The visualization of Orphadata neurology phenotypes

Daniel B. Hier[1,2]*, Raghu Yelugam[1], Michael D. Carrithers[2] and Donald C. Wunsch III[3]

[1]Applied Computational Intelligence Laboratory, Department of Electrical & Computer Engineering, Missouri University of Science & Technology, Rolla, MO, United States, [2]Department of Neurology and Rehabilitation, University of Illinois at Chicago, Chicago, IL, United States, [3]National Institute of Diabetes and Digestive and Kidney Diseases, Liver Diseases Branch, Bethesda, MD, United States

Disease phenotypes are characterized by signs (what a physician observes during the examination of a patient) and symptoms (the complaints of a patient to a physician). Large repositories of disease phenotypes are accessible through the Online Mendelian Inheritance of Man, Human Phenotype Ontology, and Orphadata initiatives. Many of the diseases in these datasets are neurologic. For each repository, the phenotype of neurologic disease is represented as a list of concepts of variable length where the concepts are selected from a restricted ontology. Visualizations of these concept lists are not provided. We address this limitation by using subsumption to reduce the number of descriptive features from 2,946 classes into thirty superclasses. Phenotype feature lists of variable lengths were converted into fixed-length vectors. Phenotype vectors were aggregated into matrices and visualized as heat maps that allowed side-by-side disease comparisons. Individual diseases (representing a row in the matrix) were visualized as word clouds. We illustrate the utility of this approach by visualizing the neuro-phenotypes of 32 dystonic diseases from Orphadata. Subsumption can collapse phenotype features into superclasses, phenotype lists can be vectorized, and phenotypes vectors can be visualized as heat maps and word clouds.

Introduction

The signs and symptoms of a disease characterize its phenotype. In addition to signs (what a physician observes in a patient) and symptoms (the complaints of a patient), a clinical phenotype can include the age at the onset of a disease, its mode of onset, its rate of progression, its mode of inheritance, and its response to treatment. Some researchers include biochemical, radiological, electrophysiological, and biosensor findings as part of the disease phenotype (1–5). Large phenotype repositories are available on the internet. The On-Line Mendelian Inheritance in Man (OMIM) has over 9,500 disease profiles (6) and Orphadata has phenotype profiles of 4,245 rare diseases (7). The Human Phenotype Ontology (HPO) draws phenotype profiles from Orphadata and OMIM so that some genetic diseases have alternative profiles from each registry (8,9). All three repositories have sophisticated search engines that retrieve phenotype features by disease or gene (1). Phenotypic features are recorded as concepts (terms) from restricted vocabularies such as the Human Phenotype Ontology (20,246 terms) (10), or the Online Mendelian Inheritance of Man ontology (99,165 terms) (11).

Neuro-phenotypes

The June 2022 release of Orphadata lists 7,261 rare diseases, with 1,740 classified as rare neurological diseases (https://www.orphadata.com/linearisation/). Orphadata provides phenotype profiles on 1,184 rare neurologic diseases (https://www.orphadata.com/phenotypes/). Neuro-phenotyping is the deep phenotyping of neurological disease (1). We have suggested that most neuro-phenotyping can be done with a restricted vocabulary of about 1,600 concepts (12). Although lists of phenotypic features for neurological diagnoses can be retrieved from Orphadata, OMIM, or HPO, these lists are difficult to visualize.

Visualizations of disease phenotypes have limitations

OMIM, Orphanet, and HPO yield lists of phenotype features of variable length, sorted by alphabetical order, feature frequency, or body system. For example, the Orphadata annotations for Dystonia Type 13 (DYT13) are:

Very frequent
- stereotypy
- torsion dystonia
- torticollis

Frequent
- limb dystonia
- dystonia
- craniofacial dystonia
- jerky head movements

Occasional
- postural tremor
- action tremor
- focal dystonia

Rare
- Generalized dystonia
- Hoarse voice

Although useful, these lists have limitations. The lists may be long. In the Orphanet dataset, 25% of the lists are more than 34 features in length. Many of these lengths are beyond the length of 7 ± 2 that is easily comprehended (13). Side-by-side comparisons of lists are difficult (Table 1). Lists of signs and symptoms from Orphadata may contain pathologies (e.g., gliosis, Lewy bodies), radiological findings (e.g., abnormal PET FDG), biochemical findings, electrophysiological findings, and modes of inheritance. Although terms in Orphadata are from the HPO-controlled vocabulary (20,246 classes) (10), redundancies, near-synonyms, hypernyms, and hyponyms populate the lists (e.g., dysarthria and slow slurred speech; bradykinesia and hypokinesia; masked facies and hypomimia, etc.) Furthermore, OMIM, Orphadata, and HPO do not provide native methods for visualization of phenotype.

