

PT II: EVALUATING THE NATURAL LANGUAGE PROCESSING ALGORITHM
ELASTEX FOR THE PHENOTYPE-GUIDED GENOMIC DIAGNOSIS PLATFORM,
GENOMEDIVER

Yalda Safaei, Sydney Maziarz and Patrick Shafer

May 2025

Submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Human Genetics
Sarah Lawrence College

ACKNOWLEDGEMENTS

We give thanks to our Sarah Lawrence thesis supervisor Laura Hercher, MS, CGC, and Dr. John Greally and Monisha Sebastin, MS, CGC at Montefiore for their supervision on our project. The Montefiore Einstein Center for Health Data Innovations (CHDI) team for randomizing and de-identifying our chart notes and to Dr. Parsa Mirhaji, Erin Henninger and team for performing the Elastex extraction. We also thank Boudewijn Aasman for performing our statistical analysis.

ABSTRACT

This study evaluates the performance of Elastex, a natural language processing (NLP) algorithm in extracting Human Phenotype Ontology (HPO) terms for integration into GenomeDiver, a phenotype-guided genomic diagnosis platform. Using a manually curated gold standard (GS) dataset derived from 100 neurodevelopmental patient records, we assessed Elastex's precision, recall, and overall effectiveness in extracting clinically relevant phenotypic information. The GS identified 538 unique HPO mapped terms, while Elastex extractions identified 527 with an average precision of 43% and a recall of 22%. Recall was notably low, highlighting Elastex's preference for specificity over breadth. Error analysis revealed common issues such as misinterpretation of negations and redundant extractions resulting in a pooled kappa value of -0.38 (agreement between GS vs. Elastex). Inter-annotator agreement was also measured using pooled kappas and was poor (-0.34 to -0.40). However, this supports that discordance between the GS and Elastex was not due to random error, but rather systematic differences in how extraction was done. This emphasizes not only the variability among different annotators, but also the inherent subjectivity in manual extraction processes even with the implementation of a standard set of guidelines. Despite its limitations, Elastex demonstrates improving potential as a scalable and reproducible tool for preliminary HPO extraction when used in conjunction with human oversight. Its integration into GenomeDiver could reduce diagnostic delays and streamline phenotypic curation by surfacing reliable candidate terms for expert validation. Future improvements to enhance context awareness and synonym recognition may further increase its clinical utility in rare disease diagnostics.

Keywords: natural language processing, human phenotype ontology, rare disease, phenotyping

TABLE OF CONTENTS

Acknowledgements	ii
Abstract	iii
Introduction	6
Methods	12
Results	20
Discussion	22
Conclusion	31
Bibliography	32

LIST OF FIGURES

1. Example of HPO mapping	16
2. Confusion matrix comparing phenotypic term extraction between the curated GS (actual) and Elastex (predicted).....	23

INTRODUCTION

Rare disease diagnosis remains one of the greatest challenges in this new era of genetic advancement. By definition, rare diseases (RDs) are conditions that affect a small portion of the population. While there is no exact consensus on a threshold, the American Orphan Drug Act defines RDs as conditions affecting fewer than 200,000 individuals in the United States. Collectively, RDs are in fact anything but rare, affecting 30 million people across the country with an estimated total economic burden of \$997 billion as of 2019 (Yang et al., 2022). The diagnostic odyssey for these individuals remains painfully long and expensive, with an average time-to-diagnosis of 4-5 years, and studies show that far longer odysseys are not infrequent (Yang et al., 2022). This places a significant financial and emotional toll on patients and families and, additionally, a significant burden on healthcare systems. Approximately 10% of the American population has experienced a struggle obtaining an accurate diagnosis, preventing them from receiving appropriate care. These statistics highlight an urgent need to recognize and address these challenges as a public health priority (Chung et al., 2022). On a global scale, initiatives such as the Rare Disease International Policy Framework, which fosters international collaboration in research and healthcare access, and the UN's *Resolution on Addressing the Challenges of Persons Living with a Rare Disease*, which advocates for integrated care models and national policy inclusion, reflect the growing recognition of this issue (Yoon et al., 2023). Efforts to improve the speed and efficacy of rare disease diagnostics are thus integral to this mission and it is in the best interest of not only the patient, but the entire healthcare team and the health systems in which they operate to investigate the pathways to facilitating diagnosis.

Currently, there are approximately 7,000 uniquely identifiable rare diseases. Of particular interest, roughly 80% of these rare diseases have a genetic basis (Marwaha et al., 2022).

Furthermore, the majority of these conditions manifest in childhood and their progressive and chronic nature often lead to significant morbidity, frequent hospitalizations, and complex, lifelong medical needs (Bogart et al., 2022). As a result, these diseases result in a poorer quality of life not just for patients, but also for caregivers who face emotional, financial, and social burdens (Delaye et al., 2022). Although the majority of these diseases have no cure, diagnosis can improve quality of life by providing accurate treatment, allowing for more informed decision-making and providing access to disease communities and researchers. Unfortunately, the simultaneous genetic heterogeneity, variable expressivity, and extensive phenotypic overlap among rare diseases makes diagnosis challenging (Diaz-Santiago et al., 2020).

In the field of rare disease research and diagnosis, the landscape of genetic testing and diagnosis of genetic disease has drastically changed with the increasing availability of Next Generation Sequencing (NGS). Whole exome and whole genome sequencing (WES/WGS) using NGS platforms have contributed to the discovery of new disease genes and disorders, and provided insight into the mechanism of already established diseases (Sullivan et al., 2023). Unexpected phenotype-genotype associations, and even phenotypes associated with specific variants have been discovered using WES/WGS data (Ross et al., 2020). As such, this type of testing is becoming increasingly commonplace both as a frontline test and as a reflex when no diagnosis can be made following traditional genetic testing methods (Wigby et al., 2024). In fact, a meta-analysis of 37 rare genetic disorders studies comparing WGS/WES to the standard chromosomal microarray analysis (CMA) showed that the former significantly improved diagnostic rates and clinical utility (Liu, Zhichao et al., 2019). However, while whole exome and genome sequencing have become a more mainstream clinical tool, challenges to the diagnosis of RDs remain.

