Leveraging a Neural-Symbolic Representation of Biomedical Knowledge to Improve Pediatric Subphenotyping

Tiffany J. Callahan MPH, PhD[1]; Lawrence E. Hunter PhD[1]; Michael G. Khan MD, PhD[2]

[1]Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[2]Department of Pediatrics, Section of Pediatric Critical Care, School of Medicine, University of Colorado School of Medicine, Aurora, CO, 80045 USA

PILOT STUDY: These results are preliminary and will be updated prior to publication

Table of Contents

ABSTRACT 3
INTRODUCTION 4
- Related Work 5
- Objectives 6

METHODS 7
- Clinical Data 7
- Patient Cohorts 7
- Representing Mechanisms of Human Disease 7
- Terminology Mapping 10
- Patient Representations 10
- Internal Evaluation 13
- External Validation 15

RESULTS 17
- Human Disease Knowledge Graph 17
- Terminology Mapping 17
- Internal Validation 18
- External Validation 46

DISCUSSION 55
CONCLUSIONS 58
REFERENCES 60

ABSTRACT

Objective: Understanding the molecular drivers of disease is a vital component of personalized medicine. Unfortunately, molecular data are not currently available in most electronic health records (EHRs). To solve this problem we created Med2Mech, a joint learning framework for inferring molecular characterizations of patients from clinical data and publicly available biomedical data.

Methods: Med2Mech was evaluated using pediatric EHR data from a subset of rare disease and other similarly medically complex patients. First, patient-level clinical embeddings were generated. Then, a PheKnowLator knowledge graph (KG) was used to generate mechanism embeddings. Finally, patient-level mechanism embeddings were derived by summing or averaging each patient's unique set of mechanism embeddings. A one-vs-the-rest multiclass classification strategy, with five cross-fold validation, was used to evaluate the discriminatory ability of the mechanism and clinical embeddings. Rare disease subphenotype differences, using both clinical and mechanism embeddings, were further investigated using K-Means, which were verified by PhD- and MD-level domain experts. As external validation, the ability to infer the genotype of the rare disease patients using an independent sample of publicly available transcriptomic data was examined.

Results: Clinical embeddings were built for four rare disease groups (n = 2,646) and 10,000 similarly complex patients using 6,382 conditions, 2,334 medications, and 272 measurements. Mechanism embeddings were generated from a PheKnowlator KG with 129,875 nodes and 3,838,935 edges. On classification, the mechanism embeddings out-performed 82.2% of the clinical embedding parameterizations. Domain expert review confirmed the mechanism embeddings produced more clinically-relevant clusters of comorbidities for each rare disease subphenotype than the clinical embeddings. External validation further demonstrated the utility of this framework by accurately inferring the genotype and phenotype of EHR-derived rare disease patients from publicly available molecular data.

Conclusion: These results illustrate the translational utility of PheKnowLator KGs and demonstrate that it is possible to derive clinically meaningful and biologically relevant patient representations from disparate sources of EHR data and expert-curated publicly available transcriptomic data.

INTRODUCTION

Deep phenotyping aims to establish highly precise and accurate genotype-phenotype associations by capturing the spectrum of phenotypic variance observed within a specific species. Deep phenotyping has been successfully applied to rare disease and genetic disorders [1–13], cancer [14–20], pregnancy [21–23], and disease-specific grant-funded research cohorts [1–12,24] using a variety of clinical and -omic data. Although limited to very small samples, specialized research cohorts, clinical trials, and N-of-1 studies [14] have recently demonstrated the value of expanding the data typically used for deep phenotyping (i.e., clinical and molecular data) with measurements from wearables [25–27], the environment [28,29], and cutting-edge speech technology [30]. The vast majority of deep computational phenotypes (CPs) have been constructed from electronic health record (EHR) data [31] with a few recent examples incorporating knowledge extracted from clinical vocabularies [32], biomedical ontologies [33–36], and the literature [37,38].

To capture the full phenotypic variance associated with a disease, researchers will need access to both individual- (e.g., multimodal data including clinical, multiomic, environmental, behavioral, psychological, and social) and population-level (e.g., longitudinal surveys and census data) data collected from databases, reported by patients, and computationally inferred as well as biomedical knowledge extracted from trusted sources of publicly available data and the literature. Unfortunately, data like these are not yet widely available for most patients seen in a hospital setting.

Combining EHR data with publicly available sources of molecular data may provide a promising solution for progressing deep computational phenotyping research until molecular data are available for the majority of patients seen in a hospital setting. Microarrays enable the simultaneous measurement of expression values for thousands of genes [39]. When mined for patterns or “profiled”, these data provide powerful insight into the etiology of complex diseases [39]. As of August 2021, two of the largest public repositories of gene expression data, the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) and the European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk/gxa/home) collectively contained data from over 163,309 experiments. While microarray data are publicly available for a wide-range of diseases, they often lack important clinical context.

In lieu of access to the wide range of data required to interrogate the full spectrum of phenotypic variance associated with human disease, the next best solution is to develop a proxy (i.e., “a placeholder [surrogate] for another object to control and get access to it” [40]) that accurately represents the knowledge stored in patient-level samples of molecular data. Ideally, this proxy would also enable the integration of clinical data, like those from an EHR, with external sources of transcriptomic data, like those in the GEO or the EBI’s Expression Atlas. Given their ability to represent complex phenomena, KGs present a promising avenue for generating proxies of human disease mechanisms. Using KGs as proxies for patient-level molecular data requires: (i) the development of a biologically accurate KG and (ii) a meaningful way to connect entities in the KG to patient-level data.

The remainder of this paper is organized as follows. First, we provide a detailed overview of

related work, focusing on the use of KGs to integrate biological and clinical data. Then, we describe the objectives of the current study and introduce Med2Mech.

Related Work

Semantic networks or KGs are commonly used in many branches of science to decipher complex processes. A KG's topology provides invaluable insight into the organization of a system [41] where biological components are usually represented as nodes and the connections between the nodes are represented as edges. Networks have been used extensively in biomedical research to represent everything from metabolomics [42], cell signaling [43], and gene regulatory networks [44] to protein-protein interactions [45], and drug interaction networks [46]. With respect to clinical data, KGs have been used to improve information retrieval [47,48], capture patient information [49], and improve the presentation of complicated results [50].

While there have been promising attempts to use KGs as proxies for integrating biomedical data with clinical data, they have largely been limited to a specific disease or application [50–52], rely on proprietary data or software [53], and/or have significant design limitations [49,54,55]. Two, highly relevant studies are detailed below.

The translational Medicine Ontology (TMO) and Knowledge Base [56] was developed in an effort to integrate disparate biological knowledge sources with EHR data. The initial development focused on creating rich semantic representations of clinical data and the molecular mechanisms underlying them. What makes this work different from what we are proposing in this paper, is the way in which EHR data are handled in order to be linked with different sources of knowledge; the TMO project created a system, which requires converting EHR data into a format that can be used by Semantic Web technologies. By utilizing this approach, this project is able to facilitate a seamless integration with other knowledge sources (i.e., other ontologies like the Gene Ontology). As far as we know this project is not actively being maintained.

Perhaps the most similar work to ours is the Propagated SPOKE Entry Vectors or PSEVs developed by Nelson et al. [53]. This project facilitated the integration of EHR concepts for 292,753 patients with heterogeneous sources of biomedical knowledge via the Hetionet KG [57]. This work differs from the work proposed in this chapter in three important ways: (i) **Knowledge Representation.** Similar to our work, this work utilized a large-scale biomedical KG. While both KGs included similar data sources, our work utilizes the PheKnowLator Ecosystem [58], which is ontologically-grounded and logically consistent [through the use of a deductive logic reasoner]. SPOKE leverages data from some ontologies, but does not utilize description logics, which makes it unable to be deductively closed or logically verified. Whether or not this difference is important requires future experimentation. Both KGs also include functionality that enables them to transform into other representations more conducive for certain downstream learning tasks; (ii) **Concept Mapping Strategy.** Unfortunately, the authors provide only high-level detail on their mapping approaches, stating that all mappings were manually created to align ICD codes to DOID identifiers, RxNorm identifiers to DrugBank identifiers, and LOINC codes to a variety of UMLS CUIs. Our approach will leverage

OMOP2OBO [59] to map clinical concepts to entities in the KG and can accommodate any of the source vocabularies and standard terminologies supported by the Observational Medical Outcomes Partnership (OMOP) common data model; and (iii) **Approach to Representing Patients.** While both methods used diagnoses, medications, and laboratory test results. The PSEVs were designed to create biologically meaningful representations of EHR concepts by combining node embeddings with an importance value derived from patients (i.e., the portion of patients in a cluster that have a given concept in their record). In contrast, our approach is designed to represent molecular entities, clinical concepts, and experimental data values like molecular signatures, and to use them to generate meaningful patient representations. While this may be possible from the PSEVs, it was not demonstrated nor discussed.

Objectives

In this paper, we introduce Med2Mech, a robust joint learning framework for inferring molecular characterizations of patients from sources of clinical and experimental data using biomedical knowledge. This framework leverages neural-symbolic representation learning of entities in a biomedical KG to generate molecular mechanism embeddings underlying human disease, which are then mapped to clinical and biological entities. The usefulness of the Med2Mech framework was evaluated using the following questions:

1. **Given different clinical domains and the availability of different embedding models, each with different parameter settings, what model and parameter setting is best for each domain when deriving patient-level representations?** Answering this question is an important step towards automating data-driven CP construction. It is also needed to establish a baseline that can be used for comparing patient representations built using data from different clinical domains as well as patient embeddings constructed from sources of non-clinical data. To address this question, clinical concept embeddings will be constructed from different combinations of clinical domains (i.e., only conditions, only medications, only measurements, and all three domains), using a reasonable variety of hyperparameter settings. Representations will be evaluated using t-SNE plots, K-Means clustering, and multiclass classification.
2. **If patients can also be represented using the molecular mechanisms that underlie their primary diagnosis, would these representations result in more meaningful subphenotypes than when constructed from only clinical data?** Answering this question is important as it will: (i) provide insight into the utility of constructing patient representations using molecular mechanisms; and (ii) help determine whether or not patients are best represented using clinical concepts or the molecular mechanisms underlying each clinical concept. For this question, molecular mechanism embeddings will be derived from a PheKnowLator KG, and then mapped to patient clinical concepts using OMOP2OBO. Representations will be built from different combinations of clinical domains and evaluated using a similar approach as in *Question 1*, with additional comparisons made to the representations generated using only clinical concepts.
3. **Can PheKnowLator be used to infer the genotype of patients with only clinical data from an independent sample of genotyped patients?** Answering this question is important as it will enable us to assess the biological relevance of the mechanisms

represented in the PheKnowLator KGs and will ultimately help to determine whether or not the KG is a suitable proxy for patient-level molecular data. Publicly available transcriptomic data will be obtained for an independent sample of genotyped patients which will be converted into the same latent space as the population of clinical patients used in *Questions 1-2* using a PheKnowLator KG. The similarity of both patient populations will be assessed using t-SNE plots, K-Means clustering, and rank-based similarity metrics.

METHODS

Clinical Data

A subset of de-identified CHCO data was extracted from the University of Colorado's Anschutz Medical Campus Health Data Compass warehouse [60]. The Children's Hospital of Colorado (CHCO) data conformed to the structure defined by the PEDSnet OMOP CDM, which is an adaptation of the OMOP CDM version 5.0 [61,62]. The current analysis utilized the Person, Condition Occurrence, Measurement, and Drug Exposure, and Visit Occurrence tables. The use of these data was approved by Colorado Multiple Institutional Review Board (COMIRB; #15-0445). Details on the preprocessing and de-identification steps applied to the EHR data used for this analysis are provided in the Supplemental Material.

Patient Cohorts

Two cohorts of pediatric patients were utilized in the current project. (i) **Case Patients**. Cases included rare disease patients with at least 10 visits and a primary diagnosis of phenylketonuria (PKU), congenital hypothyroidism (CAH), cystic fibrosis (CF), or sickle cell disease (SCD). The International Classification of DIsease (ICD) and SNOMED concepts along with the equivalent OMOP concepts used to identify the rare disease patients are shown in Table 1. These concepts were selected with the help of collaborating domain experts as it was not possible to perform a manual chart review or access genetic testing results to verify the rare disease status of these patients. Additionally, at the time of this analysis, the selected rare diseases were part of the Colorado Newborn Screen[1], which decreased the likelihood of including falsely diagnosed patients. (ii) **Control Patients.** A sample of 10,000 similarly medically complex patients with at least 10 visits and no occurrence of the diagnosis codes used to identify the rare disease patients were selected as controls.

The SQL queries used to identify cases and controls are publicly available as a GitHub Gist (https://gist.github.com/callahantiff/b72339a8d6fb7de31e1912039947605a). This GitHub Gist also includes scripts for the SQL queries used to obtain OMOP condition occurrence, drug exposure, and measurement concepts. This analysis was meant to serve as a proof-of-concept and thus all of the OMOP concepts used to identify the case and control patients were included in the final concept lists used for all experiments.

Representing Mechanisms of Human Disease

A PheKnowLator (v1.0; [58]) KG (Figure 1) was used as the proxy of human disease mechanisms (https://github.com/callahantiff/PheKnowLator/wiki/v1.0.0). This KG was built using

[1]Colorado Newborn Screen: https://www.babysfirsttest.org/newborn-screening/states/colorado

Table 1. Clinical Concepts used to Identify Rare Disease Patients.

Rare Disease	Source Vocabulary Codes	OMOP Standard Terminology Concepts
Congenital Hypothyroidism	SNOMED-CT:278503003 ICD10-CM:E03.0	Congenital hypothyroidism with diffuse goiter (omop:4081998)
	SNOMED-CT:237515009 ICD-10-CM:E03.1	Congenital hypothyroidism without goiter (omop:4130017)
	SNOMED-CT:217710005 ICD-10-CM:E00 ICD-10-CM:E00.9	Congenital iodine deficiency syndrome (omop:4314093)
Cystic Fibrosis	SNOMED-CT:190905008 ICD-10-CM:E84 ICD-10-CM:E84.9	Cystic fibrosis (omop:441267)
	SNOMED-CT:86555001 ICD-10-CM:E84.0 ICD-9-CM:277.02	Cystic fibrosis of the lung (omop:254320)
	SNOMED-CT:190909002 ICD-10-CM:E84.1	Cystic fibrosis with intestinal manifestations (omop:194325)
	SNOMED-CT:81423003 ICD-9-CM:277.00	Cystic fibrosis without meconium ileus (omop:434615)
	SNOMED-CT:86092005 ICD-10-CM:E84.11 ICD-9-CM:277.01 ICD-9-CM:277.03	Cystic fibrosis with meconium ileus (omop:193174)
	SNOMED-CT:441520002 ICD-9-CM:V83.81	Carrier of cystic fibrosis gene mutation (omop:40479565)
Phenylketonuria	SNOMED-CT:7573000 ICD-10-CM:E70.0 ICD-9-CM:270.1	Classical phenylketonuria (omop:432872)
Sickle Cell Disease	SNOMED-CT:127040003 ICD-10-CM:D57 ICD-10-CM:D57.01	Sickle cell-hemoglobin SS disease (omop:22281)
	SNOMED-CT:416180004 ICD-10-CM:D57.1 ICD-10-CM:D57.80 ICD-9-CM:282.61 ICD-9-CM:282.68	Hemoglobin SS disease without crisis (omop:30683)
	SNOMED-CT:417425009 ICD-10-CM:D57.0 ICD-10-CM:D57.00 ICD-9-CM:282.62 ICD-9-CM:282.69	Hemoglobin SS disease with crisis (omop:26942)
	SNOMED-CT:417683006 ICD-10-CM:D57.20 ICD-9-CM:282.63	Sickle cell-hemoglobin C disease without crisis (omop:443726)

Rare Disease	Source Vocabulary Codes	OMOP Standard Terminology Concepts
Sickle Cell Disease	SNOMED-CT:417517009 ICD-10-CM:D57.21 ICD-10-CM:D57.211 ICD-10-CM:D57.213 ICD-10-CM:D57.218 ICD-10-CM:D57.219 ICD-9-CM:282.64	Sickle cell-hemoglobin C disease with crisis (omop:443721)
	SNOMED-CT:417048006 ICD-10-CM:D57.40 ICD-9-CM:282.41 ICD-10-CM:D57.42	Sickle cell-thalassemia disease without crisis (omop:321263)
	SNOMED-CT:416826005 ICD-10-CM:D57.09 ICD-10-CM:D57.41 ICD-10-CM:D57.43 ICD-10-CM:D57.45 ICD-10-CM:D57.413 ICD-10-CM:D57.418 ICD-10-CM:D57.419 ICD-10-CM:D57.433 ICD-10-CM:D57.438 ICD-10-CM:D57.439 ICD-10-CM:D57.458 ICD-10-CM:D57.459 ICD-9-CM:282.42	Sickle cell-thalassemia disease with crisis (omop:443738)
	SNOMED-CT:95846001 ICD-9-CM:289.52	Red blood cell sequestration in spleen (omop:196943)
	SNOMED-CT:444108000 ICD-10-CM:D57.09 ICD-10-CM:D57.212 ICD-10-CM:D57.412 ICD-10-CM:D57.432 ICD-10-CM:D57.452 ICD-10-CM:D57.812	Acute sickle cell splenic sequestration crisis (omop:40485018)
	SNOMED-CT:372146004 ICD-10-CM:D57.01 ICD-10-CM:D57.211 ICD-10-CM:D57.411 ICD-10-CM:D57.431 ICD-10-CM:D57.451 ICD-10-CM:D57.811 ICD-9-CM:517.3	Acute chest syndrome (omop:254062)
	SNOMED-CT:16402000 ICD-10-CM:D57.3 ICD-9-CM:282.5	Sickle cells present (omop:25518)

Acronyms - ICD-CM: International Classification of Diseases - Clinical Modification; OMOP: Observational medical Outcomes partnership; SNOMED-CT: Systematized Nomenclature of Medicine - Clinical Terms.

publicly available Linked Open Data and OBO Foundry ontologies (downloaded in November 2018). The core set of ontologies included phenotypes (Human Phenotype [HP]; https://hpo.jax.org/), diseases (Human Disease Ontology [DOID]; https://disease-ontology.org/), and biological processes, molecular functions, cellular components (Gene Ontology [GO]; http://geneontology.org/). These ontologies were extended by adding relations between phenotypes, diseases, and GO biological processes, molecular functions, cellular components and new entities were added to the KG for genes, pathways, and chemicals. The Linked Open Data sources included: the Comparative Toxicogenomics Database (http://ctdbase.org/; chemical-gene, chemical-pathway, chemical-disease, gene-pathway relations), the Reactome Pathway Database (https://reactome.org/; pathway-biological processes, pathway-cellular components, and GO-molecular functions relations), and the STRING database (https://string-db.org/; gene-gene relations). Using the ELK reasoner available through Protégé (v5.1.1; https://protege.stanford.edu/), the KG was deductively closed and logically verified. Data, KGs, and node embeddings are publicly available through Zenodo (https://zenodo.org/communities/pheknowlator-ecosystem).

