1 Supplementary Method S1

1.1 Skip-gram

Formally, given a graph $G = (V, E)$ where V is a set of vertices (also known as nodes), and E is a set of paired vertices (also known as edges), let f be the mapping between each node $v \in V$ to its feature representation of size d where $f: V \rightarrow \mathbb{R}^d$. Here, *d* is a parameter that defines the dimension of the feature representation. For each node $v \in V$, $N_{S(v)} \subset V$ is a neighborhood with sampling strategy S, and the objective function maximizes the log-probability of reconstructing this neighborhood. Formally defined as:

$$\max_f \sum_{u \in V} \log P(N_s(u)|f(u)). \quad (S1)$$

Two assumptions are made to solve the optimization problem presented in **Equation S1** in polynomial time. First, the likelihood of observing a neighborhood node is independent of observing any other neighborhood node given the source node's feature representation, as:

$$P\big(N_S(u)\big|f(u)\big) = \prod_{u \in N_S(u)} P(v|f(u)). \quad (S2)$$

Second, a source node and neighbor node have a symmetric effect over each other; thus, their probabilities can be parameterized as the Softmax normalized inner product and can be written as:

$$P\big(v\big|f(u)\big) = \frac{\exp\,(f(v).f(u))}{\sum_{x \in V} \exp(f(x)\,.\,f(u))}. \quad (S3)$$

The objective function presented in **Equation S1** can be simplified with the conditional independence assumption and parameterization of the probabilities introduced in [1] as follows:

$$\max_f \sum_{u \in V} (-\log Z_u + \sum_{v\, \in N_S(u)} f(v).f(u)\,). \quad (S4)$$

Calculating the partition function for each node, $Z_u = \sum_{x \in V} exp(f(x).f(u))$, is not practical, especially for larger networks.

In Node2Vec, they approximate the objective function using the negative sampling algorithm introduced in Mikolov et al., 2013 [2]. Instead of calculating the probability of co-occurrence of all node pairs, the negative sampling algorithm attempts to increase the co-occurrence probability of the sample node with its neighbors and decrease that probability with k randomly selected nodes from the graph.

The simplified objective function presented in **Equation S4** can then find the d-dimensional representation of each node by employing the sampling strategy S and running Skip-gram with negative sampling.

Although Skip-gram revolutionized applications of NLP, it has two main drawbacks. (1) Skip-gram cannot capture polysemy, a word with multiple meanings, since it represents each word as a single vector. (2) It fails to identify compound word phrases. For example, the word "ice cream" should have a different representation than the words "ice" and "cream." As we use the HPO terms as a representation of phenotypic terms in our phenotype embedding method, we do not face the aforementioned drawbacks.

1.2 Sampling strategy

The Skip-gram model was developed for inputs of text, where a neighborhood of a word is a sliding window on its surrounding words in unstructured text. In order to make graph structures amenable to the Skip-gram model, Node2Vec introduces a biased randomized procedure that samples neighborhoods for each given node. Unlike previous studies such as [3], where the transition probability to the next node, v_n, only depends on the current node, v_c, i.e., $P(v_n|v_c)$, in a biased random walk the transition probability depends on both the current and the previous node, v_p, i.e., $P(v_n|v_c, v_p)$, formally defined as:

$$P(v_n|v_c, v_p) = \begin{cases} \dfrac{\alpha_{pq}(v_p, v_n)w(v_c, v_n)}{\sum_{v\in N(v)} \alpha_{pq}(v_p, v)w(v_c, v)} & \text{if } (v_c, v_n) \in E \\ 0 & \text{otherwise} \end{cases}, \quad \text{(S5)}$$

where $w(u, v)$ represents the edge weight between nodes u and v, $N(v)$ represents the sampled neighborhood for node v, and the bias factor α_{pq} is defined as:

$$\alpha_{pq}(v_p, v_n) = \begin{cases} \frac{1}{p} & \text{if } v_p = v_n \\ 1 & \text{if } v_p \neq v_n, (v_p, v_n) \in E \\ \frac{1}{q} & \text{if } v_p \neq v_n, (v_p, v_n) \neq E \end{cases}. \quad \text{(S6)}$$

Parameter p controls the likelihood of immediately revisiting a node in a biased random walk, where larger values, $>max(q, 1)$, encourage exploration in the graph. On the other hand, parameter q controls the exploitation criteria where larger values, >1, bias the random walk towards nodes that are closer to the previous node, v_p.

