Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Master in Innovación Digital, especialidad en Health and Medical Data Analytics

Master Thesis

Developing Ontology-Driven Knowledge Graphs to Link Genomic Variants and Clinical Phenotypes in the Prenatal Context

Author: Javier de Castro Poy

Madrid, September, 2025

This Master Thesis has been deposited in ETSI Informáticos de la Universidad Politécnica de Madrid.

Master Thesis

Master in Innovación Digital, especialidad en Health and Medical Data Analytics

Title: Developing Ontology-Driven Knowledge Graphs to Link Genomic Variants and Clinical Phenotypes in the Prenatal Context«Mes Año» /«Month Year»

Author: Javier de Castro Poy

Tutor
Supervisor:

Sergio Paraíso Medina
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

Co-Tutor
Co-supervisor:

Francisco Javier Fernández Martínez
Hospital 12 de Octubre
Grupo Genética y Herencia.

Abstract

The objective of this Master Thesis is to integrate data on genetic variants and determinations obtained from CytoGenomics tool reports (in tabular format) and NGS techniques with clinical reports. Automated extraction of phenotypic concepts (HPO) from clinical texts allows for the establishment of a correlation between genetic results and the observed clinical manifestations. This approach is based on the use of recognized standards and ontologies such as the MonDo ontology (Monarch Disease Ontology), HPO, OMIM, ORPHANET, ICD, and SNOMED, thus ensuring the interoperability of the integrated data.

The integration of these heterogeneous data is addressed through natural language processing techniques (developed in previous work with the group) and data analysis methods, which facilitate the extraction and normalization of relevant information. The developed methodology allows for detecting and categorizing variations in genetic reports and their relationship with phenotypes, paving the way for identifying patterns of gain and loss in genetic determinations.

The subsequent analysis focuses on evaluating the associations between genetic variations and clinical phenotypes, providing a comprehensive combination of the genetic dimension and clinical manifestations. This study aims to establish a framework for future research in personalized medicine and digital innovation applied to health data analysis.

Furthermore, it seeks to enhance the understanding of the genetic foundations of various pathologies, offering a tool that can be implemented in clinical systems to optimize patient diagnosis and treatment.

Table of Contents

Index of tables

Index of figures

1 Introduction

1.1 Context and motivation

In last decades, genetic prenatal diagnosis has experienced a deep transformation due to the advancement of technologies such as massive sequencing or high-resolution ultrasound techniques. However, despite this progress, the systematic integration between phenotypic data obtained during pregnancy (especially in echography) and genomic information is still limited and fragmented.

Meanwhile in postnatal context there are consolidated databases which allow genetic variants relating with clinical phenotypes, in the prenatal field this correlation is still incipient. The lack of standardization when describing fetal phenotyping, along with its poor representation in biomedical ontologies like HPO (Human Phenotype Ontology), makes the precise interpretation of genetic variants only detected during pregnancy difficult.

This gap is especially relevant if it is considered that fetal phenotype not only differs with postnatal, but also dynamically evolves during pregnancy. The lack of structured repositories which collect this kind of information prevents an adequate characterization of genetic diseases from the first manifestations, limiting the diagnosis capacity and clinical decision-making during critical development stages.

In this context, the necessity of building a specialized database able to relate prenatal phenotype with its corresponding genetic variants in a structured manner emerges. This initiative not only tries to improve the diagnosis precision of high-risk pregnancies, but also promotes the standardization of clinical data, biomedical research and laying the foundations for precision medicine from the beginning of life.

1.2 Problem statement

Despite the growing use of genomic tools in prenatal diagnosis, clinical interpretation of detected variants is still a big challenge. This is largely due to the lack of structured correlation between the phenotypic findings observed by fetal ultrasound and the genetic alterations identified. Most part of existing databases are oriented to postnatal context, limiting their application in early stages of human development.

The lack of a database which relates in a systematic manner observed phenotypes during pregnancy with genetic variants represents a critical barrier to prenatal precision medicine. The lack of this infrastructure makes clinicians face uncertainty when diagnosing, affecting clinical decision-making and pregnancy management.

1.3 Objectives

The main goal of this project is to develop a structured database which integrates in a systematically manner observed prenatal phenotype through ultrasound with associated genetic variants, using standardized biomedical ontologies and interoperable formats. This tool tries to improve diagnosis capability in prenatal context, facilitating clinical interpretation of genetic variants and promoting precision medicine in initial stages of human development.

For achieve that general purpose, the following specific objectives are presented:

- Standardized fetal phenotype description through the codification of ultrasound findings in HPO ontology terms.

- Automize clinical information extraction from Spanish medical reports, using large language models (LLMs) and name entity recognition (NER) techniques for identifying and linking relevant phenotypes.

- Integrate genomic data coming from techniques such as NGS and CNV-seq.

- Building computable objects following the Phenopacket schema from GA4GH, which allows representing in a structured and reused manner clinical and genetic information of each case.

- Design a graph-based database, using technologies such as Neo4j, allowing the visualization and exploration of relationships between patients, phenotypes, genes, variants, and diseases.

- Validate the clinical utility of the system, analyzing real cases and evaluating the model's capabilities for identifying the phenotype-genotype correlations relevant in the prenatal context.

1.4 Hypothesis

The hypothesis proposed is that the systematic integration of prenatal phenotype (described through the use of standardized terms from biomedical ontologies) with genomic data will significantly improve the interpretation of genetic variants detected during pregnancy.

In particular:

1. The fetal phenotype standardization, using resources like HPO expanded with specific fetal terminology, will facilitate the interoperability of heterogeneous clinical data and will reduce the ambiguity in the description of ultrasound findings.

2. The building of a specific database for prenatal context, which links phenotypes coming from ultrasound reports with genes and variants, will allow the identification of more precise and reusable correlations, contributing to the reduction of variants with uncertain significant and improving clinical utility of high-performance sequencing technologies.

3. The application of computational natural language processing tools and automatized learning will allow the automatized extraction of phenotypical information from clinical reports and the linking with genomic resources, thereby accelerating the diagnostic process and advancing research in early-onset rare diseases.

2 State of the art and tools

2.1 State of the art

This section presents an overview of the scientific and technical background relevant to the development of a prenatal phenotype–genotype database. The field of prenatal genetic diagnosis has evolved significantly over the past decades, driven by advances in imaging techniques, genomic technologies, and biomedical informatics. Despite these developments, the integration of phenotypic observations (particularly those identified during prenatal ultrasound examinations) with corresponding genotypic data remains limited and fragmented. The aim is to provide a comprehensive overview of the current landscape, highlight existing gaps, and establish the context in which this work is positioned.

2.1.1 Prenatal diagnosis

Prenatal diagnosis is performed by integrating information from both the phenotype and the genotype. In postnatal settings, the deep phenotyping of a patient and its relationship with their genotype has been widely implemented [1] [2]. This practice is supported by the vast amount of data compiled in publicly accessible phenotype–genotype databases for pediatric and adult populations. This wealth of curated knowledge enables more accurate diagnoses, facilitates variant interpretation, and supports evidence-based clinical decisions [3].

In contrast, the application of those methods in the prenatal context remains limited. Most of the fetal phenotype in genetic diseases is not fully understood, which is crucial when interpreting genetic variants [4]. The lack of standardized and comprehensive datasets hinders the capacity to establish precise correlations, especially for subtle or unique prenatal features that may differ from their postnatal presentation.

The establishment of an accurate diagnosis can be crucial, as it directly informs critical clinical decision-making, including the management of pregnancy progression, delivery planning, maternal healthcare, neonatal interventions, and strategies to mitigate recurrence risk in future pregnancies [5].