Terminology Mapping

OMOP2OBO (v.0; [59]) was used to map OMOP terminology concepts within each clinical domain to ontology concepts (Figure 1). Mappings were created for each of the following clinical domains: (i) **Conditions.** Systematized Nomenclature of Medicine -- Clinical Terms or SNOMED-CT standard clinical terminology concepts were mapped to HP and DOID concepts; (ii) **Medications.** RxNorm codes were mapped to Medical Subject Headings or MeSH identifiers at both the medication- and chemical compound-level; and (iii) **Measurements.** LOINC codes for measurement results outside of a normal reference range were mapped to HP and DOID concepts.

Patient Representations

As shown in Figure 2, two types of patient representations were created: (i) **Clinical Embeddings.** Embeddings were created from each patient's set of temporally-ordered visit-level EHR concepts and (ii) **Molecular Mechanism Embeddings.** These embeddings were created in two steps. First, node embeddings were created from sentences of randomly sampled paths of length n from the PheKnowLator KG. Then, patient-level molecular mechanism embeddings were assembled by summing or averaging all of the node embeddings mapped to each patient's unique set of clinical concepts. For both methods, it is important to define the following concepts: dimensions (i.e., the length of the embedding), walks (i.e., the number of runs generated for each entity), walk length (i.e., the length or number of entities included in each run), and window size (i.e., the number of entities to the right and left of the target entity, which are used as training data when learning the target entity embedding).

Clinical Embedding Algorithm

Clinical embeddings were generated using Doc2Vec [63], which creates continuous distributed paragraph vector (PV) document representations for fixed-length features from variable length text using two implementations: (i) distributed memory model (PV-DM), which is analogous to word2vec's continuous bag-of-words and (ii) distributed bag of words (PV-DBOW), which is

Figure 1. Med2Mech Knowledge Representation.
The knowledge representation which leverages clinical data and molecular mechanisms. Terminology mappings from OMOP2OBO were used to map molecular mechanism entities from to clinical concepts.

analogous to word2vec's skip-gram[2]. For each model, four different combinations of sliding window size (i.e., 5, 10, 15, 20) and three different embedding dimensions (i.e., 100, 150, 200) were examined. A bag-of-words with term frequency-inverse document frequency (BoW + TF-IDF [64] and L2 normalization was used as the baseline representation.

Molecular Mechanism Embedding Algorithm

Node and edge embeddings were generated from PheKnowLator using the Walking RDF and OWL algorithm [65]. As demonstrated in Figure 2, this method generates paths of molecular mechanism(s) from an RDF graph (e.g., `hasSymptom(asthma, cough)`, `hasPhenotype(ADBR2, cough)`, `causesOrContributesTo(mucus secretion,cough)`) using a skip-gram-like approach with 512 dimensions, 100 walks, a walk length of 20, and a sliding window length of 10, to create embeddings. Patient-level molecular

[2]From the Gensim Docs, PV-DM works by "training a neural network on the synthetic task of predicting a center word based an average of both context word-vectors and the full document's doc-vector" and PV-DBOW works by "training a neural network on the synthetic task of predicting a target word just from the full document's doc-vector" (https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html#introducing-paragraph-vector).

Figure 2. Clinical and Mapped Molecular Mechanism Embedding Overview.
This figure provides a high-level overview of the Med2Mech approaches for generating clinical and molecular mechanism embeddings for patients in an EHR. Acronyms - CHEBI: The Chemical Entities of Biological Interest Ontology; EHR: Electronic Health Record; ENTREZ: The National Center for Biotechnology Information Gene Database Identifiers; MONDO: Monarch Disease Ontology; R-HSA: Reactome Pathway Database Identifiers.

mechanism(s) embeddings were created by aggregating (i.e., summing or averaging) the set of KG node embedding mapped to each clinical concept. The KG embeddings are available on Zenodo (https://zenodo.org/record/3830982/files/embeddings.zip).

Internal Evaluation

Visualizing Patient Representations

As a qualitative evaluation, patient representations were visually examined for expected patterns (i.e., patients of the same disease type should be closer in latent space than patients of different disease types) and outliers (i.e., a few patients grouping far from all other patients of the same disease type or appearing close to patients of a different disease type). To create meaningful visualizations, patient representations were reduced to two-dimensions and visualized using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [66]. Finding optimal t-SNE hyperparameter settings can be tricky, but through experimentation, a perplexity of 50 was determined to be useful for both the KG and clinical data.

Multiclass Classification

A one-vs-the-rest multiclass classification strategy, with five cross-fold validation, was used to evaluate the discriminatory ability of the clinical embeddings and molecular mechanism embeddings. To account for class imbalance between the rare disease patient groups and control patients, Synthetic Minority Over-sampling with Edited Nearest Neighbors under-sampling was implemented [67]. The examined classifiers included logistic regression, Bernoulli Näive Bayes, and Linear Support vector machine (SVM). The following metrics, reported as the average performance across the five-folds, were used to evaluate the classifiers:

- Recall: Given a list of predicted patients, the ratio of correctly identified patients with the same diagnosis:

$$recall = \left(\frac{TP}{TP + FN}\right) \tag{1}$$

- Precision: Given a list of predicted patients, sorted by score, the ratio of patients that had the same diagnosis among all predicted patients:

$$precision = \left(\frac{TP}{TP + FP}\right) \tag{2}$$

- Accuracy: Given a list of predicted patients, the percentage of patients that had the same diagnosis:

$$accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \tag{3}$$

- F_1 Score: is the weighted average of precision and recall:

$$F_1\ Score = \left(2 * \frac{precision * recall}{precision + recall}\right) \tag{4}$$

Clustering Patient Representations

The best performing clinical and molecular embedding model parameterizations, using all

clinical domains, were assessed using K-Means [68] with a predefined k or the value of k which yielded the best performance across the following three clustering metrics:

- Adjusted Rand Index (ARI): the count of entity pairs, derived from computing all possible entity pairs assigned to the same or different clusters [69]. The version used in this study was adjusted for chance, which guaranteed a score of zero for a random assignment, regardless of the number of clusters:

$$ARI(U,V) = (RI - Expected_{RI}) / (max(RI) - Expected_{RI}), \quad (5)$$

  where U and V are separate clusters.

- Adjusted Mutual Information (AMI): is the measure of information obtained about one random variable by observing another random variable. The version used in this study was adjusted for random chance [70] in order to account for the fact that a larger number of clusters results in a higher score:

$$AMI(U,V) = [MI(U,V) - E(MI(U,V))]/[avg(H(U),H(V)) - E(MI(U,V))], \quad (6)$$

  where $U$and $V$are separate clusters.

- Cluster Purity: measures the proportion of different classes represented within the same cluster [71]:

$$\frac{1}{2} \sum_{m \in M} \max_{d \in D} |m \cap d| \quad (7)$$

All metrics were scored 0.0-1.0, where 1.0 indicated that an observed set of predefined labels and the predicted clustering assignments were identical. Descriptive metrics were recorded for each cluster and included the size (i.e., patient count), cluster breakdown (i.e., percent of each disease group out of the total patients in a given cluster), and disease breakdown (i.e., percent of each disease group out of the total patients for each disease group). Additionally, the five most frequently occurring cluster concepts and five most frequently occurring concepts unique to each cluster were reported. This output was reviewed with domain experts and used to assess the clinical relevance of the resulting clusters.

Patient representations were built using data from different combinations of clinical domains, including: only conditions, only medications, only measurements, and all clinical domains (i.e., all conditions, medications, and measurements). For clinical embeddings, a total of 96 models were constructed by varying two Doc2Vec implementations (i.e., PV-DM and PV-DBOW) using three different embedding dimensions (i.e., 100, 150, 200) and four different sliding window sizes (i.e., 5, 10, 15, 20) for each of the four clinical domains. For molecular mechanism embeddings, a total of eight models were examined by varying two aggregation methods (i.e., average and sum) for four different clinical domains. Four baseline models were examined, one for each clinical domain. This resulted in a total of 108 models that when assessed using three different classifiers and four different performance metrics generated a total of 1,296 results. All of the evaluation tasks (except classification) were applied first to rare disease patients and then

to all rare disease and control patients. This enabled the possibility of assessing the potential for using the representations in rare disease subclassification tasks as well as provided an opportunity to deeply investigate the discriminatory ability of the different embedding model parameterizations.

External Validation

The external validation aimed to evaluate the translational utility of the KG-derived molecular mechanism embeddings by determining if the genotype and phenotype of any of the CHCO rare disease patients could be inferred from an independent sample of pediatric rare disease transcriptomic data. In collaboration with a PhD-level molecular biologist, GEO was searched to identify transcriptomic data, specifically for the CHCO rare diseases (i.e., CF, CAH, PKU, SCD) within a pediatric population. Only human microarray studies with clinical and/or phenotypic data collected from pediatric populations were considered eligible for this portion of the study.

Gene Expression Processing

Gene expression data from selected studies were processed using standard R pipelines [72] and limma [73] via GitHub Gist[3], which included steps to remove missing data, filter probes with expression values below the first quantile, perform background correction and normalization, derive lists of DEGs (from case versus control comparisons), and use the average expression value in order to collapse probes by gene identifiers. Finally, as a baseline representation, each GEO patient was represented as an array of their specific molecular signature differentially expressed genes (DEG) expression values.

Mapping GEO Clinical Concepts

Clinical data provided from GEO were preprocessed and cleaned and any patients with missing clinical values were excluded. Additionally, to ensure the best possible alignment with the molecular mechanism embeddings, all relevant clinical attributes were manually mapped to PheKnowLator entity identifiers. To help with interpretation and the mapping of measurement results [if these data were provided], age-matched pediatric reference ranges were obtained from the Mayo Clinic Pediatric Testing Catalog[4].

Gene Expression Signature-Adjusted Molecular Mechanism Embeddings

Molecular mechanism embeddings from PheKnowLator were created for each patient using a similar approach as to what was done for the CHCO patients (Figure 2). First, PheKnowLator gene embeddings were obtained for each patient's DEGs. Each gene embedding was adjusted by multiplying it by its corresponding gene expression value. The final embedding for each patient was created by taking the average of the gene expression signature-adjusted molecular mechanism embeddings and all PheKnowLator KG entity-mapped clinical concept embeddings.

Exploratory Analysis

Descriptive statistics were reported for all clinical attributes, in addition to the counts of DEGs. The usefulness of representing GEO patients as an array of DEG values was examined and

[3]GitHub Gist: https://gist.github.com/callahantiff/04b4d2a653e806b974be4efed41e484a
[4]Mayo Clinic Reference Ranges: https://pediatric.testcatalog.org/show/CBC

visualized, using t-SNE plots [following the same procedures outlined for CHCO patients above], whenever possible, as a function of relevant clinical attributes.

Clustering

K-Means [68] was performed to (i) cluster patient-level molecular signature DEG expression arrays and (ii) cluster molecular mechanism embeddings from CHCO rare disease patients and the gene expression signature-adjusted molecular mechanism embeddings for GEO cases. For the DEG arrays, k was selected using a variant of the elbow method, where the "elbow" was defined as the clustering step with the largest distance growth acceleration [74]. When clustering GEO and CHCO patients, a predefined k of four (i.e., the number of examined rare disease groups) was used. Clusters were examined using the same metrics and procedures that were applied to the CHCO data including consultation with a PhD-level molecular biologist.

Patient Similarity

Two tasks were performed in an effort to demonstrate that GEO cases were closer (i.e., more similar) in vector space to CHCO patients with the same rare disease than all other CHCO rare disease and control patient groups: (i) Top-N true positive (TP) Rate and (ii) Kruskal-Wallis Nonparametric ANOVAs.

Top-N TP Rate

For each rare disease patient, the counts of TPs, defined as the presence of a CHCO patient with the same rare disease, were calculated for the top 10-100 (in increments of 10) most similar patients. This experiment was performed for rare disease patients and the combined sample of rare disease and control patients using only clinical embeddings and molecular mechanism embeddings.

Kruskal-Wallis Nonparametric ANOVAs

To determine if the mean rank cosine similarity between the GEO patients and all CHCO rare disease patients differed, a Kruskal-Wallis nonparametric ANOVA with Dunn-Bonferroni post-hoc tests was run. Eta-squared, was used as a measure of effect size [75], which ranges from 0.0-1.0, and is reported as the percentage of variance in the dependent variable explained by the independent variable. The importance of these values were interpreted using standard Cohen cut-offs [76]: < 0.10 (small), 0.11-0.39 (moderate) and ≥ 0.40 (large). For both tasks, pairwise cosine similarity was calculated between all GEO case patients and all CHCO rare disease patients. Cosine similarity was computed as the normalized dot product of two vectors X and Y:

$$K(X, Y) = < X, Y > / (||X|| * ||Y||), \quad (8)$$

Where scores ranged from -1.0 to 1.0, where -1.0 indicates strongly dissimilar vectors, 0.0 indicates independent vectors, and 1.0 indicates positive collinear vectors.

All preprocessing of data, construction of patient representations, and evaluation tasks were performed using Python 3.6.2 on a single machine with 8 cores and 16GB of RAM. Clinical embeddings were created using Gensim (v.3.8.3) [77] and node embeddings were created from

a PheKnowLator KG using Walking RDF and OWL[5] [65]. Visualizations were created using Matplotlib ([78]; v.3.3.2) and multiclass classification and K-Means clustering was performed using scikit-learn ([79]; v.0.22.1). The primary GitHub repository for this project is currently private, but all code, KGs, and mapped concept lists will be made publicly available upon publication of the work in this chapter (https://github.com/callahantiff/Med2Mech).

RESULTS

As shown in Table 2, a total of 2,646 rare disease (CAH: 760, CF: 835, PKU: 235, and SCD: 816) and 10,000 similarly medically complex control patients. The number of identified patients varied slightly by clinical domain. For rare diseases patients, all patients were available when using conditions, measurements, and data from all clinical domains. When using medications, the number of available CF, PKU, and SCD patients decreased slightly (CF: 835 to 822, PKU: 235 to 234, and SCD: 816 to 814). For controls, all patients were only available when using data from all clinical domains. The resulting set of unique features contained 259 measurement, 2,097 medication, 4,933 condition, and 7,298 all clinical domain concepts. When including controls, the resulting set of unique features included 272 measurement, 2,334 medication, 6,382 condition, and 8,988 all clinical domain concepts.

Human Disease Knowledge Graph

The PheKnowLator KG contained 129,875 nodes and 3,838,935 edges. Of these nodes and edges, 116,158 nodes and 3,593,567 edges were generated as part of the 14 edge types that were added to the base set of ontologies (i.e., GO, DOID, and HP). These nodes comprised 49,185 GO concepts, 23,776 genes, 15,019 drugs, 13,159 phenotypes, 11,124 pathways, and 3,744 diseases, and 151 relations. The counts of entities by edge type are provided in Table 3. As shown in this table, the greatest number of edges were `substanceThatTreats(chemical, disease)` (n = 1,217,160), followed by `participatesIn(chemical, pathway)` (n = 711,043), and `interactsWith(gene, gene)` (n = 594,100). The smallest number of edges were created for `hasPart(pathway, GO-Biological Process)` (n = 1,008), `realizes(pathway, GO-Molecular Function)` (n = 3,477), and `hasComponent(pathway, GO-Cellular Component)` (n = 13,421).

