Accelerating Rare Disease Diagnosis

Gursev Pirge

Agenda

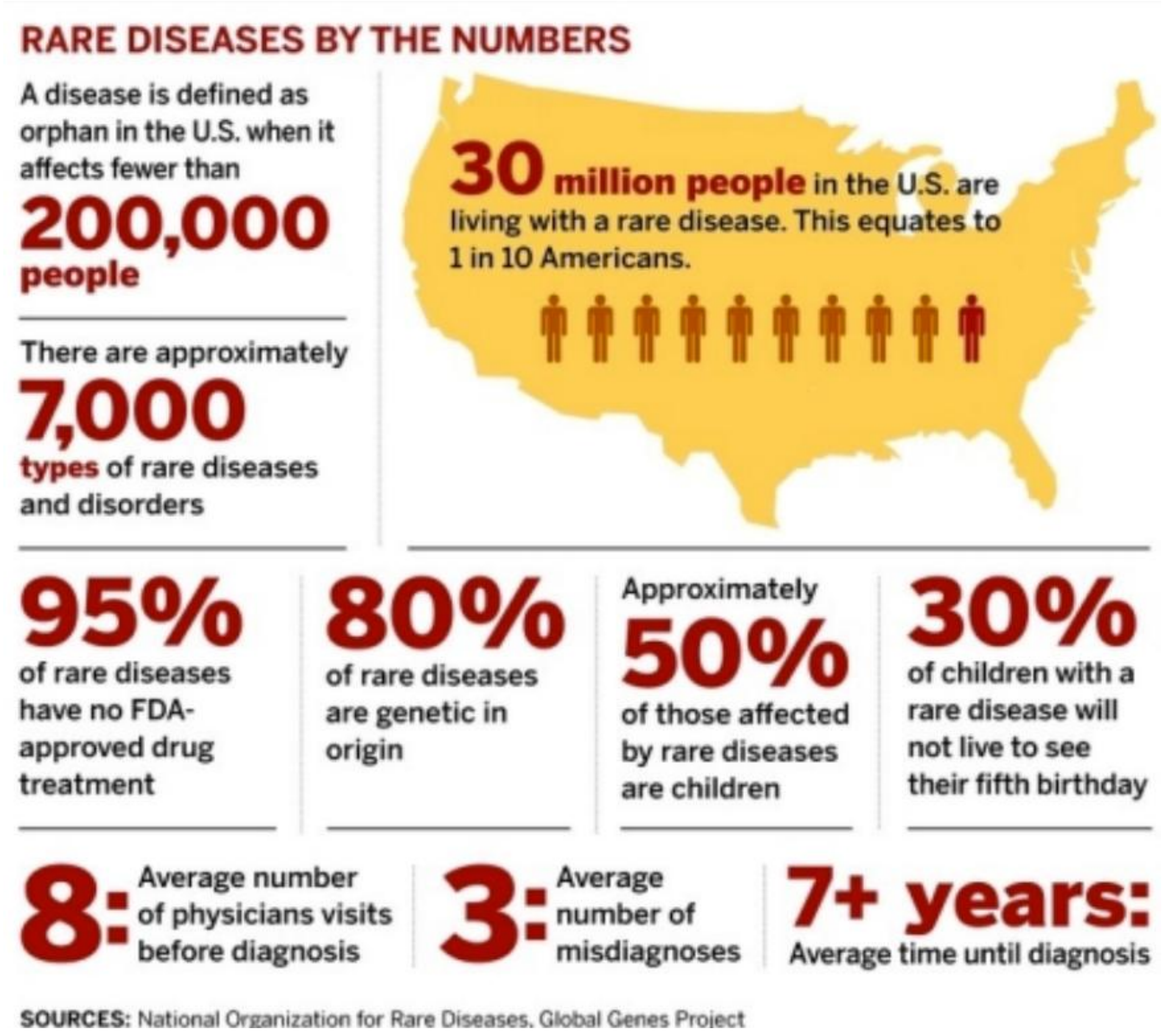
- Phenotypes
- Importance
- John Snow Labs Solution
- Using ClinPhen and LLMs
- Comparison
- Conclusion

What are Phenotypes?

In medicine, phenotypes refer to clinically observable traits including:

- Signs and Symptoms, e.g., ataxia,
- Clinical Findings, e.g., physical exam results,
- Laboratory Results, e.g., elevated glucose,
- Imaging Features, e.g., MRI findings.

Rare Disease Impact



RARE DISEASES BY THE NUMBERS

A disease is defined as orphan in the U.S. when it affects fewer than **200,000 people**

There are approximately **7,000 types** of rare diseases and disorders

30 million people in the U.S. are living with a rare disease. This equates to 1 in 10 Americans.

95% of rare diseases have no FDA-approved drug treatment

80% of rare diseases are genetic in origin

Approximately **50%** of those affected by rare diseases are children

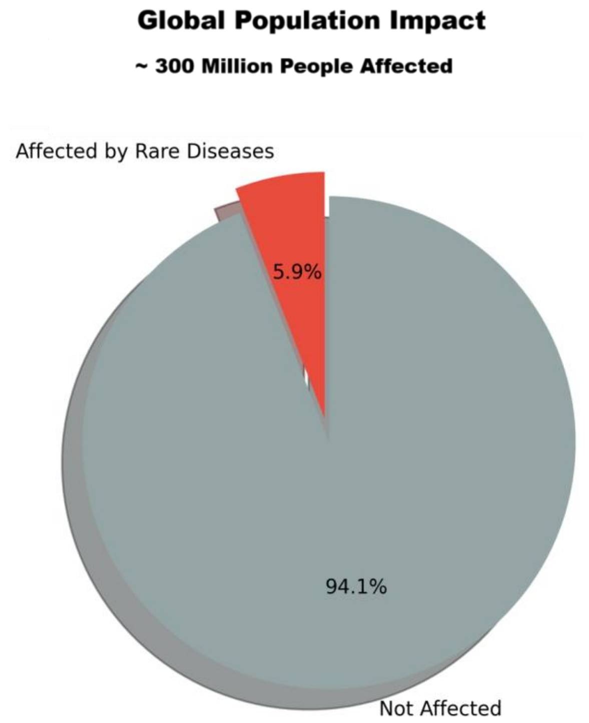
30% of children with a rare disease will not live to see their fifth birthday

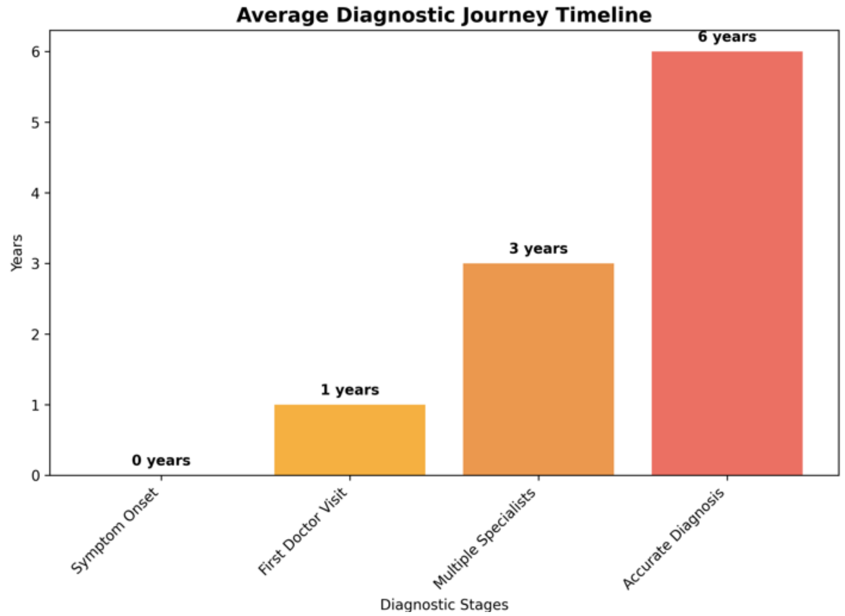
8: Average number of physicians visits before diagnosis

3: Average number of misdiagnoses

7+ years: Average time until diagnosis

SOURCES: National Organization for Rare Diseases, Global Genes Project





Human Phenotype Ontology (HPO)

A standardized vocabulary for describing phenotypic abnormalities in human diseases.