Prior work

Limited work has been done on visualizing phenotype lists retrieved from HPO, OMIM, or Orphadata. Xu et al. (14) visualized the distances between genetic diseases and their underlying phenotypes using t-SNE (stochastic neighborhood embedding) maps. The phenotype features from the OMIM dataset were used to calculate distances between genetic diseases. The t-SNE maps are a 2-dimensional representation of the distances between genetic diseases derived from multi-dimensional data. Although these t-SNE maps provide instructive information about the distances between genetic diseases, they do not reveal the details of the underlying phenotypes. Network analysis and network graphs have been used to visualize the distances between diseases based on their phenotype (15–17). However, these network diagrams do not elucidate the underlying phenotypic differences between the diseases. Several methods have been proposed to visualize disease-phenotype relationships, including radar graphs (18), co-occurrence charts (19), and sunburst diagrams (20). Cao et al. have developed visualization techniques called DICON, FacetAtlas, and SolarMap that show promise for visualizing phenotype features by disease (21–24).

An additional barrier to visualizing neurology phenotype profiles is the large number of terms in the HPO ($N = 20,390$), making the number of columns in heat maps or tables impractical. A feature reduction strategy that chunks phenotype features into a more manageable number of superclasses is needed. For example, Hier and Pearson (25) have suggested chunking problems in the electronic health record by body system to increase the readability of the problem list. Both OMIM and HPO chunk phenotype features by body system. Orphanet chunks phenotype features by feature frequency (common to rare). Yauy et al. (26) have chunked 16,600 phenotypic traits into 390 interacting symptom groups. However, the chunking of phenotype features by body system is unlikely to yield useful visualizations because dissimilar phenotypic features are grouped together. For example, chunking concepts by a nervous system category would put the unlike concepts of hypertonia, hypotonia, hyperreflexia, and hyporeflexia into the same category, a grouping of little diagnostic value. Although the chunking of phenotype concepts by body system or other schemes helps organize phenotype features, it does not reduce the number of features. Since the HPO is a hierarchical containment ontology, we have suggested that subsumption can create superclasses of phenotypic features and reduce the number of features (27,28).

Proposed approach and use case

We propose to improve the visualization of neurology phenotypes in the Orphdata dataset utilizing a combination of subsumption, vectorization, heat maps, and word clouds.

As proof of concept, we illustrate the utility of this approach with a use case that visualizes the phenotype lists of 32 dystonic diseases from Orphadata. In 1911 Oppenheim described the disease *dystonia musculorum deformans* and coined the term dystonia (29). Albanese et al. (30) defined dystonia as "a rare movement disorder characterized by sustained or intermittent muscle contractions causing abnormal, often repetitive movements, postures, or both." Since the description of dystonia by Oppenheim, many forms of dystonia have been described. Dystonia is classified along two axes: clinical and etiologic (30).

TABLE 1 The upper half of the Table shows lists of signs and symptoms for each dystonic disease from Orphadata.

Dystonias →	DYT16	DYT6	DYT27	DYT4
List of classes↓	Dystonia 16	Dystonia 6	Dystonia 27	Dystonia 4
	Limb dystonia	Dystonia	Oral dystonia	Blepharospasm
	Torticollis	Generalized dystonia	Laryngeal dystonia	Dementia
	Dysarthria	Dysarthria	Action tremor	Dysphagia
	Parkinsonism	Torticollis	Writer's cramp	Dysphonia
	Hyperreflexia	Blepharospasm	Limb dystonia	Generalized dystonia
	Dysphonia	Laryngeal dystonia	Axial dystonia	Laryngeal dystonia
	Dysphagia	Craniofacial dystonia	Focal dystonia	Abnl tongue movement
	Bradykinesia	Lingual dystonia	Postural tremor	Open mouth
	Postural tremor	Limb dystonia		Torticollis
	Orofacial dyskinesia			Gait disturbance
	Unsteady gait			Eunuchoid habitus
	Pyramidal sign			Sunken cheeks
	Lower limb pain			Involuntary movements
	Motor delay			Kyphoscoliosis
	Intellectual disability			Dysdiadochokinesis
				Respiratory distress
				Postural Tremor
Vector of superclasses ↓	Subsumption and vectorization ↓	Subsumption and vectorization ↓	Subsumption and vectorization ↓	Subsumption and vectorization ↓
Ataxia	0	0	0	1
Cognitive	1	0	0	1
Cranial nerve	1	0	0	2
Gait	1	0	0	1
Hyperkinesia	3	8	7	6
Hyperreflexia	1	0	0	0
Hypokinesia	2	0	0	0
Miscellaneous	2	0	0	4
Pain	1	0	0	0
Speech	2	1	0	1
Tremor	1	0	1	1

In the lower half of the Table, lists of classes have been converted to vectors of superclasses using subsumption governed by a lookup table. Counts are the number of times each class occurs in the superclass and is the input for the row values for the heat maps. Columns from the top half are variable length lists; columns from the bottom half are fixed length vectors.

Clinical classification is by age at onset, body distribution, the temporal pattern of symptoms, and associated phenotype features. Etiologic classification is by genetic versus non-genetic causation. Dystonia is one of the hyperkinetic movement disorders which also encompasses chorea, athetosis, hemiballismus, tics, tremors, stereotypy, myoclonus, and dyskinesia (31). Although all diseases labeled dystonia have a core symptom of dystonia, there is considerable variability in the clinical presentation (signs and symptoms) of the dystonias (29,32,33), making it an excellent use case for phenotype visualization. Furthermore, better characterization and classification of the dystonias is a major initiative of the European Reference Network for Rare Diseases, and Orphadata (34,35).