Terminology Mapping

The count of clinical concepts, by clinical domain, that were successfully mapped to entities in the PheKnowLator KG are shown in Table 4. Of the 6,832 condition concepts, 66.9% (n = 4,267) were mapped to 5,315 KG entities (1.3 entities/concept). The majority of these mappings were automatically generated (n = 2,900). Of the 2,334 medication concepts 2,327 (99.7%) were mapped to 3,259 KG entities (1.4 entities/concept). Of the 272 measurement concepts, 128 were manually mapped to 180 (47.1%) KG entities (1.4 entities/concept). All of these mappings were manually created. Of the 8,898 concepts spanning all clinical domains, 5,678 were mapped to 8,754 KG entities (1.5 entities/concept). Similar to conditions and medications,

[5]walking-rdf-and-owl: https://github.com/bio-ontology-research-group/walking-rdf-and-owl

Table 2. Counts of Patients and Features by Disease Group and Clinical Domain.

Group	Clinical Domain	Patient Count	Group Counts					Feature Count
			CF	CAH	PKU	SCD	Control	
Case	Conditions	2646	835	760	235	816	N/A	4933
	Medications	2473	822	752	91	808	N/A	2097
	Measurements	2642	834	760	234	814	N/A	259
	All Domains	2646	835	760	235	816	N/A	7289

Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; SCD: Sickle Cell Disease.

the majority of these mappings were automatically created (n = 5,360).

Internal Validation

Visualizing Baseline Patient Representations

The baseline representations (BoW + TF-IDF) for each clinical domain are visualized for rare disease patients in Figure 3 and for the combined sample of rare disease and control patients in Figure 4. For rare disease patients, representations built using only conditions (Figure 3 (A)) or all clinical domains (Figure 3 (D)) appeared to create the cleanest separation of patients by rare disease. Representations built using medications (Figure 3 (B)) appeared to create the visualization with fewest homogeneous groups and yielded several small mixed groups of two or more rare diseases. Regardless of the clinical domain, the combined sample t-SNE plots resulted in less meaningful groupings than when only using rare disease patients. When only considering the combined sample representations, those built using condition concepts (Figure 4 (A)) appeared to create slightly better groupings than all other clinical domains.

Multiclass Classification

Multiclass classification performance is presented by metric in Figure 5 (accuracy), Figure 6 (precision), Figure 7 (recall), and Figure 8 (F_1 score). Detailed results are provided in Supplemental Material Table 1. First, results are presented with respect to the baseline representations. Then, performance is described as a function of the classifier, performance metric, clinical domain, and representation parameterization. The baseline representations (BoW + TF-IDF) performed competitively in comparison to the Doc2Vec clinical representations and the averaged and summed molecular mechanism representations. The best baseline performance, across all of the metrics (0.99), was achieved when using the SVM classifier on condition concepts. This score was higher than the highest scoring Doc2Vec parameterization across any clinical domain and classifier (precision of 0.83 achieved when using the SVM

Table 3. Human Disease Knowledge Graph Edge Count by Type.

Edges	Relation	Subjects	Objects	Edges
Gene-GO Biological Process	participates in (BFO:0000056)	16,808	11776	130901
Gene-GO Molecular Function	has function (RO:0000085)	16779	4107	60756
Gene-GO Cellular Component	located in (RO:0001025)	17794	1620	73345
Pathway-GO Biological Process	has part (BFO:0000051)	1008	633	1008
Pathway-GO Molecular Function	realizes (BFO:0000055)	3420	1179	3477
Pathway-GO Cellular Component	has component (RO:0002180)	9411	98	13421
Gene-Phenotype	has phenotype (RO:0002200)	3682	6651	120288
Disease-Phenotype	has symptom (RO:0002452)	2261	5438	43848
Chemical-Gene	interacts with (RO:0002434)	8811	22995	410379
Chemical-Pathway	participates in (RO:0000056)	8886	1707	711043
Chemical-Disease	is substance that treats (RO:0002606)	14238	2564	1217160
Gene-Pathway	participates in (BFO:0000056)	10370	1860	107029
Pathway-Disease	causes or contributes to condition (RO:0003302)	1818	1649	106798
Gene-Gene	interacts with (RO:0002434)	13592	13592	594100

Acronyms - BFO: Basic Formal Ontology; GO: Gene Ontology; RO: Relation Ontology.

Table 4. OMOP Standard Terminology and Knowledge Graph Entity Alignment.

Clinical Domain	Patient Count	Group Counts					Feature Count
		CF	CAH	PKU	SCD	Control	
Conditions	12273	834	760	235	809	9635	5315
Medications	10911	818	752	88	799	8454	3259
Measurements	4994	748	638	107	567	2934	180
All Domains	12533	835	760	235	813	9890	8754

Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; SCD: Sickle Cell Disease.

classifier on medication concepts represented using the PV-DBOW representation with 150 dimensions and a sliding window size of 15). Similar as observed for the Doc2Vec representations, the best baseline representation performance, also out-performed the best molecular mechanism performance (0.99 for all performance metrics achieved when using the logistic regression classier on condition concepts represented using average aggregation). The worst performance was an F_1 score of 0.56, which was achieved when using the Näive Bayes classifier on measurement concepts. The best performing Doc2Vec representation, as mentioned above, achieved a precision of 0.83 achieved when using the SVM classifier on medication concepts represented using the PV-DBOW model (150 dimensions and a sliding window size of 15). The worst performing representation achieved an accuracy and recall of 0.50, which was achieved when using the Näive Bayes classifier on medication concepts represented using the PV-DM model (150 dimensions and a sliding window size of 20). The best performing molecular mechanism representation, as mentioned above, scored 0.99 on all performance metrics achieved when using the logistic regression classier on condition concepts represented using the average aggregation. The worst performing representation, which performed better than 82.2% of all scored Doc2Vec representation parameterizations, had an accuracy and precision of 0.81, which was achieved when using the Näive Bayes classifier on all clinical domains represented using the average aggregation.

Best Performing Classifiers

For Doc2Vec [and the BoW+TF-IDF baseline] representations, the SVM classifier resulted in the best performance for 99/100 representation parameterizations (precision = 0.83) followed by the logistic regression classier (precision = 0.82; 1/100 representations), and the Näive Bayes classifier (precision = 0.75; 0/100 representations). In contrast, for the molecular mechanism representations, the logistic regression classifier resulted in the best performance for 7/8 representation parameterizations (all metrics = 0.99) followed by the SVM classifier (all metrics = 0.95; 1/100 representations), and the Näive Bayes classifier (accuracy, precision, and recall = 0.95; 0/100 representations).

Figure 3. t-SNE Plots of BoW TF-IDF Rare Disease Baseline Representations by Clinical Domain.

This plot illustrates how baseline (Bag-of-Words + TF-IDF) rare disease patient representations built using clinical concepts differ by clinical domain: (A) only condition concepts, (B) only medication concepts, (C) only measurement concepts, and (D) all clinical domains. Colors correspond to the following groups: Cystic Fibrosis (CF; orange); Phenylketonuria (PKU; green); Congenital Hypothyroidism (CAH; purple); Sickle Cell Disease (SCD; blue). Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; SCD: Sickle Cell Disease.

Figure 4. t-SNE Plots of BoW TF-IDF Combined Sample Baseline Representations by Clinical Domain.

This plot illustrates how baseline (Bag-of-Words + TF-IDF) rare disease and control patient representations, built using concept concepts, differ by clinical domain: (A) only condition concepts, (B) only medication concepts, (C) only measurement concepts, and (D) all clinical domains. Colors correspond to the following groups: Cystic Fibrosis (CF; orange); Phenylketonuria (PKU; green); Congenital Hypothyroidism (CAH; purple); Sickle Cell Disease (SCD; blue); Control (yellow). Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; SCD: Sickle Cell Disease

Figure 5. Representation Classification Performance - Average Accuracy.

Figure presents the classification accuracy for all clinical configurations (bar plots, colored by parameter setting), molecular mechanism embeddings (red [averaged aggregation], and orange [summed aggregation] lines), and baseline representations (black dashed lines) by clinical domain: (A) conditions, (B) medications, (C) measurements, and (D) all clinical domains. Acronyms - BoW: Bag-Of-Words; PV-DBOW: Paragraph Vector Distributed Bag of Words; PV-DM: Paragraph Vector Distributed Memory; TF-IDF: Term Frequency–Inverse Document Frequency.

Figure 6. Representation Classification Performance - Average Precision.

Figure presents the classification precision for all clinical configurations (bar plots, colored by parameter setting), molecular mechanism embeddings (red [averaged aggregation], and orange [summed aggregation] lines), and baseline representations (black dashed lines) by clinical domain: (A) conditions, (B) medications, (C) measurements, and (D) all clinical domains. Acronyms - BoW: Bag-Of-Words; PV-DBOW: Paragraph Vector Distributed Bag of Words; PV-DM: Paragraph Vector Distributed Memory; TF-IDF: Term Frequency–Inverse Document Frequency.

Figure 7. Representation Classification Performance - Average Recall.

Figure presents the classification recall for all clinical configurations (bar plots, colored by parameter setting), molecular mechanism embeddings (red [averaged aggregation], and orange [summed aggregation] lines), and baseline representations (black dashed lines) by clinical domain: (A) conditions, (B) medications, (C) measurements, and (D) all clinical domains. Acronyms - BoW: Bag-Of-Words; PV-DBOW: Paragraph Vector Distributed Bag of Words; PV-DM: Paragraph Vector Distributed Memory; TF-IDF: Term Frequency–Inverse Document Frequency

Figure 8. Representation Classification Performance - Average F_1 score.

Figure presents the classification F_1 score for all clinical configurations (bar plots, colored by parameter setting), molecular mechanism embeddings (red [averaged aggregation], and orange [summed aggregation] lines), and baseline representations (black dashed lines) by clinical domain: (A) conditions, (B) medications, (C) measurements, and (D) all clinical domains. Acronyms - BoW: Bag-Of-Words; PV-DBOW: Paragraph Vector Distributed Bag of Words; PV-DM: Paragraph Vector Distributed Memory; TF-IDF: Term Frequency–Inverse Document Frequency.

Best Performance Metrics

For the Doc2Vec representations, precision, accuracy, recall, and F_1 score were the highest when using the SVM classifier on medication concepts represented using the PV-DBOW model with 150 dimensions and a sliding window size of 15 (precision: 0.84, accuracy: 0.83, recall: 0.83, and F_1 score: 0.83). For the molecular mechanism representations, accuracy, precision, recall, and F_1 score were the highest when using the logistic regression classifier on representations built from the averaged condition concepts (accuracy, precision, recall, and F_1 score: 0.99).

Clinical Domain

For the Doc2Vec representations, the best performance, regardless of classifier or performance metric, was achieved when using the PV-DBOW model with 150 dimensions and a sliding window size of 15 on medications (0.83), measurements (0.83), all clinical domain concepts (0.83), and condition concepts (0.83). The worst performance was achieved when using the PV-DM model with 150 dimensions and a sliding window size of 20 on condition concepts (0.499) or medication concepts (0.497). For molecular mechanism embeddings, the best performing clinical domains were averaged conditions (0.99). The worst performance was achieved when averaging concepts from all clinical domains (0.81).

Representation Parameterization

For Doc2Vec, on average, performance was better for the PV-DBOW models (conditions: 0.83 versus 0.78, medications: 0.83 versus 0.79, measurements: 0.83 versus 0.77, and all clinical domains: 0.83 versus 0.77). For molecular mechanisms, on average, performance was better for the summed representations when applied to measurement (0.97 versus 0.97) and all clinical domain concepts (0.99 versus 0.95). For condition (0.993 versus 0.991) and medication (0.988 versus 0.986) concepts, the averaged representations performed best.

Visualizing Clinical and Molecular Mechanism Patient Representations

The best configurations of the PV-DM models for each clinical domain are visualized using t-SNE plots for rare disease patients in Figure 9 and for the combined sample of rare disease and control patients in Figure 10. For both patient samples, the representations built using all clinical domains resulted in the tightest distinct patient groups [by rare disease] (Figure 9 (D) and Figure 10 (D), respectively), the best performance across all of the performance metrics was achieved when using medication data (M= 0.72, 0.60-0.79; Figure 9 (B) and Figure 10 (B), respectively). For all clinical domains (conditions: M= 0.71, 0.61-0.78; measurements: M= 0.71, 0.62-0.77; and all clinical domains: M= 0.71, 0.61-0.77), the best performing PV-DM models were configured with 150 dimensions and a sliding window size of five.

The best configurations of the PV-DBOW models for each clinical domain are visualized using t-SNE plots for rare disease patients in Figure 11 and for the combined sample of rare disease and control patients in Figure 12. For rare disease patients, representations built using conditions (Figure 11 (A)), medications Figure 11 (B), and all clinical domains (Figure 11 (D), resulted in the best visual representation for all rare disease groups. In contrast, for the combined sample, representations built using measurements (Figure 12 (D)) created the most

Figure 9. t-SNE Plot for Best Performing Rare Disease PV-DM Representation Parameters by Clinical Domain.

This plot illustrates how rare disease patient clinical representations built using PV-DM (150 dimensions, sliding window size of 5) differ by clinical domain: (A) only conditions, (B) only medications, (C) only measurements, and (D) all clinical domains. Colors correspond to disease groups. Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; PV-DM: Paragraph Vector Distributed Memory; SCD: Sickle Cell Disease.

Figure 10. t-SNE Plot for Best Performing Combined Sample PV-DM Representation Parameters by Clinical Domain.

This plot illustrates how rare disease and control patient clinical representations built using PV-DM (150 dimensions, sliding window size of 5) differ by clinical domain: (A) only conditions, (B) only medications, (C) only measurements, and (D) all clinical domains. Colors correspond to disease groups. Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; PV-DM: Paragraph Vector Distributed Memory; SCD: Sickle Cell Disease.

Figure 11. t-SNE Plot for Best Performing Rare Disease PV-DBOW Representation Parameters by Clinical Domain.

This plot illustrates how rare disease patient clinical representations built using PV-DBOW differ by clinical domain (dimensions, sliding window size): (A) only conditions (200, 5), (B) only medications (150, 15), (C) only measurements (150, 20), and (D) all clinical domains (150, 5). Colors correspond to disease groups. Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; PV-DBOW: Paragraph Vector Distributed Bag of Words; SCD: Sickle Cell Disease.

Figure 12. t-SNE Plot for Best Performing Combined Sample PV-DBOW Representation Parameters by Clinical Domain.
This plot illustrates how rare disease and control patient clinical representations built using PV-DBOW differ by clinical domain (dimensions, sliding window size): (A) only conditions (200, 5), (B) only medications (150, 15), (C) only measurements (150 dimensions, sliding window size of 20), and (D) all clinical domains (150, 5). Colors correspond to disease groups. Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; PV-DBOW: Paragraph Vector Distributed Bag of Words; SCD: Sickle Cell Disease.

distinct rare disease groups, even resulting in groups of control patients. The best performance across all of the performance metrics was achieved when using measurements (M= 0.79, 0.73-0.83) configured with 150 dimensions and all clinical domains (M= 0.79, 0.73-0.83), the best performing representations were sliding window length of 20. For medications (M= 0.79, 0.73-0.83) and all also built using 150 dimensions. In contrast to the medication configurations, when using conditions, which were the second best performing domain, 200 dimensions and a sliding window size of five was used (M= 0.79, 0.73-0.83).

The summed molecular mechanism embeddings for each clinical domain are visualized using t-SNE plots for rare disease patients in Figure 13 and for the combined sample of rare disease and control patients in Figure 14. The representations created for the rare disease patients displayed patterns that were very different from what was observed for the clinical embeddings. The clinical domains which resulted in the best visualizations were conditions (Figure 13 (A)) and all clinical domains (Figure 13 (D)), which was also observed for the combined sample visualizations in Figure 14 (A) and Figure 14 (D), respectively. These observations were supported by the classification results, which when applied to these domains resulted in the best performance across all of the performance metrics (conditions: M= 0.96, 0.93-0.99; all clinical domains: M= 0.96, 0.93-0.95).

The averaged molecular mechanism embeddings for each clinical domain are visualized using t-SNE plots for rare disease patients in Figure 15 and for the combined sample of rare disease and control patients in Figure 16. The clinical domains which resulted in the best visualizations were the same as what was observed for the summed molecular mechanism embeddings, for both patient samples: conditions (Figure 15 (A), Figure 16 (A)) and all clinical domains (Figure 15 (D), Figure 16 (D)). These observations were supported by the classification results, which when applied to these domains resulted in the best performance across all of the metrics (conditions: M= 0.96, 0.93-0.99; all clinical domains: M= 0.91, 0.81-0.95).

The best configuration [determined using performance on the SVM classifier] of each clinical model implementation when using data from all clinical domains was found when using 150 dimensions and a sliding window size of five. Using these settings, the PV-DM implementation achieved an average accuracy, recall, and F_1 score of 0.75 and precision of 0.77. For this configuration, the PV-DBOW implementation achieved an average accuracy, recall, and F_1 score of 0.82 and a precision of 0.83. The summed molecular mechanism embeddings created using data from all clinical domains achieved an average accuracy, precision, recall, and F_1 score of 0.95. The averaged molecular mechanism embeddings created using data from all clinical domains achieved an average accuracy, precision, recall, and F_1 score of 0.95. For the remaining *Internal Evaluation* experiments and all *External Evaluation* experiments, clinical embeddings were created using the PV-DBOW implementation with 150 dimensions and a sliding window size of five and molecular mechanism embeddings created using the average aggregation.

Figure 13. t-SNE Plot for Rare Disease Summed Molecular Mechanism Embedding by Clinical Domain.