2.1.1.1 Phenotype

Phenotype acquisition in the prenatal stage is vital for the correct diagnosis of a genetical disease, for several reasons. Firstly, some phenotypes can be attributed to certain diseases, which allow discarding other disorders. Furthermore, it orientates which diagnostic test can be more suitable for each case and it helps to correctly interpret their results.

Phenotype acquisition is based on medical imaging. The main technique used is ultrasound examination. The improvements in this technology during the past decades have permitted the obtention of the morphological characteristics of the

fetus with a high precision. This examination is often done during the second trimester (18-22 weeks), when major structures can be identifiable. However, first term scans are becoming more common, as it permits the early identification of fetus abnormalities. 3D/4D ultrasound is increasingly being used in the clinical practice, but their utility has not been establish yet, so most of the guidelines still recommend the use of 2D image [4] [6].

In addition to ultrasound, there are other complementary techniques that can also be employed. Magnetic resonance imaging (MRI) is especially useful to detect anomalies in the central nervous system and in other organs systems. Studies about the utility in postmortem fetal of MRI and computerized tomography scan (CT-Scan) modalities indicate their possible utility. Nevertheless, those modalities present several challenges, as they can be difficult to access and only few centers have the requirement equipment [4].

Prenatal phenotyping presents particular features that differ with pediatric and adult cases:

- Evolving phenotype. As the fetus develops and different anatomical structures emerge, the abnormal phenotypic characteristics naturally evolve as well. Furthermore, some characteristics can emerge after delivery. Monitoring the fetal phenotype throughout pregnancy, and re-evaluating it postnatally, is essential when interpreting prenatal exome sequencing findings [5].

- Functional characteristics. Ultrasound imaging is only useful for describing morphological structures. Nonetheless, it is unable of detecting behavioral and functional phenotypes. For example, in the Joubert syndrome, symptoms such as ataxia or intellectual disability are used for its diagnosis. However, these symptoms cannot be detected in the prenatal phase, so the diagnosis may be limited [7].

- Specific phenotype. Some characteristics can be only associated to antenatal stages, since they can lead to embryonic o fetal death and not be adequately studied [6].

- Limited fetal reports. There are many diseases for which prenatal phenotypes are neither well understood nor registered. This can lead to a biased interpretation at the gene or variant level. That is the reason why a greater emphasis should be placed on capturing and effectively sharing detailed fetal phenotypic information [3].

2.1.1.2 Genotype

Prenatal genetic diagnosis relies on the highly precise characterization of the fetal phenotype and on adequate tests, such as amniocentesis or chorionic villus sampling (CVS). It usually begins with chromosomal microarray (CMA) and/or karyotype to capture aneuploidies and copy number variants (CNVs). Another test that is used for the same purpose is cell free DNA screening.

CNV is a genome structural alteration that implies the duplication (gain) or elimination (loss) of AND segments. Those changes can really affect an important gene function, making them vital for genetic diagnosis. A correct CNVs study implies

their classification. Following the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (Clin-Gen) [8]. CNV classification is divided into pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign and benign. This classification is based on several criteria, including the number of genes affected, their known associations with diseases, comparison with established disease–phenotype correlations, population frequency, whether the variant is de novo or inherited, and whether it has been previously reported. CNV sequencing has been reported to be precisely diagnostic and has special impact on clinical decisions, such as continuing or stopping the pregnancy [9].

Another technology that is being used increasingly is Next-Generation Sequencing that allows analyzing millions of DNA fragments simultaneously. It provides whole-genome sequencing (WGS), and whole-exome sequencing (WES). Those are used for phenotype-driven strategies, which consists on using the clinical symptoms of a patient to give priority to certain genetic variants that could possibly explain the disease [1], [10]. Those strategies employes tools that will be explained the subsection 2.1.2.

2.1.2 Biomedical tools and ontologies

In the last decades the amount of information and biomedical data generated has been huge and can represent an issue when treating it [11]. In this context, the emergence of clinical terminologies and ontologies have allowed the management and integration of large data sets. Ontologies "organize the domain knowledge in the form of relevant concepts/classes and relationships among them" [12]. There are several ontologies, and each one has a particular purpose [13]. It is essential the interoperability between those ontologies, so that the data integration from heterogeneous datasets can be possible. All in all, ontologies are acquiring an important role in knowledge representation and management, data integration, natural language processing, as well as decision support for health information systems and biomedical research [12].

2.1.2.1 HPO

Human Phenotype Ontology (HPO) is a standardized bioinformatic resource, launched in 2008, that provides a structured and computationally interpretable vocabulary for describing abnormal human phenotypes found in human disease. Each term represents a specific phenotypic feature, organized hierarchically to support semantic relationships between general and specific concepts. It is available as a full Web Ontology Language (OWL) ontology, which enables robust data annotation, semantic search, and meaningful comparison across clinical and genomic datasets [14].

Because of its logical design, the HPO enables computational inference and sophisticated algorithms that support integrated genomic and phenotypic analysis in a wide range of applications, from basic research to clinical diagnostics. It has become a global standard for encoding deep phenotypic data and plays a key role in the precision medicine landscape [14].

One of the major strengths of HPO lies in its interoperability with other biomedical ontologies such as the Mondo Disease Ontology (MonDO), OMIM, Orphanet, SNOMED CT, and ICD, allowing precise integration of clinical descriptions with genomic data. This is particularly valuable in computer-assisted genetic diagnostics, rare disease research, and phenotype-driven recommendation systems.

To promote global adoption, the HPO Internationalization Effort (HPOIE) coordinates translation initiatives, aiming to overcome language barriers and ensure inclusive curation practices. HPO is being translated into several languages such as Spanish, French, Chinese or Japanese [15].

A variety of innovative software tools and systems have been developed to harness the power of the Human Phenotype Ontology (HPO) for enhanced clinical diagnostics and data integration. For example, Fenominal is a Java-based tool that automates HPO concept recognition from clinical text by addressing common challenges such as ambiguity and typographical errors, significantly improving recall in both literature and electronic health record (EHR) data [15].

Meanwhile Human Phenotype Ontology (HPO) is widely used in clinical genetics, it has limited coverage of prenatal phenotypes. As a result, fetal diagnosis becomes more challenging, the effectiveness of computational tools is reduced, and the clinical utility of exome and genome sequencing is limited, despite their growing adoption. To address this, an international collaborative effort expanded the ontology with new terms specific to prenatal development, reorganized categories such as skeletal and placental anomalies, and linked phenotypes to diagnostic methods like ultrasound. These improvements are expected to enhance diagnostic precision, discover phenotype-genotype relationships, facilitate earlier detection of genetic conditions and clinical decision-making during pregnancy [6].

HPO provides a standardized and hierarchical vocabulary for describing clinical features, facilitating interoperability across systems and the integration of phenotypic information with genomic data.

2.1.2.2 Orphanet

Orphanet, established in France in 1997, is a collaborative initiative involving a network of 41 countries. Its primary aim is to enhance the diagnosis, care, and treatment of patients with rare diseases. To achieve this, Orphanet focuses on three key objectives: increasing the visibility of rare diseases through a unique, multilingual nomenclature (ORPHA code) aligned with major international terminologies; providing high-quality, accessible information to all stakeholders via a comprehensive directory of expert services and multilingual resources; and fostering knowledge generation by integrating structured, interoperable data and working with global experts to support research and clinical practice. []

Orphanet offers several open access services, where Orphanet Rare Disease Ontology (ORDO) can be found, developed alongside European Bioinformatics Institute. It is a structured vocabulary encapsulating rare diseases. It captures the relationships between rare diseases, genes and other related information. ORDO also contains links to other biomedical ontologies, databases, and classification systems. ORDO is updated and released every six months [16].

