Leveraging Representation Learning on the Human Phenotype Ontology

Maryam Daniali[1,2], Shiva Ganesan[2,3,4], Peter D. Galer[2,3,4], Shridhar Parthasarathy[2,3,4], Edward Kim[1], Dario D. Salvucci[1], Jeffrey M. Miller[2], Ingo Helbig[2,3,4,5], Scott Haag[2]

[1]Department of Computer Science, Drexel University, Philadelphia, PA, USA; [2]Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, Philadelphia, PA, USA; [3]Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [4]The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, PA,USA; [5]Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA.

Our research focus: Converting sets of phenotypes to actionable clinical information.

The interpretation of patients' Electronic Health Records (**EHR**) is **challenging**, especially for patients that have dynamic and diverse sets of clinical features.

The Human Phenotype Ontology (**HPO**) is a standardized dictionary of phenotypic concepts and their relationships. The HPO can be represented as a Directed Acyclic Graph (DAG) with phenotypes represented as nodes and connections between phenotypes represented as edges.

Natural language processing (**NLP**) models can be used to **map** the EHR to sets of phenotypes in the HPO, making them amenable to further downstream tasks (e.g., prediction and classification tasks). To date many of these tasks are conducted using either:

Manual analysis:

- can reduce the quality of in-depth analysis.
- places a burden on a clinicians' time.

Mechanistic methods (e.g. Resnik & information coefficient):

- are based on guess-and-check approaches.
- require a large amount of computation, even with minor changes to patients' EHRs.

Deep phenotyping methods:

Most of these studies identify phenotypes based on the inheritance and structural relationships of nodes in the HPO by using graph searching algorithms. The information available in the patient corpus and the frequency of phenotypes do not play any role in these analyses.

Our method (***P**henotype **E**mbedding*):

- Using recent advancements in graph representations, we provide vector embedding of phenotypes in a latent space.
- We employ the occurrence frequency of phenotypes from a large patient corpus to shape the embedding space.
- This method provides a fast and robust setting for analyses on phenotypes.
- We believe these phenotype representations can be used to predict a variety of important patient metrics.

MATERIALS & METHODS

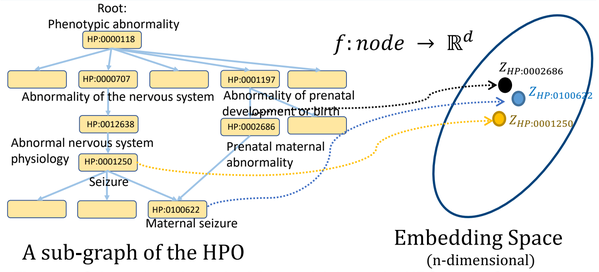


Figure 1: Schematic of the embedding model. Input: the HPO graph, Output: the embedding space for all phenotypes.

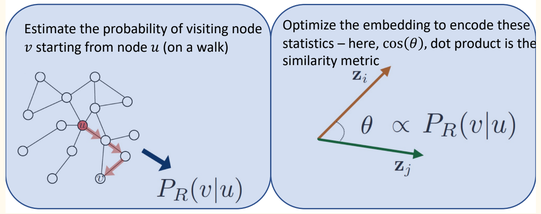
Data:

1. We have access to over 1.5 million patients' EHR as part of the Arcus Data Repository [1].
2. Arcus provides phenotype annotations extracted by the cTAKES [2] NLP system from patients' EHR notes.

Method:

Building upon the Node2Vec algorithm [3,4], we create a latent space where phenotypes that are related in the HPO get closer in the embedding space. In doing so, we

1. *compute probabilities of random walks – by incorporating weighted edges based on the frequency of the nodes.*
2. *find r biased random walks of length l starting from each graph node where we tune breadth- and depth- traversal.*
3. *put nodes that are seen in the same walk closer in the vector space, optimizing the node2vec objective function using the* Stochastic Gradient Descent algorithm.