In the case of a weak connection between nodes v_p and v_n, i.e., $0 < w(v_p, v_n) \leq 1$, Node2Vec considers it a normal connection with a bias factor of 1 rather than $1/q$. This case shows that Node2Vec does not discriminate weak connections from stronger ones and could not detect cases where the potential next node has a loose connection with the previous node.

2 Supplementary Method S2

2.1 Example of HPO embedding of phenotypic terms

Generalized-onset seizure (HP:0002197) and *Motor seizure (HP:0020219)* are represented as the children of *Seizure (HP:0001250)* in the HPO with a cumulative frequency of 0.0134 and 0.1755, respectively, in our corpus. These frequencies imply a stronger connection between *Seizure (HP:0001250)* and *Motor seizure (HP:0020219)* than Seizure (HP:0001250) and *Generalized-onset seizure (HP:0002197)*, which could be captured by our weighting mechanism.

Furthermore, *Generalized-onset seizure (HP:0002197)* has two children, *Generalized non-motor (absence) seizure (HP:0002121)*, with a frequency of 0.0057, and *Generalized-onset motor seizure (HP:0032677)*, with a frequency of 0.0079, which is also a child of *Motor seizure (HP:0020219)*. *Epileptic spasm (HP:0011097)* is another child of *Motor seizure (HP:0020219)* with the frequency of 0.1685, which indicates a more robust representation of this phenotype in our corpus.

Based on the frequency of these nodes, *Motor seizure (HP:0020219)*'s connection is much stronger to its parent, *Seizure (HP:0001250)*, implying generalization in phenotypic concepts, with $w = 0.1773$, than to either of its children, implying specialization in phenotypic concepts, with w equal to 0.1702 and 0.0096 (**Figure 3**).

These weights suggest placing *Motor seizure (HP:0020219)* and *Seizure (HP:0001250)* closer in the embedding space, resulting in a higher similarity value than *Motor seizure (HP:0020219)* to either of its children, *Epileptic spasm (HP:0011097)* or *Generalized-onset motor seizure (HP:0002197)*. The similarity values align with these suggestions and are marked in gray in **Figure 3**.

