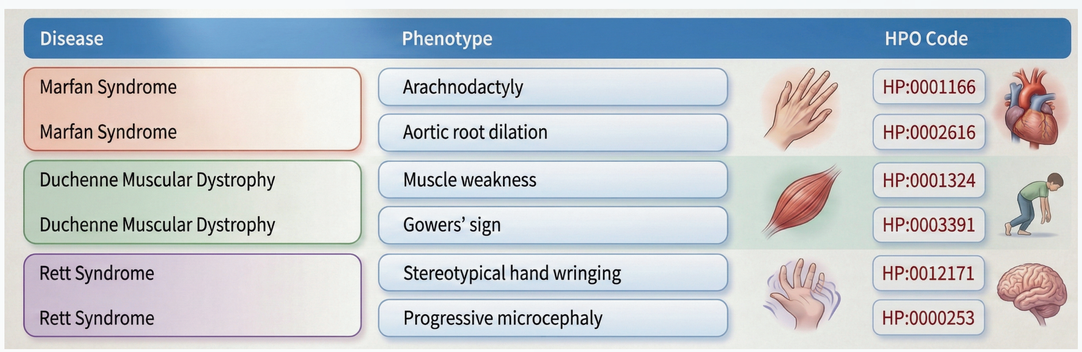
REGULATORY-GRADE HPO CODE EXTRACTION FROM UNSTRUCTURED CLINICAL NOTES: ACCELERATING RARE DISEASE DIAGNOSIS, METHODS AND RESULTS

Veysel Kocaman, PhD · David Talby, PhD

John Snow Labs Inc., Delaware, USA

Background

Rare diseases affect an estimated 300 million individuals globally, yet diagnosis often takes 5–10 years, largely due to fragmented and unstructured phenotypic information in clinical notes. Human Phenotype Ontology (HPO) provides a standardized vocabulary of over 18,000 phenotypic terms and is foundational to modern genomic diagnostics, phenomatching, and cohort discovery. However, manual extraction and normalization of phenotypes from narrative text is slow (10–20 minutes per report), inconsistent, and difficult to scale, limiting its practical use in real-world evidence (RWE) pipelines. In rare disease workflows, the key bottleneck is not the absence of phenotype information, but the lack of standardized phenotype representation; HPO provides this standardization.