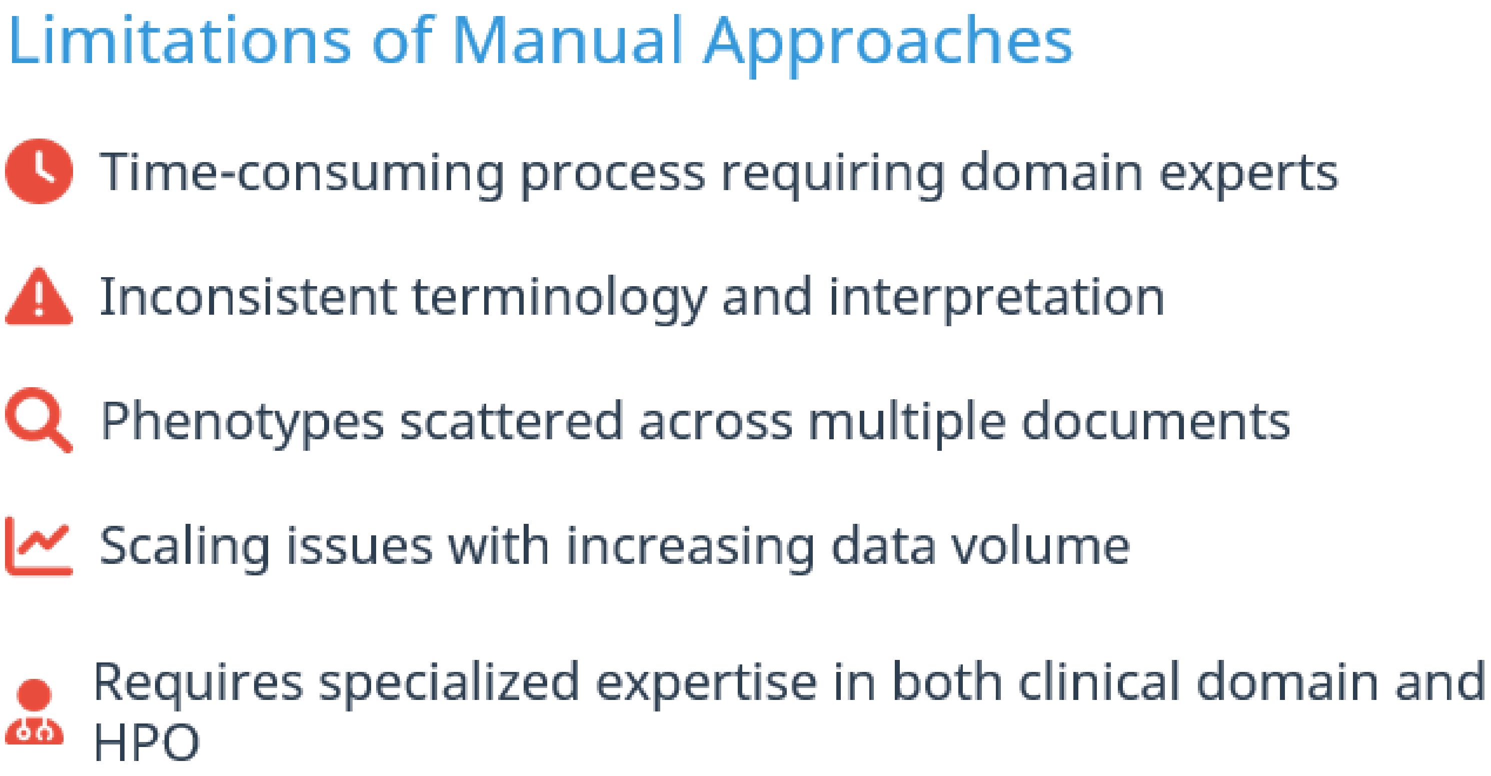
- Over 18,000 terms with unique HPO codes.
- Each HPO term describes a specific phenotypic abnormality, rather than the diseases or syndromes directly.
- Hierarchical structure organizing phenotypes.
- Global standard for genomics and rare disease research.

Rare Disease Examples

Examples of Rare Diseases with Associated Phenotypes and HPO Codes

Disease	Phenotype	HPO Code
Marfan Syndrome	Arachnodactyly (long, slender fingers)	HP:0001166
Marfan Syndrome	Aortic root dilation	HP:0001653
Duchenne Muscular Dystrophy	Muscle weakness	HP:0003324
Duchenne Muscular Dystrophy	Gowers' sign	HP:0003392
Rett Syndrome	Stereotypical hand movements	HP:0001516
Rett Syndrome	Intellectual disability	HP:0001249