We downloaded the most recent Orphadata file with phenotype annotations of 4,254 rare diseases, including 1,184 rare neurological diseases. We identified 2,946 unique HPO terms used to characterize the signs and symptoms of rare neurological diseases and created a lookup table to map each term to one of 30 superclasses based on subsumption and expert opinion. The lists of phenotypic features for 32 dystonic diseases from Orphadata were converted into 31-element vectors, with the first element of the vector being the disease name and the next 30 elements being the count of features (signs and symptoms) for each superclass. The full 32-row × 31-column matrix of the dystonic diseases can be visualized as a feature map (**Figure 2**); individual rows can be visualized as word clouds (**Figure 3B**).

Methods

Phenotype feature lists by disease (data acquisition)

An XML file with 4,254 rare disease disorders and 112,256 phenotypic annotations was downloaded (June 2022 release of Orphadata: (https://www.orphadata.com/phenotypes/). Phenotype features are coded using the HPO ontology. Orphadata defines a rare disease as affecting less than 1 in 2,000 individuals in Europe and classifies 1,184 of the diseases as rare neurological diseases. We used python to parse the XML file and create a variable-length list of phenotypic features for each disease. We retained phenotypic annotations that were clinical signs or symptoms and filtered out phenotypic annotations related to disease course (progressive, static, etc.), mode of inheritance (recessive, dominant, etc.), biochemical abnormality, radiological abnormality, pathological abnormality, or electrophysiological abnormality. Based on published literature, Orphadata classifies the frequency of each phenotypic feature from rare (1–4%) to always present (100%). We retained phenotypic features classified as occasional or higher (5–100%).

Lookup table to convert phenotype classes to superclasses (subsumption)

The HPO (10) is organized as a hierarchical subsumption ontology so that more-specific concepts in the ontology are subsumed by more general concepts (28). We identified 2,946 unique concepts that Orphadata used to phenotype neurological diseases. We collapsed these concepts into 30 superclasses using subsumption and domain expert opinion. Example class memberships and class counts are shown for each superclass below.