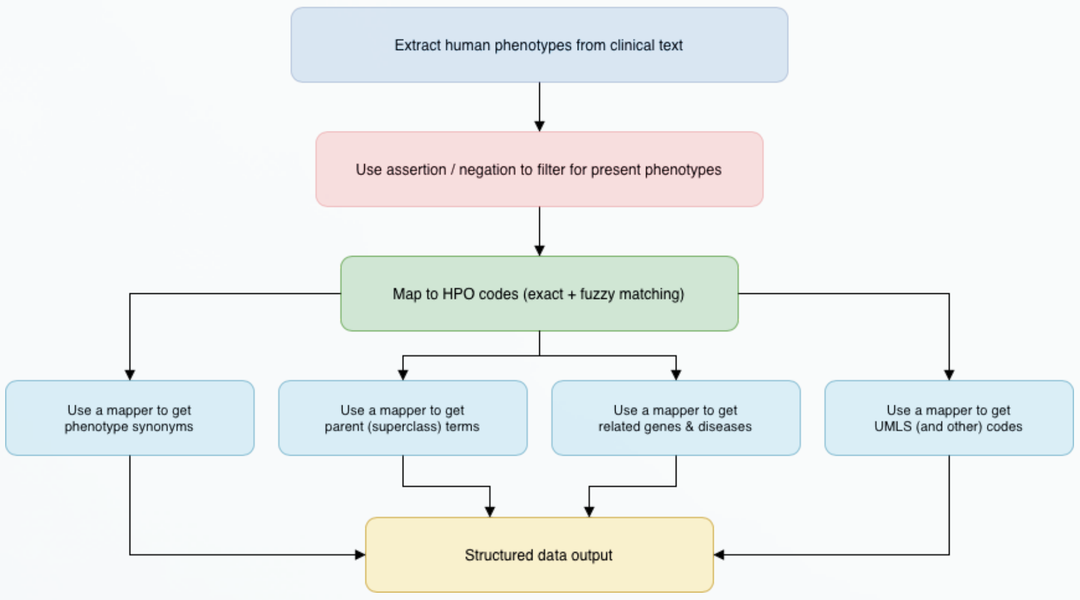


Objective

To develop and evaluate an ontology-aware clinical NLP pipeline that automatically extracts phenotype mentions from free-text notes, assigns assertion status, and normalizes mentions to HPO identifiers. We further aim to benchmark this pipeline against LLM-based and dictionary-based baselines in terms of coding accuracy and processing speed, and to assess its suitability for secure, high-throughput, real-world rare disease workflows.We approach rare disease diagnostics through an HPO lens, transforming unstructured clinical narratives into ontology-grounded phenotype profiles.

Ontology-Aware Pipeline Overview

This diagram summarizes our ontology-aware NLP pipeline for extracting and normalizing phenotypes from clinical narratives. The workflow has three stages: (1) text normalization and segmentation, (2) phenotype mention detection with assertion status classification, and (3) ontology-based concept normalization and enrichment. Detected mentions are mapped to HPO identifiers, linked to UMLS CUIs, and enriched with related genes, diseases, EOM associations, parent terms, and synonyms. The matching framework combines trie-based phrase matching with linguistic normalization and fuzzy ontology mapping to improve robustness against lexical variation while preserving clinical interpretability.



Evaluation Metrics for Phenotype Extraction

This figure defines the three binary coverage metrics used in our evaluation: HPO code coverage, chunk coverage, and synonym coverage. For each instance, a score of 1 is assigned when the gold-standard target is recovered by the system prediction (otherwise 0). Synonym coverage is computed using case-insensitive substring matching between any gold-standard synonym and the predicted text. System-level performance is reported as the mean coverage across all instances, expressed as percentages.

HPO code coverage

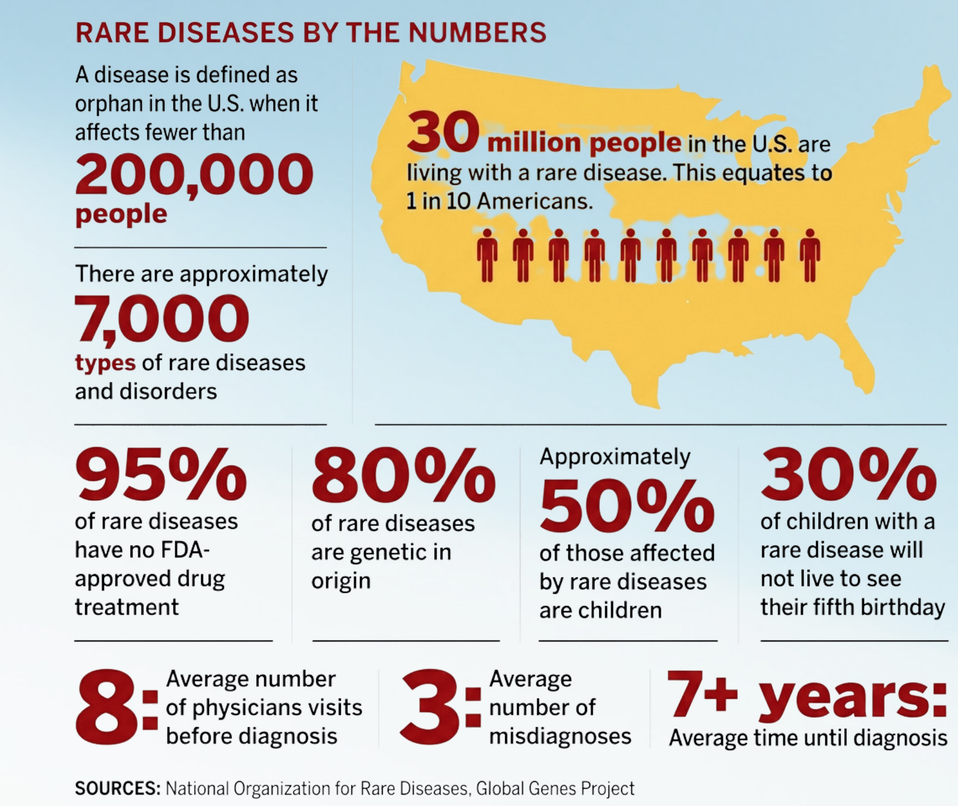
$$\text{Coverage}_{code} = \begin{cases} 1, & \text{if } GT\_\text{hpo\_code} \in \text{Predicted\_codes} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Chunk coverage

