr;13(S2):34.
107. Gan M. Correlating information contents of gene ontology terms to infer semantic similarity of gene products. Comput Math Methods Med. 2014;2014:891842.
108. Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics. 2008 Aug 4;9:327.
109. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014 Sep 3;6(252):252ra123.

110. James RA, Campbell IM, Chen ES, Boone PM, Rao MA, Bainbridge MN, et al. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Med. 2016 Feb 2;8(1):13.
111. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014 Apr 3;94(4):599–610.
112. Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: variant prioritization by genomic data fusion. Nat Methods. 2013 Nov;10(11):1083–4.
113. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1(1):13.
114. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014 Feb;24(2):340–8.
115. Borgelt C, Kruse R. Induction of Association Rules: Apriori Implementation. In: Härdle W, Rönz B, editors. Compstat [Internet]. Heidelberg: Physica-Verlag HD; 2002 [cited 2022 Jan 19]. p. 395–400. Available from: http://link.springer.com/10.1007/978-3-642-57489-4_59
116. Pilehvar MT, Bernard A, Smedley D, Collier N. PheneBank: a literature-based database of phenotypes. Wren J, editor. Bioinformatics. 2021 Nov 12;btab740.
117. Hoehndorf R, Schofield PN, Gkoutos GV. Analysis of the human diseasome using phenotype similarity between common, genetic and infectious diseases. Sci Rep. 2015 Sep;5(1):10888.
118. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, The Mouse Genome Database Group. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Research. 2015 Jan 28;43(D1):D726–36.
119. Hoehndorf R, Slater L, Schofield PN, Gkoutos GV. Aber-OWL: a framework for ontology-based data access in biology. BMC Bioinformatics. 2015 Dec;16(1):26.
120. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Research. 2019 Jan 8;47(D1):D955–62.
121. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004 Jan 1;32(90001):267D – 270.
122. Xu R, Li L, Wang Q. Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. Bioinformatics. 2013 Sep 1;29(17):2186–94.
123. American College of Physicians. American College of Physicians Journal Club [Internet]. Available from:

https://www.acponline.org/clinical-information/journals-publications/acp-journal-club
124. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence. Journal of the American Medical Informatics Association. 2009 Jan 1;16(1):25–31.
125. Kim S, Choi J. An SVM-based high-quality article classifier for systematic reviews. Journal of Biomedical Informatics. 2014 Feb;47:153–9.
126. Cohen AM, Ambert K, McDonagh M. Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. Journal of the American Medical Informatics Association. 2009 Sep 1;16(5):690–704.
127. Bian J, Abdelrahman S, Shi J, Del Fiol G. Automatic identification of recent high impact clinical articles in PubMed to support clinical decision making using time-agnostic features. Journal of Biomedical Informatics. 2019 Jan;89:1–10.
128. Collins M. Updated algorithm for the PubMed best match sort order. NLM Tech Bull [Internet]. 2017;414:e3.
129. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc. 2006 Feb;13(1):96–105.
130. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems. 1998;30(1–7):107–17.
131. Bian J, Morid MA, Jonnalagadda S, Luo G, Del Fiol G. Automatic identification of high impact articles in PubMed to support clinical decision making. J Biomed Inform. 2017 Sep;73:95–103.
132. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent data analysis. 2002;6(5):429–49.
133. Krawczyk B, Woźniak M, Schaefer G. Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing. 2014 Jan;14:554–62.
134. Zhou ZH, Liu XY. ON MULTI-CLASS COST-SENSITIVE LEARNING. Computational Intelligence. 2010 Jul 27;26(3):232–57.
135. Hasanzad M, Aghaei Meybodi HR, Sarhangi N, Larijani B. Artificial intelligence perspective in the future of endocrine diseases. J Diabetes Metab Disord. 2022 Jun;21(1):971–8.
136. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017 Jan;106(1):1–9.
137. Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. Kidney International. 2020 Apr;97(4):676–86.
138. Shen F, Zhao Y, Wang L, Mojarad MR, Wang Y, Liu S, et al. Rare disease knowledge enrichment through a data-driven approach. BMC Med Inform Decis Mak. 2019 Dec;19(1):32.
139. Ontobio. Ontobio Ontology Library [Internet]. [cited 2020 Feb 7]. Available from: https://ontobio.readthedocs.io/en/latest/index.html