It was built from the original relational database developed by Orphanet. The ontology allows an appropriate knowledge management (edition, curation, validation, and quality control) in a simpler and more accurate manner. Furthermore, the hierarchical representation of rare diseases admits the user to make search at any level [17].

2.1.2.3 MONDO

As can be seen, the previous ontologies were designed for a particular purpose, which makes them overlap and enter conflict. In order to solve those problems, the MONDO Disease Ontology (MONDO) was created. MONDO is an open, community-driven ontology that integrates key medical and biomedical terminologies by providing a logic-based structure. Some of those resources are OMIM, Orphanet, ICD, SNOMED CT and NCIt. It currently contains more than 25.000 terms (April, 2025) covering Mendelian, rare, common, complex, infectious, and neoplastic diseases [18].

Establishing equivalences between concepts from different ontologies is a complex process, where it is probable to make mistakes when only automized methods, like text coincidence, are used, due to its inability to understand the context. MONDO overcomes these limitations following an advanced computational strategy, which integrates multiple semantic characteristics, and it is validated by expert curations, obtaining precise relationships between resources [3].

Thanks to its flexible structure, transparent provenance, and regular updates, MONDO has become a reliable backbone for integrating heterogeneous disease knowledge in both clinical and research settings.

2.1.2.4 OMIM

OMIM (Online Mendelian Inheritance in Man) is an online database that comprehensively compiles information about human genes, phenotypes and the relationships between them [19]. OMIM is continuously updated and obtains its information from peer-reviewed biomedical literature. This platform is a key resource for healthcare professionals and researchers trying to obtain a better understanding of the relationship between genomics and disease.

The structure of OMIN separates genes description from phenotype description because just one gene can be associated with several clinical manifestations. Each gene entry contains information about genetic variants, function, structure and other molecular aspects, while phenotype entries include clinical features, inheritance patterns, and known molecular bases (when available). Entries are organized with fixed headings and a table of contents for easy navigation, and they include external links to other curated databases. Phenotypes with overlapping clinical features may also be grouped into *Phenotypic Series* when judged clinically relevant.

Its main utility lies in supporting clinical diagnosis and genetic research. Physicians and scientists can use OMIM to identify possible diseases associated with specific mutations, understand related symptoms, and access additional information through links to other databases such as GenBank or PubMed. Moreover, OMIM facilitates exploration of genotype-phenotype relationships, making it essential for

studies in personalized medicine, clinical genetics, and the development of treatments based on a patient's genetic profile.

2.1.2.5 SNOMED

SNOMED (Systematized Nomenclature of Medicine – Clinical Terms) is a multilingual clinical terminology standard based on ontologies that provides a common language for coding, storing, and retrieving medical information in EHRs. Its main goal is to improve healthcare quality by enabling interoperability between different health information systems. SNOMED CT encompasses a wide range of concepts related with healt, such as diseases, procedures, clinical findings, substances, and organisms, organized in a logical hierarchy that allows detailed and accurate representation of a patient's clinical condition [20].

2.1.2.6 Monarch Initiative

The Monarch Initiative is an international consortium that integrates genetic, phenotypic, and disease data across species to improve clinical diagnosis, uncover disease mechanisms, and advance translational medicine. Its approach relies on open ontologies, semantic models, and knowledge graphs that enable interoperable connections between human and model organism data. Since its inception, Monarch has developed and harmonized key ontologies such as HPO (Human Phenotype Ontology), Mondo (a unified disease ontology), GENO (for genotypes), and uPheno (for cross-species phenotype alignment), establishing a robust semantic foundation for biomedical research [18].

Among its most notable tools are Exomiser, used to prioritize genetic variants based on patient phenotype profiles; Phenotype Explorer, which enables phenotype-based comparisons across diseases and species; and Semsimian, a plugin that calculates semantic similarity between sets of ontology terms. Monarch also offers a RESTful API, a redesigned web interface, and a regularly updated knowledge graph that integrates data from over 30 biomedical sources. The development of PHENIO, an integrated ontology linking phenotypes, anatomy, chemistry, and diseases, along with the incorporation of machine learning models through the GRAPE library, has further enhanced its analytical capabilities [15].

Recent contributions include a full modernization of the platform, migration to cloud infrastructure, a redesigned interface with semantic navigation, and the launch of a ChatGPT plugin that allows natural language querying of the knowledge graph. Monarch has also begun leveraging Phenologs to infer cross-species phenotype relationships via orthologous genes and has reinforced its commitment to open science by making all data and tools freely available. These advances position Monarch as a key resource for precision medicine, rare disease research, and the exploration of complex biological mechanisms [15].

2.1.2.7 DECIPHER

Another common bioinformatic tool frequently used in clinical practice is DECIPHER. It is a consolidated key platform for clinical interpretations of genetic variants related to specific phenotypes. As it is described in [21], DECIPHER allows the collection, visualization and comparation of genomic data along with standardized phenotypic descriptions using HPO. This facilitates the comparative analysis between patients and the identification of genotype-phenotype patrons, which is vital for interpreting uncertain significance variants. Its pathogenicity can be inferred by linking them with clinical features observed in multiple individuals.

Moreover, this platform integrates predictive tools and classification rules (ACMG/Clingen), and it is connected like Matchmaker Exchange for searching similar cases at an international level. The impact of this project is reflected in thousands of scientific publications and in the discovery of new syndromes, phenotypic expansions, and previously unrecognized genetic associations, positioning it as an indispensable resource in modern genomic medicine.

2.1.3 Phenotype–genotype integration in databases

Monarch Initiative and DECIPHER are biomedical platforms where genotype-phenotype relationships can be studied, as it has already been said. Monarch Initiative allows studying different genes and their relationship with certain diseases or phenotypes, all of them standardized following ontologies (MONDO, HPO, OMIN). Meanwhile in DECIPHER, the association between genomic variants and phenotype can be examined. The intention in this section is to investigate more databases in the literature that also analyze those phenotype-genotype relationships and evaluate available prenatal data included in these resources.

In the context of genetic diagnosis of rare diseases, the development of CentoMD® stands out—a globally curated database that integrates genotypic and phenotypic information from more than 100,000 individuals, of which only 0.4% corresponds to prenatal cases. This resource, introduced by Trujillano et al. (2017) [22], enables the correlation of genetic variants with clinical manifestations through the use of standardized ontologies such as HPO, thereby facilitating both the interpretation of massive sequencing data and the dynamic reclassification of variants as new evidence emerges. Its significance lies in the fact that more than half of the recorded variants had not previously been described in the literature, underscoring its value as a complementary resource to traditional public databases in the molecular diagnosis of hereditary diseases.

GPCards is another platform that collects genotype-phenotype correlations from studies about hereditary diseases [23]. This database contains information from more than 17,000 patients and variants from 1288 genes, obtained from 2000 articles. Its design enables not only detailed exploration of clinical symptoms associated with specific variants, but also automated annotation of genetic files by cross-referencing more than 60 genomic sources, including ClinVar, OMIM, and gnomAD. However, there is no explicit mention to prenatal cases.

2.1.4 LLMs and NER extraction

2.1.4.1 LLM applied in diagnosis

In recent years, large language models (LLMs) have emerged as promising tools for clinical diagnosis. This development takes place within a context of increasing complexity in the diagnosis of rare genetic diseases, where interpreting phenotypic and genotypic data requires not only technical accuracy but also a deep understanding of clinical language. Trained LLMs with big volumes of biomedical text offer a unique capability to integrate disperse information and generate diagnostic inferences.

In Kim et al.' article [24], five LLMs were evaluated, including GPT-4 and LLaMA2 family, in prioritized genetic task based on phenotypes. Even though GPT-4 showed the best performance, its precision did not improve the traditional models such as AMELIE or Phen2Gene. This study also reveals that LLMs results were biased on popular genes and better indicators were obtained when using structural inputs (HPO terms) against natural language texts.