Manual Approach



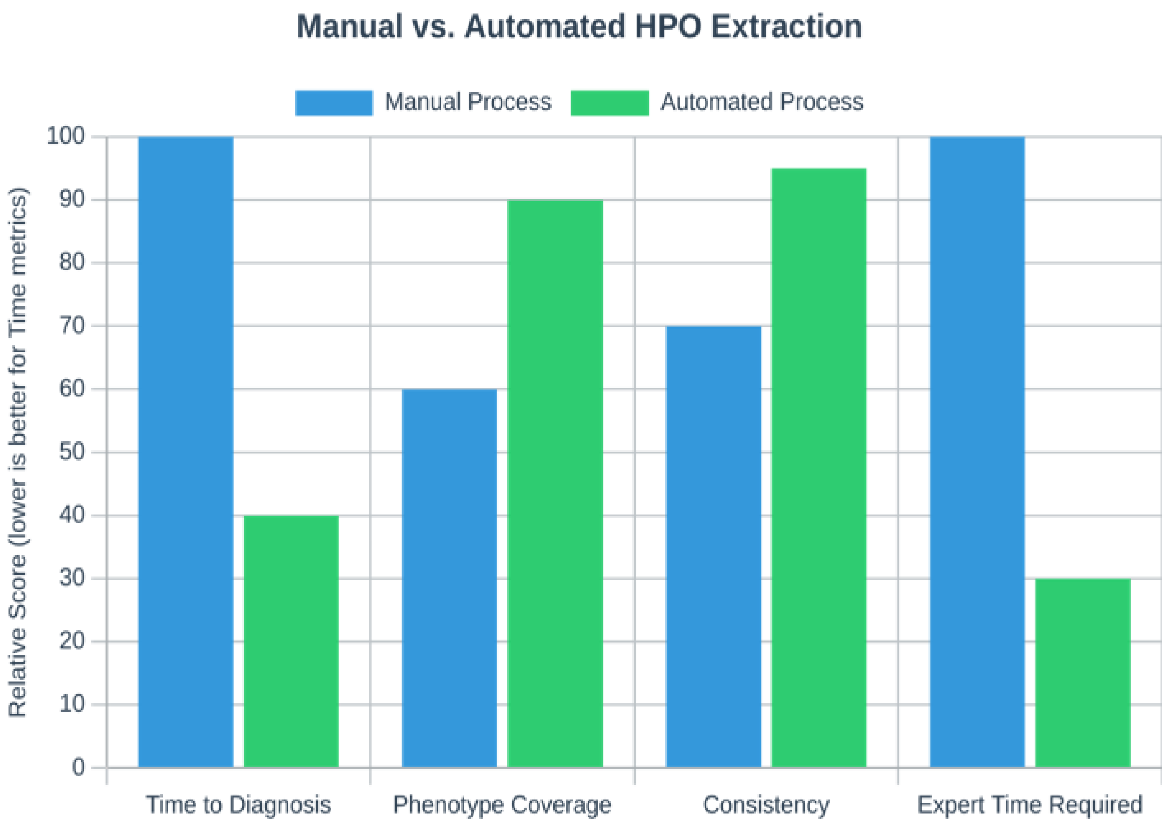
Limitations of Manual Approaches

- Time-consuming process requiring domain experts
- Inconsistent terminology and interpretation
- Phenotypes scattered across multiple documents
- Scaling issues with increasing data volume
- Requires specialized expertise in both clinical domain and HPO

Benefits of Automation

Key Advantages

- Accelerated diagnosis time for rare diseases
- Improved accuracy and consistency in phenotype identification
- Structured data for advanced analytics and research
- Reduced burden on clinical experts
- Enhanced phenotype-genotype correlation studies

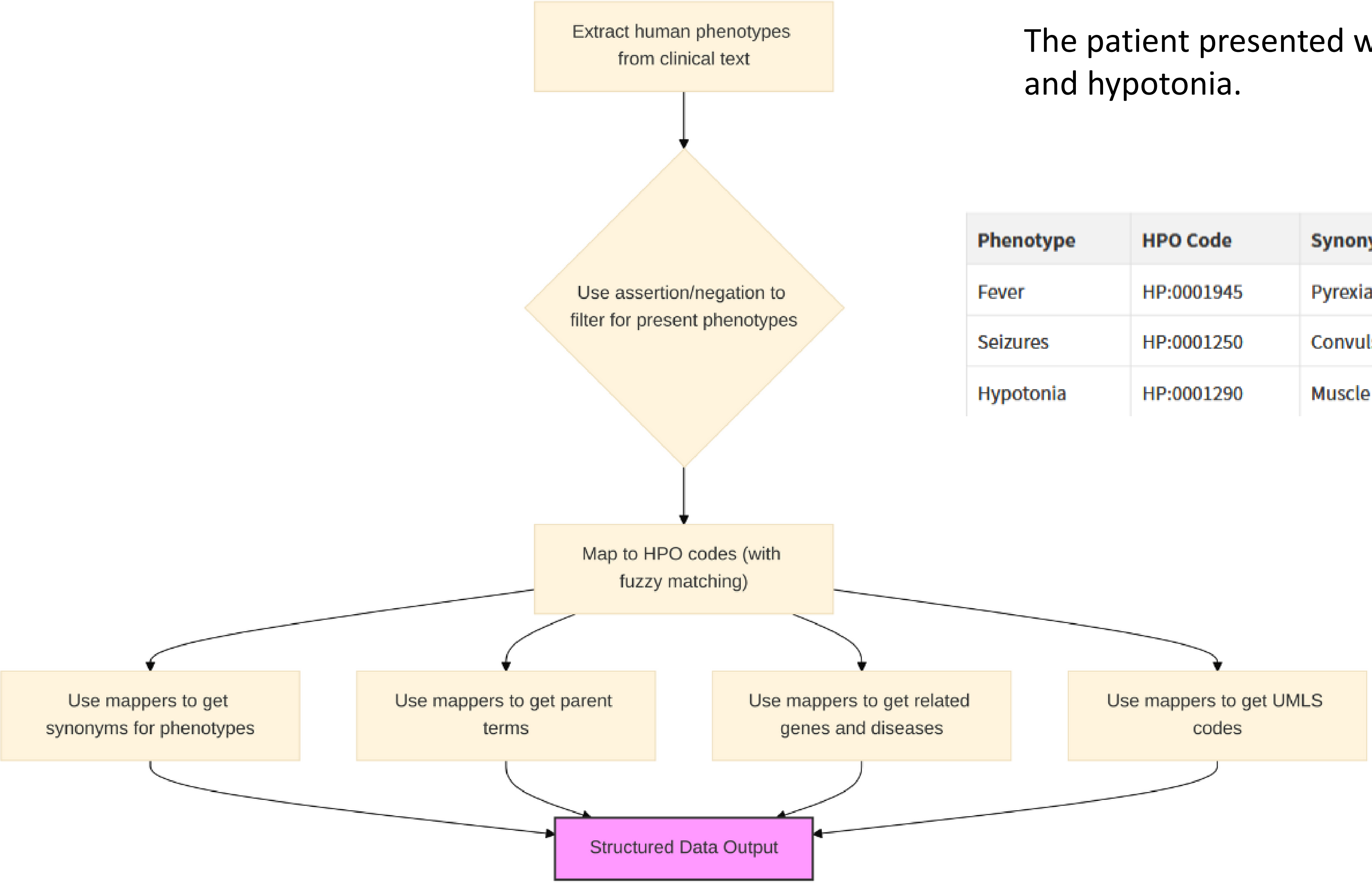


Impact of automated HPO extraction on diagnostic process

John Snow Labs Solution

- Reference Dataset: https://obofoundry.org/ontology/hp.html
- Reference Github: https://github.com/OBOFoundry/purl.obolibrary.org/
- NER models cannot solve the problem – hard to differentiate from diseases, signs, symptoms etc.
- Alternative solution – includes fuzzy matching, lemmatizer.
- Our pipeline automates the extraction of phenotypes from clinical notes, normalizes them to HPO codes to accelerate rare disease diagnosis.
- Option to map to synonyms, parent terms, genes & diseases, UMLS codes etc.

John Snow Labs Solution



Extract human phenotypes from clinical text

Use assertion/negation to filter for present phenotypes

Map to HPO codes (with fuzzy matching)

Use mappers to get synonyms for phenotypes

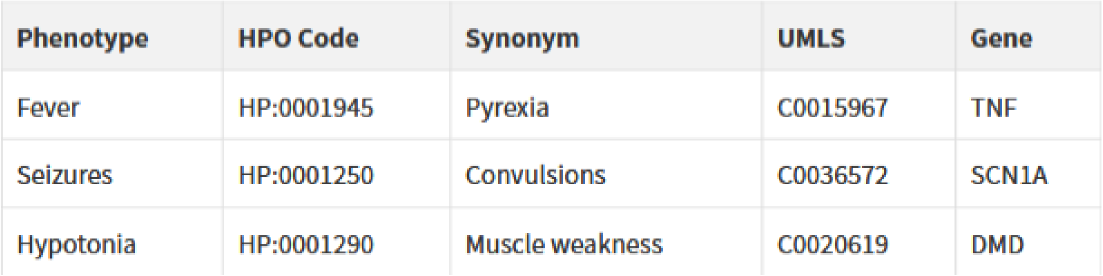
Use mappers to get parent terms

Use mappers to get related genes and diseases

Use mappers to get UMLS codes

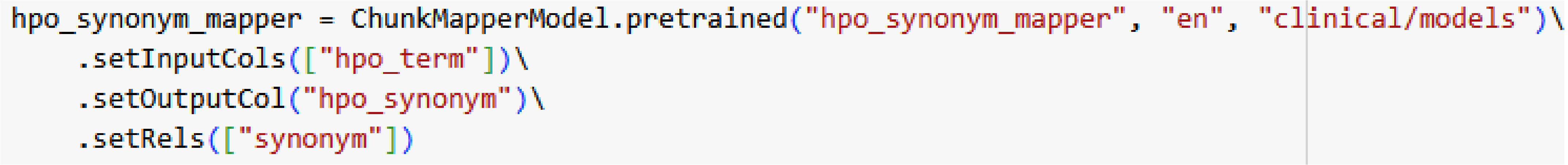
Structured Data Output

The patient presented with fever, seizures, and hypotonia.