3 Supplementary Figure S1

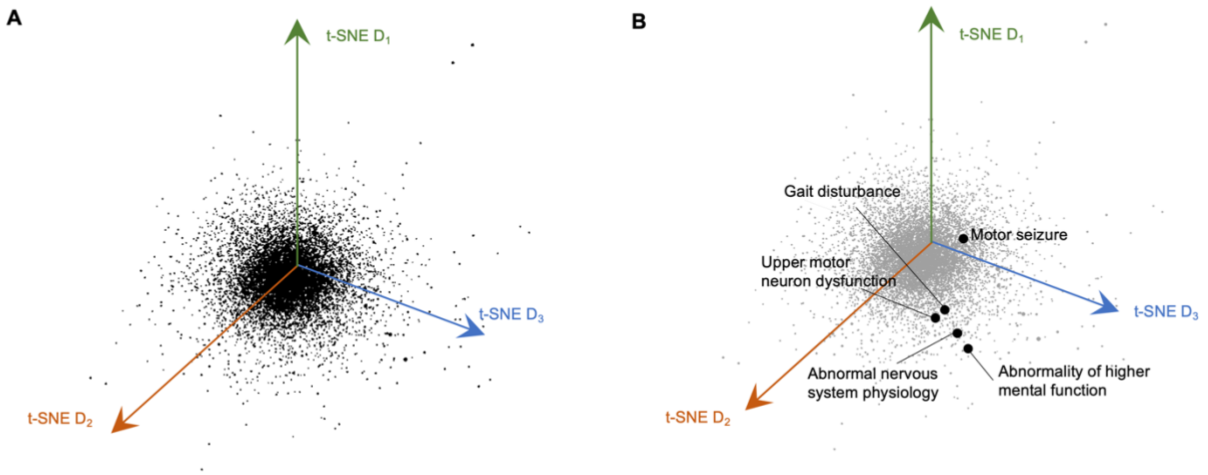

Figure S1. The HPO can be represented in a lower-dimensional space using the t-SNE algorithm, where similarities and differences between the phenotypes are preserved. (A) A 3D representation of all phenotypes in the embedding space using the t-SNE algorithm with 32 iterations. The axes are based on the first three t-SNE dimensions. We can see a change in the 3D representation compared to PCA 3D space (**Figure 5**). **(B)** The five closest phenotypes to *Seizure (HP:0001250)* are marked in the 3D space. While the closest phenotypes in PCA 3D space and t-SNE 3D space match, t-SNE requires hyper-parameter tuning, making it more costly compared to PCA.

4 Supplementary Figure S2

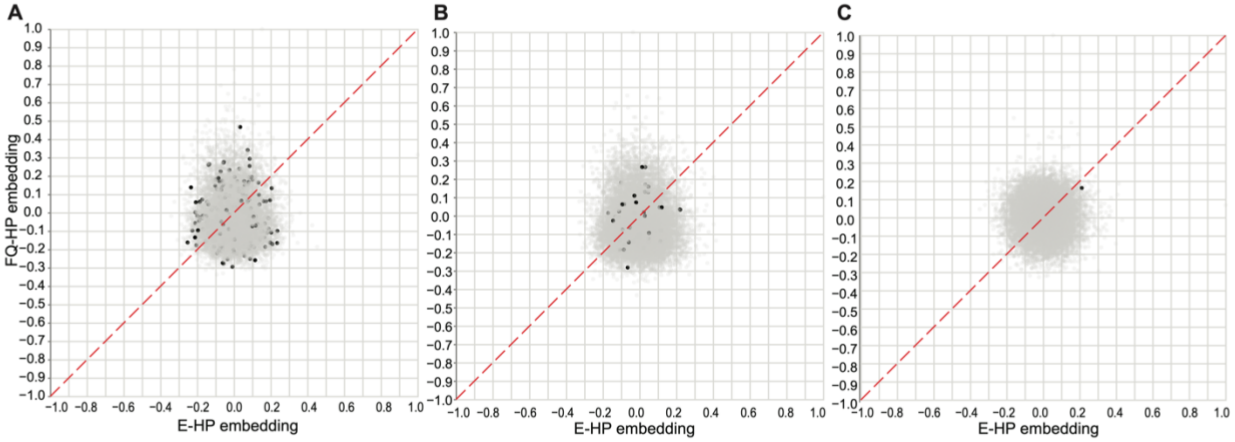

Figure S2. Different patterns of the cosine similarity value change using the E-HP and FQ-HP embedding. Pair-wise cosine similarity changes between a reference phenotype, **(A)** *Seizure (HP:0001250)*, **(B)** *Neurodevelopmental abnormality (HP:0012759)*, and **(C)** *Infantile spasms (HP:0012469)*, with all other phenotypes in the HPO using the FQ-HP and E-HP embedding methods. Points marked in black represent the cosine similarity values with phenotypes that are descendants of the reference phenotype. Points marked in grey are the cosine similarity values of non-descendant phenotypes.