Because they cast a wide net, WES/WGS return a large number of variants of uncertain significance (VUS) (Shashi et al., 2014) and often there are discordant interpretations of the variants identified (Sullivan et al., 2023). These methods often also generate multiple candidate genes that complicate the search for causative variants (Eilbeck et al., 2018). While VUS classification is typically improved through family studies, functional assays, and model organism research, deep phenotyping has also been shown to improve the efficiency of clinical WES analysis and increase diagnostic yield by way of gene prioritization (Wright et al., 2018; Son et al., 2018). To address this ongoing issue in the analysis of sequencing data, various machine learning models have been developed that target the information available in the electronic health record (EHR) (Roman-Naranjo et al., 2023). For example, GenomeDiver, a program developed at the Montefiore Einstein College of Medicine, presents clinicians with a targeted list of phenotypic features that would drive differentiation between candidate genes (Pearson et al., 2021). This program aims to streamline communication and improve diagnostic rates by increasing the quantity and applicability of phenotypic information provided to laboratories. This facilitates communication between the laboratory and the clinician and is a first step that could be made significantly more effective, efficient and scalable by the use of machine learning to automate the process of extracting potentially relevant data from the EHR (Roman-Naranjo et al., 2023).

Patient EHRs are a great resource to clinicians, however their characteristic non-standardized and often difficult-to-follow narrative format significantly limits the access to meaningful healthcare information that could impact a diagnosis (Iroju et al., 2015). EHRs typically have large volumes of data containing inconsistencies and extraneous or ambiguous details which can make pinpointing relevant phenotypic features challenging (Miandoab, et al.,

2023). Similarly, clinicians may ignore seemingly unrelated phenotypes, which might have provided relevant information for the purposes of ordering genetic tests and variant interpretation, and therefore informed and improved patient care (Vally et al., 2023). In addition, clinicians, frequently rushed and potentially inexperienced, may fail to extract the most relevant and clinically significant information. These challenges are mirrored in laboratory findings as well. Technicians may struggle to prioritize determination of key clinical features when interpreting test results (Iroju et al., 2015).

One effort to standardize descriptions of phenotypic information is the system of human phenotype ontology (HPO) terms, developed by Dr. Peter Robinson and the HPO consortium (Robinson et al., 2008). HPO is described as a universal standardized vocabulary of atypical phenotypes for over 7000 diseases (Gargano et al., 2023). The components of HPO terms are increasingly an industry standard for use in computational deep phenotyping for international rare disease organizations, clinical labs, and natural language processing (NLP) software tools, including those used by GenomeDiver (Köhler et al., 2017). GenomeDiver's fully automated usage of NLP aims to integrate deep phenotyping into its interface in order to optimize the use of EHR information to establish genotype-phenotype correlations. This would permit providers and laboratories to more quickly and accurately establish genes of interest and increase the likelihood of diagnosing RDs and the implementation of targeted care (Pearson et al., 2021).

NLP can thus be vital in enhancing the accessibility and integration of phenotypic information (Pendergrass and Crawford 2018). NLP is defined as a computational technique that can be used to analyze naturally occurring texts at multiple linguistic levels to achieve human-like processing capabilities for different applications (Iroju, O. G., & Olaleke, J. O., 2015). Fundamental NLP components include syntax, semantics, pragmatics, and a range of

other tasks such as text classification, information extraction, and sentiment analysis -- a list that showcases the complexity required of NLP programs in processing various language formats.

The utility of NLP in healthcare depends on its effectiveness in analyzing unstructured data from clinical notes, medical records, and biomedical literature. (Iroju, O. G., & Olaleke, J. O., 2015).

These tasks are fundamental for identifying key medical terms in clinical texts.

Additional NLP techniques in healthcare include Named Entity Recognition (NER) and Information Extraction (IE). These techniques are crucial for processing unstructured medical data, structuring relevant information, and improving data management. ClinPhen, an NLP developed by researchers at Stanford University, has demonstrated success in extracting HPO terms (Deisseroth et al., 2019). For example, in one study, investigators used patient records from an individual with Marfan syndrome to successfully extract phenotypes such as "aortic root dilatation," "ectopia lentis," and "tall stature" – all hallmark features of this condition. These terms were then sorted based on relevance, allowing the tool to suggest Marfan syndrome as a likely diagnosis (Deisseroth et al., 2019). This study demonstrated the NLP ClinPhen's ability to accurately extract and prioritize clinically relevant phenotypes.

While these NLP techniques may enhance processing speed, there are still many challenges that have been identified such as the need for continued refinement of these techniques to balance efficiency with accuracy in clinical applications (Iroju, O. G., & Olaleke, J. O., 2015). Limitations include variability in documentation practices, differences in clinical language, and incomplete records which can introduce error into the systems (Deisseroth et al., 2019). Additionally, there is the issue of interoperability in heterogeneous health information systems (HIS) which are often diverse in structure, function, and data standards (Miandoab, et al., 2023).

Interoperability refers to the ability of different systems, devices, or applications to work together seamlessly, exchanging information and using that information effectively (Paljoki et al., 2024). NLP tools rely on accessing consistent, structured data to extract meaningful phenotypic information. Lack of interoperability, particularly in EHRs, can hinder the effectiveness of NLP tools (Miandoab, et al., 2023). There are several key challenges to achieving interoperability in healthcare information systems. These challenges include technical issues such as incompatible data formats and standards, organizational barriers between labs, hospitals and corporations such as differing policies and workflows, and semantic discrepancies where the same data might be interpreted differently across systems (Paljoki et al., 2024). When data is fragmented across incompatible systems or stored in inconsistent formats, NLP tools may struggle to interpret the data correctly, leading to incomplete or inaccurate phenotypic extraction (Miandoab, et al., 2023). Improving interoperability is crucial for ensuring that NLP tools can access and accurately interpret comprehensive, high-quality data, which is essential for applications like Genome Diver that rely on precise phenotypic information for genomic analysis.

To mitigate this, NLP paired with human manual extraction can be used to refine the data from EHRs. Manual extraction can recognize and reconcile discrepancies in terminology to ensure extracted data adheres to standardized vocabulary (i.e. HPO), helping reduce the variability in EHR data that might disrupt NLP processes (Hier et al., 2024). A manual extractor can also interpret these ambiguous medical notes, nuanced languages, and implicit references that NLP may misinterpret or miss altogether. Using real patient EHR with RD, this method can be employed to create a “gold standard”, which is then used as a reference to compare the quality of the machine learning’s extraction tool. Since many gene variants can result in similar

phenotypic consequences, evaluating NLP extraction of patient chart terms can offer crucial information on the interplay of genes and disease manifestations (Robinson et al., 2008). However, manual extraction can be labor intensive, time consuming, and subject to bias. Additionally, the use of multiple extractors introduces variability (Wei & Denny, 2015). Since there is currently no proven NLP extraction method routinely used by clinicians and diagnostic laboratories, we investigated the clinical utility of the NLP “Elastex” created by the Montefiore Einstein Center for Health Data Innovation (CHDI) by comparing its performance in the extraction of HPO terms to our “gold standard” of human extraction in a sample of 100 de-identified neurodevelopmental notes from NYCKidSeq patient records (Odgis et al., 2021). The goal of this study was to refine this diagnostic tool and further explore its capabilities as an integral component of the GenomeDiver platform.