1. alertness (53 terms) delirium, drowsy, somnolence
2. ataxia (62 terms) asynergia, clumsiness, dystaxia
3. atrophy (69 terms) muscle atrophy, atrophy, limb fasciculations
4. behavior (238 terms) apathy, anxiety, delusions
5. cognitive (202 terms) agnosia, apraxia, forgetfulness
6. cranial nerve (203 terms) ageusia, hyperacusis, facial diplegia
7. dysautonomia (35 terms) hypohidrosis, orthostatic syncope, dysautonomia
8. eye movements (272 terms) upgaze palsy, nystagmus, hypometric saccades
9. fatigue (26 terms) muscle fatigue, fatigable weakness, fatigue
10. gait (110 terms) ataxic gait, falls, unsteady gait
11. head (263 terms) microcephaly, macrocephaly, increased head size
12. hyperkinesia (157 terms) dyskinesia, dystonia, hyperkinesia
13. hyperreflexia (58 terms) increased reflexes, clonus, hyperreflexia
14. hypertonia (58 terms) increased muscle tone, rigidity, spasticity
15. hypokinesia (66 terms) bradykinesia, akinesia, hypomimia
16. hyporeflexia (43 terms) areflexia, hyporeflexia, absent ankle reflex
17. hypotonia (19 terms) decreased tone, muscle flaccidity, limb hypotonia
18. other muscle (119 terms) myokymia, muscle hypertrophy, myotonia
19. neck (48 terms) stiff neck, neck rigidity, meningismus
20. pain (145 terms) pain, arm pain, allodynia
21. seizure (358 terms) seizure, tonic-clonic seizure, febrile seizure
22. sensory (192 terms) hyperesthesia, dysesthesia, hypesthesia
23. skin (194 terms) cafe au lait spots, petechiae, rash
24. sleep (48 terms) cataplexy, narcolepsy, hypersomnia
25. speech_language (116 terms) dysarthria, aphasia, echolalia
26. sphincter (67 terms) urinary incontinence, constipation, enuresis
27. tremor (48 terms) tremor, resting tremor, action tremor
28. vision (450 terms) achromatopsia, scotoma, optic atrophy
29. weakness (159 terms) proximal weakness, foot drop, triceps weakness
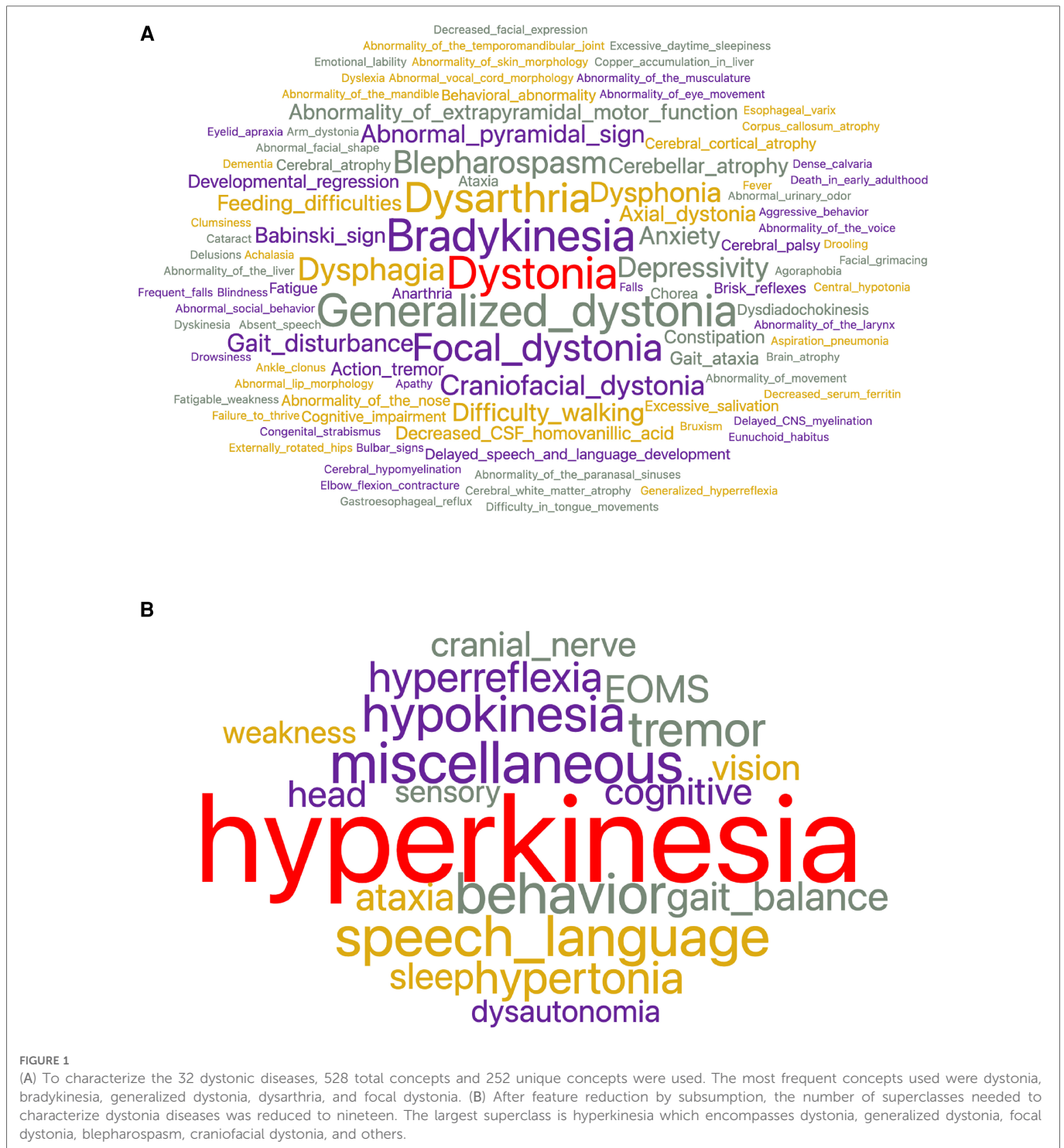30. miscellaneous (618 terms) nausea, vomiting, bradycardia

We used python to assign each phenotypic feature (sign or symptom) to one of the thirty superclasses based on the lookup table (see Table 1 for an illustration of how individual phenotype features were mapped to superclasses). The lookup table is available in the Supplementary Materials.

Vectorization (conversion of phenotype lists to phenotype vectors)

Variable-length lists of phenotypic features were converted into vectors of fixed length 31 elements. The first element of the list was the disease label, and the following 30 elements were the counts of features in each of the 30 superclasses based on the lookup table. When the phenotype is represented as a vector, phenotypes can be compared by distance metrics. Furthermore, the magnitude of each element in the phenotype vector carries additional information that allows comparisons between diseases. For example, one disease with hyperkinetic features dystonia, chorea, and athetosis would have a hyperkinesia superclass value of $n = 3$, whereas a disease with only dystonia would have a hyperkinesia superclass value of $n = 1$. Such weightings could be useful in distinguishing between phenotypes of similar diseases.

Visualization (creation of heat maps and word clouds

Heat maps and word clouds were based on the phenotype vectors generated by python. Heat maps were created using the *heat map widget* from Orange (36). The score mapped for each superclass was the count of the phenotype features subsumed by that class. When a superclass had no features assigned to it, that superclass was dropped from the heat map. Word clouds were produced using the *word cloud widget* from Orange. Word size in the word cloud reflected the frequency of phenotypic features for a group of diseases (Figure 1B) or a single disease (Figure 3B).

FIGURE 1

(A) To characterize the 32 dystonic diseases, 528 total concepts and 252 unique concepts were used. The most frequent concepts used were dystonia, bradykinesia, generalized dystonia, dysarthria, and focal dystonia. (B) After feature reduction by subsumption, the number of superclasses needed to characterize dystonia diseases was reduced to nineteen. The largest superclass is hyperkinesia which encompasses dystonia, generalized dystonia, focal dystonia, blepharospasm, craniofacial dystonia, and others.