5 Supplementary Method S3

We formally define Info Score as,

$$Info\ Score(\,m, \tau_1, \tau_2) = \frac{P_{99(\tau_1)} - P_1(\tau_1)}{P_{99}(\tau_2) - P_1(\tau_2)}\ , \qquad (S7)$$

where m is the main phenotype, and $P_k(\tau)$ is the k^{th} percentile of the pair-wise similarity values between m and all other phenotypes using the embedding technique τ. See Y and X in **Figure 6** representing the numerator and denominator of **Equation S7** with τ_1 = FQ-HP and τ_2 = E-HP, respectively. If the Info Score is very close to 1, there is no significant change in the cosine similarity values using the method τ_1 over τ_2. However, as an Info Score moves toward values smaller or greater than 1, it shows a significant change in the cosine similarity values and vector representations of the phenotypes.

6 Supplementary Table S1

#	Scenario	Reference HPO	Reference Name	Candidate1 HPO	Candidate1 Name	Candidate2 HPO	Candidate2 Name
1	S1	HP:0000708	Behavioral abnormality	HP:0007018	Attention deficit hyperactivity disorder	HP:0000736	Short attention span
2	S1	HP:0012638	Abnormal nervous system physiology	HP:0007185	Loss of consciousness	HP:0011443	Abnormality of coordination
3	S1	HP:0002011	Morphological central nervous system abnormality	HP:0001287	Meningitis	HP:0012503	Abnormality of the pituitary gland
4	S1	HP:0012443	Abnormality of brain morphology	HP:0100659	Abnormality of the cerebral vasculature	HP:0002060	Abnormal cerebral morphology
5	S1	HP:0009830	Peripheral neuropathy	HP:0000763	Sensory neuropathy	HP:0007021	Pain insensitivity
6	S1	HP:0001257	Spasticity	HP:0002313	Spastic paraparesis	HP:0006983	Slowly progressive spastic quadriparesis
7	S1	HP:0100852	Abnormal fear/anxiety-related behavior	HP:0000740	Episodic paroxysmal anxiety	HP:0025253	Claustrophobia
8	S1	HP:0012443	Abnormality of brain morphology	HP:0002139	Arrhinencephaly	HP:0002951	Partial absence of cerebellar vermis
9	S1	HP:0100851	Abnormal emotion/affect behavior	HP:0040082	Happy demeanor	HP:0031588	Unhappy demeanor
10	S1	HP:0009830	Peripheral neuropathy	HP:0007267	Chronic axonal neuropathy	HP:0002495	Impaired vibratory sensation
11	S2	HP:0001289	Confusion	HP:0031258	Delirium	HP:0002060	Abnormal cerebral morphology
12	S2	HP:0002315	Headache	HP:0002076	Migraine	HP:0100033	Tics
13	S2	HP:0011446	Abnormality of higher mental function	HP:0031258	Delirium	HP:0001297	Stroke
14	S2	HP:0011446	Abnormality of higher mental function	HP:0002464	Spastic dysarthria	HP:0012444	Brain atrophy
15	S2	HP:0009830	Peripheral neuropathy	HP:0003477	Peripheral axonal neuropathy	HP:0040148	Cortical myoclonus
16	S3	HP:0012759	Neurodevelopmental abnormality	HP:0001297	Stroke	HP:0000763	Sensory neuropathy
17	S3	HP:0012638	Abnormal nervous system physiology	HP:0030692	Brain neoplasm	HP:0100006	Neoplasm of the central nervous system
18	S3	HP:0100852	Abnormal fear/anxiety-related behavior	HP:0000613	Photophobia	HP:0002870	Obstructive sleep apnea
19	S3	HP:0003269	Sudanophilic leukodystrophy	HP:0012164	Asterixis	HP:0025454	Abnormal CSF metabolite level
20	S3	HP:0004372	Reduced consciousness/confusion	HP:0007029	Cerebral berry aneurysm	HP:0010830	Impaired tactile sensation