RESULTS

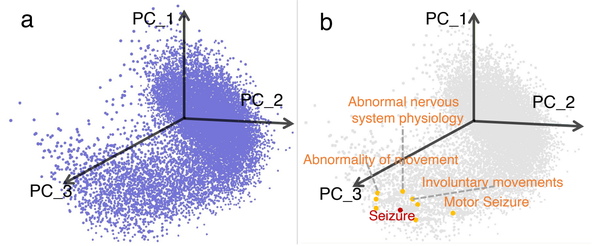


Figure 2: We used PCA to visualize the original (128D) embedding space. a) A 3D representation phenotypes in the space. b) Closest phenotypes to Seizure in the 3D space.

Table 1: List of ten closest phenotypes to a selected phenotype in the original space, 128D

a) Selected phenotype: Seizure

Nearest phenotypes to Seizure	Cosine distance
Abnormal nervous system physiology	0.185
Abnormality of movement	0.250
Reduced consciousness/confusion	0.313
Gait disturbance	0.342
Motor seizure	0.348
Abnormality of higher mental function	0.359
Behavioral abnormality	0.363
Autistic behavior	0.395
Upper motor neuron dysfunction	0.397
Involuntary movements	0.406

b) Selected phenotype: Neurodevelopmental abnormality

Nearest phenotypes to Neurodevelopmental abnormality	Cosine distance
Neurodevelopmental delay	0.319
Abnormality of higher mental function	0.415
Delayed speech and language development	0.422
Abnormal nervous system physiology	0.428
Reduced consciousness/confusion	0.454
Abnormality of the nervous system	0.464
Global developmental delay	0.494
Behavioral abnormality	0.508
Seizure	0.514
Headache	0.518

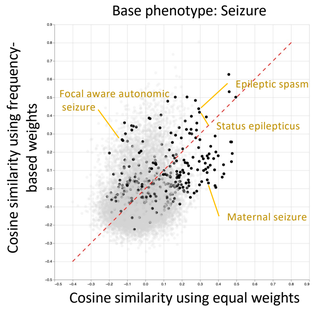


Figure 4: A comparison of the effects of incorporating phenotypes frequencies in the embedding space. Phenotypes under Seizure in the HPO DAG (i.e., the children nodes) are marked in black. Many of the rare phenotypes from this category move closer to Seizure using the frequency-based approach vs. the standard Node2Vec algorithm. Conversely, phenotypes such as Maternal seizure, which belong to a different sub-graph (see Figure 1) and therefore have a weak connection to Seizure, move farther away.

CONCLUSIONS

- Using an embedding algorithm, we transform phenotypes into a n-dimensional vector space, where phenotypes that are close to each other in the HPO graph are closer in the embedding space.
- We use phenotype occurrence frequency to capture the relationships, moving rare terms closer to more common parent terms.
- This technique helps us translate patients EHR records to a vector-based representation that is ready for use in various downstream tasks.
- We believe that applying this method to common predictive problems will improve and accelerate important research and clinical processes in pediatrics health care.
- In the future, we plan on employing the co-occurrence frequency of phenotypes in patient notes to augment the edges in the HPO DAG, allowing nodes that co-occur but are far apart to be directly connected.

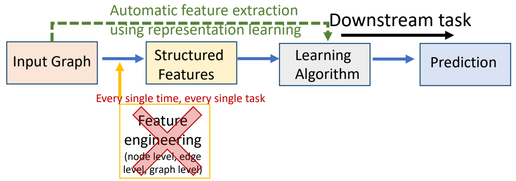


Figure 5: Schematics of the benefits of using graph representation learning

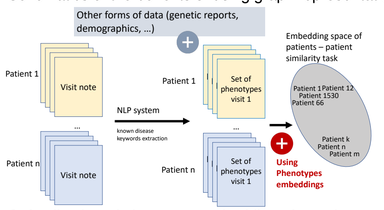


Figure 6: Application of phenotype embeddings for patient representation

REFERENCES

[1] Arcus Data Team. 2021. Deidentified Arcus Data Repository. Arcus at Children's Hospital of Philadelphia. Accessed on 2021/10/22.

[2] http://ctakes.apache.org

[3] Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864).

[4] Liu, R., Hirn, M., & Krishnan, A. (2021). Accurately Modeling Biased Random Walks on Weighted Graphs Using Node2vec+. *arXiv preprint arXiv:2109.08031*.