Meanwhile, Liang et al. (2024) [25] presents GeneT, a LLMs system specifically fitted to identify pathogenic variants causing rare genetic diseases. GeneT showed a substantial improvement, reducing the number of candidates from more than 400 to less than 10 per sample. Furthermore, its integration int he iGeneT platform allowed the genetic analysis time to reduce from 60 to only 3 minutes.

2.1.4.2 Text extraction

The automatic extraction of clinical information from EHRs has evolved in a significant way thanks to the advancement of Natural Language Processing (NLP) and the use of LLM.

In Moreno-Barera et al. (2025) [26] the use of Named Entity Recognition (NER) is studied. NER is an NLP task which consists on automatically identify text fragments that can refer to a certain thing and then classify it in its correspondent category. Here it is used for eliminating protect health information in EHRs written in Spanish. This work uses recurrent architecture (BiLSTM, BiLSTM-CRF) and Transformer models (XLM-R, RoBERTa-BNE, XLM-R-Galén) and evaluates their performance using real corpus (Galén) and synthetic (MEDDOCAN). The main challenge faced in this project is the lack of tagged corpus in Spanish, which limits the generalization.

In second place, Baddour et al. [27] analyzes the phenotype extraction codified with ontologies like HPO. Biomedical Entity Linking (BEL) is defined as "finding and connecting biomedical concepts, terms, and entities mentioned in medical texts to their matching entries in structured databases or referenced ontologies". It remarks on the importance of linking those medical terms found to the HPO to improve the diagnosis. A clear structure is defined that goes from span detection, candidates' retrieval and ranking. The analysis of PhenoBERT revealed limitations, as it does not always detect relevant spans and relies heavily on explicit matches, leading to performance drops on more realistic datasets with implicit phenotype references. Integrating LLM-based span detectors improves recall, though the gain is partly

offset by later filtering stages. Notably, many out-of-gold-standard spans generated by the LLM were relevant, highlighting both the potential of LLMs to support annotation and the incompleteness of current datasets. This underscores the need to enrich existing resources and to explore LLM-based approaches for candidate selection, while also assessing new methods such as Retrieval Augmented Generation (RAG) and updated evaluation protocols.

The Master's Thesis by Luis Couto Seller (UPM, 2024) [28] also addresses the problem of phenotype entity recognition, but here is done in Spanish clinical texts. Similarly to the previous paper, the terms are linked to the HPO. This challenge is crucial for achieving semantic interoperability in healthcare systems, as most phenotypic information in electronic health records (EHRs) is recorded in free-text format. Moreover, HPO is not fully translated into Spanish, which limits the availability of annotated data and the direct applicability of deep learning models originally developed for English.

The work proposes a hybrid model that combines dictionary-based search techniques with deep learning models built on BERT-like architectures trained on Spanish biomedical data. The system involves constructing an extended HPO-based dictionary through lemmatization and data augmentation (round-trip translation), along with a deep learning module capable of generalizing beyond explicit dictionary terms. The results show a strong balance between precision and recall (precision 0.70, recall 0.76, F1=0.73), outperforming dictionary-only methods.

This thesis represents a pioneering contribution to clinical text processing in Spanish, demonstrating the feasibility of hybrid approaches for phenotype entity recognition. Its results lay the groundwork for the development of automated systems for phenotype coding, contributing to clinical data standardization, improved diagnostic accuracy, and the advancement of biomedical research, particularly in the field of rare diseases.

2.1.5 Representation formats and interoperability: Phenopacket Schema

Interoperability is one of the main issues when talking about biomedical research. Almost every center could have its own protocols and formats to save information, which have been recollected from their patients. This becomes a valuable problem when data from different partners is needed to develop investigations. This field of research is also affected by this obstacle. Multiple EHR written by different clinicians can reference the same disease or phenotype, although they may be described in several ways.

To address this problem, the Global Alliance for Genomics and Health (GA4GH) have designed the Phenopacket Schema, a standard for "sharing disease and phenotype information that characterizes an individual person, linking that individual to detailed phenotypic descriptions, genetic information, diagnoses and treatments". By integrating with other GA4GH data and technical standards, the Phenopacket Schema will support efficient data exchange and establish a framework for computational analysis of disease and phenotype information, ultimately advancing diagnostics and research across diverse disorders, from cancer to rare diseases [29].

The Phenopacket Schema is designed to support FAIR principles (findable, accessible, interoperable, and reusable). They are characterized for being human and machine-interpretable, facilitating integration into genomic diagnostic tools and patient

stratification algorithms. Furthermore, its structure is aimed at being interoperable between people, organizations and systems; easy to integrate with data repositories, RHRs and knowledge bases; and flexible about which terminologies or ontologies must be used.

The technology used is based on Protocol Buffers (protobuf) from Google, which allows representing data in binary, JSON or YAML formats. Moreover, it guarantees efficiency, flexibility and compatibility with different programming languages such as Java, Python or C++.

The article by Danis et al. (2025) [30] offers a significant contribution through the creation of Phenopacket Store—an extensive collection of 6,668 Phenopackets derived from published clinical reports. This dataset encompasses 475 Mendelian and chromosomal diseases, linked to 423 genes and over 3,800 unique pathogenic variants. Unlike other repositories that aggregate data, Phenopacket Store focuses on individual case-level information, enabling a granular and computable representation of phenotypes, diagnoses, and genetic variants.

Moreover, the article introduces pyphetools, a Python library designed to streamline the creation of Phenopackets from tabular data extracted from medical literature. This methodological approach not only enables efficient curation of phenotypic data but also establishes best practices for annotating clinical cases using ontologies such as HPO. The resulting collection becomes a valuable resource for researchers and software developers, offering a standardized dataset that can be used to test gene prioritization algorithms, perform phenotypic similarity analyses, and explore genotype-phenotype correlations. Taken together, this work reinforces the role of phenopackets as a key tool in precision medicine and rare disease research.

Continuing the analysis of tools that support the GA4GH Phenopacket schema, the article by Danis et al. (2023) [31] focuses on the development of *phenopacket-tools*, a Java library designed to efficiently build, convert, and validate phenopackets in a standardized way. This tool addresses the need for technical infrastructure that ensures the quality and consistency of computable clinical data, especially in contexts like genetic diagnosis and precision medicine.

Phenopacket-tools provides concise builders, predefined functions for ontological concepts (such as organs, sample types, age of onset, etc.), and validators that check both syntax and user-defined compliance requirements. It also includes functionality to convert phenopackets from version 1 to version 2 of the schema—an essential feature given the evolution of the standard and its adoption by various international consortia.

This work complements the initiative presented in the article on *Phenopacket Store*, by offering a technical tool that enables programmatic generation and validation of phenopackets. While Phenopacket Store focuses on curating clinical data from medical literature, phenopacket-tools facilitates its implementation in computational workflows—from data entry to analysis. The library also promotes the use of open, well-established ontologies such as HPO, Mondo, NCIT, and LOINC, reinforcing interoperability.

2.2 Rationale

The state-of-the-art analysis shows a significant gap between the advancement accomplished in the postnatal context against the prenatal one, in the domain of phenotype-genotype correlation and their application in clinical practice. Consolidated databases exist in pediatrics and adulthood and are widely used to support diagnosis and variant interpretation; the prenatal landscape remains fragmented and poorly resourced. The lack of standardized repositories capable of integrating phenotypic observations obtained through ultrasound or other imagine techniques with genomic data considerably limits the diagnosis capacity and the decision-making during pregnancy.