#	Scenario	Reference HPO	Reference Name	Candidate1 HPO	Candidate1 Name	Candidate2 HPO	Candidate2 Name
21	S4	HP:0100851	Abnormal emotion/affect behavior	HP:0031589	Suicidal ideation	HP:0000737	Irritability
22	S4	HP:0009830	Peripheral neuropathy	HP:0003401	Paresthesia	HP:0003474	Sensory impairment
23	S4	HP:0001288	Gait disturbance	HP:0040083	Toe walking	HP:0031955	Antalgic gait
24	S4	HP:0012638	Abnormal nervous system physiology	HP:0001308	Tongue fasciculations	HP:0004372	Reduced consciousness/confusion
25	S4	HP:0011446	Abnormality of higher mental function	HP:0032588	Hand apraxia	HP:0004372	Reduced consciousness/confusion
26	S5	HP:0002011	Morphological central nervous system abnormality	HP:0001264	Spastic diplegia	HP:0012447	Abnormal myelination
27	S5	HP:0002527	Falls	HP:0001289	Confusion	HP:0001254	Lethargy
28	S5	HP:0001288	Gait disturbance	HP:0100022	Abnormality of movement	HP:0001257	Spasticity
29	S5	HP:0001289	Confusion	HP:0002315	Headache	HP:0012447	Abnormal myelination
30	S5	HP:0001276	Hypertonia	HP:0004372	Reduced consciousness/confusion	HP:0004305	Involuntary movements
31	S6	HP:0012651	Abasia	HP:0008153	Periodic hypokalemic paresis	HP:0010829	Impaired temperature sensation
32	S6	HP:0012657	Abnormal brain positron emission tomography	HP:0032812	Neonatal electro-clinical non-motor seizure	HP:0012082	Cerebellar Purkinje layer atrophy
33	S6	HP:0007159	Fluctuations in consciousness	HP:0002536	Abnormal cortical gyration	HP:0010549	Weakness due to upper motor neuron dysfunction
34	S6	HP:0100660	Dyskinesia	HP:0012096	Intracranial epidermoid cyst	HP:0002392	EEG with polyspike wave complexes
35	S6	HP:0012487	Cerebellopontine angle arachnoid cyst	HP:0003380	Decreased number of peripheral myelinated nerve fibers	HP:0032787	Focal impaired awareness sensory seizure
36	S6	HP:0000722	Obsessive-compulsive behavior	HP:0010636	Schizencephaly	HP:0032865	Myoclonic absence status epilepticus
37	S6	HP:0040197	Encephalomalacia	HP:0007206	Hemimegalencephaly	HP:0410014	Abnormality of ganglion
38	S6	HP:0012076	Borderline personality disorder	HP:0031947	Tongue tremor	HP:0011195	EEG with focal sharp slow waves
39	S6	HP:0040292	Left hemiplegia	HP:0031166	Eyelid myokymia	HP:0045052	Abnormality of the brachial nerve plexus
40	S6	HP:0002188	Delayed CNS myelination	HP:0012194	Episodic hemiplegia	HP:0040331	Focal hypointensity of cerebral white matter on MRI
41	S6	HP:0004614	Spina bifida occulta at S1	HP:0006930	Frontoparietal cortical dysplasia	HP:0030221	Sweet craving