This plot illustrates how rare disease patient molecular mechanism representations built using summed aggregation differ by clinical domain (512 dimensions, sliding window size of 10): (A) only conditions, (B) only medications, (C) only measurements, and (D) all clinical domains. Colors correspond to the following groups: Cystic Fibrosis (CF; orange); Phenylketonuria (PKU; green); Congenital Hypothyroidism (CAH; purple); Sickle Cell Disease (SCD; blue).

Figure 14. t-SNE Plot for Combined Sample Summed Molecular Mechanism Embedding by Clinical Domain.

This plot illustrates how rare disease and control patient molecular mechanism representations built using summed aggregation differ by clinical domain (512 dimensions, sliding window size of 10): (A) only conditions, (B) only medications, (C) only measurements, and (D) all clinical domains. Colors correspond to the following groups: Cystic Fibrosis (CF; orange); Phenylketonuria (PKU; green); Congenital Hypothyroidism (CAH; purple); Sickle Cell Disease (SCD; blue); control (yellow).

Figure 15. t-SNE Plot for Rare Disease Averaged Molecular Mechanism Embedding by Clinical Domain.

This plot illustrates how rare disease patient molecular mechanism representations built using averaged aggregation differ by clinical domain (512 dimensions, sliding window size of 10): (A) only conditions, (B) only medications, (C) only measurements, and (D) all clinical domains. Colors correspond to the following groups: Cystic Fibrosis (CF; orange); Phenylketonuria (PKU; green); Congenital Hypothyroidism (CAH; purple); Sickle Cell Disease (SCD; blue).

Figure 16. t-SNE Plot for Combined Sample Averaged Molecular Mechanism Embedding by Clinical Domain.

This plot illustrates how rare disease and control patient molecular mechanism representations built using averaged aggregation differ by clinical domain (512 dimensions, sliding window size of 10): (A) only conditions, (B) only medications, (C) only measurements, and (D) all clinical domains. Colors correspond to the following groups: Cystic Fibrosis (CF; orange); Phenylketonuria (PKU; green); Congenital Hypothyroidism (CAH; purple); Sickle Cell Disease (SCD; blue); control (yellow).

K-Means Clustering

T-SNE plots for each embedding type are shown in Figure 17 (A) and Figure 17 (B), respectively. Results from applying K-Means to the clinical embeddings using a k of six are visualized in Figure 17 (C) and described in Table 5. Results from applying K-Means to the molecular mechanism embeddings using a k of nine are visualized in Figure 17 (D) and described in Table 6. For each table, two types of feature lists are reported: (i) the top five most frequently occurring clinical concepts and (i) the top-five most frequently occurring clinical concepts unique to each cluster. Each table also includes two ways of measuring the homogeneity of a cluster with respect to the disease type it represents, which are bolded and highlighted (i.e., colors are consistent with those used in the t-SNE plots: **CAH**, **CF**, **PKU**, **SCD**): (i) **Cluster Breakdown:** the percent of patients from each rare disease group out of the total patients in that cluster; (ii) **Disease Breakdown:** the percent of rare disease patients for each disease group out of the total number of patients in that disease group. Taken together, these metrics provide a means for interpreting the distribution of a cluster with respect to each disease (e.g., 73% of the patients in a specific cluster have CF [cluster breakdown], which represents 45% of the total sample of CF patients [disease breakdown]). Clinical and molecular mechanism embeddings achieved similar performance when evaluated using the clustering metrics (Clinical: 0.52 ARI, 0.55 AMI, 0.83 purity; Clinical: 0.38 ARI, 0.529 AMI, 0.86 purity).

Clinical Clusters

As demonstrated in Table 5, the clustering results revealed three SCD clusters and one CF, CAH, and PKU cluster.

- **Cluster Zero.** 96.2% (n = 256) of the patients in this cluster had SCD, which represented 31.4% of all SCD patients. This cluster was defined by 3,437 concepts, of which 430 were unique to patients in this cluster. This cluster had the highest frequency of hemoglobin SS disease without crisis (omop:30683; n = 818) and sickle cell-hemoglobin C disease with crisis (omop:443721; n = 42). High frequency concepts unique to this cluster included: Lyrica (omop:19123163; n = 1864), hemophilia B (omop:436678; n = 540), and Mucopolysaccharidosis (omop:433446; n = 474). Sickle cell-specific terms unique to this cluster included hereditary spherocytosis (omop:24909; n = 98), hereditary persistence of fetal hemoglobin (omop:4098015; n = 2), hemoglobin E/beta thalassemia disease (omop:4120455; n = 2), acute chest syndrome (omop:254062; n = 2), sickle cell-thalassemia disease with crisis (omop:443738; n = 2), and acute sickle cell splenic sequestration crisis (omop:40485018; n = 2).
- **Cluster One.** 100% (n = 648) of the patients in this cluster had CF, which represented 77.6% of the CF patients. This cluster was defined by 3,867 concepts, of which 576 were unique to patients in this cluster. This cluster contained no counts of any of the concepts used to define the other rare disease groups. High frequency concepts unique to this cluster included: Nephrocaps (omop:40233393; n = 420), Beconase 0.042 MG (omop:36250144; n = 254), and Urecholine (omop:40162842; n = 252).

Figure 17. K-Means Clustering of the Best Performing Clinical and Molecular Mechanism Embeddings for Rare Disease Patients using Clinical Domains.

The best performing configuration for each model clinical and molecular mechanism model when applied to all clinical domains for rare disease patients (dimensions, sliding window size): (A) PV-DBOW (150, 5), (B) averaged molecular mechanism embeddings. K-Means clustering results are presented in (C) clinical embeddings (k = 6), and (D) molecular mechanism embeddings (k = 9). Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; PV-DBOW: Paragraph Vector Distributed Bag of Words; SCD: Sickle Cell Disease.

Table 5. Clustering Results for the Best PV-DBOW Configuration.

Total Patients by Disease Group: CAH (n = 760); CF (n = 835); PKU (n = 235); SCD (n = 816)		
0	Cluster Breakdown	**Number of Patients: 266** **SCD: 96.24%**; CF: 1.88%; CAH: 1.88%
	Disease Breakdown	**SCD: 31.37%**; CF: 0.60%; CAH: 0.66%
	Top Features (n = 3,437)	Lyrica (omop:19123163): 1,864 Creon (omop:40166158): 1,540 Tobramycin 60 MG/ML Inhalant Solution (omop:902806): 1,272 Topiramate 100 MG Oral Tablet (omop:742304): 784 Hemoglobin (omop:3000963): 648
	Unique Features (n = 430)	Lyrica (omop:19123163): 1,864 Hereditary factor IX deficiency disease (omop:436678):540 Mucopolysaccharidosis (omop:433446): 474 Moyamoya disease (omop:378774): 330 Fenofibrate 54 MG Oral Tablet (omop:1551839): 274
1	Cluster Breakdown	**Number of Patients: 648** **CF: 100%**
	Disease Breakdown	**CF: 77.60%**
	Top Features (n = 3,867)	Dilaudid (omop:40173138): 1,636 Gabapentin 400 MG Oral Capsule (omop:19077549): 540 Levetiracetam 500 MG Oral Tablet (omop:711589): 508 Sildenafil 25 MG Oral Tablet (omop:1316268): 496 Marinol (omop:19029811): 436
	Unique Features (n = 576)	Nephrocaps (omop:40233393): 420 Beconase 0.042 MG (omop:36250144): 254 Urecholine (omop:40162842): 252 Escitalopram 20 MG Oral Tablet (omop:19098166): 136 Corgard (omop:19034018): 120
2	Cluster Breakdown	**Number of Patients: 593** **CAH: 97.13%;** SCD: 1.52%; CF: 1.18%; PKU: 0.17%
	Disease Breakdown	**CAH: 75.79%**; SCD: 1.10%; CF: 0.84%; PKU: 0.43%
	Top Features (n = 3,358)	Neurontin (omop:797437): 522 Keppra (omop:19102835): 386 Cholecalciferol 13300 UNT/ML Oral Solution (omop:19095250): 384 sildenafil 25 MG Oral Tablet (omop:1316268): 320
	Unique Features (n = 341)	Ancobon (omop:19034615): 174 Xalatan (omop:954692): 162 Cardiac transplant disorder (omop:320299): 156 Prune belly syndrome (omop:201397): 116 Nephrotic syndrome (omop:195314): 112

Total Patients by Disease Group: CAH (n = 760); CF (n = 835); PKU (n = 235); SCD (n = 816)		
3	Cluster Breakdown	**Number of Patients: 343** **SCD: 51.31%;** CF: 33.24%; CAH: 12.83%; PKU: 2.62%
	Disease Breakdown	**SCD: 21.57%**; CF: 13.65%; CAH: 5.79%; PKU: 3.83%
	Top Features (n = 4,138)	Sildenafil 25 MG Oral Tablet (omop:1316268): 3,284 Ursodiol 250 MG Oral Tablet (omop:40165331): 1,032 Buspar (omop:40167432): 728 Mesalamine 400 MG Delayed Release Oral Tablet (omop:19078859): 678 Vimpat (omop:40222474): 632
	Unique Features (n = 615)	Sildenafil 25 MG Oral Tablet (omop:1316268): 3,284 Ursodiol 250 MG Oral Tablet (omop:40165331): 1,032 Buspar (omop:40167432): 728 Mesalamine 400 MG Delayed Release Oral Tablet (omop:19078859): 678 Vimpat (omop:40222474): 632
4	Cluster Breakdown	**Number of Patients: 416** **PKU: 51.68%**; CAH: 24.04%; CF: 12.26%; SCD: 12.02%
	Disease Breakdown	**PKU: 91.49%**; CAH: 13.16%; CF: 6.11%; SCD: 6.13%
	Top Features (n = 3,590)	Potassium Chloride 1.33 MEQ/ML Oral Solution (omop:19049335): 1,154 Paralytic syndrome on one side of the body from CVA (omop:43530742): 828 Bumex (omop:19029374): 662 Sodium Chloride 234 MG/ML Injectable Solution (omop:19079551): 656 Phenytoin sodium 100 MG Extended Release Oral Cap (omop:40163208): 644
	Unique Features (n = 421)	Phenytoin sodium 100 MG Extended Release Oral Cap (omop:40163208): 644 Congenital stenosis of aortic valve (omop:314457): 500 Dilantin (omop:40163230): 282 5p partial monosomy syndrome (omop:440518): 254 Riboflavin 50 MG Oral Capsule (omop:19062880): 224
5	Cluster Breakdown	**Number of Patients: 380** **SCD: 85.53%;** CAH: 9.21%; CF: 2.63%; PKU: 2.63%
	Disease Breakdown	**SCD: 39.83%**; CAH: 4.61%; PKU: 4.26%; CF: 1.20%
	Top Features (n = 2,517)	Triamcinolone Acetonide 1 MG/ML Topical Cream (omop:40228087): 496 Calcium Carbonate 250 MG/ML Oral Suspension (omop:19074522): 340 Advair (omop:42902351) : 334 Disorder of acoustic nerve (omop:439842): 304 Speech and language develop delay due to hearing loss (omop:40480431): 290
	Unique Features (n = 214)	Erythromycin 0.02 MG/MG Topical Gel (omop:19076956): 20 Ocular laceration without intraocular prolapse (omop:380575): 18 Cervical spina bifida with hydrocephalus (omop:376210): 18 Zarontin (omop:750145): 18 Light-for-dates with signs of fetal malnutrition (omop:72726): 16

Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; SCD: Sickle Cell Disease.
Note. Physiological measurements like blood pressure and respiratory rate, were excluded from the top features.

- **Cluster Two.** 97.1% (n = 576) of the patients in this cluster had CAH, which represented 75.8% of CAH patients. This cluster was defined by 3,358 concepts, of which 341 were unique to patients in this cluster. High frequency concepts unique to this cluster included: Ancobon (omop:19034615; n = 174), Xalatan (omop:954692; n = 162), and cardiac transplant disorder (omop:320299; n = 156).
- **Cluster Three.** 51.3% (n = 176) of the patients in this cluster had SCD, which represented 21.6% of all SCD patients. This cluster was defined by 3,358 concepts, of which 341 were unique to patients in this cluster. The only sickle cell-related concept identified in this cluster was Sickle cell trait (omop:25518; n = 40). High frequency concepts unique to this cluster included: sildenafil 25 MG Oral Tablet (omop:1316268; n = 3,284), Ursodiol 250 MG Oral Tablet (omop:40165331; n = 1,032), and Buspar (omop:40167432; n = 728).
- **Cluster Four** 51.7% (n = 215) of the patients in this cluster had PKU, which represented 91.5% of PKU patients. This cluster was defined by 3,590 concepts, of which 421 were unique to patients in this cluster. High frequency concepts unique to this cluster included: Phenytoin sodium 100 MG Extended release oral cap (omop:40163208; n = 644), congenital stenosis of aortic valve (omop:314457; n = 500), and Dilantin (omop:40163230; n = 282).
- **Cluster Five.** 85.5% (n = 325) of the patients in this cluster had SCD, which represented 39.8% of SCD patients. This cluster was defined by 2,517 concepts, of which 214 were unique to patients in this cluster. Sickle-cell related concepts identified in cluster included sickle cell-thalassemia disease without crisis (omop:321263; n = 2), heterozygous thalassemia (omop:4023612; n = 2), and thalassemia (omop:30978; n = 2). High frequency concepts unique to this cluster included: erythromycin 0.02 MG/MG Topical Gel (omop:19076956; n = 20), ocular laceration without intraocular prolapse (omop:380575; n = 18), and cervical spina bifida with hydrocephalus (omop:376210; n = 18).

Molecular Mechanism Clusters

As demonstrated in Table 6, the clustering results revealed three SCD clusters, two CF clusters (both contained only CF patients), three CAH clusters (one contained only CAH patients), and one PKU cluster (containing only PKU patients).

- **Cluster Zero.** 68.2% (n = 247) of the patients in this cluster had SCD, which represented 30.4% of SCD patients. This cluster was defined by 1,074 concepts, of which 42 were unique to patients in this cluster. This cluster contained concepts used to define all of the rare disease groups (CAH: 12, PKU: 8, CF: 6). With respect to SCD codes, this cluster contained a higher frequency of hemoglobinopathy concepts (n = 8) than the other SCD clusters. High frequency concepts unique to this cluster included: diabetic autonomic neuropathy (DOID:11503; n = 44), bicornuate uterus (HP:0000813; n = 40), and aphakia (HP:0007707; n = 22).
- **Cluster One.** 100% (n = 322) of the patients in this cluster had SCD, which represented

38.6% of SCD patients. This cluster was defined by 1,329 concepts, of which 73 were unique to patients in this cluster. This cluster also contained 22 concepts used to define CF patients, which was the second highest count of these CF concepts. With respect to SCD codes, this cluster contained two thalassemia and hemoglobinopathy codes. High frequency concepts unique to this cluster included: nadolol (MeSH:D009248; n = 120), chloramphenicol (MeSH:D002701; n = 94), and ziprasidone (MeSH:C092292; n = 70).

- **Cluster Two.** 65.5% (n = 253) of the patients in this cluster had CAH, which represented 33.3% of CAH patients. This cluster was defined by 1,879 concepts, of which 306 were unique to patients in this cluster. This cluster contained concepts used to define all of the rare disease groups (PKU: 18, CF: 2, SCD: 36). With respect to CAH codes, this cluster was tied for the highest frequency of codes with Cluster Seven. High frequency concepts unique to this cluster included: hemophilia B (DOID:12259; n = 540), cri-du-chat syndrome (DOID:12580; n = 254), and dysostosis multiplex (HP:0000943; n = 226).
- **Cluster Three.** 84.7% (n = 419) of the patients in this cluster had SCD, which represented 51.5% of SCD patients. This cluster was defined by 1,610 concepts, of which 131 were unique to patients in this cluster. This cluster contained concepts used to define all of the rare disease groups (CAH: 4, PKU: 16, CF: 2). With respect to sickle cell codes, this cluster contained the highest frequency of sickle cell anemia concepts (n = 176) than the other SCD clusters. High frequency concepts unique to this cluster included: dabigatran (MeSH:D000069604; n = 244), histiocytosis (DOID:3405; HP:0100727; n = 152), and Evans' syndrome (DOID:8931; n = 124).
- **Cluster Four.** 100% (n = 215) of the patients in this cluster had PKU, which represented 91.5% of PKU patients. This cluster was defined by 319 concepts, of which four were unique to patients in this cluster. This cluster did not contain any concepts used to define any of the other rare disease groups. High frequency concepts unique to this cluster included: pramipexole (MeSH:C061333; n = 6), secondary Parkinson disease (DOID:13548; n = 4), and ropinirole (MeSH:C046649; n = 2).
- **Cluster Five.** 100% (n = 338) of the patients in this cluster had CF, which represented 40.5% of CF patients. This cluster was defined by 684 concepts, of which nine were unique to patients in this cluster. This cluster contained two concepts used to define SCD patients. High frequency concepts unique to this cluster included: x-linked ichthyosis (DOID:1700; n = 10), royal jelly (MeSH:C058787; n = 4), and xerostomia (HP:0000217; n = 4).
- **Cluster Six.** 60.2% (n = 59) of the patients in this cluster had SCD, which represented 7.3% of SCD patients. This cluster was defined by 681 concepts, of which 13 were unique to patients in this cluster. This cluster contained concepts used to define all of the rare disease groups (CAH: 6, PKU: 6, CF: 10). With respect to sickle cell codes, this cluster only contained four. High frequency concepts unique to this cluster included: cervical spina bifida (HP:0005857; n = 18), temporomandibular joint crepitus (HP:0012479; n = 6), and Alosetron (MeSH:C090840; n = 6).