This gap is particularly relevant given that the fetal phenotype not only differs from the postnatal one but also evolves dynamically throughout gestation. The lack of systematic documentation of these features and their limited representation in biomedical ontologies, such as HPO, hinders both the accurate characterization of variants and the identification of clinical patterns associated with rare diseases. Despite recent efforts to extend HPO into the prenatal domain and initiatives such as the Monarch Initiative, DECIPHER, or CentoMD®, the data available remain limited in number, coverage, and granularity.

Furthermore, the automated analysis of clinical information through NLP techniques and large language models represents an emerging but still underdeveloped opportunity in this field. Although progress has been made in clinical entity extraction and ontology-based term linking, the absence of multilingual and specifically prenatal corpora restricts their real applicability in medical environments.

In this context, the development of a database focused on the prenatal domain, integrating phenotypes described in ontological terms with associated genetic variants, is clearly justified. Such a resource would not only facilitate the interpretation of genomic tests in high-risk pregnancies but also promote clinical data standardization, scientific knowledge advancement, and the implementation of precision medicine at the earliest stages of life. Ultimately, this work seeks to address an unmet need in both clinical practice and biomedical research: the availability of a computational, interoperable, and prenatal-oriented resource capable of establishing reliable phenotype–genotype correlations.

2.3 Tools and technologies

The development of this project has been carried out entirely in Python, chosen for its versatility, extensive ecosystem of scientific libraries, and strong community support in the biomedical and data science domains.

To enrich clinical and genomic information extracted from the reports provided, Application Programming Interfaces (APIs) were used during this project. They allow direct access to biomedical ontologies and resources. Through these APIs, it was possible to automatically annotate information and retrieve metadata that ensure the standardization, interoperability and semantic consistency.

A key part during the implementation of the project has been the Phenopackets library. This provides a predefined data and classes model, aligned with GA4GH Phenopacket schema. The library enforces a strict object-oriented structure, with elements such as Phenopacket, Subject, PhenotypicFeature, GenomicInterpretation, and VariantInterpretation, among others. Each class includes datatypes and specific restrictions that should be followed to ensure valid and interoperable representations. This not only guarantees that the result is standardized, but also facilitates the transformation to JSON format, allowing it to be reused in different platforms.

For the storage, querying, and visualization of the resulting data, the project relies on Neo4j, a graph database management system optimized for representing entities and their relationships. It was chosen due to its ability to model complex biomedical relations in a natural way, such as links between patients, phenotypes and genes. For the creation of the database, Cypher was used. This is Neo4j's declarative query language. It allows creating and querying nodes and relations through its syntaxis, characterized by its legibility and flexibility. Its integration inside the Python module served as a bridge between the standardized data representation (Phenopacket) and its storage in the graph-based database.

Furthermore, the project also used Neo4j Browser and Neo4j Aura (cloud service), which provide a visual, user-friendly interface for database inspection and exploration. These tools were fundamental, not only for the validation of the database, but also to visualize in an intuitive way the graph structure.

Supporting technologies included version control through Git, ensuring traceability of the codebase and collaborative development, as well as widely used Python libraries for handling JSON data, file management, and interaction with web services (e.g., requests). Together, these tools formed a cohesive technological ecosystem enabling the transformation of unstructured clinical and genomic data into a structured, standardized, and explorable graph database.

3 Methodology

3.1 Project Structure

The methodology of this project is divided into several phases (Figure 1) designed to ensure a systematic and reproducible workflow. Each phase is meant to complete a specific objective and the sum of all of them will give as a result the building of a prenatal phenotype–genotype database. The phases that constitute the work are:

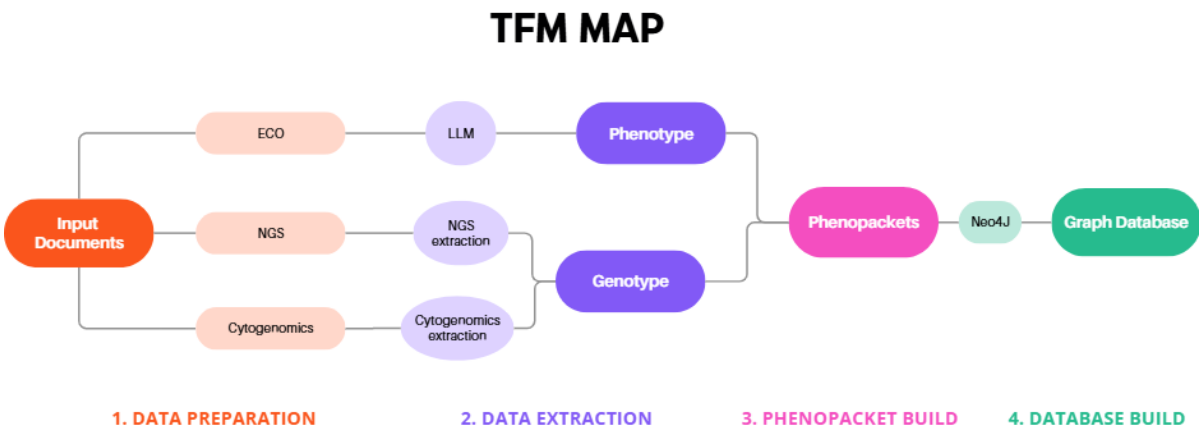

Figure 1. Project structure and its phases

1. **Data pre-processing**: Raw clinical and genomic data is treated for posterior analysis. This includes cleaning, normalization, and transformation into interoperable formats. This step ensures the transformation of the raw data into formats that can be operable.

2. **Data extraction**: phenotypic and genotypic features are detected in the pre-processed data. A LLM model will be applied to clinical text to obtain the phenotype described following the HPO ontology. For genomic data, variants are extracted and classified according to international standards.

3. **Phenopacket Builder**: Integration of phenotypic and genotypic information into standardized Phenopacket objects, according to the GA4GH Phenopacket Schema. This phase guarantees interoperability, reproducibility, and machine-readability of the data.

4. **Graph Database Builder**: Conversion from the Phenopacket schema to a graph-based database where they can be saved. The database is designed to ensure scalability and semantic relationships.

This structured workflow provides a clear roadmap from raw data handling to the creation of a specialized and interoperable resource for prenatal phenotype–genotype associations. The complete Python code is available in the GitHub repository [1].

3.2 Data pre-processing

3.2.1 Data collection

The clinical and genomic data that has been collected for this project were obtained from Hospital 12 de Octubre (Madrid). All documents were written in Spanish. The information was fully anonymized prior to analysis. Data collection is the initial step for following processing. Three main types of documents were included:

- **Echography documents:** EHRs, clinical case reports and prenatal ultrasound reports describing the main morphological features of the fetus. These documents serve as the source for phenotypic data extraction.

- **NGS documents:** Files containing the results of exome sequencing studies. These reports specify which genes are affected and the nature of the detected alterations. The extracted genotypic information includes affected genes, genomic coordinates, variant nomenclature, zygosity, and pathological classification.

- **Cytogenomics CNV:** Text files generated by Cytogenomics software designed for the analysis of microarray data. These documents provide information about copy number variants (CNVs) (previously described in the subsection 2.1.1.). It includes the affected chromosome, cytoband, impacted genes, and the size of the variant.

3.2.2 Data Cleaning and Normalization

3.2.2.1 Echography documents

The ultrasound reports collected represent phenotypic data and they are stored as PDF files. These files, as well as in a typical EHR, contain free-text descriptions which make them non-machine-readable due to the formatting artifacts, inconsistent spacing, and character encoding issues. All those problems need to be corrected before further processing. To solve this problem a python function is implemented, `extract_text_spacy` (Figure 2). This function is able of extract and clean the text.