#	Scenario	Reference HPO	Reference Name	Candidate1 HPO	Candidate1 Name	Candidate2 HPO	Candidate2 Name
42	S6	HP:0006989	Dysplastic corpus callosum	HP:0008757	Unilateral vocal cord paralysis	HP:0012046	Areflexia of upper limbs
43	S6	HP:0002549	Deficit in phonologic short-term memory	HP:0032664	Adversive status epilepticus	HP:0032821	Neonatal electro-clinical tonic seizure
44	S6	HP:0011099	Spastic hemiparesis	HP:3000061	Abnormality of infra-orbital nerve	HP:0007187	Focal lissencephaly
45	S6	HP:0002445	Tetraplegia	HP:0004423	Cranium bifidum occultum	HP:0006961	Jerky head movements
46	S6	HP:0008757	Unilateral vocal cord paralysis	HP:0007375	Abnormality of the septum pellucidum	HP:0032044	Decreased vigilance
47	S6	HP:0011182	Interictal epileptiform activity	HP:0010530	Palatal myoclonus	HP:0030890	Hyperintensity of cerebral white matter on MRI
48	S6	HP:0031629	Impaired tandem gait	HP:0032699	Focal cognitive seizure with dysgraphia/agraphia	HP:0032161	Coccidioidal meningitis
49	S6	HP:0002352	Leukoencephalopathy	HP:0003398	Abnormal synaptic transmission at the neuromuscular junction	HP:0032920	Focal impaired awareness manual automatism seizure
50	S6	HP:0006742	Congenital neuroblastoma	HP:0032393	Diffuse ribbon-like subcortical heterotopia	HP:0003009	Enhanced neurotoxicity of vincristine
51	S6	HP:0006850	Hypoplasia of the ventral pons	HP:0007330	Frontal encephalocele	HP:0041056	Hot cross bun sign
52	S6	HP:0002282	Gray matter heterotopia	HP:0012228	Tension-type headache	HP:0007206	Hemimegalencephaly
53	S6	HP:0002524	Cataplexy	HP:0002392	EEG with polyspike wave complexes	HP:0032777	Focal impaired awareness autonomic seizure with pallor/flushing
54	S6	HP:0000762	Decreased nerve conduction velocity	HP:0002516	Increased intracranial pressure	HP:0003384	Peripheral axonal atrophy
55	S6	HP:0011097	Epileptic spasm	HP:0000748	Inappropriate laughter	HP:0011158	Focal sensory seizure with auditory features
56	S6	HP:0007105	Infantile encephalopathy	HP:0032506	Alien limb phenomenon	HP:0011749	Adrenocorticotropic hormone excess
57	S6	HP:0031843	Bradyphrenia	HP:0006999	Basal ganglia gliosis	HP:0001305	Dandy-Walker malformation
58	S6	HP:0000745	Diminished motivation	HP:0002070	Limb ataxia	HP:0012898	Abnormal lower-limb motor evoked potentials
59	S6	HP:0011960	Substantia nigra gliosis	HP:0010626	Anterior pituitary agenesis	HP:0006790	Cerebral cortex with spongiform changes
60	S6	HP:0003406	Peripheral nerve compression	HP:0031589	Suicidal ideation	HP:0100316	Hirano bodies