Table 6. Clustering Results for the Best Molecular Mechanism Configuration.

Total Patients by Disease Group: CAH (n = 760); CF (n = 816); PKU (n = 235); SCD (n = 813)		
0	Cluster Breakdown	**Number of Patients: 362** **SCD: 68.23%;** CAH: 16.85% ; CF: 13.26%; PKU: 1.66%
	Disease Breakdown	**SCD: 30.38%**; CAH: 8.03%; CF: 5.75%; PKU: 2.55%
	Top Features (n = 1,074)	sensorineural hearing impairment (DOID:10003; HP:0000407): 472 cerebral palsy (HP:0100021): 294 bosentan (MeSH:C086232): 142 muscular dystrophy (DOID:9884; HP:0003560): 102 torticollis (HP:0000473): 94
	Unique Features (n = 42)	diabetic autonomic neuropathy (DOID:11503): 44 bicornuate uterus (HP:0000813) :40 aphakia (HP:0007707): 22 dermatomyositis (DOID:10223): 20 unicameral bone cyst (HP:0012064): 18
1	Cluster Breakdown	**Number of Patients: 322** **CF: 100%**
	Disease Breakdown	**CF: 38.56%**
	Top Features (n = 1,329)	sildenafil citrate (MeSH:D000068677): 496 pyridoxine (MeSH:D011736): 420 tramadol (MeSH:D014147): 308 ehlers-danlos syndrome (DOID:13359): 268 allergic bronchopulmonary aspergillosis (DOID:13166): 198
	Unique Features (n = 73)	nadolol (MeSH:D009248): 120 chloramphenicol (MeSH:D002701): 94 ziprasidone (MeSH:C092292): 70 nicotine (MeSH:D009538): 70 herpes simplex encephalitis (HP:0012302): 40
2	Cluster Breakdown	**Number of Patients: 386** **CAH: 65.54%**; SCD: 22.28%; CF: 11.14%; PKU: 1.04%
	Disease Breakdown	**CAH: 33.29%**; SCD: 10.58%; CF: 5.15%; PKU: 1.70%
	Top Features (n = 1,879)	buspirone (MeSH:D002065): 728 duloxetine hydrochloride (MeSH:D000068736): 612 dorzolamide (MeSH:C062765): 594 trichloroacetaldehyde (MeSH:C021100): 562 hemophilia b (DOID:12259): 540
	Unique Features (n = 306)	hemophilia b (DOID:12259): 540 cri-du-chat syndrome (DOID:12580): 254 dysostosis multiplex (HP:0000943): 226 flucytosine (MeSH:D005437): 168 latanoprost (MeSH:C072042): 162

Total Patients by Disease Group: CAH (n = 760); CF (n = 816); PKU (n = 235); SCD (n = 813)		
3	Cluster Breakdown	**Number of Patients: 495** **SCD: 84.65%**; CF: 10.91%; CAH: 3.23%; PKU: 1.21%
	Disease Breakdown	**SCD: 54.54%**; CF: 6.47%; PKU: 2.55%; CAH: 2.11%
	Top Features (n = 1,610)	mesalamine (MeSH:D019804): 678 bilateral sensorineural hearing impairment (HP:0008619): 464 labetalol (MeSH:D007741): 352 moyamoya disease (DOID:13099): 330 tacrolimus (MeSH:D016559): 310
	Unique Features (n = 131)	dabigatran (MeSH:D000069604): 244 histiocytosis (DOID:3405; HP:0100727): 152 evans' syndrome (DOID:8931): 124 multiple joint dislocation (HP:0012095): 70 foscarnet (MeSH:D017245): 56
4	Cluster Breakdown	**Number of Patients: 215** **PKU: 100.0%**
	Disease Breakdown	**PKU: 91.49%**
	Top Features (n = 319)	nystatin (MeSH:D009761): 100 disease of metabolism (DOID:0014667): 84 levodopa (MeSH:D007980): 46 acyclovir (MeSH:D000212): 46 generalized dystonia (DOID:0050835): 32
	Unique Features (n = 4)	pramipexole (MeSH:C061333): 6 secondary parkinson disease (DOID:13548): 4 ropinirole (MeSH:C046649): 2 selegiline (MeSH:D012642): 2
5	Cluster Breakdown	**Number of Patients: 338** **CF: 100.0%**
	Disease Breakdown	**CF: 40.48%**
	Top Features (n = 684)	fluticasone (MeSH:D000068298): 404 insulin aspart (MeSH:D061267): 196 nafcillin (MeSH:D009254): 108 citric acid (MeSH:D019343): 100 sodium citrate (MeSH:C102006): 100
	Unique Features (n = 9)	x-linked ichthyosis (DOID:1700): 10 royal jelly (MeSH:C058787): 4 xerostomia (HP:0000217): 4 flunisolide (MeSH:C007734): 2 glutathione (MeSH:D005978): 2

Total Patients by Disease Group: CAH (n = 760); CF (n = 816); PKU (n = 235); SCD (n = 813)		
6	Cluster Breakdown	**Number of Patients: 98** **SCD: 60.2%;** CF: 28.57% ; CAH: 8.16%; PKU: 3.06%
	Disease Breakdown	**SCD: 7.26%**; CF: 3.35% ; CAH: 1.05% ; PKU: 1.28%
	Top Features (n = 681)	agnosia (DOID:4090): 130 apraxia (HP:0002186): 130 chronic rhinitis (HP:0002257): 74 cefotaxime (MeSH:D002439): 70 disease of metabolism (DOID:0014667): 68
	Unique Features (n = 13)	cervical spina bifida (HP:0005857): 18 temporomandibular joint crepitus (HP:0012479): 6 alosetron (MeSH:C090840): 6 impulsivity (HP:0100710): 4 borderline personality disorder (DOID:10930; HP:0012076): 2
7	Cluster Breakdown	**Number of Patients: 368** **CAH: 98.64%**; CF: 0.54%; SCD: 0.54%; PKU: 0.27%
	Disease Breakdown	**CAH: 47.76%**; CF: 0.24%; SCD: 0.25%; PKU: 0.43%
	Top Features (n = 848)	cerebral palsy (HP:0100021): 508 hydrochlorothiazide (MeSH:D006852): 306 diazoxide (MeSH:D003981): 156 abnormality of the nervous system (HP:0000707): 132 down syndrome (DOID:14250): 120
	Unique Features (n = 18)	hyperlordosis (HP:0003307): 64 perineal hypospadias (HP:0000051): 18 hordeolum internum (HP:0010608): 14 fourth cranial nerve palsy (HP:0007011): 6 erythroderma (HP:0001019): 4
8	Cluster Breakdown	**Number of Patients: 59** **CAH: 100.0%**
	Disease Breakdown	**CAH: 7.76%**
	Top Features (n = 110)	torticollis (HP:0000473): 28 psychosis (HP:0000709): 20 down syndrome (DOID:14250): 16 severe short stature (HP:0003510): 12 short stature (HP:0004322): 12
	Unique Features (n = 1)	blue sclerae (HP:0000592): 4

Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; PKU: Phenylketonuria; SCD: Sickle Cell Disease.

- **Cluster Seven.** 98.7% (n = 363) of the patients in this cluster had CAH, which represented 47.8% of CAH patients. This cluster was defined by 848 concepts, of which 18 were unique to patients in this cluster. This cluster contained concepts used to define cystic fibrosis patients (n = 2). With respect to CAH codes, this cluster was tied for the highest frequency of codes with *Cluster Two*. High frequency concepts unique to this cluster included: hyperlordosis (HP:0003307; n = 64), perineal hypospadias (HP:0000051; n = 18), and hordeolum internum (HP:0010608; n = 14).
- **Cluster Eight.** 100% (n = 59) of the patients in this cluster had CAH, which represented 7.8% of the CAH patients. This cluster was defined by 110 concepts, of which only one was unique to patients in this cluster. This cluster did not contain concepts used to define any of the other rare disease groups. The only concept unique to this cluster was blue sclerae (HP:0000592; n = 4).

External Validation

Our domain expert reviewed GEO for studies that provided molecular and clinical information for any of the CHCO rare diseases within a pediatric population. After reviewing the available data, a SCD study titled "The genomic architecture of sickle cell disease in children from West Africa" (GSE35007; [80])[6] was selected. This study aimed to examine the clinical heterogeneity of SCD in a sample of 311 West African children aged 0.2-9.5 years. Of the 311 patients, 250 had either the HbSS or HbSC phenotype and the remaining 61 were age-matched controls (noted as having a normal HbAA or HbAS phenotype). For all patients, whole blood gene expression and eQTL analyses was performed via Illumina HumanHT-12 V4.0 expression beadchips. In addition to the molecular data, this study collected several clinical attributes including red and white blood cell counts, mean corpuscular volume, and SCD severity score. The SCD severity score was designed to predict the risk of death within five years and ranged from 0.0 (least severe) to 1.0 (most severe) [81].

Descriptives

Patients were 54.5% male with an average age of 4.3 years (0.2-9.5). After cleaning the data we were left with 198 SCD patients (n = 51 HbSC phenotype [DOID:2859]; n = 147 HbSS phenotype [DOID:109231) and 61 controls. For detailed patient-level phenotype information, including the ontology terms that were mapped able to be mapped to the three provided laboratory tests (i.e., mean corpuscular volume, white blood cell count, and red blood cell count) and each SCD phenotype, see Supplemental Material Table 2. On average, there were more SCD patients with the HbSS phenotype that had higher counts of white blood cells (HP:0001974; 55 versus 2), abnormal red blood cell counts (HP:0001877; 100 versus 9 patients), and were at a higher average risk of death from SCD within five years than HbSC patients (0.35 versus 0.22). Note that at the time the PheKnowLator v1.0 KG was built, the DOID did not contain separate terms for HbSC and HBSS. As a result, embeddings for these concepts were not used when creating the GEO patient embeddings.

After processing the transcriptomic data, 4,909 HbSC DEGs (HbSC versus Control adjusted p <

[6]GSE35007: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35007.

0.05) and 9,625 HbSS DEGs (HbSS versus Control adjusted p < 0.05) were identified. Each patient's molecular signature decomposed into a two-dimensional space is shown in Figure 18. The distance between the average control patient and each SCD patient was significantly associated with a clinical measure of SCD severity (r_s = 0.40, p < 0.0001), such that the molecular signatures of the most severe SCD patients were the least similar to those of the controls. These results provide preliminary support for the clinical utility of patient-level representations of gene expression data.

To gain a deeper understanding of how well the DEG value arrays might enhance the PheKnowLator gene embeddings, the HbSS and HbSC gene expression signature-adjusted molecular mechanism embeddings were visualized using a t-SNE plots (HbSS t-SNE plot shown in Figure 19 (A)). For both SCD phenotypes, embeddings were clustered using K-Means with a (k = 9). The clustering results for the HbSS phenotype are shown in Figure 19 (B)**.** The genes assigned to each cluster were functionally annotated using Topp Gene [82]. For HbSS, Topp Gene returned a total of 78,017 annotations (M= 8,668.6, p < 0.05 or Q-value Bonferroni < 0.05). For HbSC, Topp Gene returned a total of 95,052 annotations (M= 10,561.3, p < 0.05 or Q-value Bonferroni < 0.05). For both phenotypes, each cluster included unique significantly enriched (p < 0.001) disease annotations to SCD-associated DEGs and known SCD clinical features. A single example for each cluster is overlaid on the clustering results in Figure 19 (B).

Clustering

Clustering was performed using the best molecular mechanism embeddings trained using data from clinical domains for CHCO rare disease patients and the GEO SCD gene expression signature-adjusted molecular mechanism embeddings. T-SNE plots for each embedding type are shown in Figure 20 (A). Results from applying K-Means to the clinical embeddings using a k of four are visualized in Figure 20 (B) and described in Table 7. Similar to the clinical and molecular mechanism embedding clustering results, this table contains two types of feature lists: (i) the top five most frequently occurring clinical concepts and (ii) the most frequently occurring clinical concepts unique to each cluster. This table also includes two ways of measuring the homogeneity of a cluster with respect to the disease type it represents, which are bolded and highlighted (i.e., colors are consistent with those used in the t-SNE plots: CHCO patients [**CAH**, **CF**, **PKU**, **SCD**] and GEO [**GEO-HbSS**, **GEO- HbSC**]): (i) **Cluster Breakdown:** the percent of patients from each rare disease group out of the total patients in that cluster; (ii) **Disease Breakdown:** the percent of rare disease patients for each disease group out of the total number of patients in that disease group. Taken together, these metrics provide a means for interpreting the distribution of a cluster with respect to each disease (e.g., 73% of the patients in a specific cluster have CF [cluster breakdown], which represents 45% of the total sample of CF patients [disease breakdown]). The clustering performance metrics were 0.40 ARI, 0.47 AMI, and 0.70 purity. As demonstrated in Table 7 and visualized in Figure 20 (B), the clustering results revealed two SCD clusters and one CF and CAH cluster.

- **Cluster Zero.** 49.2% (n = 454) of the CHCO patients in this cluster had SCD, which represented 55.8% of the CHCO SCD patients. Interestingly, this cluster also contained all of the GEO HbSS patients (n = 147). This cluster was defined by 2,119 concepts, of

Figure 18. t-SNE of GEO (GSE35007) Sickle Cell Disease Phenotypes Versus Control DEG Expression Arrays.
The figure visualizes the DEG expression signatures from sickle cell disease patients with HbSC (pink) and HbSS (blue) phenotypes and controls (yellow).

Figure 19. GEO (GSE35007) HbSS Gene Expression Signature-Adjusted Molecular Mechanism Embeddings.
(A) a t-SNE representation of gene embeddings from the PheKnowLator Knowledge Graph for HbSS patient differentially expressed genes (n = 9625). (B) K-Means clusters (n = 9) with primary annotations for HbSS upregulated differentially expressed genes.

Figure 20. Visualizing CHCO Rare Disease Patient Molecular Mechanism Embeddings and GEO (GSE35007) Gene Expression Signature-Adjusted Molecular Mechanism Embeddings. (A) t-SNE plot of the rare disease patient molecular mechanism embeddings overlaid with PheKnowLator aligned GEO HbSS (red triangles) and GEO HbSC (yellow triangles) Gene Expression-Adjusted Molecular Mechanism Embeddings. (B) K-Means results for four clusters. Acronyms - GEO: Gene Expression Omnibus.

Table 7. PheKnowLator Aligned GEO (GSE35007) Phenotype Clustering Results.

Total Patients by Disease Group: CAH (n = 760); CF (n = 835); PKU (n = 235); SCD (n = 813); GEO-HbSS (n = 147); GEO-HbSC (n = 51)		
0	Cluster Breakdown	**Number of Patients: 923** **SCD: 49.19%**; CAH: 21.34%; CF: 9.43%; PKU: 4.12%; **GEO-HbSS: 15.93%**
	Disease Breakdown	**SCD: 55.84%**; CAH: 25.92%; PKU: 16.17%; CF: 10.42%; **GEO-HbSS: 100.0%**
	Top Features (n = 2,119)	buspirone (MesH:D002065): 728 mesalamine (MesH:D019804): 678 dorzolamide (MesH:C062765): 594 down syndrome (DOID:14250): 568 trichloroacetaldehyde (MesH:C021100): 562
	Unique Features (n = 494)	hemophilia b (DOID:12259): 540 brain cancer (DOID:1319): 364 moyamoya disease (DOID:13099): 330 autosomal recessive polycystic kidney disease (DOID:0110861): 274 cri-du-chat syndrome (DOID:12580): 254
1	Cluster Breakdown	**Number of Patients: 660** **CF: 98.51%**; CAH: 1.04%; PKU: 0.15%; SCD: 0.3%
	Disease Breakdown	**CF: 79.04%**; CAH: 0.92%; PKU: 0.43%; SCD: 0.25%
	Top Features (n = 1,408)	sildenafil citrate (MesH:D000068677): 496 pyridoxine (MesH:D011736): 420 tramadol (MesH:D014147): 308 ehlers-danlos syndrome (DOID:13359): 268 sorbitol (MesH:D013012): 172
	Unique Features (n = 97)	nadolol (MesH:D009248): 120 chloramphenicol (MesH:D002701): 94 thioctic acid (MesH:D008063): 88 ziprasidone (MesH:C092292): 70 duane retraction syndrome (DOID:12557; HP:0009921): 70
2	Cluster Breakdown	**Number of Patients: 683** **CAH: 71.89%**; PKU: 27.53%; CF: 0.29%; SCD: 0.29%
	Disease Breakdown	**PKU: 80.0%**; **CAH: 64.61%**; SCD: 0.25%; CF: 0.24%
	Top Features (n = 1,048)	nadolol (MesH:D009248): 120 chloramphenicol (MesH:D002701): 94 thioctic acid (MesH:D008063): 88 ziprasidone (MesH:C092292): 70 duane retraction syndrome (DOID:12557; HP:0009921): 70
	Unique Features (n = 39)	hyperlordosis (HP:0003307): 64 subvalvular aortic stenosis (HP:0001682): 26 perineal hypospadias (HP:0000051): 18 optic nerve coloboma (DOID:11975; HP:0000588): 18 hordeolum internum (HP:0010608): 14

Total Patients by Disease Group: CAH (n = 760); CF (n = 835); PKU (n = 235); SCD (n = 813)		
3	Cluster Breakdown	**Number of Patients: 565** **SCD: 62.83%**; CF: 15.22%; CAH: 11.5%; PKU: 1.42%; **GEO-HbSC: 9.03%**
	Disease Breakdown	**SCD: 43.67%**; CF: 10.3%; CAH: 8.55%; PKU: 3.4%; **GEO-HbSC: 100.0%**
	Top Features (n = 1,294)	cerebral palsy (HP:0100021): 294 bosentan (MesH:C086232): 142 agnosia (DOID:4090): 130 apraxia (HP:0002186): 130 congenital ptosis (DOID:0060261; HP:0007970): 116
	Unique Features (n = 67)	aphakia (HP:0007707): 22 dermatomyositis (DOID:10223): 20 unicameral bone cyst (HP:0012064): 18 cervical spina bifida (HP:0005857):1 8 corneal edema (DOID:11030): 16

Acronyms - CAH: Congenital Hypothyroidism; CF: Cystic Fibrosis; GEO: Gene Expression Omnibus; PKU: Phenylketonuria; SCD: Sickle Cell Disease.

which 494 were unique to the patients in this cluster. This cluster contained concepts used to define all of the rare disease groups (CAH: 2, PKU: 10, CF: 2). With respect to sickle cell codes, this cluster contained 12.6 times more HbSS codes than Cluster Three (176 versus 14). High frequency concepts unique to this cluster included: hemophilia b (DOID:12259; n = 540), brain cancer (DOID:1319; n = 364), and Moyamoya disease (DOID:13099; n = 330).