Results

As our use case, we examined the phenotype profiles of 32 disease variants of dystonia in Orphadata. Phenotype profiles were lists of features (see **Table 1** for examples of DYT4, DYT6, DYT16, and DYT27). Feature lists ranged from 5 to 48 elements, with a mean of 18.4 features ± 10.5. The 252 unique features in the phenotype lists were reduced by subsumption into one of the 19 available 30 superclasses (**Table 1** and **Figure 1A,B**). This allowed

visualization of the entire dystonia disease set of 32 variants as a heat map (**Figure 2**). This heat map allows an easy distinction of pure dystonia (e.g., DYT25 and DYT26) from dystonias with sensory loss (e.g., autosomal dominant dopa-responsive dystonia), cognitive impairment (e.g., DYT4) and hypokinesia (e.g., adult-onset dystonia-parkinsonism). Individual rows in the heat map (**Figure 3A** can be further visualized with word clouds which emphasize phenotypic differences between the dystonia variants (see **Figure 3B** for word clouds of DYT4, DY6, DYT16, and DYT 27.)
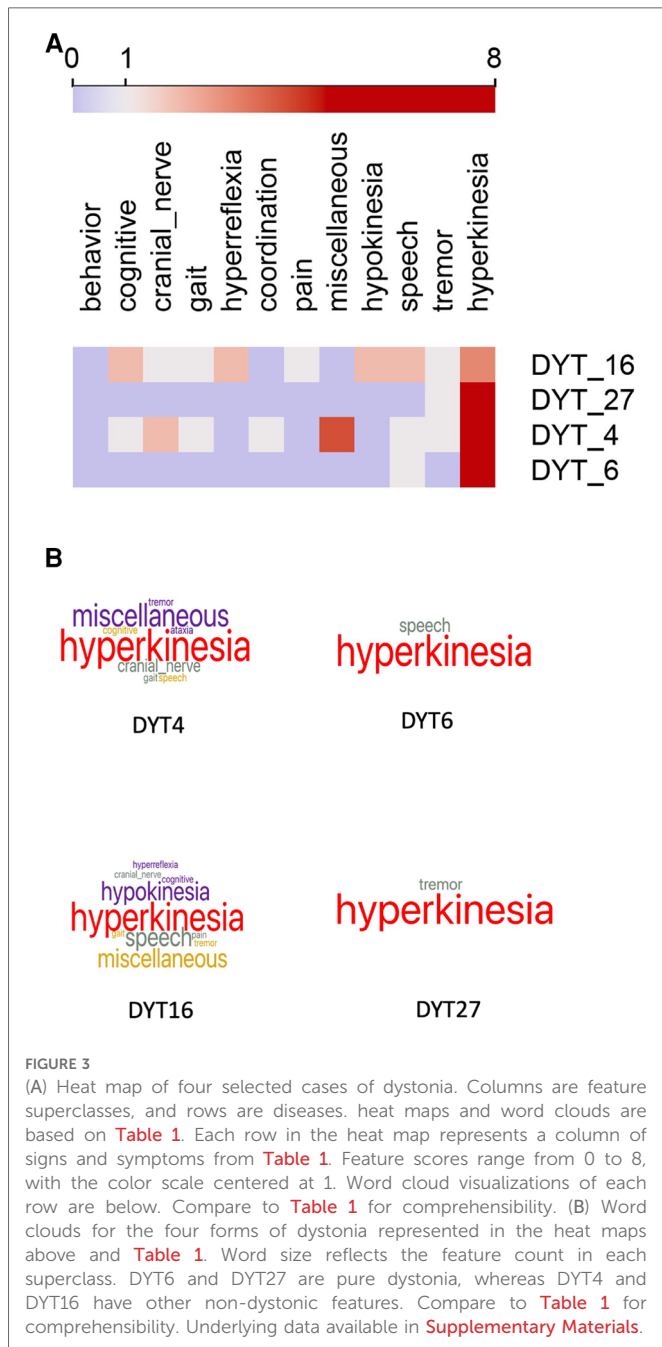
FIGURE 2

Feature map of 32 dystonias from Orphadata. Each row is a different variant of dystonia. Each column is one of 19 phenotype superclasses. Counts in columns range from 0 to 8. The color scale is centered at 1. Rows and columns are clustered by hierarchical clustering with Ward linkage. Distances between columns are by Pearson correlation coefficient. Distances between rows are by Euclidean distance. Hyperkinesia is the most frequent feature, followed by tremor, behavior, hypokinesia, speech_language, and miscellaneous (See word cloud in Figure 1B). Data underlying this table is available in the Supplementary Materials.