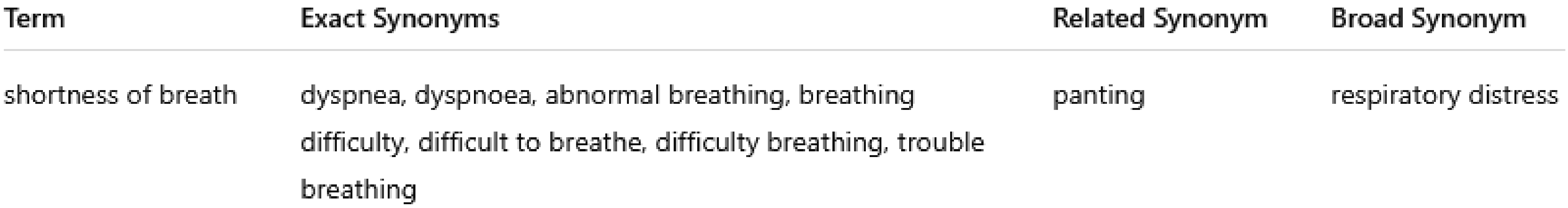


Phenotype	HPO Code	Synonym	UMLS	Gene
Fever	HP:0001945	Pyrexia	C0015967	TNF
Seizures	HP:0001250	Convulsions	C0036572	SCN1A
Hypotonia	HP:0001290	Muscle weakness	C0020619	DMD

Synonym Mapper



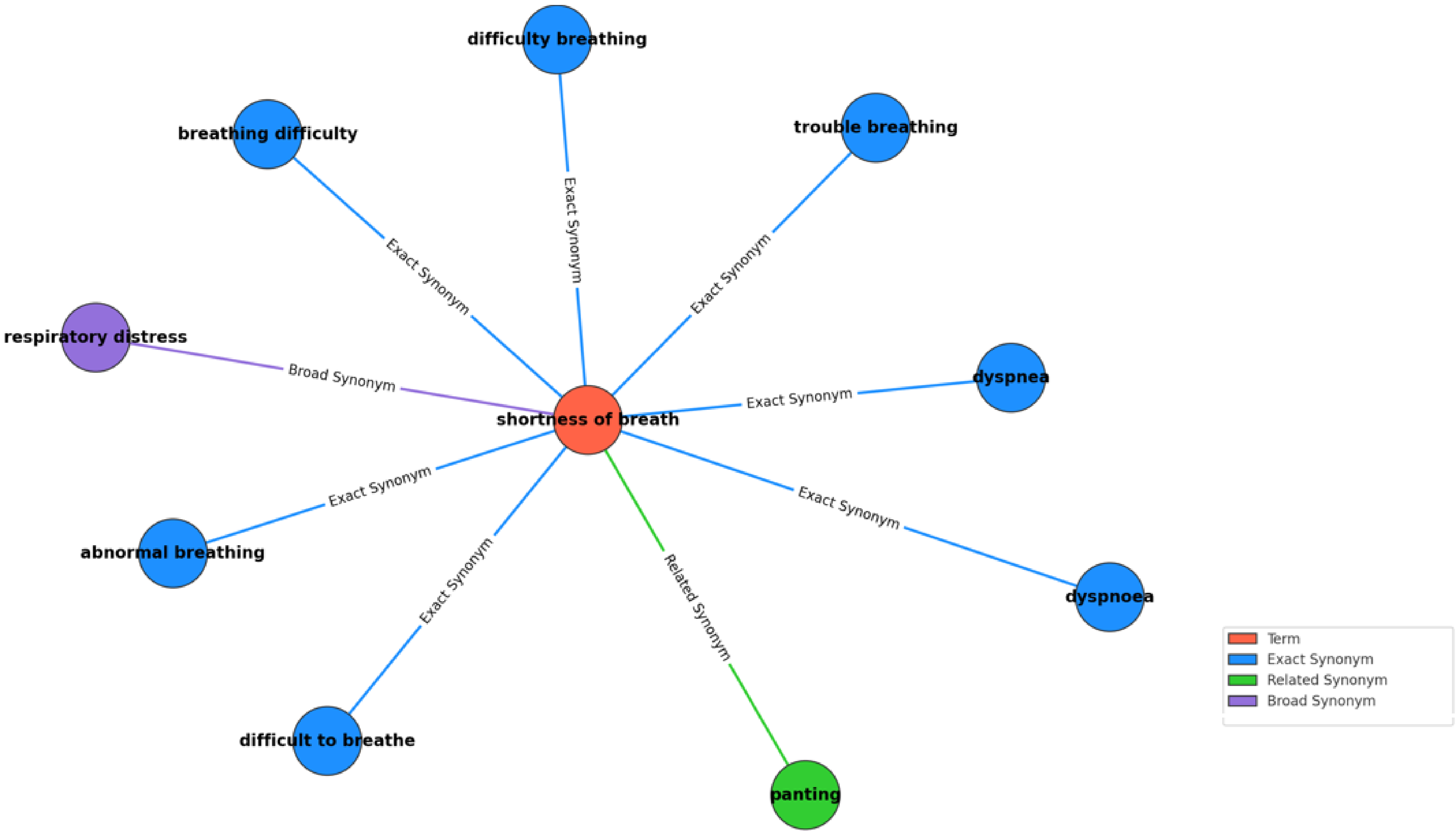
```
hpo_synonym_mapper = ChunkMapperModel.pretrained("hpo_synonym_mapper", "en", "clinical/models")\
    .setInputCols(["hpo_term"])\
    .setOutputCol("hpo_synonym")\
    .setRels(["synonym"])
```



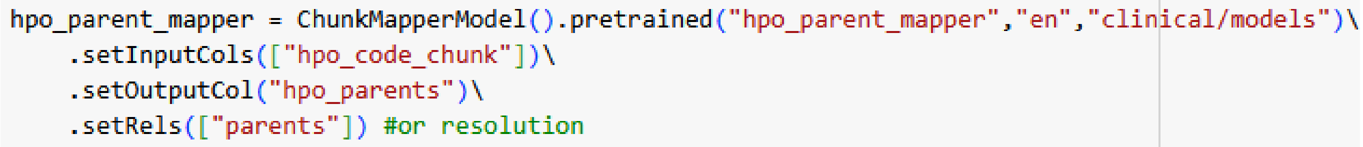
Term	Exact Synonyms	Related Synonym	Broad Synonym
shortness of breath	dyspnea, dyspnoea, abnormal breathing, breathing difficulty, difficult to breathe, difficulty breathing, trouble breathing	panting	respiratory distress