[1] https://github.com/javierdcp20/tfm-phenopacket-pipeline

The function uses the `spaCy` library, specifically the Spanish language model (`es_core_news_sm`). This allows obtaining the raw text. Once extracted, the text undergoes a series of normalization steps, including:

- Removal of zero-width spaces and non-breaking spaces.
- Replacement of line breaks with single spaces to produce continuous text.
- Collapsing of multiple consecutive spaces into a single one.
- Substitution of common PDF encoding artifacts (e.g., (cid:XX) patterns) with their correct character sequences, ensuring better readability and semantic accuracy.

Once the text is extracted and cleaned it is saved as a .txt file, so it can be reused. This process ensures that the clinical text describing morphological features observed with ultrasound is stored in an interpretable format, so it can be human and machine-readable. This step is essential for the for the following task (natural language processing (NLP), phenotype entity recognition, and ontology-based annotation).

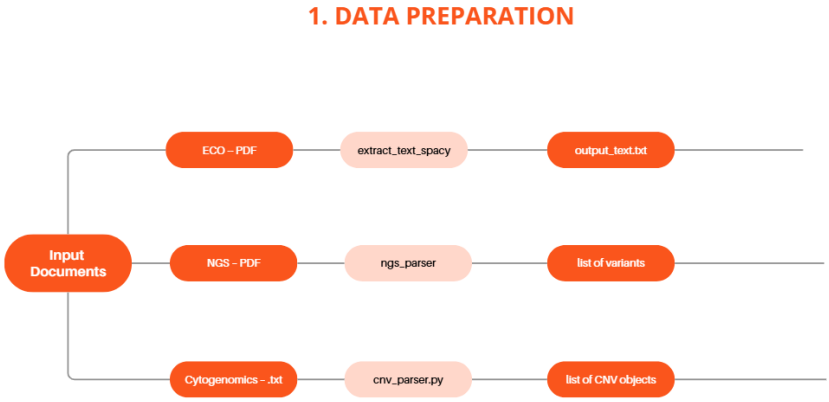

Figure 2. First phase. Data preparation. It is shown which Python modules are applied to each input document and the outputs obtained from them.

3.2.2.2 NGS documents

Genomic data is obtained from next-generation sequencing (NGS), specifically from exome study. Data has been provided in PDF reports, which include variant interpretation tables. Each row represents an alteration in the genome, and each one includes information about the gene affected, genomic position, variant, zygosity and pathological classification. These tables, in a similar way to the phenotypic data, often contain irregular formatting, inconsistent headers, and line breaks that complicate downstream analysis.

To ensure structured and standardized information, the Python function `ngs_parser` has been implemented (Figure 2). NGS variant tables are extracted, cleaned, and arranged in a machine-readable format by this function. The function parses the PDF and finds tables that correspond to different interpretations using `pdfplumber`

library. After being found, the subsequent procedures for cleaning and normalization are implemented:

- Header normalization: Column names (e.g.,) are standardized by removing line breaks and unnecessary spaces.

- Row parsing: Each table row is converted into a dictionary, ensuring that every field is associated with its corresponding header.
- Field-specific cleaning: For key fields such as *Genomic Position*, *Variant*, and *Classification*, line breaks are removed and spacing is normalized to guarantee consistency and readability.

- Structured output: The cleaned data are stored as a list of dictionaries, where each dictionary represents a variant with its attributes (gene, genomic coordinates, variant type, zygosity, and classification).

This approach transforms unstructured tables into a structured dataset, enabling easier integration with external genomic resources and facilitating variant annotation pipelines. By automating the cleaning process, the function ensures data quality, minimizes human error, and produces standardized variant representations suitable for phenotype–genotype correlation studies.

3.2.2.3 CNV files

CNVs data are reported in text files generated by Cytogenomics software. These files usually contain both metadata and information of each variant. In the same way that has been explained with the other documents, it is needed to prepare those files, so that they can be used for the posterior processing. To address this, a Python module (`cnv_parser.py`) has been developed in order to standardize CNV information into structured objects (Figure 2).

The pipeline consists of the following steps:

- Table detection and loading: the input file is scanned to locate the header line (e.g., *Chromosome, Start, Stop, Cytoband, Gene Name, Type*). After that, the data are imported into a Pandas Data Frame, filtering for valid chromosome entries (e.g., chr1, chrX).

- Row parsing: each CNV variant is transformed into an instance of the CNV class. This ensures that all attributes are consistently stored and validated.

- Normalization and validation:

    - Chromosome identifiers are normalized to the format chrN.

    - The CNV type is validated and restricted to *loss*, *gain*, or *amplification*.

    - Numerical values are explicitly cast into their proper data types.

    - Gene names are saved on a list.

- Structured output: The result is a list of CNV objects, each one of them with their own identifier. These objects are characterized for being machine-readable and they represent each CNVs, facilitating integration with phenotypic data and the posterior analysis.

3.3 Data extraction

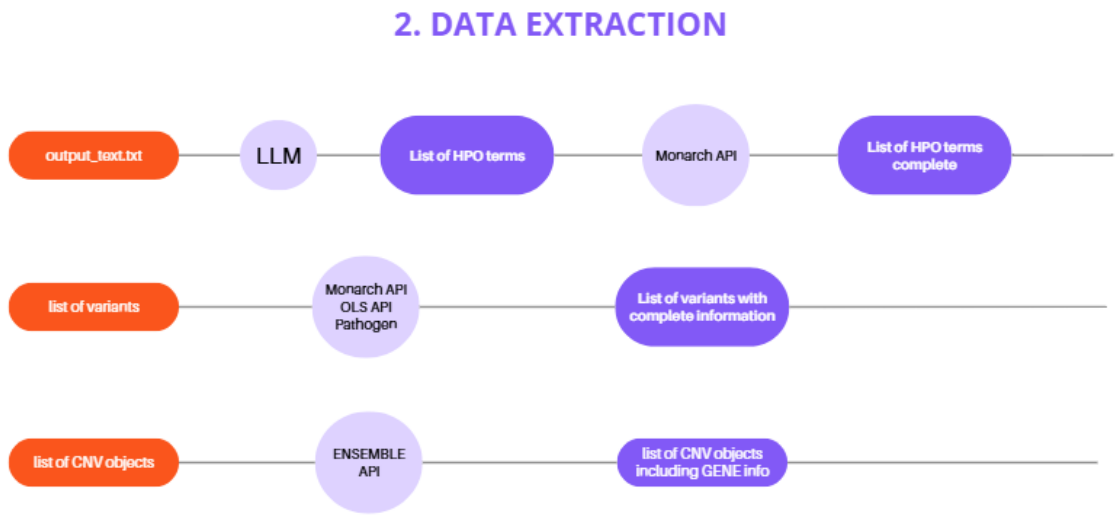

Figure 3. Phase 2. Data extraction. An LLM model is going to be applied to the preprocessed text. The results obtained from NGS and CNV reports are going to be querying in several database APIs to obtain their complete information.

3.3.1 HPO terms obtention

An essential part of this project is the extraction of the HPO terms from Spanish EHR, in this case from the echography documents (Figure 3). To accomplish this purpose the model presented in [28] was used. This model is a hybrid system, employing two main modules for the HPO phenotype annotation. The first module is "Dictionary Matching" and it searches for exact or approximated coincidences between the text and a dictionary that saves the HPO terms. The second module is founded on a deep-learning model based on BioBERT, which is able of identify linguistic variants and synonyms not presented in the dictionary. The final result is a combination of the outputs from both modules.

The HPO dictionary building was made taking exclusively the branch *Phenotypic Abnormality* from HPO and incorporating not only the official terms in Spanish but also validated synonyms by experts and its lemmatized forms using spaCy. Due to de lack of annotated terms in Spanish, data augmentation techniques were implemented using inverse translation (Spanish – English – Spanish) through Helsinki-NLP models, which allowed generating new synonyms variants and enrich the dictionary with additional instances. The training dataset for the second module was built using positive examples (tags and synonyms from HPO) and negatives (n-grams extracted from PubMed biomedical texts, which did not correspond to

phenotypes). This guarantees a balanced and representative corpus for posterior training.