Discussion

Rich and detailed information on the phenotypes of neurological diseases is held in online repositories such as OMIM, HPO, and Orphadata. Detailed phenotypic data is available for download and can be used to gain insights into the inter-relationships between genes, disease, and phenotypes. Nonetheless, the visualization of the phenotypes retrieved as lists remains problematic. We identified several limitations to the visualization of disease phenotypes that included:

1. Phenotype feature lists are long.
2. Too many of the phenotype features are near synonyms, hyponyms, or hypernyms.
3. The number of unique features is large.

4. Side-by-side comparisons of phenotypes are difficult.
5. Phenotype lists of signs and symptoms are co-mingled with radiological, pathological, biochemical, and electrophysiological findings.

To address these limitations, we proposed restricting our attention to visualizing the phenotypes of rare neurological diseases in Orphadata ($N = 1,184$). We mapped each of the 4,505 unique features used to describe signs and symptoms in Orphadata into one of 30 superclasses (see list in the Methods section). This allowed us to convert phenotype lists of variable length to vectors of fixed length (31 elements), in which the first element of the vector was the disease label and the next 30 elements were the count of features for each of the 30 superclasses. This process of converting a list to a vector is illustrated in Table 1 for DYT4,

FIGURE 3
(A) Heat map of four selected cases of dystonia. Columns are feature superclasses, and rows are diseases. heat maps and word clouds are based on Table 1. Each row in the heat map represents a column of signs and symptoms from Table 1. Feature scores range from 0 to 8, with the color scale centered at 1. Word cloud visualizations of each row are below. Compare to Table 1 for comprehensibility. (B) Word clouds for the four forms of dystonia represented in the heat maps above and Table 1. Word size reflects the feature count in each superclass. DYT6 and DYT27 are pure dystonia, whereas DYT4 and DYT16 have other non-dystonic features. Compare to Table 1 for comprehensibility. Underlying data available in Supplementary Materials.

DYT6, DYT16, and DYT27. Only 11 of the 30 superclasses were needed to represent these four dystonias. Once phenotype lists are converted to vectors, a group of diseases can be represented as a matrix. For example, 32 dystonic diseases from Orphadata can be converted to a matrix with 32 rows (each row a disease) and 20 columns (each column a superclass of phenotypic features plus one column for the disease label) and then visualized as a heat map (Figure 2). For easy readability, individual rows (diseases) in the heat maps can be converted to word clouds to visualize better the phenotype (Figure 3B).

We have addressed limitation (1) (long feature lists) by using subsumption to collapse 4,505 phenotypic classes into 30 neurological superclasses. This subsumption of numerous phenotypic features into 30 superclasses also addressed limitation

(2) (too many near-synonyms) and limitation (3) (too many unique features). Once phenotype lists of variable length are converted to vectors of fixed length, side-by-side comparisons of diseases become feasible through the use of heat maps and word clouds (Figures 3A,B); addressing limitation (4). Another advantage of vectorization is that it allows the calculation of distances between phenotypes using standard distance metrics such as cosine and Euclidean. Figure 2 demonstrates the clustering of rows (dystonic diseases) using the Euclidean distance. We filtered out biochemical, radiological, electrophysiological, and pathological features to address limitation (5) (thus, limiting the phenotype to signs and symptoms.)

This work has some significant limitations. First, collapsing granular phenotype features into superclasses by subsumption involves information loss. The superclasses retain no laterality information (left-sided versus right-sided weakness, etc.) The superclasses retain no topographical information (proximal versus distal weakness, etc.) The high information value of some granular phenotype features, such as impaired vertical gaze (a sign of progressive supranuclear palsy) or internuclear ophthalmoplegia (a sign of multiple sclerosis), is lost when the granular features are collapsed into the superclass of abnormal eye movements. Second, our current process of collapsing phenotype concepts into superclasses requires a manually constructed lookup table that assigns each concept to a superclass. Errors can be made in assigning concepts to superclasses. We are looking at ways to improve the subsumption process that collapses ontology concepts into superclasses. Third, heat map scales are non-linear. For each superclass score, we counted the number of features in that superclass. For example, a disease phenotype with the term *hemiparesis* would have a superclass score of 1 for weakness. In contrast, a disease phenotype with terms *arm weakness* and *leg weakness* would have a superclass score of 2. Furthermore, we did not weight phenotype features by importance. In building the features maps, a more general concept like *hyperreflexia* carries the same weight as a more limited concept such as *increased biceps reflex*. We are exploring whether normalization or other transformations of the underlying data would improve the utility of the heat maps. Fourth, the size and granularity of the superclasses were not uniform. For example, the vision superclass subsumed 450 concepts and had many different types of visual impairment, whereas the fatigue superclass subsumed only 26 concepts and reflected the concept of fatigue alone. Fifth, our selection of thirty superclasses was somewhat arbitrary and subject to modification. Although the selection of the thirty superclasses reflected domain expert opinion and the underlying structure of the ontologies, other useful partitions of the ontology into superclasses are possible. For example, chorea or dystonia could have been distinct superclasses instead of subsumed into hyperkinesia. Speech (e.g., dysarthria) and language disorders (e.g., aphasia) could have been separate superclasses. Sixth, the superclasses were restricted to neurological terms and neurological diseases. As a result, the heat maps will not be useful in visualizing the phenotypes of non-neurological diseases. Furthermore, the heat maps will not adequately visualize important non-neurological signs and symptoms of diagnostic value (such as Kayser-Fleisher rings for Wilson's disease (37)). Although true pathognomonic signs and