Synonym Mapper

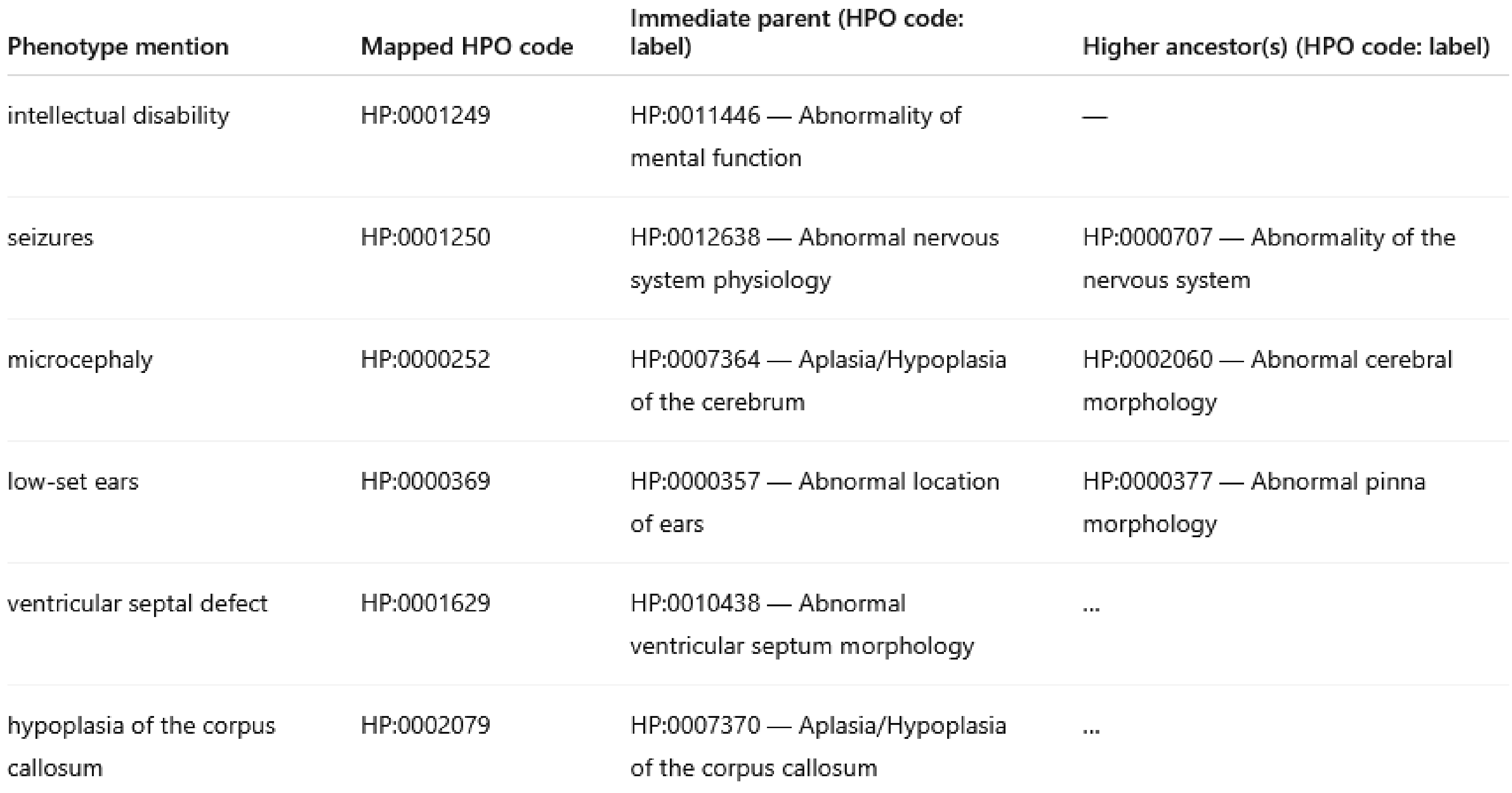
Term	Exact Synonyms	Related Synonym	Broad Synonym
shortness of breath	dyspnea, dyspnoea, abnormal breathing, breathing difficulty, difficult to breathe, difficulty breathing, trouble breathing	panting	respiratory distress



Parent Mapper

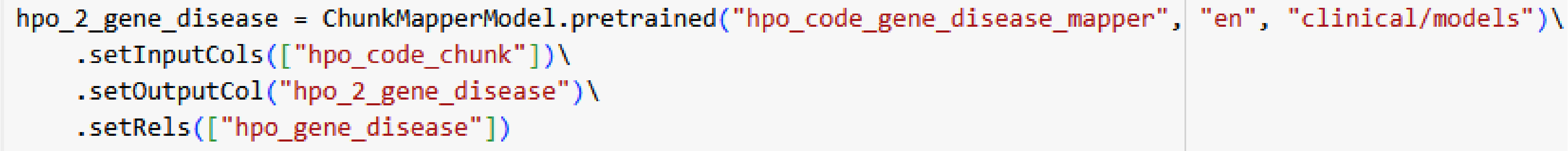


```
hpo_parent_mapper = ChunkMapperModel().pretrained("hpo_parent_mapper","en","clinical/models")\
    .setInputCols(["hpo_code_chunk"])\
    .setOutputCol("hpo_parents")\
    .setRels(["parents"]) #or resolution
```