#	Scenario	Reference HPO	Reference Name	Candidate1 HPO	Candidate1 Name	Candidate2 HPO	Candidate2 Name
61	S6	HP:0012285	Abnormal hypothalamus physiology	HP:0011178	Alpha-EEG	HP:0031885	Hyperglycorrhachia
62	S6	HP:0006886	Impaired distal vibration sensation	HP:0002457	Abnormal head movements	HP:0007258	Severe demyelination of the white matter
63	S6	HP:0012171	Stereotypical hand wringing	HP:0003438	Absent Achilles reflex	HP:0001325	Hypoglycemic coma
64	S6	HP:0002200	Pseudobulbar signs	HP:0032886	Focal impaired awareness cognitive seizure with expressive dysphasia/aphasia	HP:0001067	Neurofibromas
65	S6	HP:0012486	Myelitis	HP:0011443	Abnormality of coordination	HP:0011400	Abnormal CNS myelination
66	S6	HP:0002313	Spastic paraparesis	HP:0032771	Focal autonomic seizure with lacrimation	HP:0030797	Reduced volume of central subdivision of bed nucleus of stria terminalis
67	S6	HP:0002118	Abnormality of the cerebral ventricles	HP:0030202	Favorable response of weakness to acetylcholine esterase inhibitors	HP:0011179	Beta-EEG
68	S6	HP:0010795	Cerebellar glioma	HP:0004336	Myelin outfoldings	HP:0030708	Myeloschisis
69	S6	HP:0003388	Easy fatigability	HP:0045007	Abnormal substantia nigra morphology	HP:0006960	Choroid plexus calcification
70	S6	HP:0032892	Infection-related seizure	HP:0011291	EEG with central sharp slow waves	HP:0001285	Spastic tetraparesis
71	S6	HP:0025170	Neuronal/glioneuronal neoplasm of the central nervous system	HP:0009733	Glioma	HP:0011184	EEG with hyperventilation-induced generalized epileptiform discharges
72	S6	HP:0002323	Anencephaly	HP:0002143	Abnormality of the spinal cord	HP:0032681	Focal aware cognitive seizure
73	S6	HP:0006957	Loss of ability to walk	HP:0006863	Severe expressive language delay	HP:0032046	Focal cortical dysplasia
74	S6	HP:0100565	Hydromyelia	HP:0030048	Colpocephaly	HP:0030746	Intraventricular hemorrhage
75	S6	HP:0030858	Addictive behavior	HP:0012503	Abnormality of the pituitary gland	HP:0010625	Anterior pituitary dysgenesis
76	S6	HP:0000719	Inappropriate behavior	HP:0001483	Eye poking	HP:0007027	Poorly formed metencephalon
77	S6	HP:0002511	Alzheimer disease	HP:0100034	Motor tics	HP:0030217	Limb apraxia
78	S6	HP:0011167	Focal tonic seizure	HP:0010829	Impaired temperature sensation	HP:0020098	Herpes encephalitis
79	S6	HP:0001335	Bimanual synkinesia	HP:0005462	Calcification of falx cerebri	HP:0010528	Prosopagnosia

#	Scenario	Reference HPO	Reference Name	Candidate1 HPO	Candidate1 Name	Candidate2 HPO	Candidate2 Name
80	S6	HP:0200147	Neuronal loss in basal ganglia	HP:0007807	Optic nerve compression	HP:0032801	Focal impaired awareness cognitive seizure with memory impairment
81	S7	HP:0006846	Acute encephalopathy	HP:0012638	Abnormal nervous system physiology	HP:0000707	Abnormality of the nervous system
82	S7	HP:0032267	Empty delta sign	HP:0002143	Abnormality of the spinal cord	HP:0002011	Morphological central nervous system abnormality
83	S7	HP:0002600	Hyporeflexia of lower limbs	HP:0001265	Hyporeflexia	HP:0012638	Abnormal nervous system physiology
84	S7	HP:0002491	Spasticity of facial muscles	HP:0002493	Upper motor neuron dysfunction	HP:0011442	Abnormal central motor function
85	S7	HP:0007206	Hemimegalencephaly	HP:0012443	Abnormality of brain morphology	HP:0000707	Abnormality of the nervous system
86	S8	HP:0500089	Optic nerve sheath meningioma	HP:0002011	Morphological central nervous system abnormality	HP:0010553	Oculogyric crisis
87	S8	HP:0000750	Delayed speech and language development	HP:0012759	Neurodevelopmental abnormality	HP:0004423	Cranium bifidum occultum
88	S8	HP:0011451	Congenital microcephaly	HP:0000252	Microcephaly	HP:0100703	Tongue thrusting
89	S8	HP:0007354	Amyotrophic lateral sclerosis	HP:0002011	Morphological central nervous system abnormality	HP:0030218	Punding
90	S8	HP:0007360	Aplasia/Hypoplasia of the cerebellum	HP:0002011	Morphological central nervous system abnormality	HP:0000741	Apathy
91	S9	HP:0002372	Normal interictal EEG	HP:0001311	Abnormal nervous system electrophysiology	HP:0032689	Focal cognitive seizure with dissociation
92	S9	HP:0040326	Hypoplasia of the olfactory bulb	HP:0012639	Abnormal nervous system morphology	HP:0032784	Focal aware autonomic seizure with palpitations/tachycardia/bradycardia/asystole
93	S9	HP:0030303	Hypoplastic anterior commissure	HP:0030301	Abnormality of the anterior commissure	HP:0002134	Abnormality of the basal ganglia
94	S9	HP:0011174	Focal hyperkinetic seizure	HP:0007359	Focal-onset seizure	HP:0005788	Abnormal cervical myelogram
95	S9	HP:0012492	Cerebral artery stenosis	HP:0002011	Morphological central nervous system abnormality	HP:0010829	Impaired temperature sensation