METHODS

This study is a continuation of the unpublished work *Evaluating Natural Language Processing Algorithms for the Phenotype-Guided Genomic Diagnosis Platform, GenomeDiver*, expanding on the foundation established by Rutter, Kan, and Rosales (2024). In this phase, we evaluate the performance of the refined Elastex algorithm against the gold standard, with modifications to the data collection process to enhance consistency and increase interrater agreement. These adjustments included the use of three extractors instead of two, with all extractors involved in the adjudication process, informed by insights from the initial study (Rutter et al., 2024) and ongoing internal assessments.

Data Collection

Note Identification

Notes for 100 NYCKidSeq participants seen at Montefiore were extracted from their medical records and de-identified for this study by the Montefiore Einstein Center for Health Data Innovation (CHDI). To reduce variability of phenotypic information across different chart notes for more accurate comparison, we looked solely at cases of neurodevelopmental disease. Participants had neurodevelopmental notes of varying types including outpatient notes (OP), inpatient notes (IP), physician letters, and history and physical exam notes (H&P). Only the most recent note was provided for each participant to avoid redundancy.

Creating the Gold Standard Dataset

Three principal investigators (YS, SM, PS) reviewed and annotated each note according to a predetermined set of rules as outlined below. These rules were iteratively refined by all three principal investigators over the course of the extraction process. The annotators each extracted terms from 2/3 of the notes so that each note always had two reviewers, and entered them into separate spreadsheets. Once completed, all investigators reviewed the terms extracted from the notes and identified any instances of disagreement, adjudicating as a whole in cases of disagreement to create the final “gold standard dataset” against which the terms extracted by the NLP systems would be compared.

General Term Extraction Rules

During the extraction, <https://hpo.jax.org/app/> was used to navigate HPO and terms were identified using the following protocol:

- All possible phenotypes which could be mapped to standard HPO terms were extracted.

- Ex: “neurological weakness” would not be included because it is not a standard phenotype term defined in the HPO.
- **Every** adjective/disease/condition found in each note was subject to a search in the HPO database for that term plus any potential alternate descriptor.
 - Ex: hypertonic muscles could be logged as hypertonia
- Only phenotypes that could be mapped to standard HPO concepts descending from “Phenotypic abnormality (HP:0000118)” were recorded in our database.
 - Ex: “Fetal narcotic exposure” could not be included because it belongs to the category “Past medical history (HP:0032443)”
- Of all corresponding terms found in the HPO, the most specific was recorded to strive for the finest possible level of granularity.
- Each HPO term was recorded only once per patient, even if mentioned multiple times.
- When searching for a term for the first time, the term was typed in exactly as it was written in the chart note.
 - Ex: “diet, multiple bladder diverticula with” - type in “multiple bladder diverticula” first
- If no corresponding HPO terms were found, extractors then looked for more general terms.
 - Ex: “bladder diverticuli”
- If the condition was described in language suggesting chronicity or more than one episode, extractors attempted to find a term that expressed a multiple or recurring phenotype.

- Ex: “seizure disorder” would be documented as “recurrent seizures” rather than “seizure”
- If the patient was described as having a bacterial infection, extractors documented the name of the bacteria and attempted to determine the outcome. If information on the outcome was not available in the notes, extractors did the following:
 - First, they would search for HPO terms associated with the name of the bacteria
 - Ex: “clostridium difficile colitis”
 - If no terms were available, they removed the name of the bacteria and searched only by condition.
 - Ex: “klebsiella pyelonephritis” should be documented under “pyelonephritis”
- Surgical procedures were not considered for extraction.
- Medication related side effects were not considered for extraction.
- Symptoms secondary to a feature were not considered for extraction.
 - Ex: In relation to a seizure episode “shaking, drooling, eye rolling” were not considered as separate terms if “seizure” was already extracted.
- Features found on imaging were searched for in the HPO database, and, if identified, those terms were added to the database.
- If no corresponding HPO term was found for a phenotypic term in the note, the term was discarded.

Description of NLP Systems

The NLP system evaluated in this study was Elastex. Elastex is a clinical text understanding and natural language processing platform, capable of collecting, indexing, and

processing more than 150 different types of clinical notes and reports (including imaging, pathology, microbiology, procedure, and surgery notes) developed by the Center for Health Data Innovations (CHDI) at Montefiore Medical Center/Albert Einstein School of Medicine.

CHDI ran Elastex to automatically extract HPO terms from the de-identified chart notes. Investigators compared the HPO terms from the gold standard dataset to the HPO terms returned by Elastex. Because of the relatedness of the terms in the HPO, the presence of an HPO term in the gold standard dataset implies the presence of all the ancestor phenotypes (Figure 1).