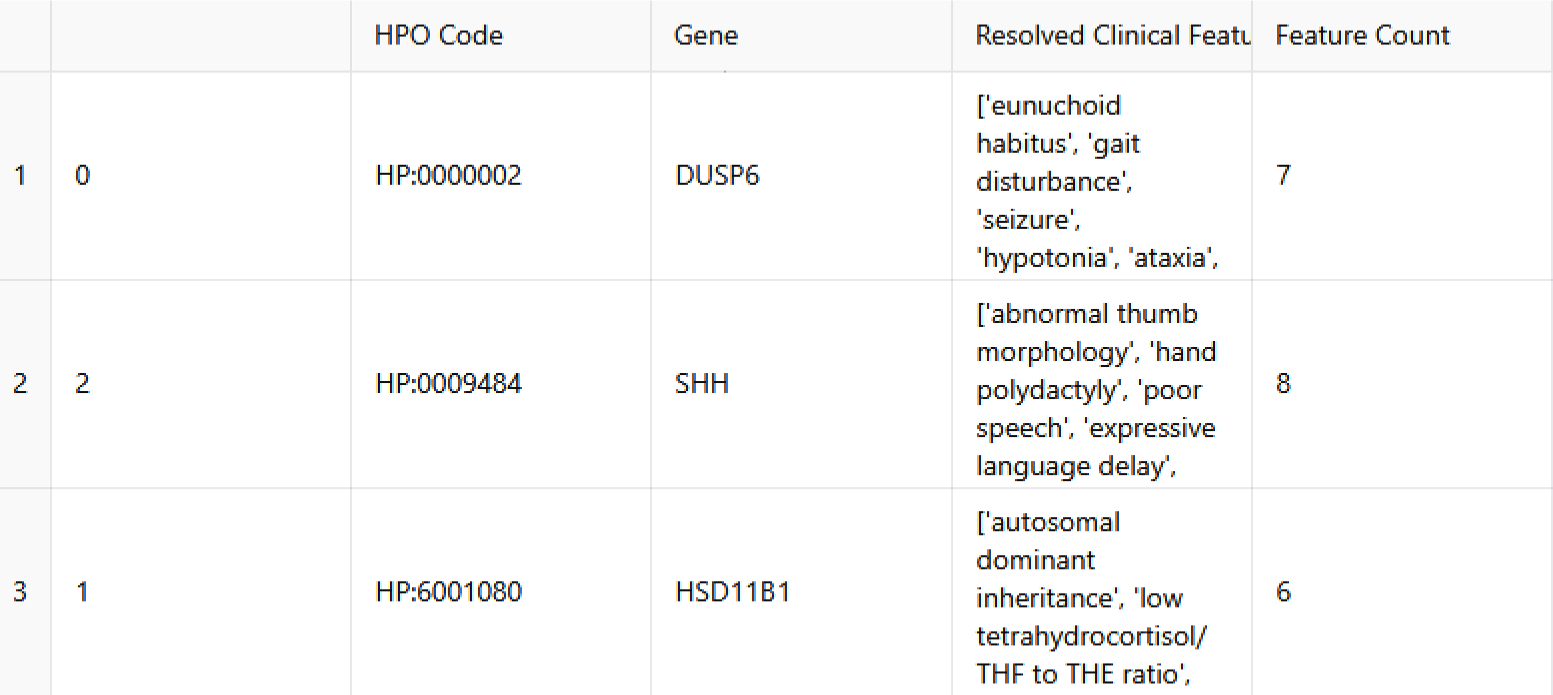


Phenotype mention	Mapped HPO code	Immediate parent (HPO code: label)	Higher ancestor(s) (HPO code: label)
intellectual disability	HP:0001249	HP:0011446 — Abnormality of mental function	—
seizures	HP:0001250	HP:0012638 — Abnormal nervous system physiology	HP:0000707 — Abnormality of the nervous system
microcephaly	HP:0000252	HP:0007364 — Aplasia/Hypoplasia of the cerebrum	HP:0002060 — Abnormal cerebral morphology
low-set ears	HP:0000369	HP:0000357 — Abnormal location of ears	HP:0000377 — Abnormal pinna morphology
ventricular septal defect	HP:0001629	HP:0010438 — Abnormal ventricular septum morphology	...
hypoplasia of the corpus callosum	HP:0002079	HP:0007370 — Aplasia/Hypoplasia of the corpus callosum	...

Gene-Disease Mapper

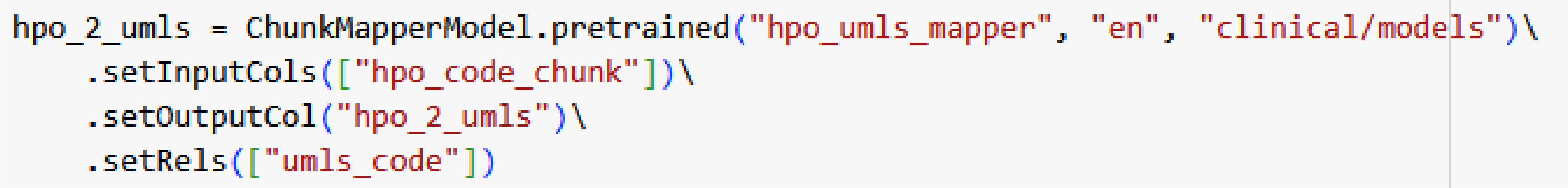


```
hpo_2_gene_disease = ChunkMapperModel.pretrained("hpo_code_gene_disease_mapper", "en", "clinical/models")\
    .setInputCols(["hpo_code_chunk"])\
    .setOutputCol("hpo_2_gene_disease")\
    .setRels(["hpo_gene_disease"])
```

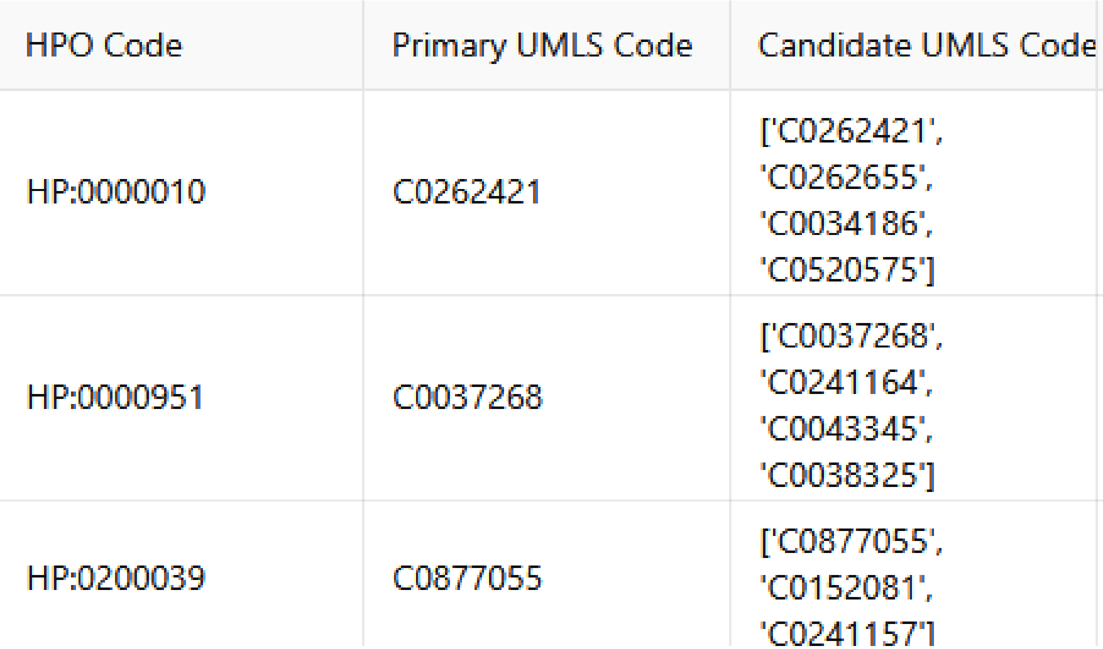


		HPO Code	Gene	Resolved Clinical Featu	Feature Count
1	0	HP:0000002	DUSP6	['eunuchoid habitus', 'gait disturbance', 'seizure', 'hypotonia', 'ataxia',	7
2	2	HP:0009484	SHH	['abnormal thumb morphology', 'hand polydactyly', 'poor speech', 'expressive language delay',	8
3	1	HP:6001080	HSD11B1	['autosomal dominant inheritance', 'low tetrahydrocortisol/ THF to THE ratio',	6

UMLS Mapper



```
hpo_2_umls = ChunkMapperModel.pretrained("hpo_umls_mapper", "en", "clinical/models")\
    .setInputCols(["hpo_code_chunk"])\
    .setOutputCol("hpo_2_umls")\
    .setRels(["umls_code"])
```



HPO Code	Primary UMLS Code	Candidate UMLS Code
HP:0000010	C0262421	['C0262421', 'C0262655', 'C0034186', 'C0520575']
HP:0000951	C0037268	['C0037268', 'C0241164', 'C0043345', 'C0038325']
HP:0200039	C0877055	['C0877055', 'C0152081', 'C0241157']