- **Cluster One.** 98.5% (n = 660) of the CHCO patients in this cluster had CF, which represented 79% of the CHCO CF patients. This cluster was defined by 1,408 concepts, of which 97 were unique to the patients in this cluster. This cluster contained concepts used to define two other rare disease groups (CAH: 12, SCD: 2). With respect to CF-specific codes, this cluster contained 3.7 times more CF concepts used to define the CF rare disease groups than any other clusters containing CF patients. High frequency concepts unique to this cluster included: nadolol (MesH:D009248; n = 120), chloramphenicol (MesH:D002701; n = 94), thioctic acid (MesH:D008063; n = 88), ziprasidone (MesH:C092292; n = 70), and duane retraction syndrome (DOID:12557; HP:0009921; n = 70).
- **Cluster Two.** 71.9% (n = 491) of the CHCO patients in this cluster had CAH, which represented 64.6% of the CHCO CAH patients. This cluster was defined by 1,048 concepts, of which 39 were unique to the patients in this cluster. This cluster contained concepts used to define two other rare disease groups (PKU: 8, CF: 2). High frequency concepts unique to this cluster included: hyperlordosis (HP:0003307; n = 64), subvalvular aortic stenosis (HP:0001682; n = 26), and optic nerve coloboma (DOID:11975; HP:0000588; n = 18).

- **Cluster Three.** 62.8% (n = 355) of the CHCO patients in this cluster had SCD, which represented 43.7% of the CHCO SCD patients. Interestingly, this cluster also contained all of the GEO HbSC patients (n = 51). This cluster was defined by 1,294 concepts, of which 67 were unique to the patients in this cluster. This cluster contained concepts used to define all of the rare disease groups (CAH: 12, PKU: 8, CF: 6). With respect to sickle cell codes, this cluster contained twice as many HbSC-specific codes than HbSS-specific codes (30 versus 14). High frequency concepts unique to this cluster included: aphakia (HP:0007707; n = 22), dermatomyositis (DOID:10223; n = 20), unicameral bone cyst (HP:0012064; n = 18), cervical spina bifida (HP:0005857; n = 18), and corneal edema (DOID:11030; n = 16).

For the SCD clusters, two clusters contained all of the GEO HbSS and HbSC patients (i.e., Zero and Three, respectively). Cluster Zero, which contained 100% of the GEO HbSS patients (n = 147), contained 55.8% (n = 454) of the CHCO patients with SCD. For these patients had 714 (M = 16.6, 2-108) occurrences of one or more of the SCD codes used to identify HbSC-related diagnosis and 2,560 (M= 59.5, 2-396) of the SCD codes used to identify HbSS-related diagnosis. Upon closer examination, it was discovered that all HbSC patients (n = 43) also had at least two HbSS diagnoses. The count, range, and means of SCD-related codes (ranked from highest to lowest frequency) for this cluster included: Hemoglobin SS disease without crisis (omop:30683; n = 18,536 total, 2-818, M= 96), Sickle cell-hemoglobin SS disease (omop:22281; n = 4,174 total, 2-176, M= 23.1), Hemoglobin SS disease with crisis (omop:26942; 3,390 total, 2-210, M= 21.2), Sickle cells present (omop:25518; 1,408 total, 2-50, M= 6.6), Sickle cell-thalassemia disease without crisis (omop:321263; 1,102 total, 2-322, M= 34.4), Sickle cell-hemoglobin C disease with crisis (omop:443721; 714 total, 2-108, M= 16.6), Acute chest syndrome (omop:254062; 446 total, 2-28, M= 5.6), Red blood cell sequestration in spleen (omop:196943; 220 total, 2-26, M= 5.8), Sickle cell-thalassemia disease with crisis (omop:443738; 166 total, 2-30, M= 8.3), and Acute sickle cell splenic sequestration crisis (omop:40485018; 10 total, M= 2). As observed from the frequency of the HbSS codes, especially when compared to the frequency of these codes in Cluster Three, this cluster likely represented the more severe HbSS patients.

Cluster Three, which contained 100% of the GEO HbSC patients (n = 51), contained 43.7% (n = 355) CHCO patients with SCD. For these SCD patients, all SCD codes were HbSS-related. The count, range, and means of SCD-related codes (ranked from highest to lowest frequency) for this cluster included: Sickle cells present (omop:25518; 1,730 total, 2-50, M= 6.8), Hemoglobin SS disease without crisis (omop:30683; n = 344 total, 2-58, M= 11.1), Sickle cell-hemoglobin SS disease (omop:22281; n = 210 total, 2-40, M= 7), Hemoglobin SS disease with crisis (omop:26942; 104 total, 2-26, M= 8), Sickle cell-thalassemia disease without crisis (omop:321263; 38 total, 2-32, M= 12.7), Red blood cell sequestration in spleen (omop:196943; 22 total, 2-4, M= 2.4), Acute chest syndrome (omop:254062;126 total, 2-4, M = 2.4), and Sickle cell-thalassemia disease with crisis (omop:443738; 2 total, 2, M= 2). As previously mentioned, this cluster likely contained patients with a less severe HbSS phenotype than the HbSS patients in Cluster Zero.

Patient Similarity

Two tasks were performed in an effort to demonstrate that each GEO SCD phenotype was closer (i.e., more similar) in vector space to the CHCO SCD patients than all other CHCO rare disease and control patient groups.

Top-N TP Rate

The results of this task are shown in Table 8. For each of the following comparisons, the counts of TPs, defined as the percent of SCD CHCO patients within the top 10-100 (in increments of 10) most similar patients, were recorded:

I. **CHCO SCD Patients - Clinical Embeddings.** When only including rare disease patients, the TP rates ranged from 0.7-0.8 (M= 0.70), which were higher than the TPRs that were obtained when using all rare disease and control patients (0.6-0.6, M= 0.6).

II. **CHCO SCD Patients - Molecular Mechanism Embeddings.** When only including rare disease patients TP rates ranged from 0.7-0.9 (M= 0.8), which were higher than the TPRs that were obtained when using all rare disease and control patients (0.4-0.5, M= 0.5).

III. **CHCO and GEO SCD Patients - Molecular Mechanism Embeddings.** In contrast to the prior comparisons, including the GEO SCD patients decreased performance. When only including rare disease patients TPRs ranged from 0.5-0.7 (M= 0.6), which were higher than the TP rates that were obtained when using all rare disease and control patients (0-0.1, M= 0.03).

IV. **CHCO and GEO HbSS Patients - Molecular Mechanism Embeddings.** When only including rare disease patients TPRs ranged from 0.5-0.7 (M= 0.6), which were higher than the TP rates that were obtained when using all rare disease and control patients (0-0.1, M= 0.03); and (iv) *CHCO and GEO HbSC Patients - Molecular Mechanism Embeddings*. When only including rare disease patients TPRs ranged from 0.5-0.7 (M = 0.6), which were higher than the TPRs that were obtained when using all rare disease and control patients (0-0.1, M= 0.04).

Kruskal-Wallis Nonparametric ANOVAs

To determine if the mean rank cosine similarity between the SCD GEO patients and all CHCO rare disease patients differed, a Kruskal-Wallis nonparametric ANOVA with Dunn-Bonferroni post hoc tests was run.

- **HbSC Patients.** The Kruskal-Wallis test provided evidence of a significant difference (χ^2 (3) = 27,820.5, $p < 0.001$) between the mean ranks of at least one pair of CHCO SCD disease groups. The Eta-squared effect size value (η^2 = 0.21) met Cohen's criteria (η^2 > 0.11) to be considered a "moderate" effect [76]. Dunn's pairwise tests were carried out for the four pairs of groups. There was strong evidence (ps < 0.001, adjusted using the

Table 8. True Positive Rate for the Top-N Most Similar Sickle Cell Disease Patients.

RARE DISEASE

Configuration	Top-N Patients									
	10	20	30	40	50	60	70	80	90	100
SCD Clinical Data (CHCO)	0.75	0.73	0.72	0.71	0.71	0.70	0.69	0.69	0.68	0.68
SCD Molecular Mechanisms (CHCO)	0.90	0.84	0.81	0.79	0.77	0.76	0.75	0.74	0.73	0.72
All SCD Molecular Mechanisms (CHCO + GEO)	0.70	0.60	0.53	0.53	0.52	0.57	0.55	0.57	0.58	0.59
HbSS Molecular Mechanisms (CHCO + GEO)	0.70	0.60	0.53	0.52	0.52	0.57	0.56	0.56	0.58	0.59
HbSC Molecular Mechanisms (CHCO + GEO)	0.70	0.60	0.55	0.53	0.52	0.57	0.54	0.58	0.58	0.60
COMBINED										
Configuration	**Top-N Patients**									
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
SCD Clinical Data (CHCO)	0.63	0.63	0.62	0.62	0.62	0.61	0.60	0.60	0.59	0.59
SCD Molecular Mechanisms (CHCO)	0.54	0.51	0.49	0.48	0.47	0.46	0.45	0.44	0.44	0.43
All SCD Molecular Mechanisms (CHCO + GEO)	0.00	0.05	0.03	0.05	0.04	0.03	0.04	0.04	0.03	0.03
HbSS Molecular Mechanisms (CHCO + GEO)	0.00	0.05	0.03	0.05	0.04	0.03	0.04	0.04	0.03	0.03
HbSC Molecular Mechanisms (CHCO + GEO)	0.00	0.05	0.03	0.05	0.04	0.03	0.04	0.04	0.03	0.03

Acronyms - CHCO: Children's Hospital of Colorado; GEO: Gene Expression Omnibus; KG: Knowledge Graph.

Bonferroni correction) of a significant difference between all groups. The median similarity score between the HbSC GEO patients and the CHCO SCD patients was the highest 0.3 (SD = 0.5), followed by CAH (Md = 0.3, SD = 0.04), CF (Md = 0.2, SD = 0.04), and PKU (Md = 0.003, SD = 0.1).

- **HbSS Patients.** The Kruskal-Wallis test provided evidence of a significant difference (χ^2 (3) = 80,760.3, p < 0.001) between the mean ranks of at least one pair of the SCD rare disease groups. The Eta-squared effect size value (η^2 = 0.21) meets Cohen's criteria (η^2 > 0.11) to be considered a "moderate" effect [76]. Dunn's pairwise tests were carried out for the four pairs of groups. There was strong evidence (ps < 0.001, adjusted using the Bonferroni correction) of a difference between all groups. The median similarity score between the HbSS GEO patients and the CHCO SCD patients was the highest 0.27 (SD = 0.05), followed by CAH (Md = 0.26, SD = 0.04), CF (Md = 0.25, SD = 0.04), and PKU (Md = 0.003, SD = 0.11).

DISCUSSION

Translating deep computational phenotypes from research to clinical practice requires precise patient characterization. Precisely characterizing patients requires access to a wide range data

and expertise capable of capturing the patient-specific phenotypic variance observed with human disease. EHRs are one of the richest sources of individual- and population-level patient clinical data. While these data contain rich contextualized multimodal *snapshots* into human disease, they do not yet systematically integrate molecular data. Given the abundance of expert-curated publicly available molecular experiments, like those stored in GEO, the next best solution is to develop a proxy or surrogate representation of the molecular and pathophysiological knowledge present in sources of linked clinical and molecular data. Ideally, this proxy would also enable patient-level integration of clinical data, like those from an EHR, with transcriptomic data from GEO or the EBI Expression Atlas.

This goal of the work described in this chapter was to examine the clinical utility of a large-scale biomedical KG, carefully and expertly constructed to represent the molecular mechanisms underlying human disease and to use it to enable the bidirectional translation of disparate and incomplete (i.e., patient-level clinical data with no molecular data or patient-level molecular data no clinical context) sources of biomedical data into a shared mechanistic space offering new opportunities to integrate a wide variety of data from different EHRs, patients, and molecular entities. To accomplish this goal, we designed experiments to answer three specific questions.

Question One: Given different clinical domains and the availability of different embedding models, each with different parameter settings, what model and parameter setting is best for each domain when deriving patient-level representations?
For the first experiment, we aimed to determine the best representation for patients using only EHR data. For this task, embeddings were created from each patient's set of temporally-ordered visit-level EHR concept sentences. Embeddings were created from the two configurations of Doc2Vec and compared to a baseline model (BoW + TF-IDF) on a rare disease subclassification task when using different clinical domains (i.e., conditions [n = 6,382], medications [n = 2,334], measurements [n = 272], and all clinical domains [n = 8,988]). The best performance from over 100 comparisons (precision = 0.83) was found when using the SVM classifier on medication concepts represented using the PV-DBOW representation with 150 dimensions and a sliding window size of 15. Interestingly, the best baseline model when applied to condition concepts, outperformed this, scoring 0.99 across all of the performance metrics, when using the SVM classifier. While the baseline model performed as well or better than several of the clinical and mechanism embedding models, this performance must be weighted appropriately with respect to the value of the mechanistic hypotheses. When taking this into consideration, the more complex models, which may have been less performant, produced more relevant hypotheses when using only clinical data or molecular mechanisms.

Question Two: If patients can also be represented using the molecular mechanisms that underlie their primary diagnosis, would these representations result in more meaningful subphenotypes than when constructed from only clinical data?
For the second experiment, patients were represented using node embeddings learned from a PheKnowLator KG-based proxy of molecular mechanisms underlying human disease (i.e., 129,875 nodes and 3,838,935 edges). Using the OMOP2OBO semantic alignment algorithm, 66.9% of condition, 99.7% of drug exposure ingredient, and 47.1% of measurement result concepts were successfully aligned to at least one KG node embedding. When applied to the

same rare disease subclassification task, the best performance of the molecular mechanism embeddings (0.99 for all performance metrics), which was also out-performed by the baseline representation, was achieved when using the logistic regression classier on condition concepts created by averaging each patient's mapped node embeddings. The worst performing molecular mechanism representation, which performed better than 82.2% of all Doc2Vec parameterizations, achieved an accuracy and precision of 0.81 when using the Näive Bayes classifier on the averaged node embeddings for all clinical domains. While some configurations of the summed molecular mechanism embeddings performed similarly to the averaged mechanism embeddings, their visualizations were not intuitive. Interestingly, the molecular mechanism embeddings out-performed nearly all of the Doc2Vec implementations when applied to 2,115 fewer condition, 7 fewer medication, and 143 fewer measurement result concepts.

The best clinical configuration [determined using performance on the SVM classifier] from both experiments (clinical: all clinical domains using the PV-DBOW model with 150 dimensions and a sliding window size of five [accuracy, recall, and F_1 score = 0.82, precision = 0.83]; molecular: mechanism embeddings created from all clinical domains [accuracy, precision, recall, and F_1 score of 0.95]) were then evaluated using K-Means clustering. The molecular mechanism embeddings generated a greater number of clinically relevant (nine versus six) and pure disease clusters (two CF clusters, two CAH clusters, and one PKU cluster versus one pure CF cluster) than the clinical embeddings. When investigating the clinical utility of the most frequent clinical concepts (total and unique) for each cluster, the molecular mechanism embeddings generated a larger number of subphenotypes than the clinical embeddings. Both methods generated multiple clusters for SCD, which were examined by a clinical domain expert.