$$\text{Coverage}_{chunk} = \begin{cases} 1, & \text{if } GT\_\text{Chunk} \in \text{Predicted\_chunks} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

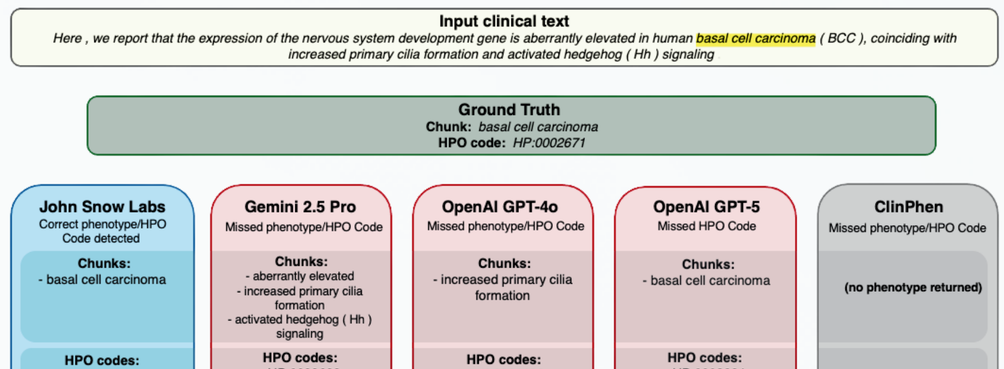
Synonym coverage

$$\text{Coverage}_{synonym} = \begin{cases} 1, & \exists\, s \in GT\_\text{synonyms} : \; s \subseteq \text{Predicted\_text} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$