symptoms are rare in neurology (1,38–40), the heat maps lack the granularity to show pathognomonic signs. Furthermore, the current heat maps do not support a *drill down* to the underlying granular phenotype features. Although we used Orange to create the heat maps, suitable heat maps are also available in python, and R. Other heat map color schemes are available and may give better visualizations. The Orphadata phenotype datasets are undergoing revisions and improvements. Some diseases are phenotyped more completely than others. Although the dataset is curated, omissions, errors, and discrepancies can still occur. Finally, a similar analysis could have been done with phenotypic annotations from the OMIM or HPO datasets.

Despite these limitations, combining feature reduction by subsumption with vectorization of phenotype lists followed by visualization by heat maps and word clouds offers a robust method to explore neurology phenotypes. Subsumption permits the reduction of thousands of ontological concepts into a reduced number of phenotype superclasses. Vectorization allows the conversion of variable-length phenotype feature lists into superclass vectors of fixed length. Matrices of superclass vectors allow the side-by-side comparison of disease phenotypes as heat maps. Individual rows in the heat maps can be visualized with word clouds, providing an easy-to-grasp representation of a disease phenotype.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: www.orphadata.com/phenotypes/ and in Supplementary Materials.

Author contributions

Concept by DBH and RY. Data analysis by DBH and RY. Data interpretation by DBH, MDC, RY, and DCW III. Writing, revision, and approval by DBH, MDC, RY, and DCW III. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2023.1064936/full#supplementary-material

References

1. Hier D, Yelugam R, Azizi S, Wunsch III D. A focused review of deep phenotyping with examples from neurology. *Eur Sci J.* (2022) 18:4–19. doi: 10.19044/esj.2022.v18n4p4

2. Gupta AS. Digital phenotyping in clinical neurology. *Semin Neurol.* (2022) 42:48–59. doi: 10.1055/s-0041-1741495

3. Delude CM. Deep phenotyping: the details of disease. *Nature.* (2015) 527:S14–5. doi: 10.1038/527S14a

4. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping, genomic data. *Nature.* (2018) 562:203–9. doi: 10.1038/s41586-018-0579-z

5. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat.* (2012) 33:777–80. doi: 10.1002/humu.22080

6. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: online mendelian inheritance in man (OMIM®), an online catalog of human genes, genetic disorders. *Nucleic Acids Res.* (2015) 43:D789–98. doi: 10.1093/nar/gku1205

7. Maiella S, Rath A, Angin C, Mousson F, Kremp O. Orphanet and its consortium: where to find expert-validated information on rare diseases. *Rev Neurol.* (2013) 169 (Suppl 1):S3–S8. doi: 10.1016/s0035-3787(13)70052-3

8. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet.* (2015) 97:111–24. doi: 10.1016/j.ajhg.2015.05.020

9. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* (2014) 42:D966–74. doi: 10.1093/nar/gkt1026

10. Kohler S, Robinson P. Human Phenotype Ontology (2022). Available from: https://bioportal.bioontology.org/ontologies/HP

11. McKusick-Nathans Institute for Genetic Medicine. Online Mendelian Inheritance in Man (2022). Available from: https://bioportal.bioontology.org/ontologies/OMIM

12. Hier DB, Brint SU. A neuro-ontology for the neurological examination. *BMC Med Inform Decis Mak.* (2020) 20:1–9. doi: 10.1186/s12911-020-1066-7

13. Miller GA. The magic number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev.* (1956) 63:91–7. doi: 10.1037/h0043158

14. Xu W, Jiang X, Hu X, Li G. Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization. *BMC Med Genomics.* (2014) 7:1–9. doi: 10.1186/1755-8794-7-S2-S1

15. Emmert-Streib F, Tripathi S, Simoes RDM, Hawwa AF, Dehmer M. The human disease network: opportunities for classification, diagnosis, and prediction of disorders and disease genes. *Syst Biomed*. (2013) 1:20–8. doi: 10.4161/sysb.22816

16. Wei DH, Kang T, Pincus HA, Weng C. Construction of disease similarity networks using concept embedding and ontology. *Stud Health Technol Inform*. (2019) 264:442. doi: 10.3233/SHTI190260

17. Köhler S, Doelken SC, Rath A, Aymé S, Robinson PN. Ontological phenotype standards for neurogenetics. *Hum Mutat*. (2012) 33:1333–9. doi: 10.1002/humu.22112

18. Clementz BA, Trotti RL, Pearlson GD, Keshavan MS, Gershon ES, Keedy SK, et al. Testing psychosis phenotypes from bipolar–schizophrenia network for intermediate phenotypes for clinical application: biotype characteristics and targets. *Biol Psychiatry*. (2020) 5:808–18. doi: 10.1016/j.bpsc.2020.03.011