140. Köhler S, Øien NC, Buske OJ, Groza T, Jacobsen JOB, McNamara C, et al. Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics. Current Protocols in Human Genetics [Internet]. 2019 Sep [cited 2022 Jan 14];103(1). Available from: https://onlinelibrary.wiley.com/doi/10.1002/cphg.92
141. Skim. Skim App [Internet]. Available from: https://skim-app.sourceforge.io
142. Singh V. Replace or Retrieve Keywords In Documents at Scale. arXiv:171100046 [cs] [Internet]. 2017 Nov 9 [cited 2021 Dec 8]; Available from: http://arxiv.org/abs/1711.00046
143. Honnibal, Matthew, Montani, Ines. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
144. Williamson KA, FitzPatrick DR. The genetic architecture of microphthalmia, anophthalmia and coloboma. Eur J Med Genet. 2014 Aug;57(8):369–80.
145. Lines M, Hartley T, MacDonald SK, Boycott KM. Mandibulofacial Dysostosis with Microcephaly. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2022 Jan 23]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK214367/
146. Balasubramanian M, Schirwani S. ASXL3-Related Disorder. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2022 Jan 23]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK563693/
147. Contreras-Capetillo SN, Vilchis-Zapata ZH, Ribbón-Conde J, Pinto-Escalante D. Global developmental delay and postnatal microcephaly: Bainbridge-Ropers syndrome with a new mutation in ASXL3. Neurologia (Engl Ed). 2018 Sep;33(7):484–6.
148. Dinwiddie DL, Soden SE, Saunders CJ, Miller NA, Farrow EG, Smith LD, et al. De novo frameshift mutation in ASXL3 in a patient with global developmental delay, microcephaly, and craniofacial anomalies. BMC Med Genomics. 2013 Sep 17;6:32.
149. Taylor CM, Smith R, Lehman C, Mitchel MW, Singer K, Weaver WC, et al. 16p11.2 Recurrent Deletion. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2022 Jan 23]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK11167/
150. McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JAS, et al. 22q11.2 deletion syndrome. Nat Rev Dis Primers. 2015 Nov 19;1:15071.
151. Karakülah G, Dicle O, Koşaner O, Suner A, Birant ÇC, Berber T, et al. Computer based extraction of phenoptypic features of human congenital anomalies from the digital literature with natural language processing techniques. Stud Health Technol Inform. 2014;205:570–4.

152. Bamshad MJ, Nickerson DA, Chong JX. Mendelian Gene Discovery: Fast and Furious with No End in Sight. Am J Hum Genet. 2019 Sep 5;105(3):448–55.
153. Kafkas Ş, Hoehndorf R. Ontology based text mining of gene-phenotype associations: application to candidate gene prediction. Database [Internet]. 2019 Jan 1 [cited 2022 Jan 30];2019. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baz019/5365528
154. Li S, Guo Z, Ioffe JB, Hu Y, Zhen Y, Zhou X. Text mining of gene–phenotype associations reveals new phenotypic profiles of autism-associated genes. Sci Rep. 2021 Dec;11(1):15269.
155. Chen C. Python Implementation of Rank Biased Overlap [Internet]. [cited 2021 Jun 1]. Available from: https://pypi.org/project/rbo/
156. National Library of Medicine. Fielded MetaMap Indexing (MMI) Output Explained [Internet]. 2015 [cited 2021 Nov 30]. Available from: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/MMI_Output_2016.pdf
157. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020 Sep 17;585(7825):357–62.
158. Hypothesis. Hypothesis annotation software [Internet]. [cited 2021 Mar 9]. Available from: https://web.hypothes.is
159. van den Berg C. Super Fast String Matching in Python [Internet]. [cited 2021 Mar 12]. Available from: https://bergvca.github.io/2017/10/14/super-fast-string-matching.html
160. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. arXiv:12010490 [cs] [Internet]. 2018 Jun 5 [cited 2021 Dec 28]; Available from: http://arxiv.org/abs/1201.0490
161. Kaur S, Christodoulou J. MECP2 Disorders. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2022 Feb 2]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1497/
162. Campbell J. Automated PubMed Search script [Internet]. Available from: https://github.com/biomedicalinformaticsgroup/gene2phenotype/blob/main/1.gene_search_to_pmids.ipynb
163. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422–3.
164. Campbell J. Automated PubMed Metadata script [Internet]. Available from: https://github.com/biomedicalinformaticsgroup/gene2phenotype/blob/main/2.pmids_to_medline_to_df.ipynb
165. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009 Jul 10;10:213.

166. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018 Nov 1;34(21):3711–8.
167. Pham B, Jovanovic J, Bagheri E, Antony J, Ashoor H, Nguyen TT, et al. Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow. Syst Rev. 2021 Dec;10(1):156.
168. Koras K, Juraeva D, Kreis J, Mazur J, Staub E, Szczurek E. Feature selection strategies for drug sensitivity prediction. Sci Rep. 2020 Jun 10;10(1):9377.
169. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus. 2016;5(1):1608.
170. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016 Nov;5(4):221–32.
171. van Der Burgt I, Brunner H. Genetic heterogeneity in Noonan syndrome: evidence for an autosomal recessive form. Am J Med Genet. 2000 Sep 4;94(1):46–51.
172. Holub K, Hardy N, Kallmes K. Toward Automated Data Extraction According to Tabular Data Structure: Cross-sectional Pilot Survey of the Comparative Clinical Literature. JMIR Form Res. 2021 Nov 24;5(11):e33124.
173. Khusro S, Latif A, Ullah I. On methods and tools of table detection, extraction and annotation in PDF documents. Journal of Information Science. 2015 Feb;41(1):41–57.
174. Anand R, Paik HY, Wang C. Integrating and querying similar tables from PDF documents using deep learning. arXiv:190104672 [cs] [Internet]. 2019 Jan 15 [cited 2022 Feb 4]; Available from: http://arxiv.org/abs/1901.04672
175. Aitken S, Firth HV, McRae J, Halachev M, Kini U, Parker MJ, et al. Finding Diagnostically Useful Patterns in Quantitative Phenotypic Data. The American Journal of Human Genetics. 2019 Nov;105(5):933–46.
176. Ross KE, Natale DA, Arighi C, Chen SC, Huang H, Li G, et al. Scalable Text Mining Assisted Curation of Post-Translationally Modified Proteoforms in the Protein Ontology. CEUR Workshop Proc. 2016 Aug;1747:http://ceur-ws.org/Vol-1747/BIT103_ICBO2016.pdf.
177. Wang J, Zhou X, Zhu J, Zhou C, Guo Z. Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinformatics. 2010 Dec;11(1):290.
178. Kyrimi E, McLachlan S, Dube K, Neves MR, Fahmi A, Fenton N. A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. Artif Intell Med. 2021 Jul;117:102108.
179. Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics. 2012 Oct 1;28(19):2502–8.
180. Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, et al. Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. The American Journal of Human Genetics. 2018 Jul;103(1):58–73.

181. Ganesan S, Galer PD, Helbig KL, McKeown SE, O'Brien M, Gonzalez AK, et al. A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. Genetics in Medicine. 2020 Dec;22(12):2060–70.
182. Hully M, Lo Barco T, Kaminska A, Barcia G, Cances C, Mignot C, et al. Deep phenotyping unstructured data mining in an extensive pediatric database to unravel a common KCNA2 variant in neurodevelopmental syndromes. Genetics in Medicine. 2021 May;23(5):968–71.
183. Slater LT, Karwath A, Williams JA, Russell S, Makepeace S, Carberry A, et al. Towards similarity-based differential diagnostics for common diseases. Computers in Biology and Medicine. 2021 Jun;133:104360.
184. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015 Feb 20;347(6224):1257601.
185. Aitken S, Firth HV, Wright CF, Hurles ME, FitzPatrick DR, Semple CA. IMPROVE-DD: Integrating Multiple Phenotype Resources Optimises Variant Evaluation in genetically determined Developmental Disorders [Internet]. Genetic and Genomic Medicine; 2022 May [cited 2022 Sep 23]. Available from: http://medrxiv.org/lookup/doi/10.1101/2022.05.20.22275135
186. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015 Dec;10(12):2004–15.