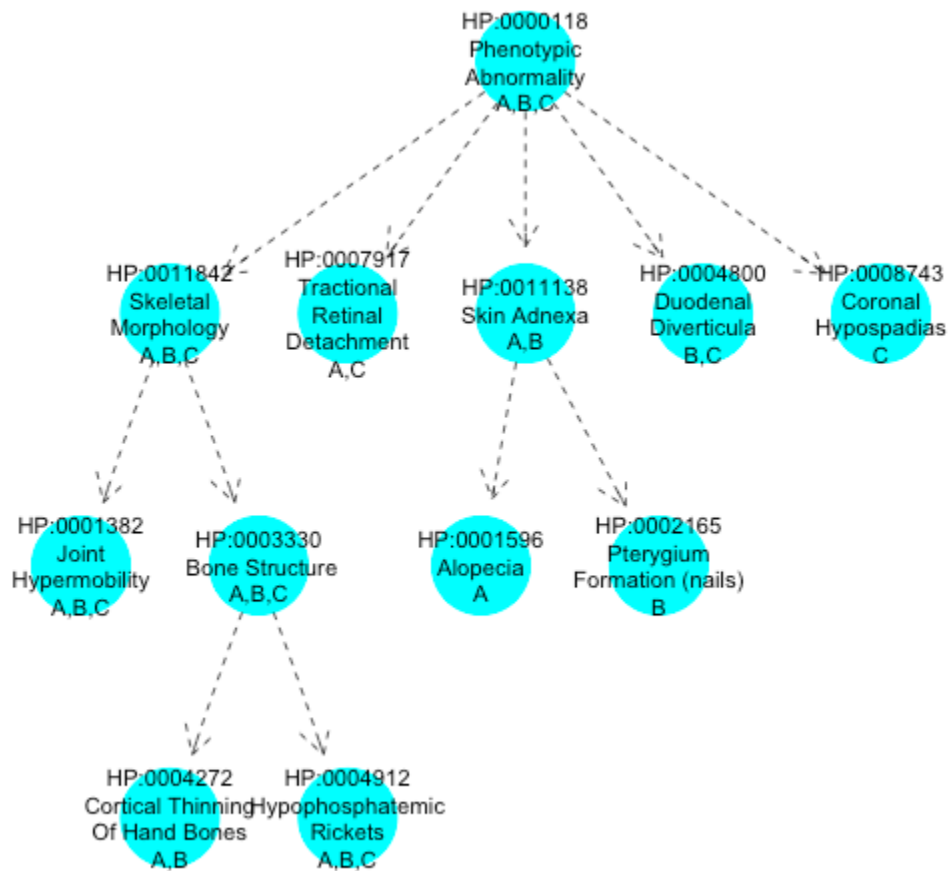


Figure 1. Example of HPO mapping

Comparison of Elastex to the Gold Standard Dataset

The gold standard dataset for each note was compared to the list of NLP-extracted terms to identify false positives, false negatives, and true positives (or concordance). True negatives were not assessable because the breadth of potential phenotypic terms was not possible to quantify.

There were two types of true positives (TP): exact matches and partial matches. False positives (FP) were defined as the terms only present in the set found by Elastex. For the purpose of assessing exact matches, terms that were found to be partial matches would be classified as a false positive. For assessments of all matches, partial matches would be considered true positives, leading to a smaller number of false positives. False negatives (FN) were defined as the terms only present in the gold standard dataset, and not found by Elastex.

Only taking into account <u>exact</u> matches		Gold Standard Dataset	
		Present	Absent
Elastex	Present	True Positive (TP) # terms (<u>exact HPO ID match</u>) in gold standard dataset found by Elastex	False Positive (FP) # of terms found by Elastex that were not in gold standard dataset
	Absent	False Negative (FN) # terms in gold standard dataset NOT found by Elastex	True Negative (TN) # terms NOT in gold standard data and # terms NOT found by Elastex

Table 1. Elastex evaluation classifications taking into account exact HPO term matches

Taking into account <u>partial</u> matches		Gold Standard Dataset	
		Present	Absent
Elastex	Present	True Positive (TP) # terms (<u>not an exact HPO ID match, but an ancestor</u>)	# of terms found by Elastex that were not in gold standard dataset

		<u>or daughter term</u>) in gold standard dataset found by Elastex	
	Absent	False Negative (FN) # terms in gold standard dataset NOT found by Elastex	True Negative (TN) # terms NOT in gold standard data and # terms NOT found by Elastex

Table 2. Elastex evaluation classifications taking into account exact HPO term matches

Example gold standard dataset	Terms found by Elastex	
cat	calico cat	<u>Exact matches:</u> TP = 2: banana, mansion FP = 3: calico cat, strawberry, boat FN = 2: cat, book
banana	banana	
mansion	mansion	<u>Partial matches:</u> TP = 3: calico cat, banana, mansion FP = 2: strawberry, boat FN = 1: book
book	strawberry	
	boat	

Table 3. Example of the differences between Elastex evaluation taking into account exact matches, versus partial matches

Recall	$TP / (TP + FN)$	Proportion of terms in the gold standard dataset found by Elastex
Precision	$TP / (TP + FP)$	Proportion of the terms extracted by Elastex that are in the gold standard dataset
F-measures	$(1 + \beta_2) * (\text{precision} * \text{recall}) / (\beta_2 * \text{precision} + \text{recall})$	Scores how well Elastex balances precision and recall

Table 4. Performance measure equations

Data Analysis

Inter-Annotator Agreement

Percent agreement for each HPO term extracted was computed and averaged across the pair. Cohen's Kappa was computed for each HPO term across all notes to assess interrater agreement corrected for chance agreement, which was then be pooled to quantify overall agreement (McHugh 2012). Using a pooled Kappa for sparsely used HPO terms was determined to be more accurate than averaging the Kappa from each use (De Vries, et. al. 2008). Pooled Kappa was determined using the following equation:

$$\text{Pooled Kappa: } \frac{P_o - P_e}{1 - P_e},$$

where $\underline{P_o}$ is the average observed agreement across all HPO terms, $\underline{P_e}$ is the average expected agreement across all HPO terms. When Kappa was undefined for any particular HPO term, it was excluded from the analysis.

Performance evaluation metrics

Recall (also known as sensitivity), precision (also known as positive predictive value), and f-measure were used to assess the performance of Elastex (Table 4). An NLP system with perfect precision would not identify any terms that were not in the gold standard dataset, while an NLP system with perfect recall would identify all the terms in the gold standard dataset.

For each note, the number of true positives, false negatives, and false positives were scored and totaled. The recall and precision were calculated from the totals. From the recall and precision, the f-measure was calculated.

Error Analysis

An error analysis was performed to categorize the nature of the mismatches between Elastex and the gold standard dataset. False positive and false negative mismatches from Elastex were examined and classified.

RESULTS

The pooled Kappa value between Elastex and the GS was -0.38 (Table 5). The pooled Kappa values between SM and YS, YS and PS, and PS and YS were -0.34, -0.39, and -0.40, respectively. Overall, there was 33% agreement among annotators on the exact terms extracted from the note. When the same term was extracted, it was mapped to the same HPO term and code 81% of the time.

Comparison	Pooled Kappa
Elastex vs. GS	-0.38
SM vs. YS	-0.34
YS vs. PS	-0.39
PS vs. SM	-0.40

Table 5. Pooled Kappa agreement

Elastex vs. GS

The GS identified 1,166 total HPO terms across all notes, including 538 unique terms. At least 1 HPO term was identified from each note, with an average of 11.98 HPO terms per note. Elastex identified 6,999 total HPO terms (including redundancies), of which 527 were unique terms, and produced an average of 32.33 HPO terms per note (Table 6). The total number of

HPO terms extracted by Elastex (6,999), the GS (1,166), or both was 8,165. The average number of identical terms extracted by Elastex and the GS per note was 16.4, with the remaining HPO terms being discordant.

Agreement was defined as the presence or absence of a particular HPO term in both the GS and Elastex data sets. The pooled Kappa value was -0.38, which indicates that the observed agreement between Elastex and the GS was lower than expected by chance, reflecting systematic disagreement between the automated extraction and the GS. The recall (sensitivity) and precision (PPV) were calculated to be 22% and 43%, respectively. This means that the GS contains 22% of the HPO ID's selected by Elastex, while 43% of terms included in the GS selected an HPO ID that was selected by Elastex as well. The F-measure was calculated to be 0.23.

	Total number of unique HPO terms identified	Total number of notes with no HPO term identified (coded with N/A)	Average number of HPO terms identified per note
Elastex	527	0	32.33
GS review	538	0	11.98

Table 6. Data comparison of NLP extractions between Elastex and the GS review.