19. Glueck M, Gvozdik A, Chevalier F, Khan A, Brudno M, Wigdor D. Phenostacks: cross-sectional cohort phenotype comparison visualizations. *IEEE Trans Vis Comput Graph*. (2016) 23:191–200. doi: 10.1109/TVCG.2016.2598469

20. Glueck M, Hamilton P, Chevalier F, Breslav S, Khan A, Wigdor D, et al. Phenoblocks: phenotype comparison visualizations. *IEEE Trans Vis Comput Graph*. (2015) 22:101–10. doi: 10.1109/TVCG.2015.2467733

21. Cao N, Sun J, Lin YR, Gotz D, Liu S, Qu H. Facetatlas: multifaceted visualization for rich text corpora. *IEEE Trans Vis Comput Graph*. (2010) 16:1172–81. doi: 10.1109/TVCG.2010.154

22. Cao N, Gotz D, Sun J, Lin YR, Qu H. SolarMap: multifaceted visual analytics for topic exploration. *2011 IEEE 11th International Conference on Data Mining*. IEEE (2011). p. 101–10. Available from: https://doi.org/10.1109/ICDM.2011.135

23. Cao N, Gotz D, Sun J, Qu H. DICON: interactive visual analysis of multidimensional clusters. *IEEE Trans Vis Comput Graph*. (2011) 17:2581–90. doi: 10.1109/TVCG.2011.188

24. Gotz D, Sun J, Cao N. Multifaceted visual analytics for healthcare applications. *IBM J Res Dev*. (2012) 56:6–1. doi: 10.1147/JRD.2012.2199170

25. Hier DB, Pearson J. Two algorithms for the reorganisation of the problem list by organ system. *BMJ Health Care Inform*. (2019) 26:e100024. doi: 10.1136/bmjhci-2019-100024

26. Yauy K, Duforet-Frebourg N, Testard Q, Beaumeunier S, Audoux J, Simard B, et al. Learning phenotypic patterns in genetic disease by symptom interaction modeling. *medRxiv* (2022). Available from: https://doi.org/10.1101/2022.07.29.22278181

27. Wunsch III DC, Hier DB. Subsumption is a novel feature reduction strategy for high dimensionality datasets. *Eur Sci J*. (2022) 18:20–33. doi: 10.19044/esj.2022.v18n4p20

28. Wunsch DC, Hier DB. Subsumption reduces dataset dimensionality without decreasing performance of a machine learning classifier. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE (2021). p. 1618–21. Available from: https://doi.org/10.1109/EMBC46164.2021.9629897

29. Grütz K, Klein C. Dystonia updates: definition, nomenclature, clinical classification, and etiology. *J Neural Transm*. (2021) 128:395–404. doi: 10.1007/s00702-021-02314-2

30. Albanese A, Bhatia K, Bressman SB, DeLong MR, Fahn S, Fung VS, et al. Phenomenology, classification of dystonia: a consensus update. *Mov Disord*. (2013) 28:863–73. doi: 10.1002/mds.25475

31. Jankovic J. Treatment of hyperkinetic movement disorders. *Lancet Neurol*. (2009) 8:844–56. doi: 10.1016/s1474-4422(09)70183-8

32. Lange LM, Junker J, Loens S, Baumann H, Olschewski L, Schaake S, et al. Genotype–phenotype relations for isolated dystonia genes: MDSgene systematic review. *Mov Disord*. (2021) 36:1086–103. doi: 10.1002/mds.28485

33. di Biase L, Di Santo A, Caminiti ML, Pecoraro PM, Di Lazzaro V. Classification of dystonia. *Life*. (2022) 12:206. doi: 10.3390/life12020206

34. Centen LM, Pinter D, van Egmond ME, Graessner H, Kovacs N, Koy A, et al. Dystonia management across Europe within ERN-RND: current state, future challenges. *J Neurol*. (2022) 1–13. doi: 10.1007/s00415-022-11412-4. [Epub ahead of print]

35. Graessner H, Brunelle A, Reinhard C, Hermanns S, Post A. European reference network for rare neurological diseases–ERN-RND. *Information Brochure* (2020). Available from: https://www.ern-rnd.eu/

36. Demšar J, Curk T, Erjavec A, Č G, Hočevar T, Milutinovič M, et al. Orange: data mining toolbox in python. *J Mach Learn Res*. (2013) 14:2349–53. doi: 10.5555/2567709.2567736

37. Finelli PF. Kayser-Fleischer ring: hepatolenticular degeneration (Wilson's disease). *Neurology*. (1995) 45:1261–2. doi: 10.1212/wnl.45.7.1261

38. Janeway EG. Limitations of pathognomonic signs and symptoms. *J Am Med Assoc*. (1884) 3:116–20. doi: 10.1001/jama.1884.02390540004001a

39. Barrows HS, Bennett K. The diagnostic (problem solving) skill of the neurologist: experimental studies and their implications for neurological training. *Arch Neurol*. (1972) 26:273–7. doi: 10.1001/archneur.1972.00490090099009

40. Chimowitz MI, Logigian EL, Caplan LR. The accuracy of bedside neurological diagnoses. *Ann Neurol*. (1990) 28:78–85. doi: 10.1002/ana.410280114