The deep-learning module was purposed as a classification problem, where each term from the HPO ontology was considered an independent class. A BERT model trained in Spanish biomedical context (bsc-bio-es) was used. Furthermore, an additional class was included (HP: None) to identify terms that do not correspond to phenotype.

After preprocessing the text, both modules are runed parallelly. Then, the result integration occurs. Each identify concept by the dictionary gets a confident value of 1, meanwhile the ones which were identified by the deep learning model keeps the probability assigned by the *SoftMax* function from the model. In case of overlaps, the concept with the highest score is prioritized, or all options are retained if the codes differ and occupy different positions in the text.

The LLM gives as output a list with the HPO terms. However, to obtain the complete information of those terms, the Monarch API was used. Integrated into the Python workflow, the query was made, and it was possible to obtain the labels and descriptions of each term.

3.3.2 NGS

In the NGS documents, most of the information was already obtained during the data preprocessing phase. However, although the genes affected and their zygosity were already included in the documents, it is needed to obtain information for its standardization (Figure 3). For genes the Monarch API was used again, obtaining its id (HGNC:13797 for example), symbol, description and alternates ids (other database codes). In case of zygosity, the OLS API was used, and the id and label was obtained (for example, id: "GENO:0000135" and label: "heterozygous".

3.3.3 CNV

This case is similar to the NGS one. Most part of the information can already be acquired during the preprocessing. However, the genes information was obtained using Ensembl API, which retrieves the symbol, the id, the description and its location (chromosome, start and end position) (Figure 3).

3.4 Phenopacket builder

The Phenopacket Schema, following its definition in the documentation [32], "*represents an open standard for sharing disease and phenotype information to improve our ability to understand, diagnose, and treat both rare and common diseases. A Phenopacket links detailed phenotype descriptions with disease, patient, and genetic information, enabling clinicians, biologists, and disease and drug researchers to build more complete models of disease*".

A Phenopacket builder was developed during this project. It is vital to guarantee that the phenotypic and genomic information previously collected can be represented in

a standardized and interoperable format, following the GA4GH specificities. This module can operate as an integration layer, where the heterogeneous data obtained from clinical documents is transformed into structured objects (Phenopackets). Those are ready for their analysis and storage in the database. Given the distinct nature of the data, genotype and phenotype are built in different processes.

3.4.1 Phenopacket structure

The Phenopacket schema is structured into several elements at different levels, which are defined by specific datatypes and linked through attributes. This hierarchical design allows the representation from clinical concepts at high level, such as a patient or phenotype description; to granular information, like the onset age of the phenotype or a disease severity. Furthermore, each element is well-defined thanks to the use of biomedical ontologies like HPO, MONDO or others; and at the same time, it is computationally precise because of the use of certain datatypes. This structure (figure 4) ensures that clinical and genomic data can be represented in a standardized and interoperable way.

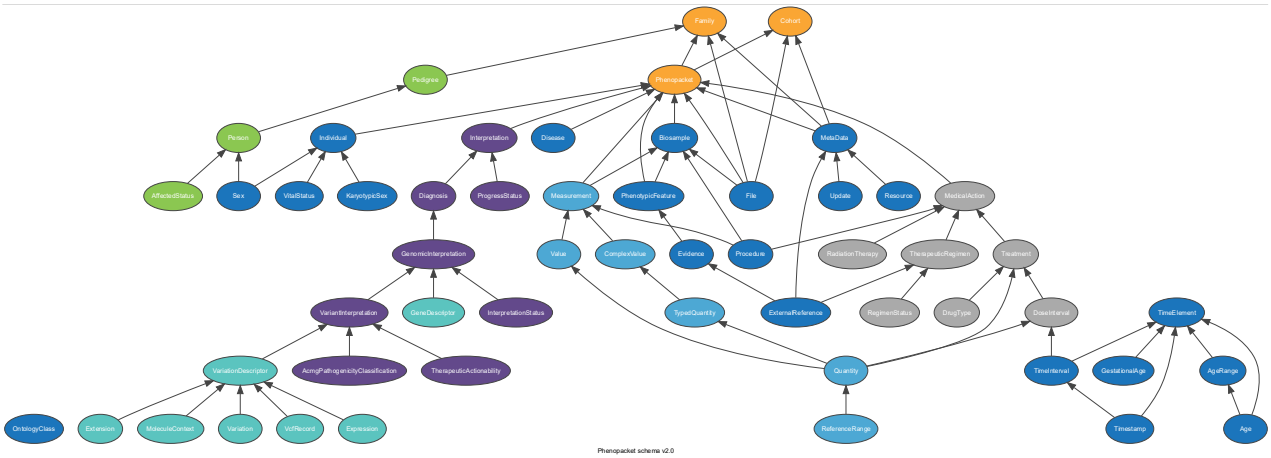

Figure 4. Phenopacket schema

As can be seen in the figure, there are a lot of elements in this complex structure, however some of them are only going to be used for achieving our purposes. In particular, this work focuses on those components that are directly relevant to the prenatal setting, namely the *Subject*, the *Phenotypic Features* extracted from ultrasound reports, the *Genomic Interpretations* derived from genetic testing, and, when available, the *Disease* block to capture suspected or confirmed diagnoses.

In first place, the Subject block, represented by the *Individual* element, describes the clinical entity, a human being or other organism, in this case the evaluated fetus. Basic patient's information is included here, for example, sex or gestational age. These results are vital in a prenatal context because temporality and embryotic development can affect the interpretation of phenotypic findings.

Phenotypic Features block is used to describe the subject's phenotype. Each characteristic is codified using HPO, which ensures semantic consistency and avoids clinical text ambiguity. Furtehrmore, it is permitted to include complementary data, such as the gestational age when the anomaly was observed or the modifiers presence which describes the clinical curse.

Another fundamental block is Genomic Interpretations, which is included in the block Interpretation. Information about genetic variants collected in NGS or CNV studies is stored here. Inside this block other elements can be found: Variant Interpretation, Gene Descriptor and Interpretation Status. The Gene Descriptor is used for defining the characteristics of the gene affected. Variant Interpretation indicates the variant and its genomic location (following standardized nomenclatures like HGVS). Moreover, the clinical interpretation is also gathered indicating if it is pathogenic or not, following guides like ACMG.

The Disease component allows including associated clinic diagnostics, codified in ontologies such as MONDO, OMIM or Orphanet. Although in the prenatal context, a confirmed diagnosis is not always available, this block is vital when from both genomic and phenotypic findings allow inferring or suspect of a concrete disease.

The modular Phenopacket structure offers the necessary flexibility in order to represent partial or complete clinical reports by adapting at different granular levels. This is especially useful in the prenatal context, where the information available can be incomplete or susceptible to change during pregnancy. Using those guidelines, it is possible for this information to be represented in a standardized format allowing its computational analysis and its integration with international repositories.