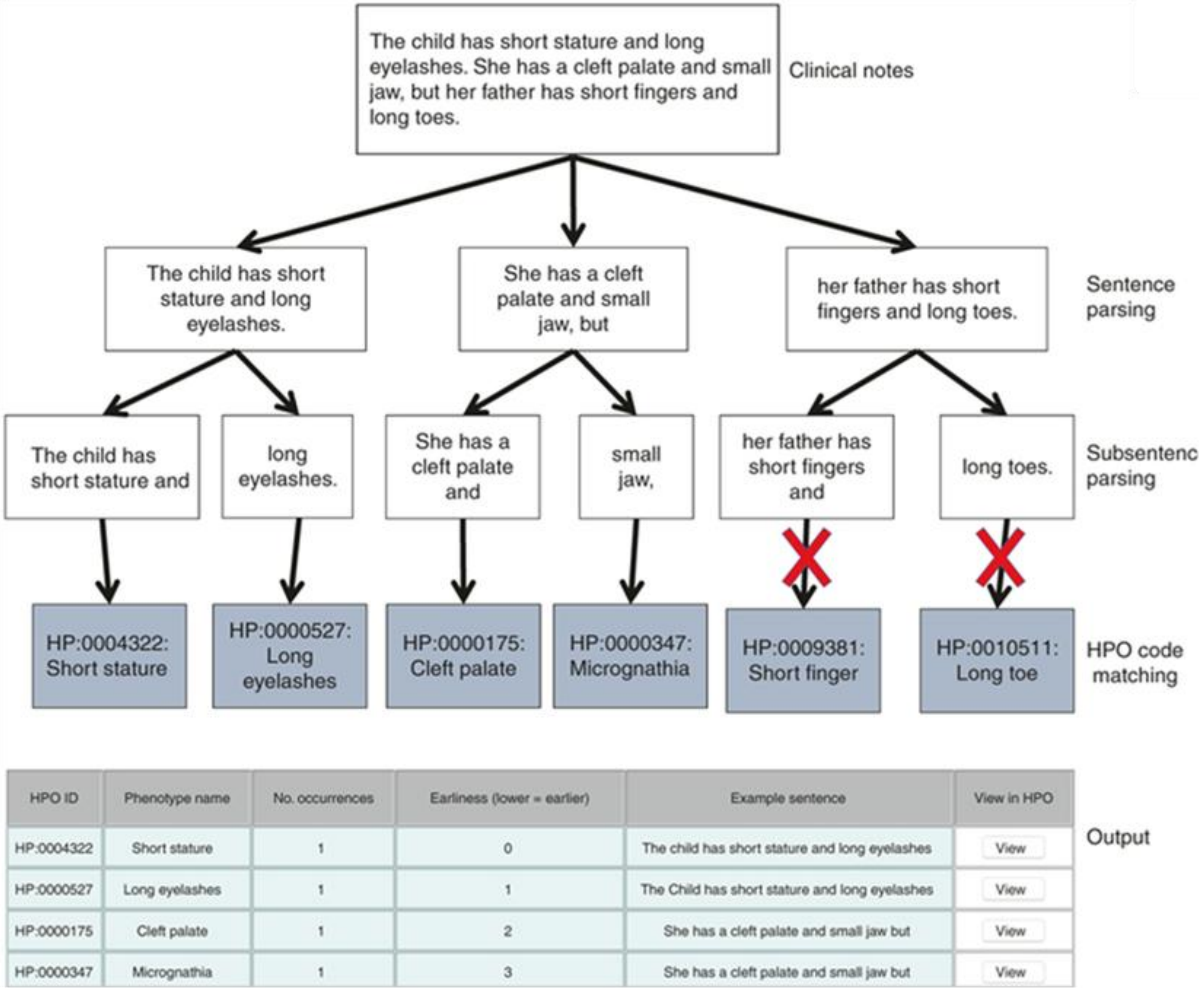
ClinPhen Library

ClinPhen library is a tool designed to automatically extract Human Phenotype Ontology **(HPO)** terms from clinical notes or free-text patient records.

Shortcomings:

- Mainly looks for phenotype mentions in text and maps them to HPO.
- Problem with handling negation, temporality, or uncertainty **well** (e.g., *“no seizures observed”* may still be marked as **HP:0001250 — Seizure**).
- Strong at capturing explicit HPO terms, but it can miss **synonyms** or implicit descriptions unless the wording matches its dictionary closely.

ClinPhen Library



The child has short stature and long eyelashes. She has a cleft palate and small jaw, but her father has short fingers and long toes.

Clinical notes

The child has short stature and long eyelashes.

She has a cleft palate and small jaw, but

her father has short fingers and long toes.

Sentence parsing

The child has short stature and

long eyelashes.

She has a cleft palate and

small jaw,

her father has short fingers and

long toes.

Subsentenc parsing

HP:0004322: Short stature

HP:0000527: Long eyelashes

HP:0000175: Cleft palate

HP:0000347: Micrognathia

HP:0009381: Short finger

HP:0010511: Long toe

HPO code matching

Output

HPO ID	Phenotype name	No. occurrences	Earliness (lower = earlier)	Example sentence	View in HPO
HP:0004322	Short stature	1	0	The child has short stature and long eyelashes	View
HP:0000527	Long eyelashes	1	1	The Child has short stature and long eyelashes	View
HP:0000175	Cleft palate	1	2	She has a cleft palate and small jaw but	View
HP:0000347	Micrognathia	1	3	She has a cleft palate and small jaw but	View

Using LLMs

LLMs can handle complex and unstructured texts.

Shortcomings:

- Computational resources.
- Risk of Hallucination: May generate incorrect information, such as inventing a phenotype not mentioned in the text.
- Can misinterpret negation (e.g., "no chest pain"), family history, or hypothetical scenarios, leading to critical errors.
- Inaccurate HPO Code Assignment: May select an HPO code that is incorrect, too general, or too specific, even if the phenotype itself is correctly identified.

Comparison

Dataset: The dataset was derived from the NER training data for HPO and GENE, but only HPO terms were kept.

Entries marked as Absent or Associated with Someone Else were removed during assertion review.

Each HPO term was then matched with official HPO JSON data to add corresponding HPO codes and synonyms.

The resulting dataset includes original text, the extracted HPO term (target chunk), its HPO code, and all associated synonyms.

Comparison

Evaluation Criteria: Results from LLM and ClinPhen were considered correct if they contained the target chunk.

For ClinPhen, the output is provided as phenotypes, which introduces some complexity.

In certain cases, the exact target chunk is returned, while in others, a corresponding synonym appears.

Consequently, the coverage metrics for ClinPhen should be interpreted as capturing both the target chunks and their associated synonyms.