When using the clinical embeddings, three potential subphenotypes were identified (i.e., Clusters Zero, Three, and Five in Table 5). Cluster Zero contained 31.4% of SCD patients (n = 256) and was defined by unique concepts that have well-established associations with SCD, which demonstrates the ability of our approach to generate interesting mechanistic hypotheses. For example, Lyrica (omop:19123163; n = 1,864) or pregabalin, which was the most frequent concept unique to this cluster. Lyrica is a gabapentinoid which is commonly prescribed to SCD patients in order to treat neuropathic pain [83]. Another frequent concept unique to this cluster was mucopolysaccharidosis (omop:433446; n = 474), which causes the accumulation of high levels of microparticles in the body due to the inability of the body to break down long chain sugar molecules [84]. Studies investigating microparticles in SCD have found that HbSC patients tend to have lower levels in their bloodstream compared to HbSS patients [85]. Cluster Three contained 21.6% of SCD patients (n = 176) and compared to Cluster Zero, which was defined by unique condition and medication concepts, this cluster was primarily defined by medication concepts. The most frequently occurring unique medication was sildenafil 25 MG Oral Tablet (omop:1316268; n = 3,284) and Ursodiol 250 MG Oral Tablet (omop:40165331; n = 1,032). Sildenafil citrate or Viagra has been shown to prevent priapism in SCD by increasing the degradation of serum nitric oxide [86]. Ursodiol or ursodeoxycholic acid is commonly prescribed to SCD patients suffering from hepatobiliary complications like hepatic crisis, cirrhosis, or cholelithiasis [87]. Finally, Cluster Five contained 39.8% of SCD patients (n = 325) and was defined by frequent concepts unique to this cluster like Erythromycin 0.02 MG/MG Topical Gel (omop:19076956; n = 20) and Zarontin (omop:750145; n = 18). Erythromycin is

frequently used to provide pneumococcal infection prophylaxis in SCD patients [88]. Finally, SCD patients are ten times more likely to suffer from seizures than the general population and Zarontin or ethosuximide is commonly prescribed to treat absence seizures [89].

There were also three SCD subphenotypes identified by the molecular mechanism embeddings (i.e., Cluster Zero, Three, and Six in Table 6). Cluster Zero contained 30.4% of SCD patients (n = 247) with the most frequent concepts unique to this cluster being diabetic autonomic neuropathy (DOID:11503; n = 44), aphakia (HP:0007707; n = 22), and dermatomyositis (DOID:10223; n = 20). With respect to diabetic autonomic neuropathy, similar to SCD, this disorder has been shown to significantly affect the autonomic nervous system and both disorders are associated with priapism [90–92]. With respect to aphakia, which is a disorder characterized by the absence of a lens of an eye, is most commonly caused during cataract removal surgery [93]; cataracts are frequently associated with SCD [94]. Finally, although rare, there are case studies that indicate coexistence of dermatomyositis and SCD [95]. Cluster Three contained 51.5% of SCD patients (n = 419) with the most frequent concepts unique to this cluster being dabigatran (MeSH:D000069604; n = 244) and histiocytosis (DOID:3405, HP:0100727; n = 152). Dabigatran or Pradaxa is an anticoagulant which is prescribed to SCD patients experiencing venous thromboembolisms [96]. Further, SCD patients often experience swollen lymph nodes as a result of chronic infection [97] and histiocytosis [98]. Cluster Six contained 7.3% of SCD patients (n = 59) with the most frequent concepts unique to this cluster being cervical spina bifida (HP:0005857; n = 18) and temporomandibular joint crepitus (HP:0012479; n = 6). Patients with SCD are at an increased risk of folate deficiency due to increased erythropoiesis [99] and folate deficiency is known to cause neural tube defects like cervical spina bifida [100]. There is a direct association between temporomandibular joint crepitus and SCD [101,102].

Question Three: Can PheKnowLator be used to infer the genotype of patients with only clinical data from an independent sample of genotyped patients?

For the third experiment, we provided additional proof of the translational utility of PheKnowLator KGs by demonstrating how it enabled the seamless integration of EHR patient data and transcriptomic data from different patient populations. Using a sample of pediatric SCD patients obtained from GEO, we translated them into the same latent space as the CHCO SCD patients by averaging each patient's gene expression signature-adjusted molecular mechanism embeddings. The clinical relevance of these embeddings were found to be significantly associated with a clinical measure of SCD severity such that the molecular signatures of the most severe SCD patients were the least similar to those of the controls. Perhaps the most compelling finding was the perfect split of the GEO HbSC and HbSS patients into two separate SCD patient clusters. When investigating the primary diagnoses of the CHCO SCD kids closest to each SCD phenotype (and subsequently, to infer their genotype), we found that forty-three patients, with at least one HbSC diagnosis, appeared in the same cluster as the HbSS GEO patients. The occurrence of the HbSC patients in this cluster could be the result of incorrectly assigned HbSS codes (i.e., diagnosis codes assigned without blood test confirmation of the HbSS or HbSC phenotype). The incorrect assignment of diagnosis codes for pediatric HbSC

patients has been observed in the literature when using diagnosis codes to determine the genotype of these patients [103].

Although only preliminary, these results highlight Med2Mech's ability to automatically generate biologically relevant and clinically meaningful clusters using only clinical data, proxies of molecular mechanisms, and transcriptomic data. To the best of our knowledge, there are very few organizations or researchers performing this type of work. In fact, to the best of our knowledge, we are the first to use an open source KG built entirely on publicly available data to create clinically meaningful connections between independent samples of pediatric rare disease patients.

Limitations and Future Works

There are several important limitations of this work that should be addressed in the future. First, our analysis was limited to a small set of pediatric rare diseases. While pediatric rare diseases can be incredibly complex to characterize, accurately diagnose, and treat, it is important to understand how our method works when utilized with more common and complex disorders like asthma and diabetes. Additionally, our evaluation was limited to a single embedding algorithm for both the clinical and molecular mechanism embeddings. Exploring more complex methods, like transformers, could provide additional specificity and improve subphenotyping. Similarly, we did not investigate the clustering results for the best performing clinical or molecular mechanism embeddings when created from the individual clinical domains. Given the observed differences in multiclass classification, it will be important to investigate whether or not the molecular mechanism embedding performance differs by disease, when using different clinical domains. The use of an early version of PheKnowLator KGs (v1.0.0) and OMOP2OBO mappings (v0.0.0), could have reduced the overall performance. For example, the version of the DOID used in this KG did not contain a concept for *Hemoglobin C Disease* (DOID:2859) and the concept for *Sickle Cell Anemia* (DOID:10923) included both HbSC and HbSS phenotypes in its definition. This lack of specificity was determined to be insufficient for the type of subphenotyping we performed and thus those concepts were not included in the GEO SCD embeddings. For OMOP2OBO, the current version of the mappings (v1.0) are significantly more expressive and comprehensive with respect to the ontologies they utilize and clinical entities that they cover (e.g., measurement results are mapped to phenotypes, cell types, anatomical entities, organisms, chemicals, and proteins versus only phenotypes in v0.0.0). Rerunning this analysis with updated versions of PheKnowLator and OMOP2OBO will be crucial for validating this method and each of these respective algorithms.

CONCLUSIONS

This work describes preliminary results from leveraging neural-symbolic representation learning of a biomedical KG to generate patient-level mechanism(s) embeddings. Our findings demonstrate superior performance of these embeddings when compared to embeddings learned using only clinical data. Clustering the embeddings from the best performing models resulted in several promising rare disease subphenotypes. Finally, by successfully aligning the genotypes and phenotypes of CHCO SCD patients who only had clinical data to GEO SCD patients with rich transcriptomic data, but limited clinical information, we demonstrated the utility

of Med2Mech to serve as a clinically meaningful and biologically relevant data translator.

REFERENCES

1. Dorsey ER, Omberg L, Waddell E, Adams JL, Adams R, Ali MR, et al. Deep Phenotyping of Parkinson's Disease. J Parkinsons Dis [Internet]. 2020 Jan 1 [cited 2021 Jul 6];10(3):855–73. Available from: https://content.iospress.com:443/download/journal-of-parkinsons-disease/jpd202006?id=journal-of-parkinsons-disease%2Fjpd202006

2. Georgiou M, Robson AG, Singh N, Pontikos N, Kane T, Hirji N, et al. Deep Phenotyping of PDE6C-Associated Achromatopsia. Invest Ophthalmol Vis Sci [Internet]. 2019 Dec 2 [cited 2021 Jul 6];60(15):5112–23. Available from: https://iovs.arvojournals.org/arvo/content_public/journal/iovs/938292/i1552-5783-60-15-5112.pdf

3. Fassihi H, Sethi M, Fawcett H, Wing J, Chandler N, Mohammed S, et al. Deep phenotyping of 89 xeroderma pigmentosum patients reveals unexpected heterogeneity dependent on the precise molecular defect. Proc Natl Acad Sci U S A [Internet]. 2016 Mar 1 [cited 2021 Jul 6];113(9):E1236–45. Available from: https://www.pnas.org/content/113/9/E1236.abstract

4. Russo RS, Gambello MJ, Murphy MM, Aberizk K, Black E, Lindsey Burrell T, et al. Deep phenotyping in 3q29 deletion syndrome: recommendations for clinical care. Genet Med [Internet]. 2021 Feb 9 [cited 2021 Jul 6];23(5):872–80. Available from: https://www.nature.com/articles/s41436-020-01053-1

5. Daich VM, Jani P, Zein WM, D'Souza P, Wolfe L, Chisholm J, et al. The peroxisomal disorder spectrum and Heimler syndrome: Deep phenotyping and review of the literature. Am J Med Genet C Semin Med Genet [Internet]. 2020 Sep [cited 2021 Jul 6];184(3). Available from: https://pubmed.ncbi.nlm.nih.gov/32866347/

6. Mei C, Fedorenko E, Amor DJ, Boys A, Hoeflin C, Carew P, et al. Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. Eur J Hum Genet [Internet]. 2018 Feb 14 [cited 2021 Jul 6];26(5):676–86. Available from: https://www.nature.com/articles/s41431-018-0102-x

7. Droogmans G, Swillen A, Van Buggenhout G. Deep Phenotyping of Development, Communication and Behaviour in Phelan-McDermid Syndrome. MSY [Internet]. 2019 [cited 2021 Jul 6];10(6):294–305. Available from: https://www.karger.com/Article/Abstract/503840

8. Benign or not benign? Deep phenotyping of liver Glycogen Storage Disease IX. Mol Genet Metab [Internet]. 2020 Nov 1 [cited 2021 Jul 6];131(3):299–305. Available from: http://dx.doi.org/10.1016/j.ymgme.2020.10.004

9. Mak E, Bickerton A, Padilla C, Walpert MJ, Annus T, Wilson LR, et al. Longitudinal trajectories of amyloid deposition, cortical thickness, and tau in Down syndrome: A deep-phenotyping case report. Alzheimer's & dementia (Amsterdam, Netherlands) [Internet]. 2019 Nov 25 [cited 2021 Jul 6];11. Available from: https://pubmed.ncbi.nlm.nih.gov/31909173/

10. Mishra R, Jain V, Gupta D, Saxena R, Kulshreshtha S, Ramprasad VL, et al. Robinow Syndrome and Brachydactyly: An Interplay of High-Throughput Sequencing and Deep

Phenotyping in a Kindred. MSY [Internet]. 2020 [cited 2021 Jul 6];11(1):43–9. Available from: https://www.karger.com/Article/Abstract/505506

11. Welsink-Karssies MM, Ferdinandusse S, Geurtsen GJ, Hollak CEM, Huidekoper HH, Janssen MCH, et al. Deep phenotyping classical galactosemia: clinical outcomes and biochemical markers. Brain Commun [Internet]. 2020 Jan 29 [cited 2021 Jul 6];2(1). Available from: https://academic.oup.com/braincomms/article-pdf/2/1/fcaa006/33639818/fcaa006.pdf

12. Shim Y, Go YJ, Kim SY, Kim H, Hwang H, Choi J, et al. Deep Phenotyping in 1p36 Deletion Syndrome. Ann Child Neurol [Internet]. 2020 Aug 24 [cited 2021 Jul 6];28(4):131–7. Available from: http://www.annchildneurol.org/journal/view.php?doi=10.26815/acn.2020.00108

13. Spedicati B, Cocca M, Palmisano R, Faletra F, Barbieri C, Francescatto M, et al. Natural human knockouts and Mendelian disorders: deep phenotyping in Italian isolates. Eur J Hum Genet [Internet]. 2021 Mar 16 [cited 2021 Jul 6];1–10. Available from: https://www.nature.com/articles/s41431-021-00850-9

14. Yurkovich JT, Tian Q, Price ND, Hood L. A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. Nat Rev Clin Oncol [Internet]. 2019 Oct 16 [cited 2021 Jul 6];17(3):183–94. Available from: https://www.nature.com/articles/s41571-019-0273-6

15. Papadimitriou K, Tsakirakis N, Malandrakis P, Vitsos P, Metousis A, Orologas-Stavrou N, et al. Deep Phenotyping Reveals Distinct Immune Signatures Correlating with Prognostication, Treatment Responses, and MRD Status in Multiple Myeloma. Cancers [Internet]. 2020 Nov 4 [cited 2021 Jul 6];12(11):3245. Available from: https://www.mdpi.com/2072-6694/12/11/3245

16. Christopoulos P, Bozorgmehr F, Brückner L, Chung I, Krisam J, Schneider MA, et al. Brigatinib versus other second-generation ALK inhibitors as initial treatment of anaplastic lymphoma kinase positive non-small cell lung cancer with deep phenotyping: study protocol of the ABP trial. BMC Cancer [Internet]. 2021 Jun 28 [cited 2021 Jul 6];21(1):1–10. Available from: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-021-08460-w

17. Sirinukunwattana K, Aberdeen A, Theissen H, Sousos N, Psaila B, Mead AJ, et al. Improving the diagnosis and classification of Ph-negative myeloproliferative neoplasms through deep phenotyping [Internet]. bioRxiv. 2019 [cited 2021 Jul 6]. p. 762013. Available from: https://www.biorxiv.org/content/10.1101/762013v1.abstract

18. Nagaoka K, Shirai M, Taniguchi K, Hosoi A, Sun C, Kobayashi Y, et al. Original research: Deep immunophenotyping at the single-cell level identifies a combination of anti-IL-17 and checkpoint blockade as an effective treatment in a preclinical model of data-guided personalized immunotherapy. Journal for Immunotherapy of Cancer [Internet]. 2020 [cited 2021 Jul 6];8(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7583806/

19. Song TH, Cao M, Min J, Im H, Lee H, Lee K. Deep Learning-Based Phenotyping of Breast Cancer Cells Using Lens-free Digital In-line Holography [Internet]. bioRxiv. 2021 [cited 2021 Jul 6]. p. 2021.05.29.446284. Available from: https://www.biorxiv.org/content/10.1101/2021.05.29.446284v1.abstract

20. Kuai R, Ochyl LJ, Bahjat KS, Schwendeman A, Moon JJ. Designer vaccine nanodiscs for personalized cancer immunotherapy. Nat Mater [Internet]. 2017 Apr [cited 2021 Jul 6];16(4). Available from: https://pubmed.ncbi.nlm.nih.gov/28024156/

21. Paquette AG, Hood L, Price ND, Sadovsky Y. Deep Phenotyping During Pregnancy for Delivery of Predictive and Preventive Medicine. Sci Transl Med [Internet]. 2020 Jan 22 [cited 2021 Jul 6];12(527). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7268934/

22. Davidson L, Boland MR. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. Brief Bioinform [Internet]. 2021 Jan 6 [cited 2021 Jul 6]; Available from: https://academic.oup.com/bib/advance-article-pdf/doi/10.1093/bib/bbaa369/35460043/bbaa369.pdf

23. OBGYN [Internet]. [cited 2021 Jul 6]. Available from: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/pd.5616

24. Kennedy SH, Victora CG, Craik R, Ash S, Barros FC, Barsosio HC, et al. Deep clinical and biological phenotyping of the preterm birth and small for gestational age syndromes: The INTERBIO-21 st Newborn Case-Control Study protocol. Gates Open Research [Internet]. 2018 [cited 2021 Jul 6];2. Available from: https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC6545521/

25. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med [Internet]. 2011 Mar [cited 2021 Jul 6];8(2). Available from: https://pubmed.ncbi.nlm.nih.gov/21695041/

26. Plasqui G, Westerterp KR. Physical activity assessment with accelerometers: an evaluation against doubly labeled water. Obesity [Internet]. 2007 Oct;15(10):2371–9. Available from: http://dx.doi.org/10.1038/oby.2007.281

27. Wong LJ, Buckingham BA, Kunselman B, Istoc E, Leach J, Purvis R. Extended use of a new continuous glucose monitoring system with wireless data transmission in children with type 1 diabetes mellitus. Diabetes Technol Ther [Internet]. 2006 Apr [cited 2021 Jul 6];8(2). Available from: https://pubmed.ncbi.nlm.nih.gov/16734544/

28. Accumulated environmental risk determining age at schizophrenia onset: a deep phenotyping-based study. The Lancet Psychiatry [Internet]. 2014 Nov 1 [cited 2021 Jul 6];1(6):444–53. Available from: http://dx.doi.org/10.1016/S2215-0366(14)70379-7

29. Lakerveld J, Wagtendonk A, Vaartjes I, Karssenberg D. Deep phenotyping meets big data: the Geoscience and hEalth Cohort COnsortium (GECCO) data to enable exposome studies in The Netherlands. Int J Health Geogr [Internet]. 2020 Nov 13 [cited 2021 Jul 6];19(1):1–16. Available from: https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-020-00235-z

30. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform [Internet]. 2020 Apr 1 [cited 2021 Jul 6];104:103362. Available from: http://dx.doi.org/10.1016/j.jbi.2019.103362

31. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected

national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med [Internet]. 2016 Jul;71:57–61. Available from: http://dx.doi.org/10.1016/j.artmed.2016.05.005

32. Alawad M, Gao S, Shekar MC, Hasan SMS, Christian JB, Wu X-C, et al. Integration of Domain Knowledge using Medical Knowledge Graph Deep Learning for Cancer Phenotyping [Internet]. 2021 [cited 2021 Jul 6]. Available from: http://arxiv.org/abs/2101.01337

33. Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, Carmody LC, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. npj Digital Medicine [Internet]. 2019 May 2 [cited 2021 Jul 6];2(1):1–9. Available from: https://www.nature.com/articles/s41746-019-0110-4

34. 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. Am J Hum Genet [Internet]. 2018 Jul 5;103(1):58–73. Available from: http://dx.doi.org/10.1016/j.ajhg.2018.05.010

35. Tang X, Chen W, Zeng Z, Ding K, Zhou Z. An ontology-based classification of Ebstein's anomaly and its implications in clinical adverse outcomes. Int J Cardiol [Internet]. 2020 Apr 26; Available from: http://dx.doi.org/10.1016/j.ijcard.2020.04.073

36. Shen F, Wen A, Liu H. Subgrouping Rare Disease Patients Leveraging the Human Phenotype Ontology Embeddings. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) [Internet]. 2020. p. 169–72. Available from: http://dx.doi.org/10.1109/CBMS49503.2020.00039

37. Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS. An information model for computable cancer phenotypes. BMC Med Inform Decis Mak [Internet]. 2016 Sep 15 [cited 2021 Jun 30];16(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27629872/

38. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One [Internet]. 2018 Feb 15 [cited 2021 Jul 6];13(2):e0192360. Available from: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0192360&type=printable

39. Kim K, Zakharkin SO, Allison DB. Expectations, validity, and reality in gene expression profiling. J Clin Epidemiol [Internet]. 2010 Sep;63(9):950–9. Available from: http://dx.doi.org/10.1016/j.jclinepi.2010.02.018

40. Raj P, Raman AC, Subramanian H. Architectural Patterns: Uncover Essential Patterns in the Most Indispensable Realm of Enterprise Architecture [Internet]. Packt Publishing; 2017. 468 p. Available from: https://play.google.com/store/books/details?id=SEHvtAEACAAJ

41. Ma'ayan A. Introduction to network analysis in systems biology. Sci Signal [Internet]. 2011 Sep 6;4(190):tr5. Available from: http://dx.doi.org/10.1126/scisignal.2001965

42. Samal A, Singh S, Giri V, Krishna S, Raghuram N, Jain S. Low degree metabolites explain

essential reactions and enhance modularity in biological networks. BMC Bioinformatics [Internet]. 2006 Mar 8;7:118. Available from: http://dx.doi.org/10.1186/1471-2105-7-118

43. Bernabò N, Barboni B, Maccarrone M. The biological networks in studying cell signal transduction complexity: The examples of sperm capacitation and of endocannabinoid system. Comput Struct Biotechnol J [Internet]. 2014 Aug;11(18):11–21. Available from: http://dx.doi.org/10.1016/j.csbj.2014.09.002

44. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Front Cell Dev Biol [Internet]. 2014 Aug 19;2:38. Available from: http://dx.doi.org/10.3389/fcell.2014.00038

45. Bultinck J, Lievens S, Tavernier J. Protein-protein interactions: network analysis and applications in drug discovery. Curr Pharm Des [Internet]. 2012;18(30):4619–29. Available from: http://dx.doi.org/10.2174/138161212802651562

46. Lee M, Park K, Kim D. Interaction network among functional drug groups. BMC Syst Biol [Internet]. 2013 Oct 16;7 Suppl 3:S4. Available from: http://dx.doi.org/10.1186/1752-0509-7-S3-S4

47. Liu Z, Peng E, Yan S, Li G, Hao T. T-know: a knowledge graph-based question answering and infor-mation retrieval system for traditional chinese medicine. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations [Internet]. 2018. p. 15–9. Available from: https://www.aclweb.org/anthology/C18-2004.pdf

48. Ruan T, Huang Y, Liu X, Xia Y, Gao J. QAnalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. BMC Med Inform Decis Mak [Internet]. 2019 Apr 1;19(1):82. Available from: http://dx.doi.org/10.1186/s12911-019-0798-8

49. Liu X, Jin J, Wang Q, Ruan T, Zhou Y, Gao D, et al. PatientEG Dataset: Bringing Event Graph Model with Temporal Relations to Electronic Medical Records [Internet]. arXiv [cs.CY]. 2018. Available from: http://arxiv.org/abs/1812.09905

50. Shang J, Xiao C, Ma T, Li H, Sun J. Gamenet: Graph augmented memory networks for recommending medication combination. In: Proceedings of the AAAI Conference on Artificial Intelligence [Internet]. 2019. p. 1126–33. Available from: https://www.aaai.org/ojs/index.php/AAAI/article/view/3905

51. Aziguli, Zhang Y, Xie Y, Xu Y, Chen Y. Structural technology research on symptom data of Chinese medicine. In: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom) [Internet]. 2017. p. 1–4. Available from: http://dx.doi.org/10.1109/HealthCom.2017.8210797

52. Santos A, Colaço AR, Nielsen AB, Niu L, Geyer PE, Coscia F, et al. Clinical Knowledge Graph Integrates Proteomics Data into Clinical Decision-Making [Internet]. bioRxiv. 2020 [cited 2021 Jul 22]. p. 2020.05.09.084897. Available from: https://www.biorxiv.org/content/10.1101/2020.05.09.084897v1.abstract

53. Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable

embeddings. Nat Commun [Internet]. 2019 Jul 10;10(1):3045. Available from: http://dx.doi.org/10.1038/s41467-019-11069-0

54. Yu Y, Wang Y, Xia Z, Zhang X, Jin K, Yang J, et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D1090–101. Available from: http://dx.doi.org/10.1093/nar/gky1042

55. Su C, Hou Y, Guo W, Chaudhry F, Ghahramani G, Zhang H, et al. CBKH: The Cornell Biomedical Knowledge Hub. medRxiv [Internet]. 2021; Available from: https://www.medrxiv.org/content/10.1101/2021.03.12.21253461v1.abstract

56. Luciano JS, Andersson B, Batchelor C, Bodenreider O, Clark T, Denney CK, et al. The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. J Biomed Semantics [Internet]. 2011 May 17;2 Suppl 2:S1. Available from: http://dx.doi.org/10.1186/2041-1480-2-S2-S1

57. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife [Internet]. 2017 Sep 22;6. Available from: http://dx.doi.org/10.7554/eLife.26726

58. Callahan TJ, Tripodi IJ, Hunter, Lawrence E, Baumgartner Jr., William A. The Phenotype Knowledge Translator (PheKnowLator) Ecosystem [Internet]. 2021. (Zenodo Community). Available from: https://zenodo.org/communities/pheknowlator-ecosystem

59. Callahan TJ. OMOP2OBO [Internet]. 2021. (Zenodo Community). Available from: https://zenodo.org/communities/omop2obo

60. Health Data Compass [Internet]. [cited 2021 Jul 20]. Available from: https://www.healthdatacompass.org/

61. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc [Internet]. 2012 Jan;19(1):54–60. Available from: http://dx.doi.org/10.1136/amiajnl-2011-000376

62. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc [Internet]. 2014 Jul;21(4):602–6. Available from: http://dx.doi.org/10.1136/amiajnl-2014-002743

63. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: International Conference on Machine Learning [Internet]. 2014 [cited 2018 Sep 6]. p. 1188–96. Available from: http://proceedings.mlr.press/v32/le14.html

64. Rajaraman A, Ullman JD. Data Mining. In: Mining of Massive Datasets [Internet]. Cambridge University Press; 2011 [cited 2021 Jun 13]. p. 1–17. Available from: https://www.cambridge.org/core/books/mining-of-massive-datasets/data-mining/E5BFF4C1DD5A1FB946D616D619B373C2

65. Alshahrani M, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N, Hoehndorf R. Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics [Internet]. 2017 Sep 1;33(17):2723–30. Available from: http://dx.doi.org/10.1093/bioinformatics/btx275

66. van der Maaten L, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. J Mach Learn Res. 9:2579–605.

67. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl [Internet]. 2004 Jun 1;6(1):20–9. Available from: https://doi.org/10.1145/1007730.1007735

68. Lloyd SP. Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM. IEEE Trans Inf Theory. 1957;18.

69. Steinley D. Properties of the Hubert-Arable Adjusted Rand Index. Psychol Methods [Internet]. 2004 Sep;9(3):386–96. Available from: https://psycnet.apa.org/fulltext/2004-17801-007.pdf

70. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2009 [cited 2021 Jun 13]. p. 1073–80. (ICML '09). Available from: https://doi.org/10.1145/1553374.1553511

71. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval [Internet]. Cambridge University Press; 2008. 482 p. Available from: https://play.google.com/store/books/details?id=GNvtngEACAAJ

72. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. limma: Linear Models for Microarray and RNA-Seq Data User's Guide [Internet]. Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research; 2002 Dec. Available from: http://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf

73. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res [Internet]. 2015 Apr 20;43(7):e47. Available from: http://dx.doi.org/10.1093/nar/gkv007

74. Thorndike RL. Who belongs in the family? Psychometrika [Internet]. 1953 Dec 1;18(4):267–76. Available from: https://doi.org/10.1007/BF02289263

75. Richardson JTE. Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review [Internet]. 2011 Jan 1;6(2):135–47. Available from: https://www.sciencedirect.com/science/article/pii/S1747938X11000029

76. Cohen J. Statistical Power Analysis for the Behavioural Sciences [Internet]. Academic Press, New York; 1969. 415 p. Available from: https://play.google.com/store/books/details?id=ojHWcQAACAAJ

77. Gensim: Topic Modelling for Humans [Internet]. [cited 2021 Aug 23]. Available from: https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html

78. Hunter. Matplotlib: A 2D Graphics Environment. 2007 May 1;9:90–5. Available from: http://dx.doi.org/10.1109/MCSE.2007.55

79. scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation [Internet].

[cited 2021 Jun 13]. Available from: https://scikit-learn.org/

80. Quinlan J, Idaghdour Y, Goulet J-P, Gbeha E, de Malliard T, Bruat V, et al. Genomic architecture of sickle cell disease in West African children. Front Genet [Internet]. 2014 Feb 14;5:26. Available from: http://dx.doi.org/10.3389/fgene.2014.00026

81. Sebastiani P, Nolan VG, Baldwin CT, Abad-Grau MM, Wang L, Adewoye AH, et al. A network model to predict the risk of death in sickle cell disease. Blood [Internet]. 2007 Oct 1;110(7):2727–35. Available from: http://dx.doi.org/10.1182/blood-2007-04-084921

82. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res [Internet]. 2009 Jul;37(Web Server issue):W305–11. Available from: http://dx.doi.org/10.1093/nar/gkp427

83. Schlaeger JM, Molokie RE, Yao Y, Suarez ML, Golembiewski J, Wilkie DJ, et al. Management of Sickle Cell Pain Using Pregabalin: A Pilot Study. Pain Manag Nurs [Internet]. 2017 Dec;18(6):391–400. Available from: http://dx.doi.org/10.1016/j.pmn.2017.07.003

84. Muenzer J. Overview of the mucopolysaccharidoses. Rheumatology [Internet]. 2011 Dec;50 Suppl 5:v4–12. Available from: http://dx.doi.org/10.1093/rheumatology/ker394

85. Garnier Y, Ferdinand S, Etienne-Julan M, Elana G, Petras M, Doumdo L, et al. Differences of microparticle patterns between sickle cell anemia and hemoglobin SC patients. PLoS One [Internet]. 2017 May 10;12(5):e0177397. Available from: http://dx.doi.org/10.1371/journal.pone.0177397

86. Claudino MA, Franco-Penteado CF, Corat MAF, Gimenes AP, Passos LAC, Antunes E, et al. Increased cavernosal relaxations in sickle cell mice priapism are associated with alterations in the NO-cGMP signaling pathway. J Sex Med [Internet]. 2009 Aug;6(8):2187–96. Available from: http://dx.doi.org/10.1111/j.1743-6109.2009.01337.x

87. Gardner K, Suddle A, Kane P, O'Grady J, Heaton N, Bomford A, et al. How we treat sickle hepatopathy and liver transplantation in adults. Blood [Internet]. 2014 Apr 10;123(15):2302–7. Available from: http://dx.doi.org/10.1182/blood-2013-12-542076

88. Ndefo UA, Maxwell AE, Nguyen H, Chiobi TL. Pharmacological management of sickle cell disease. P T [Internet]. 2008 Apr;33(4):238–43. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19750169

89. Prengler M, Pavlakis SG, Boyd S, Connelly A, Calamante F, Chong WK, et al. Sickle cell disease: ischemia and seizures. Ann Neurol [Internet]. 2005 Aug;58(2):290–302. Available from: http://dx.doi.org/10.1002/ana.20556

90. Wooton AK, Melchior LM. Diabetic autonomic neuropathy resulting in sexual dysfunction. Nurse Pract [Internet]. 2018 Nov;43(11):39–45. Available from: http://dx.doi.org/10.1097/01.NPR.0000546447.94226.53

91. Vinik AI, Maser RE, Mitchell BD, Freeman R. Diabetic autonomic neuropathy. Diabetes Care [Internet]. 2003 May;26(5):1553–79. Available from: http://dx.doi.org/10.2337/diacare.26.5.1553

92. Crane GM, Bennett NE Jr. Priapism in sickle cell anemia: emerging mechanistic understanding and better preventative strategies. Anemia [Internet]. 2011;2011:297364. Available from: http://dx.doi.org/10.1155/2011/297364

93. Lundström M, Brege KG, Florén I, Lundh B, Stenevi U, Thorburn W. Postoperative aphakia in modern cataract surgery: part 2: detailed analysis of the cause of aphakia and the visual outcome. J Cataract Refract Surg [Internet]. 2004 Oct;30(10):2111–5. Available from: http://dx.doi.org/10.1016/j.jcrs.2004.01.044

94. Babalola OE, Danboyi P, Abiose AA. Hereditary congenital cataracts associated with sickle cell anaemia in a Nigerian family. Trop Doct [Internet]. 2000 Jan;30(1):12–4. Available from: http://dx.doi.org/10.1177/004947550003000107

95. Fraint E, Leffler M, Schultz CL. Coexisting Juvenile Dermatomyositis and Sickle Cell Disease: Maintaining a High Degree of Suspicion. J Pediatr [Internet]. 2017 Jul;186:189–91. Available from: http://dx.doi.org/10.1016/j.jpeds.2017.03.038

96. Roberts MZ, Gaskill GE, Kanter-Washko J, Kyle TR 3rd, Jones BC, Bohm NM. Effectiveness and safety of oral anticoagulants in patients with sickle cell disease and venous thromboembolism: a retrospective cohort study. J Thromb Thrombolysis [Internet]. 2018 May;45(4):512–5. Available from: http://dx.doi.org/10.1007/s11239-018-1637-y

97. Saito N, Nadgir RN, Flower EN, Sakai O. Clinical and radiologic manifestations of sickle cell disease in the head and neck. Radiographics [Internet]. 2010 Jul;30(4):1021–34. Available from: http://dx.doi.org/10.1148/rg.304095171

98. Jaffe ES. Histiocytoses of lymph nodes: biology and differential diagnosis. Semin Diagn Pathol [Internet]. 1988 Nov;5(4):376–90. Available from: https://www.ncbi.nlm.nih.gov/pubmed/3064218

99. Dixit R, Nettem S, Madan SS, Soe HHK, Abas ABL, Vance LD, et al. Folate supplementation in people with sickle cell disease. Cochrane Database Syst Rev [Internet]. 2016 Feb 16;2:CD011130. Available from: http://dx.doi.org/10.1002/14651858.CD011130.pub2

100. Pei P, Cheng X, Yu J, Shen J, Li X, Wu J, et al. Folate deficiency induced H2A ubiquitination to lead to downregulated expression of genes involved in neural tube defects. Epigenetics Chromatin [Internet]. 2019 Nov 13;12(1):69. Available from: http://dx.doi.org/10.1186/s13072-019-0312-7

101. Caracas M da S, Jales SP, Jales Neto LH, da Silva Castro JC, Suganuma LM, Fonseca GHH, et al. Temporomandibular joint arthritis in sickle cell disease: a case report. Oral Surg Oral Med Oral Pathol Oral Radiol [Internet]. 2013 Feb;115(2):e31–5. Available from: http://dx.doi.org/10.1016/j.oooo.2012.05.018

102. Braimah RO, Oladejo T, Olarinoye TO, Adetoye AO, Osho PO. A multidisciplinary approach to the management of temporomandibular joint ankylosis in a sickle-cell anemia patient in a resource-limited setting. Ann Maxillofac Surg [Internet]. 2016 Jan;6(1):130–4. Available from: http://dx.doi.org/10.4103/2231-0746.186124

103. Snyder AB, Lane PA, Zhou M, Paulukonis ST, Hulihan MM. The accuracy of hospital ICD-9-CM codes for determining Sickle Cell Disease genotype. J Rare Dis Res Treat

[Internet]. 2017 Jul 28;2(4):39–45. Available from:
https://www.ncbi.nlm.nih.gov/pubmed/29202133