A confusion matrix table was generated to visualize and summarize the performance of Elastex:

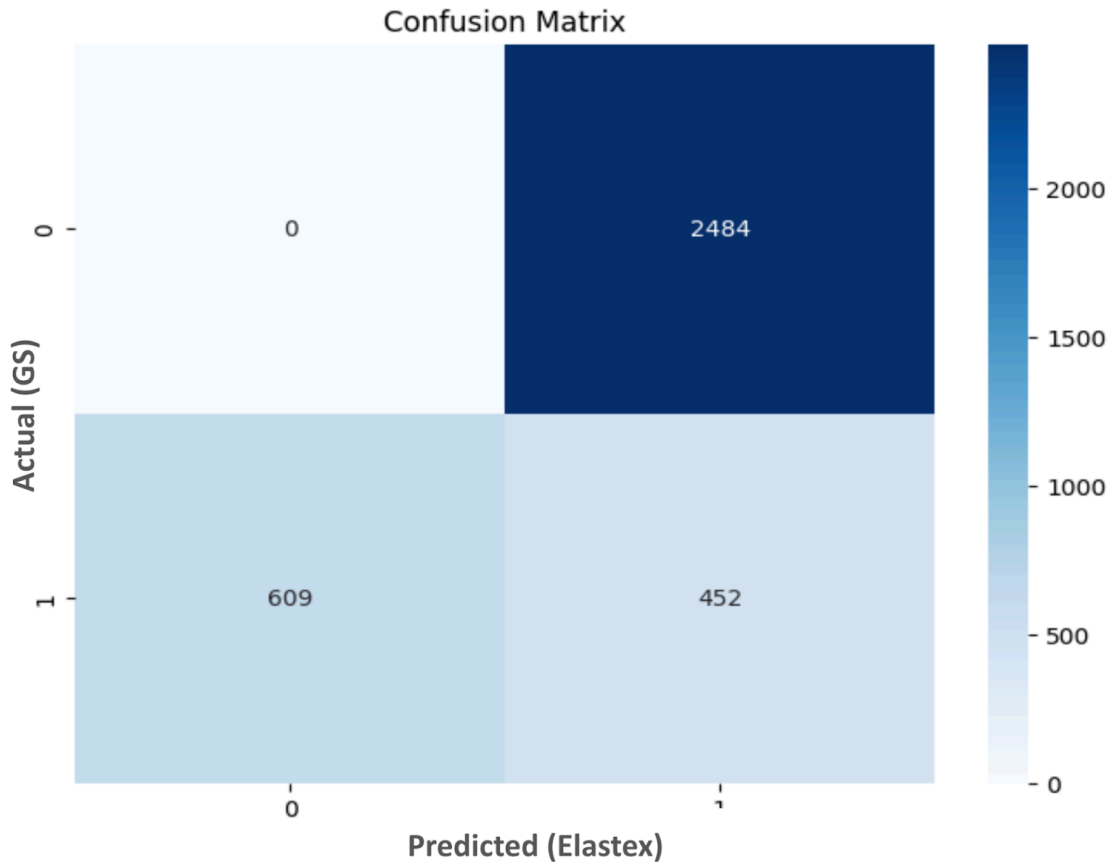


Figure 2. Confusion matrix comparing phenotypic term extraction between the manually curated GS (actual) and Elastex (predicted). By comparing the predicted values from Elastex to the actual gold standard (GS), we see 2484 false positives in the top right, 609 false negatives in the bottom left, 0 true negatives in the top left, and 452 true positives in the bottom right.

DISCUSSION

Key Findings

In this study, we aimed to assess the performance of the NLP system Elastex against a manually defined gold standard extracted from rare disease patient notes. By providing insight into the potential for scaling and automating GenomeDiver, we hope to increase its utility in rare disease diagnosis. The gold standard review identified 538 unique HPO mapped terms from 100

neurodevelopmental patient notes with an average of 11.98 terms per note. Elastex extractions identified 527 unique HPO terms across all notes, and an average of 32.33 HPO terms identified per note. This discrepancy indicates that while Elastex is capable of identifying a similar breadth of phenotypic concepts, it tends to over-extract terms per note, introducing redundancy and less clinically relevant information.

Precision was used to evaluate the accuracy of Elastex, defined as the proportion of terms identified by Elastex that were also found in the GS. When comparing Elastex to the GS, precision was found to be 0.43 (43%), meaning that less than half of the GS curated terms were captured by Elastex. This indicates that the recall of Elastex is limited in comparison to human review, suggesting that Elastex may rely heavily on syntactic patterns, missing context dependent terms or complex phenotypic descriptions that require clinical interpretation. As a result, Elastex may underperform where diagnostic accuracy relies on extracting subtle signs such as features that are not explicitly stated but implied through clinical context. However, given the low inter-annotator agreement, this value should be interpreted with caution. The terms included in the GS are likely incomplete and inconsistent, and thus cannot be used as a truth dataset for assessing precision.

The pooled kappa value of -0.38 indicates that agreement between Elastex and the GS was worse than what would be expected by chance, suggesting consistent mismatches in term extraction. While it does not necessarily undermine its potential, this finding highlights the importance of pairing NLP tools like Elastex with human review to ensure accurate clinical interpretation. It suggests that, in its current form, Elastex should be used as a support tool rather than a stand-alone solution.

False Positives and False Negatives

False positives occurred where the tool misunderstood abbreviations or misinterpreted contextual clues. For example, certain terms related to family history and vague abbreviations were extracted in contexts that did not warrant proposed phenotypic relevance. This suggests that while Elastex can extract explicitly stated terms (seizures or other neurodevelopmental traits), it can sometimes struggle with nuanced language that may require deeper semantic understanding.

We also observed that some HPO terms (609) were not captured (i.e. false negatives). This is likely due to Elastex's failure to recognize acronyms or uncommon synonyms as well as poor interpretation of context. This indicates that improved performance may depend on future modifications to increase synonym recognition and expansion of abbreviated terms. See Fig. 2 for a visual summary of these results.

Strengths and Weaknesses of Elastex

Elastex demonstrated consistency in its ability to extract phenotypic terms from the 100 clinical notes, with 452 terms identified by both Elastex and the GS. The tool reliably identified phenotypes that were judged to be clinically accurate and contextually appropriate in most cases. Elastex also applied uniform extraction, reducing variability in interpretation and maximizing replicability across data sets as compared to human extractors. This suggests that Elastex could serve as a reliable *foundation* for scalable, reproducible phenotyping with manual oversight. In clinical settings where time is limited and reproducibility matters, the ability to capture essential terms, even if imperfect, is a key strength over more subjective manual extraction.