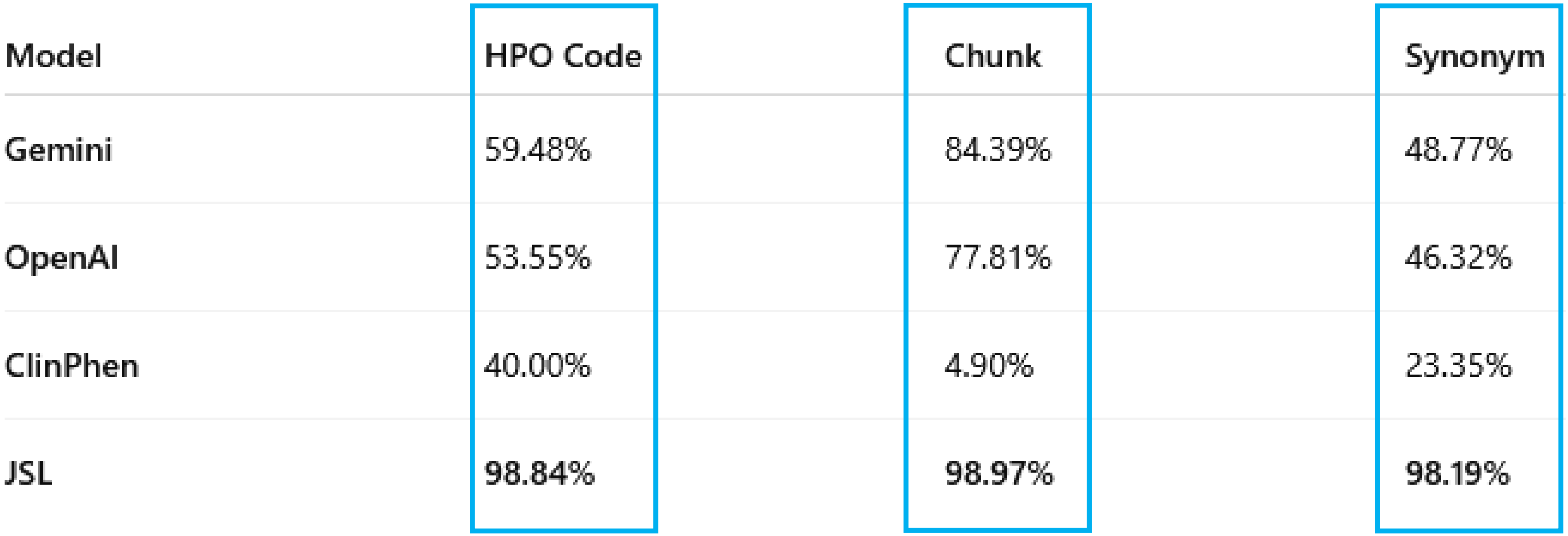
Comparison

Assertion Handling: When obtaining scores from LLM and ClinPhen, the corresponding assertions were examined.

Any entity labeled as ***Absent*** or ***Associated with Someone Else*** was not considered valid HPO terms.

ClinPhen operates in a similar manner; if a term is not related to the patient, ClinPhen does not identify it as an HPO term.

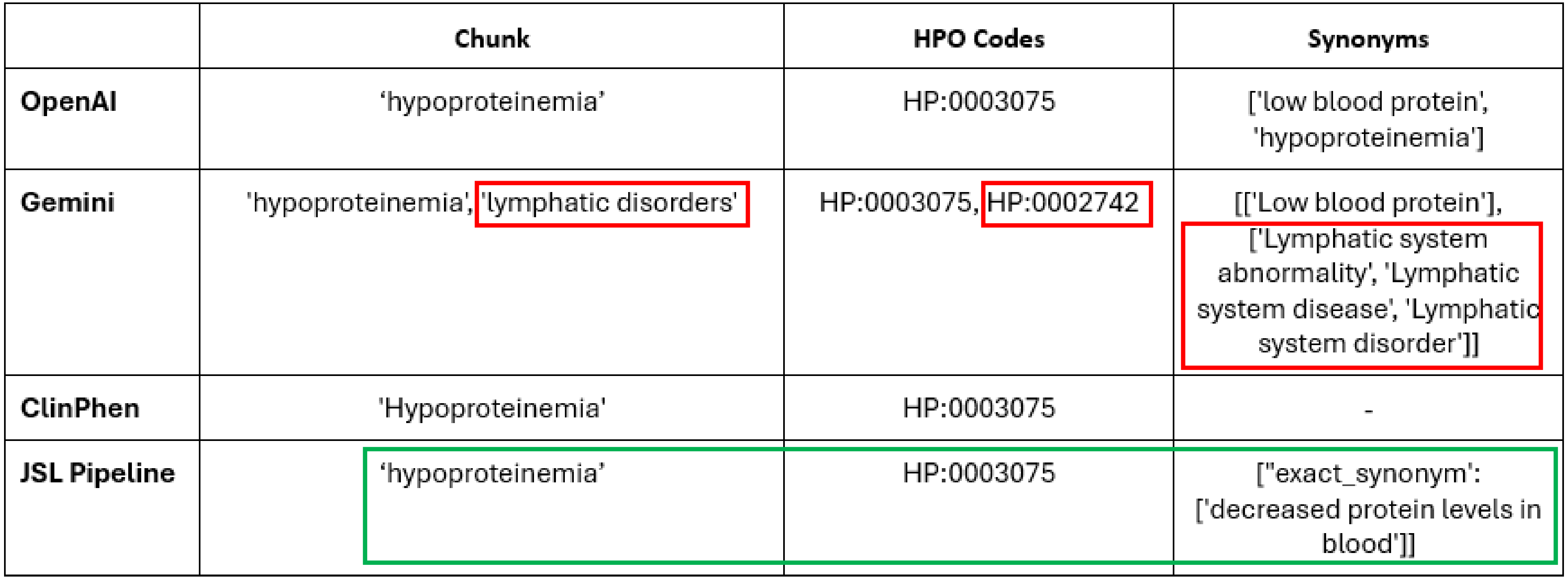
Results



Model	HPO Code	Chunk	Synonym
Gemini	59.48%	84.39%	48.77%
OpenAI	53.55%	77.81%	46.32%
ClinPhen	40.00%	4.90%	23.35%
JSL	**98.84%**	**98.97%**	**98.19%**

Example

A 17-year-old adolescent boy with a history of hypoproteinemia underwent regulation of apoptotic process-NEB PET/MRI to evaluate possible lymphatic disorders suggested by FDG PET/phosphatidylcholine biosynthetic process imaging .



	Chunk	HPO Codes	Synonyms
OpenAI	'hypoproteinemia'	HP:0003075	['low blood protein', 'hypoproteinemia']
Gemini	'hypoproteinemia', 'lymphatic disorders'	HP:0003075, HP:0002742	[['Low blood protein'], ['Lymphatic system abnormality', 'Lymphatic system disease', 'Lymphatic system disorder']]
ClinPhen	'Hypoproteinemia'	HP:0003075	-
JSL Pipeline	'hypoproteinemia'	HP:0003075	["exact_synonym': ['decreased protein levels in blood']]

Example

During the clinical evaluation, the patient exhibited fused nails and teeth, a finding consistent with a rare congenital syndrome affecting ectodermal structures.

	Chunk	HPO Codes	Synonyms
OpenAI	'fused nails', 'fused teeth'	HP:0010622, HP:0011069	-
Gemini	'fused nails', 'fused ... teeth'	HP:0010573, HP:0000695	['Fused nails', 'Nail fusion'], ['Dental fusion', 'Synodontia']
ClinPhen	‘Fused nails’	HP:0011312	-
JSL Pipeline	‘fused nails’	HP:0011312	‘fused nails’

Conclusion

- Phenotypes are critical for understanding and diagnosing rare diseases,
- HPO provides a standardized vocabulary with over 18,000 terms,
- Automated extraction significantly improves the diagnostic process,
- John Snow Labs' pipeline transforms unstructured text into actionable insights with high accuracy and speed.

TL; DR: Accelerating rare disease diagnosis through AI-powered phenotype extraction.