Discovering Novel Patterns of Autoinflammation via Phenotype-First Unsupervised Machine Learning

Ryan S. Laird[1], Amanda K. Ombrello[1], Patrycja Hoffmann[1], Daniel L. Kastner[1]

[1] National Human Genome Research Institute, Bethesda MD

Introduction

In recent years the integration of large-scale biomedical knowledge sources and machine learning has accelerated patient stratification in rare disease. However, such projects often focus on known diseases within large and phenotypically diverse cohorts.

Our research team studies a subset of rare diseases characterized by dysregulation of the innate immune system, termed *autoinflammation*. Many of these disorders have highly overlapping phenotypic profiles, which often clouds downstream analysis and diagnosis.

In order to better understand clinical variation within our cohort, we leverage the Human Phenotype Ontology (HPO) to guide two unsupervised machine learning pathways. As a proof-of-concept we attempt to cluster a subset of 54 patients in our cohort with Deficiency of Adenosine Deaminase 2 (DADA2), a disease known to be clinically heterogeneous.

General HPO Structure

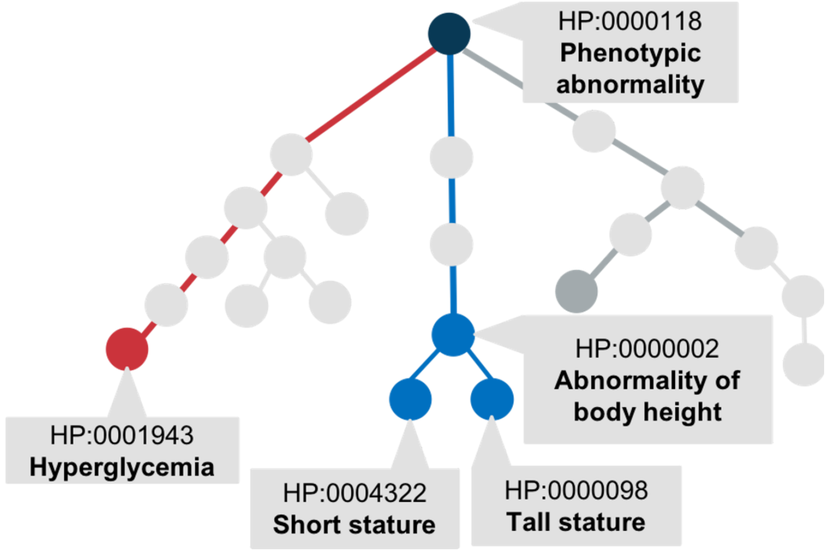

Figure 1: The HPO provides a structured vocabulary of phenotypic abnormalities encountered in human disease, and has become the *de facto* standard for studying the complex relationship between genetic variants and clinical presentation. . https://hpo.jax.org/app/tools/loinc2hpo

Methods

Data Generation & Preprocessing

Source code for analysis will be hosted on GitHub.

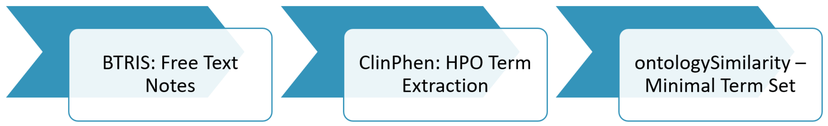

Figure 2: A BTRIS query was used to gather all clinical notes for 54 DADA2 patients on our protocol. A natural language processing tool, ClinPhen, was used to extract HPO term occurance from the corpus. To avoid redundancy, R package `ontologySimilarity` was applied to generate a minimal term set per patient.

Clustering Approaches

Each patient was treated as a binary vector of their minimal term set (e.g. Arthritis: 0, Ischemic Stroke: 1, …).

Table 1: Comparison of Patient Clustering Approaches

Approach	Patient Similarity Metric	Clustering Method	Pros	Cons
Ontology Similarity	Lin	Leiden	Easier to interpret, rewards granular data	No connections between organ systems
HPO2Vec	Average Term Embedding	PCA, UMAP, HDBSCAN	Closer to expert opinion	Not directly interpretable

HPO2Vec

To enrich the HPO with cross organ system associations, a new edge was drawn between any two phenotypes if they shared a common Orphanet disease annotation.

Figure 3: An example Node2Vec embedding that depicts node homophily. https://snap.stanford.edu/node2vec

Results

Patient Clusters

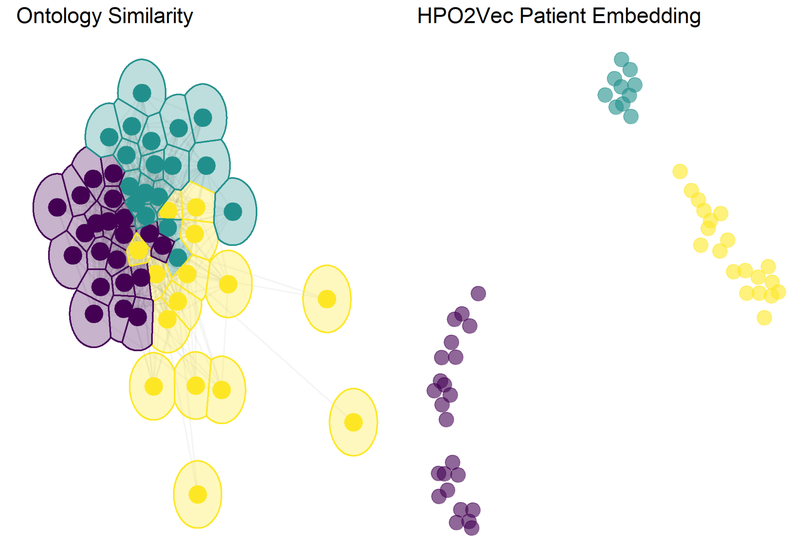

Figure 4: Leiden clustering by ontology similarity; HDBSCAN Clustering by Average HPO2Vec Embedding Space.

Cluster Interpretation

Table 2: Phenotypic Comparison of Patient Clusters

Clusters	Frequent Terms	Qualitative Category
Onto Sim		
1	Stroke, Arthritis	inflam
2	Abdominal Pain, Neutropenia	marrow
3	Red Cell Aplasia, Vasculitis	other
HPO2Vec		
1	Stroke, Arthritis	inflam
2	Abdominal Pain, Neutropenia	marrow
3	Red Cell Aplasia, Vasculitis	other

Discussion

As with any unsupervised method, perhaps the most difficult step is interpretation. Future work includes a deeper dive into the clinical nature of each respective cluster, and a more robust comparison of direct ontology similarity vs HPO term embedding.

If patient clusters are clinically relevant, we hope to apply these methods to better stratify undifferentiated patients that make up ~60% of our cohort.