However, Elastex did exhibit some vulnerabilities with regards to context misinterpretation and language processing. For example, terms like "Family History" were extracted when explicitly negated (i.e. "no family history of"), and therefore excluded from the GS. This is not only an issue of error, but of limited depth in language processing. Elastex can

recognize a term, but not always its true meaning in context, and it misses clinically relevant phenotypes consistently extracted by human annotators. This weakness is demonstrated by the -0.38 kappa value and low precision when compared to the GS (0.43). These errors can introduce unwanted noise into phenotypic profiles that is potentially misleading, especially if the aforementioned false positives are not manually reviewed. Furthermore, the F1 score of 0.226 reflects both low precision and poor recall. Together, these findings suggest that the tool may fall short with regards to *true* clinical understanding.

In addition, the repetitive extraction of identical terms within a single note highlights the fact that Elastex is currently unable to reduce the redundancy found in the medical record. Rather than interpreting repeated references as clinical emphasis or documentation habits, Elastex treats them as distinct data points. This demonstrates a surface level reading of text which might have the effect of increasing the apparent phenotypic complexity.

In situations where multiple HPO terms might apply (i.e. “expressive language delay vs. language impairment”), Elastex sometimes defaulted to broader terms. While technically not incorrect, this may reflect a tendency to generalize where specificity is possible. This can be a pitfall in genomic diagnosis where more specific terms can guide variant prioritization, thus limiting Elastex’s clinical value.

InterAnnotator Discrepancies and GS Construction

Among our three principal investigators, inter-annotator agreement was poor. To reiterate, a pooled kappa coefficient of 1 indicates perfect agreement, meaning that raters agree on every term, while a kappa of 0 suggests agreement no better than chance. Although the predetermined guidelines used for manual extraction as described in the methods were an iteration of Part I of this project, the level of agreement decreased significantly. In Part I, pooled Kappa between two

annotators was calculated to be 0.67 whereas the present study resulted in pooled Kappa values of -0.39, -0.34, and -0.40 for PS vs. YS, SM vs. YS, and PS vs. YS, respectively. Negative values indicate that although the same set of predetermined guidelines was used, each annotator was systematically inconsistent in applying these guidelines. Annotators extracted the same term from the note in 33% of cases, and when the same term was extracted, it was mapped to the same HPO term and code 81% of the time. The low term-level agreement is likely a result of ambiguous terms and numerous different terms describing the same feature within one note. Overall, these values suggest that agreement between annotators could be improved through further refinement of the guidelines as well as tailoring rules to the specific note type. As discussed, this project had only one note type and one note per patient as opposed to the previous group, which used a diversified data set and 4 notes per patient. We had hypothesized that minimizing variability in the dataset would lead to an increase in agreement and therefore higher Kappa values, however, this was not the case.

Many possible sources introduced variability in manual review and annotator interpretation. Disagreements between annotators were often the result of redundancy in the HPO database such as “poor speech” (HP:0002465) vs “incomprehensible speech” (HP:0002546) when searching “unclear speech.” As well, several accurate HPO descriptors may return for one query. For example, “infantile colic” was extracted as both “tearfulness” (HP:0033705) and “irritability” (HP:0000737), both of which describe the phenotype.

Sometimes the use of non-clinical language in notes, particularly when quoting the patient’s parents, resulted in differing levels of specificity. For example, “has never been able to make full sentences” was extracted as “language impairment” (HP:0002463) by one annotator and “expressive language delay” (HP:0002474) by the other. Though this distinction might

warrant specialized evaluation from speech pathology, the agreed upon term was “expressive language delay” as it more accurately reflected what was found in the note. This potentially highlights overinterpretation, resulting in extraction of HPO terms that may not accurately reflect the patient’s symptoms. Furthermore, the severity of certain symptoms when not made explicit also created disagreement, e.g. “sleep abnormality” (HP:0002360) vs “insomnia” (HP:0100785) for “lack of sleep.” This pattern emerged across many notes, another example being “visual impairment” (HP:0000505) for “wears glasses” while the other annotator did not extract this term at all.

In fact, the hierarchical structure of HPO terms was a common reason for disagreement among annotators. Although both annotators would match the extracted term to the lineage of the same parent term, one would often be a level above the other. As in an example we used earlier, “insomnia” is a descendant term of “sleep abnormality,” but HPO IDs do not capture their relatedness and thus functionally they are no different than any other two unrelated terms. This decreases the level of agreement between annotators, though context suggests that this is not the same as non-agreement. To address this issue, the previous group mentioned aggregating HPO terms to their parent term, however, this operates under the assumption that the more specific terms are reliably being extracted under the same hierarchy.

The variable extraction of terms related to seizure and epilepsy from phenotypic information, particularly EEG findings, illustrates another challenge that reduced concordance between annotators. Because we were extracting from neurodevelopmental notes, these were terms we saw frequently. However, as annotators, we lacked the expertise to parse out what was a relevant phenotype and so we were forced to resort to a “best guess.” For example, the text “frequent spikes and polyspikes in the posterior quadrant regions” was extracted as both “EEG

with generalized spikes” (HP:0012000) and “EEG with irregular generalized spike and wave complexes” (HP:0001326). Adjudication in these instances often resulted in selecting the umbrella HPO term “EEG abnormality” (HP:0002353) to avoid introducing error as we did not have the background knowledge to choose between the more granular terms. Inclusion and exclusion of allergies was also an area of frequent disagreement, as some annotators included all allergies, while others did not include them unless it seemed likely to yield a genetic diagnosis.

Though the interpretation of context and its use to inform clinically relevant information is one of the key advantages of the gold standard in many situations, the above variability may be perceived as a liability relative to the internal consistency of Elastex. Furthermore, the use of Elastex in conjunction with clinician expertise may alleviate some of the confusion introduced by manual extraction, particularly in cases where redundant terms are produced, or specialized knowledge is required. By providing a standardized, reproducible approach to extracting HPO terms, Elastex minimizes the risk of overinterpretation and ensures that term selection is based on algorithmic consistency rather than subjective judgment.

Clinical Utility and Application in GenomeDiver

Although Elastex’s precision was relatively low (43%), this characteristic may actually support integration into clinician-facing tools like GenomeDiver. GenomeDiver relies on accurate HPO terms to guide variant prioritization and facilitate dialogue between clinicians and diagnostic laboratories (Pearson et al., 2021). While Elastex is not currently suitable for fully automated use, this study showed that it can quickly identify a broad set of candidate phenotypes for clinician review without overwhelming them with irrelevant terms, supporting a more streamlined manual review process. This allows for more efficient downstream curation and fits

GenomeDiver's interactive setting where users are expected to confirm or reject candidate phenotypes.