#	Scenario	Reference HPO	Reference Name	Candidate1 HPO	Candidate1 Name	Candidate2 HPO	Candidate2 Name
96	S10	HP:0002893	Pituitary adenoma	HP:0010853	EEG with periodic lateralized epileptiform discharges	HP:0010661	Absence of the third cerebral ventricle
97	S10	HP:0000845	Growth hormone excess	HP:0005341	Autonomic bladder dysfunction	HP:0002470	Nonprogressive cerebellar ataxia
98	S10	HP:0032656	Febrile status epilepticus	HP:0011145	Symptomatic seizures	HP:0007097	Cranial nerve motor loss
99	S10	HP:0001483	Eye poking	HP:0002073	Progressive cerebellar ataxia	HP:0030180	Oppenheim reflex
100	S10	HP:0031097	Abnormal thyroid-stimulating hormone level	HP:0032914	Focal aware perseverative automatism seizure	HP:0025040	Thalamic edema

Table S1. List of the 100 trios used for evaluation against expert opinion.

7 Supplementary Method S4

7.1 Agreement level calculation

Referring to how we generated the dataset, each expert indicated their prioritization of candidate 1, candidate 2, or none (in case of uncertainty or tie). Thus, the minimum/chance level agreement is when only 5 out of the 13 experts (38.46%) have the same choice. When 6, 7, or 8 experts (≤61.53%) agree, we refer to it as “fair-level” agreement. In the case of an agreement between 9 or 10 experts (≤76.92%), we categorize the agreement level as “substantial.” Finally, we refer to it as “high-level” agreement if more than 11 experts (>76.92%) agree. **Figure 7B** (in the main manuscript) displays the distribution of agreement levels for the 100 trios.

8 Supplementary Table S2

Table S2 contains the frequency and propagated frequency of phenotypic terms in our corpus that have a propagated occurrence of 4 or more (see **Table_S2.csv**).

9 Supplementary Method S5

9.1 Time complexity of the patient similarity algorithms in a dynamic setting

Suppose we want to calculate the similarity score between individuals P_1 and P_2. Let P_1 and P_2 have n and m phenotypes, respectively. As a result, we need to calculate phenotypic similarity for $n \times m$ phenotypic pairs.

Now, assume there is a new diagnosis for P_2 with a newly observed phenotype. When calculating the phenotypic similarity between P_1 and P_2 using $Sim_{max,Resnik}$, we need to call the MICA algorithm n times that has $O\left(n \times (E + V)\right)$time complexity, where V represents the number of vertices (15,371) and E represents the number of edges in HPO. The number of edges in a DAG, such as HPO, is of $O(V^2)$. This is under the assumption that we have stored all intermediate maximum values calculated in the base case. Otherwise, $n \times (m + 1)$ calls for MICA would have been needed.

On the other hand, when using $Sim_{max,Emb.}$ in calculating patient similarity with a new phenotype record, we only need to call the cosine similarity algorithm n times, resulting in a time complexity of $O(n)$.

10 References

1. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.
2. Mikolov, T., et al., *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems, 2013. **26**.
3. Perozzi, B., R. Al-Rfou, and S. Skiena. *Deepwalk: Online learning of social representations*. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014.