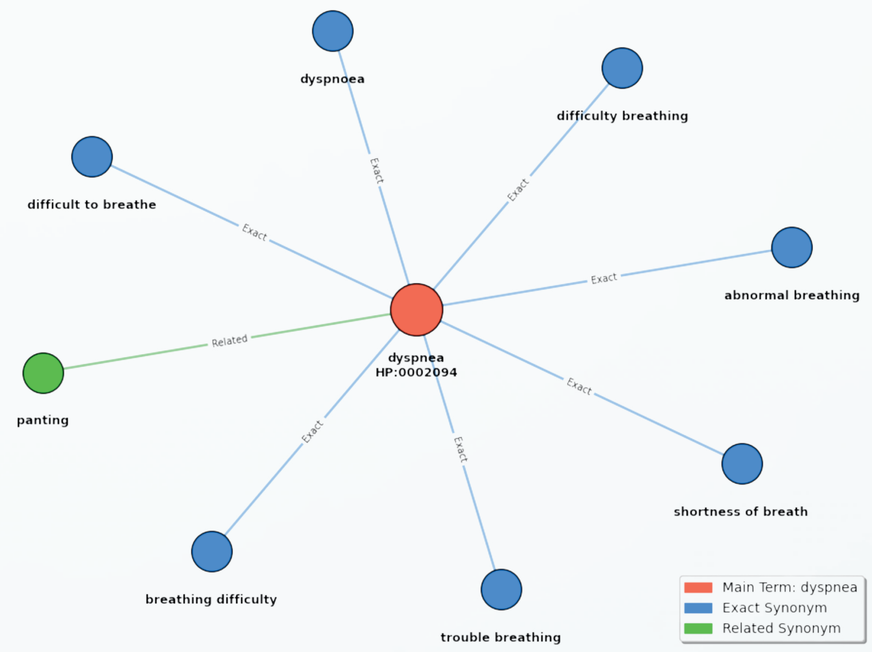
Qualitative Error Analysis

This figure shows a representative qualitative comparison of phenotype extraction and HPO normalization across systems. In the example, the clinical text contains a true disease phenotype mention (basal cell carcinoma) together with mechanistic biological descriptions. The ontology-grounded pipeline correctly identifies the disease mention and maps it to the appropriate HPO concept. In contrast, LLM-based systems exhibit common failure modes, including prioritizing biologically plausible surface phrases, assigning unrelated or overly broad HPO codes, and inconsistent grounding. The dictionary-based baseline fails to return a phenotype in this case. Overall, the comparison highlights that ontology-aware pipelines provide more reliable, clinically aligned phenotype grounding than general-purpose language models.



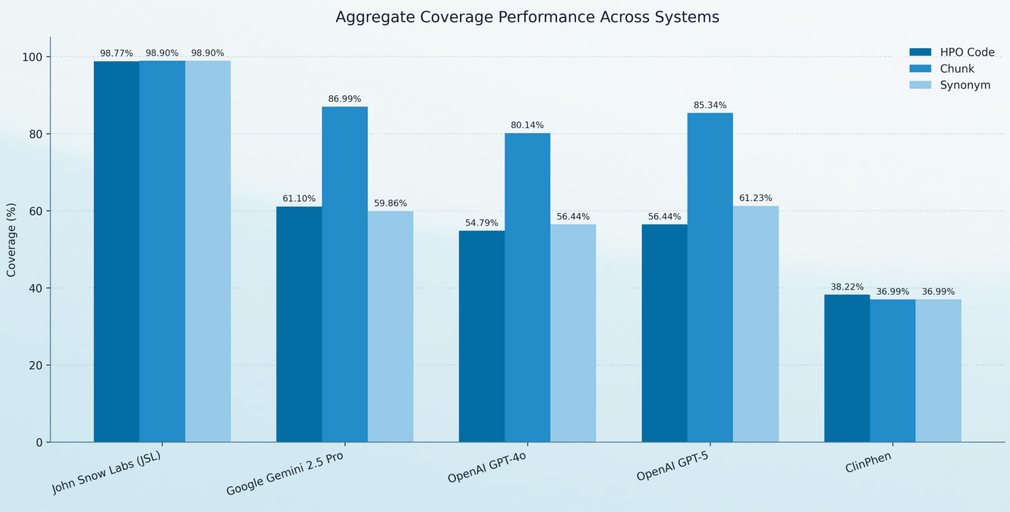
Example Synonym Network of an HPO Term

This figure presents an example synonym network for the HPO term "dyspnea." The central node represents the main clinical concept, while surrounding nodes show alternative expressions grouped by semantic relation: exact synonyms (blue), related synonyms (green). The network demonstrates how different patient- and clinician-used terms can be normalized to a single phenotype concept while preserving semantic granularity.



Results

Aggregate results show that the ontology-aware JSL pipeline clearly outperforms LLM-based systems and ClinPhen across coverage metrics, with the largest gap at ontology-level HPO code grounding.



Runtime comparison per 100 documents shows substantially higher throughput for the CPU-based ontology pipeline (JSL) than API-dependent LLM systems, with ClinPhen performance affected by non-batched single-document execution.