Although Elastex did not identify more unique HPO terms when compared to the GS, the total number of HPO terms generated was higher due to redundancy rather than overgeneration of unrelated information. Within GenomeDiver's structure, this redundancy may be advantageous as surfacing a relevant but less specific term allows clinicians to adjust specificity as appropriate, rather than forcing early commitments to more detailed descriptors which may not be fully supported by context. Thus, while precision remains an area for improvement, emphasizing sensitivity over specificity may be acceptable for GenomeDiver's intended clinical workflow

Additionally, while redundancy in term extraction is seemingly inefficient in a research setting, it could have practical utility in clinical tools such as GenomeDiver. Multiple mentions of the same term could serve as a signal of clinical relevance highlighting features that warrant closer attention (Borchert et al., 2024). If term weighting was processed afterwards or filtered at the interface level, repetition could prove as a way to prioritize certain phenotypes and indicate which ones are more persistent and draw the most clinical attention.

The negative pooled kappa value of -0.38 demonstrates systematic disagreement between Elastex and the GS, which further reinforces the need for human oversight in interpretation. Despite Elastex's inability to understand complex negations and linguistic processing, GenomeDiver's design allows the clinician to be the final arbiter of phenotypic accuracy. Because of this, Elastex may be the foundation for collaborative interpretation rather than an endpoint. Therefore, Elastex's ability to extract standardized phenotypic terms and

GenomeDiver's inclusion of clinical judgement allows for a balance that may ultimately reduce diagnostic delays and improve the effectiveness of rare disease workflows.

Limitations

While this study focused on assessing the optimization of Elastex, some limitations remain. A key challenge is that, despite improvements, NLP still struggles to fully replicate the nuanced decision-making of human extractors. The advantage of the gold standard, built through manual extraction, is its ability to interpret clinical context—considering implicit relationships between terms, prioritizing relevant findings, and filtering out extraneous details. This explains why precision dropped from 0.9 to 0.43 when compared to the GS suggesting that while terms extracted by Elastex are typically correct, it fails to identify a large portion of clinically relevant phenotypes such as those that are context dependent or inconsistently documented.

However, the GS is not without its own limitations, as subjectivity in manual extraction, dependency on medical knowledge, and variability in clinician judgment can introduce inconsistencies as shown through poor inter-annotator agreement. Additionally, the lack of exact term matching between NLP-extracted and manually curated terms complicates direct comparisons, as minor differences in phrasing can still carry clinical significance. Time constraints also limited a deeper investigation into false positives and false negatives, preventing a more granular assessment of areas for further refinement.

Future Directions

Future directions should focus on improving NLP precision through enhanced context-aware processing, more sophisticated filtering mechanisms, and deeper integration of semantic relationships between terms (Iroju et al., 2015). One promising approach is refining how NLP models evaluate term relevance in clinical settings—prioritizing phenotypes based on

their likelihood of being genetically informative, much like our gold standard approach (Kim et al., 2023). Integrating these enhancements into GenomeDiver could streamline diagnostic workflows by ensuring that only the most clinically meaningful terms are surfaced for interpretation. Further evaluation of NLP-extracted terms in real-world clinical applications will also be essential to fine-tuning performance and ensuring that NLP tools contribute effectively to rare disease diagnosis.

Conclusion

Despite low precision (0.43) due to the aforementioned limitations and poor agreement with manual extraction (-0.38), Elastex demonstrated potential as an NLP tool for initial HPO term extraction. Our findings suggest that both manual and algorithmic curation have shortcomings, albeit potentially compatible ones. Elastex, when paired with clinician review, may effectively replace manual extraction methods similar to part I of this project. Integrating NLP tools like Elastex into clinical software, such as GenomeDiver, could enhance phenotyping, streamline diagnostic workflows by informing genetic testing, and improve communication between laboratories and clinicians. Ultimately, these advancements may help shorten diagnostic journeys for patients, reduce the burden of rare disease diagnosis, and improve patient outcomes. Further refinement and validation of Elastex within real-world clinical workflows will be essential to maximizing its utility in genomic medicine.

BIBLIOGRAPHY

Bello, S. M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D. F., Botas, P., Boztug, K., Čady, J., ... Robinson, P. N. (2024). The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic acids research*, 52(D1), D1333–D1346.

<https://doi.org/10.1093/nar/gkad1005>

Bogart, K., Hemmesch, A., Barnes, E. et al. Healthcare access, satisfaction, and health-related quality of life among children and adults with rare diseases. *Orphanet J Rare Dis* 17, 196 (2022).

<https://doi.org/10.1186/s13023-022-02343-4>

Borchert, F., Llorca, I., & Schapranow, M.-P. (2024). Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. *Database: The Journal of Biological Databases and Curation*, 2024, baae067. <https://doi.org/10.1093/database/baae067>

Chung, C. C., Chu, A. T., & Chung, B. H. (2022). Rare disease emerging as a global public health priority. *Frontiers in Public Health*, 10.

Deisseroth, C. A., Birgmeier, J., Bodle, E. E., Kohler, J. N., Matalon, D. R., Nazarenko, Y., Genetti, C. A., Brownstein, C. A., Schmitz-Abe, K., Schoch, K., Cope, H., Signer, R., Undiagnosed Diseases Network, Martinez-Agosto, J. A., Shashi, V., Beggs, A. H., Wheeler, M. T., Bernstein, J. A., & Bejerano, G. (2019). ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genetics in medicine : official journal of the American College of Medical Genetics*, 21(7), 1585–1593.

<https://doi.org/10.1038/s41436-018-0381-1>

Delaye J, Cacciatore P and Kole A (2022) Valuing the “Burden” and Impact of Rare Diseases: A Scoping Review. *Front. Pharmacol.* 13:914338. doi: 10.3389/fphar.2022.914338

De Vries, H., Elliott, M. N., Kanouse, D. E., & Teleki, S. S. (2008). Using pooled kappa to summarize interrater agreement across many items. *Field Methods*, 20(3), 272–282.

<https://doi.org/10.1177/1525822X08317166>

Díaz-Santiago, E., Jabato, F. M., Rojano, E., Seoane, P., Pazos, F., Perkins, J. R., & Ranea, J. A. G. (2020). Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. *PLoS genetics*, 16(10), e1009054.

<https://doi.org/10.1371/journal.pgen.1009054>

Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. *Nature reviews. Genetics*, 18(10), 599–612.

<https://doi.org/10.1038/nrg.2017.52>

Gargano, M. A., Matentzoglou, N., Coleman, B., Addo-Lartey, E. B., Anagnostopoulos, A. V., Anderton, J., Avillach, P., Bagley, A. M., Bakštein, E., Balhoff, J. P., Baynam, G.,

Hier, D. B., Carrithers, M. D., Do, T. S., & Obafemi-Ajayi, T. (2024). Efficient Standardization of Clinical Notes using Large Language Models. arXiv preprint arXiv:2501.00644.

Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science*, 7(8), 44–50.

Kim J, Yang J, Wang K, Weng C, Liu C. Assessing the utility of large language models for phenotype-driven gene prioritization in rare genetic disorder diagnosis. *The American Journal of Human Genetics*. 2024;115(4):789–802. doi:10.1016/j.ajhg.2024.03.012

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gouridine, J. P., Gargano, M., Harris, N. L., Matentzoglou, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., ... Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research*, 47(D1), D1018–D1027. <https://doi.org/10.1093/nar/gky1105>

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., Brudno, M., Buske, O. J., Chinnery, P. F., Cipriani, V., Connell, L. E., Dawkins, H. J., DeMare, L. E., Devereau, A. D., de Vries, B. B., Firth, H. V., ... Robinson, P. N. (2017). The Human Phenotype Ontology in 2017. *Nucleic acids research*, 45(D1), D865–D876. <https://doi.org/10.1093/nar/gkw1039>

Liu, Z., et al. (2019). Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: Where are we? *Trends in Genetics*, 35(11), 852–867. <https://doi.org/10.1016/j.tig.2019.08.006>

Marwaha, S., Knowles, J. W., & Ashley, E. A. (2022). A guide for the diagnosis of rare and undiagnosed disease: Beyond the exome. *Genome Medicine*, 14(1).

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276–282.

Odgis, J. A., Gallagher, K. M., Suckiel, S. A., Donohue, K. E., Ramos, M. A., Kelly, N. R., Bertier, G., Blackburn, C., Brown, K., Fielding, L., Lopez, J., Aguiniga, K. L., Maria, E., Rodriguez, J. E., Sebastin, M., Teitelman, N., Watnick, D., Yelton, N. M., Abhyankar, A., Abul-Husn, N. S., ... Kenny, E. E. (2021). The NYCKidSeq project: study protocol for a

randomized controlled trial incorporating genomics into the clinical care of diverse New York City children. *Trials*, 22(1), 56. <https://doi.org/10.1186/s13063-020-04953-4>

Palojoki S, Lehtonen L, Vuokko R. Semantic Interoperability of Electronic Health Records: Systematic Review of Alternative Approaches for Enhancing Patient Information Availability. *JMIR Med Inform*. 2024 Apr 25;12:e53535. doi: 10.2196/53535. PMID: 38686541; PMCID: PMC11066539.

Pearson, N. M., Stolte, C., Shi, K., Beren, F., Abul-Husn, N. S., Bertier, G., Brown, K., Diaz, G. A., Odgis, J. A., Suckiel, S. A., Horowitz, C. R., Wasserstein, M., Gelb, B. D., Kenny, E. E., Gagnon, C., Jobanputra, V., Bloom, T., & Grealley, J. M. (2021). GenomeDiver: A platform for phenotype-guided medical genomic diagnosis. *Genetics in Medicine*, 23(10). <https://doi.org/10.1038/s41436-021-01219-5>

Pendergrass, S. A., & Crawford, D. C. (2019). Using Electronic Health Records To Generate Phenotypes For Research. *Current protocols in human genetics*, 100(1), e80. <https://doi.org/10.1002/cphg.80>

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5), 610–615. <https://doi.org/10.1016/j.ajhg.2008.09.017>

Roman-Naranjo, P., Parra-Perez, A. M., & Lopez-Escamez, J. A. (2023). A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases. *Journal of Biomedical Informatics*, 143. <https://doi.org/10.1016/j.jbi.2023.104429>

Ross, J. P., Dion, P. A., & Rouleau, G. A. (2020). Exome sequencing in genetic disease: recent advances and considerations. *F1000Research*, 9, F1000 Faculty Rev-336. <https://doi.org/10.12688/f1000research.19444.1>

Shashi, V., McConkie-Rosell, A., Rosell, B., Schoch, K., Vellore, K., McDonald, M., Jiang, Y.-H., Xie, P., Need, A., & Goldstein, D. B. (2014). The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 16(2), 176–182. <https://doi.org/10.1038/gim.2013.99>

Smedley, D., & Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Medicine*, 7, 81. <https://doi.org/10.1186/s13073-015-0199-2>

Son, J. H., Xie, G., Yuan, C., Ena, L., Li, Z., Goldstein, A., Huang, L., Wang, L., Shen, F., Liu, H., Mehl, K., Groopman, E. E., Marasa, M., Kiryluk, K., Gharavi, A. G., Chung, W. K., Hripcsak, G., Friedman, C., Weng, C., & Wang, K. (2018). Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *The American Journal of Human Genetics*, 103(1), 58–73. <https://doi.org/10.1016/j.ajhg.2018.05.010>

Sullivan, J. A., Schoch, K., Spillmann, R. C., & Shashi, V. (2023). Exome/Genome Sequencing in Undiagnosed Syndromes. *Annual Review of Medicine*, 74, 489–502. <https://doi.org/10.1146/annurev-med-042921-110721>

Torab-Miandoab, A., Samad-Soltani, T., Jodati, A., & Rezaei-Hachesu, P. (2023). Interoperability of heterogeneous health information systems: a systematic literature review. *BMC medical informatics and decision making*, 23(1), 18. <https://doi.org/10.1186/s12911-023-02115-5>

Vally, Z. I., Khammissa, R. A. G., Feller, G., Ballyram, R., Beetge, M., & Feller, L. (2023). Errors in clinical diagnosis: a narrative review. *The Journal of international medical research*, 51(8), 3000605231162798.

Wigby, K.M., Brockman, D., Costain, G. et al. Evidence review and considerations for use of first line genome sequencing to diagnose rare genetic disorders. *npj Genom. Med.* 9, 15 (2024). <https://doi.org/10.1038/s41525-024-00396-x>

Wright, M., Menon, V., Taylor, L., Shashidharan, M., Westercamp, T., & Ternent, C. A. (2018). Factors predicting reclassification of variants of unknown significance. *American Journal of surgery*, 216(6), 1148–1154. <https://doi.org/10.1016/j.amjsurg.2018.08.008>

Yang, G., Cintina, I., Pariser, A. et al. The national economic burden of rare disease in the United States in 2019. *Orphanet J Rare Dis* 17, 163 (2022). <https://doi.org/10.1186/s13023-022-02299-5>

Yoon, S., Lee, M., Jung, H. et al. Prioritization of research engaged with rare disease stakeholders: a systematic review and thematic analysis. *Orphanet J Rare Dis* 18, 363 (2023). <https://doi.org/10.1186/s13023-023-02892-2>