3.4.2 Technical implementation

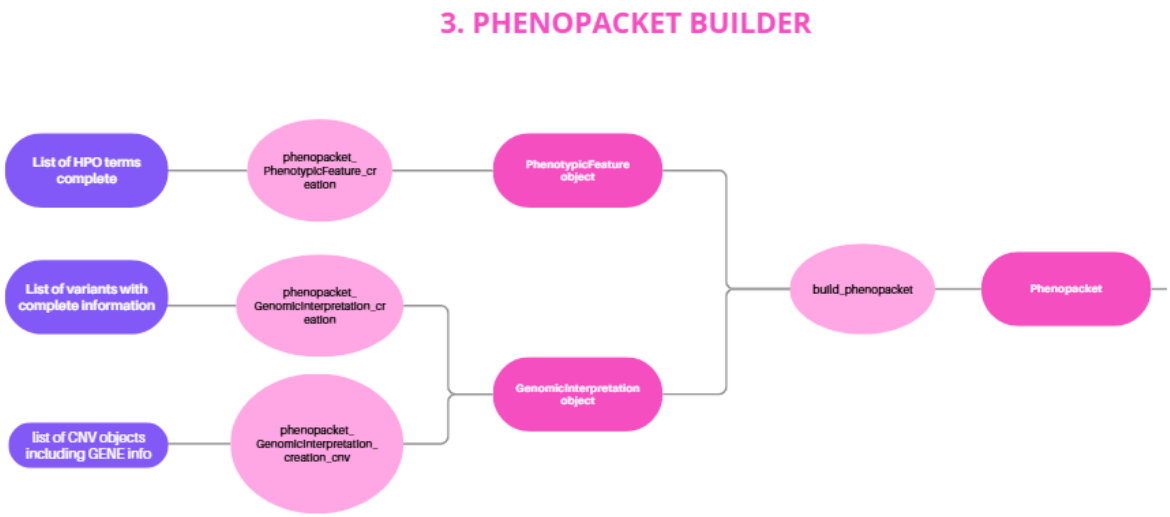

Figure 5. Phase 3. Phenopacket Builder. Phenotype and genotype information will be structured in Phenopacket blocks, applying the correct Python module to each one of them. Then, both blocks will be integrated into a single Phenopacket object.

The practical implementation of Phenopacket was accomplished by designing a Python module, `builder.py` (Figure 5), whose purpose is to build automatically Phenopacket objects from phenotypic characteristics and the genomic interpretations previously obtained. For this task the official python library developed by GA4H was used. It permits the creation of different classes and structures that configure the schema. Additionally, auxiliar libraries were employed,

like google.protobuf for the conversion of objects into JSON representations and standard Python modules.

The construction workflow starts with the creation of the Individual object, which represents the subject under study. In the prenatal context, here it is included basic information like, gestational age or sex, which are essential for the correct clinical interpretation of those findings.

The next object is Phenotypic Features, which is in charge of representing the clinical characteristics observed in the patient. The input for creating this block is the HPO terms obtained using the LLM, explained in 3.3.1. The goal is to transform clinical information into a standardized structure that can be integrated with genomic data, as has already been said.

The HPO terms are introduced to that function as dictionaries whose keys are value_id, name and description. From those a list of PhenotypicFeature objects is built. Each one of these objects are composed of an *OntologyClass* which encompasses the unique term identifier (for example: HP:0001629) with its standardized tag (for example: "*Ventricular septal defect*"). Furthermore, it is possible to add additional information like clinic description, onset age or severity, which enriches the representation. This PhenotypicFeature list is going to be integrated into the Phenopacket.

The next component is the genotype, which is represented in the Genomic Interpretations object. This element provides structured information about the genetic variants detected and constitutes a fundamental step for the subsequent integration with the echography findings. In a similar manner to the Phenotypic Features, while executing `builder.py` several functions from the module `genomics.py` will run, depending on which data input has been introduced, if it is a NGS study or a Cytogenomics report.

If the input is a NGS document, the function `phenopacket_GenomicInterpretation_NGS` will be executed. This function receives the preprocessed data explained in 3.3.2 (NGS documents). From this data the GenomicInterpretation objects are created, which at the same time contain the following information:

1. Genomic context: represented by the object GeneDescription, which includes the unique gene identifier in HGNC, its official symbol and descriptions or alternative identifiers. This guarantees that the genes are annotated in a standardized format that can be used in different biomedical databases.

2. Specific variant: described as a VariantDescription object that uses normalized expressions like HGVS, for coding the exact genomic alteration. Moreover, the descriptor can include information about zygosity also standardized using biomedical ontologies.

3. Clinical variant classification: the VariantDescriptor object handles the VariationDescriptor and the clinical interpretation of the pathogeny following the American College of Medical Genetics and Genomics (ACMG) criteria. The function ngs_2_AcmgPathogenicityClassification is implemented, which translates categories from EHR to normalized terms of ACMG standard.

In the other case, where the input is the CNV report, the function `phenopacket_GenomicInterpretation_creation_cnv` is executed. Apart from the annotation of the implied genes, the CopyNumber object is built, which defines the genomic location of the alteration and the number of observed copies. This permits to represent deletions, duplications or other relevant alterations.

The Disease element is used for representing suspected or confirmed pathologies in the clinical case. Those diseases are codified using the MONDO ontology. The Disease element is part of a larger block, Diagnosis, where the GenomicInterpretations are also included. Diagnosis, at the same time is inside of the Interpretation object.

Lastly, the MetaData component is added. It contains information about when the packet was created (date and time) and references about the ontologies used. This block is fundamental to ensure data tradability and reproducibility, because it allows identifying the version of the ontologies used.

Once all the components have been defined, all of them are integrated into a Phenopacket object, which plays the main container role. This object will be transformed into JSON format, and it will be saved. In this way, each clinical case is represented in a standardized file that can be easily shared, computationally analyzed, or incorporated into a structured database, such as the one developed later in this work.

3.5 Database builder

The final step is the building of a graph-based database (Figure 6). The decision to create this type of database was made due to the highly interconnected nature of the biomedical data involved, where the phenotypic characteristics and the genomic variants should be represented not only as isolated entities, but also through their relationships. For generate it, Neo4J was used, a graph-oriented database management system designed to store and query information represented through nodes and relationships.

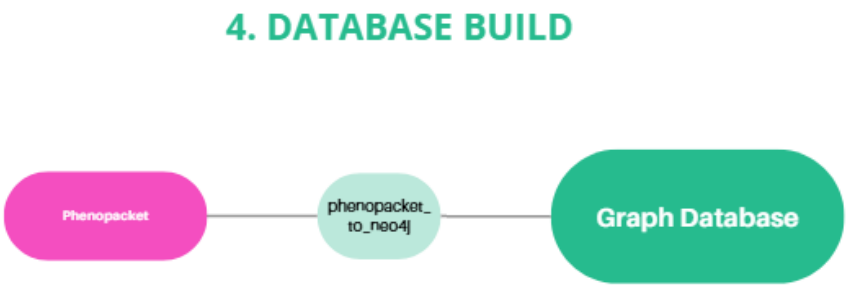

Figure 6. Phase 4. Database Build. The Phenopacket object will be processed by the Python module phenopacket_to_neo4j and the graph-based database will be obtained.

3.5.1 Graph model design

The first task to be completed, for designing the graph model, is the nodes and relationships definition. Nodes are the main entities of the database, and they correspond to an individual record, for example, a thing or a fact. Nodes are classified by one or more labels. Labels indicate to which category a node belongs to, and they are useful for organizing and consulting them. Relationships represent the link between nodes, and they reflect interactions, associations or dependencies between entities. Finally, both elements are described by their properties, additional information that can describe both elements.

The definition of nodes and relationships was based on the Phenopacket schema. Therefore, the labels assigned to the nodes correspond directly to the blocks defined in the schema diagram (Figure 4). The properties of each node are those attributes described in the Phenopacket documentation that capture intrinsic characteristics of the element itself, while attributes that indicate a connection to another element are instead modeled as relationships. In the following image the nodes definition is shown (Figure 7).

To distinguish whether a block should be modeled as a node by itself or as a property, it is needed to revise the attributes in the documentation. If the field refers to a simple datatype (e.g., a number or a string) that only describes a characteristic of the entity, it is a property. In contrast, if it refers to a datatype that, at the same time, admits several subfields or can expand into additional information, it is modeled as a node.

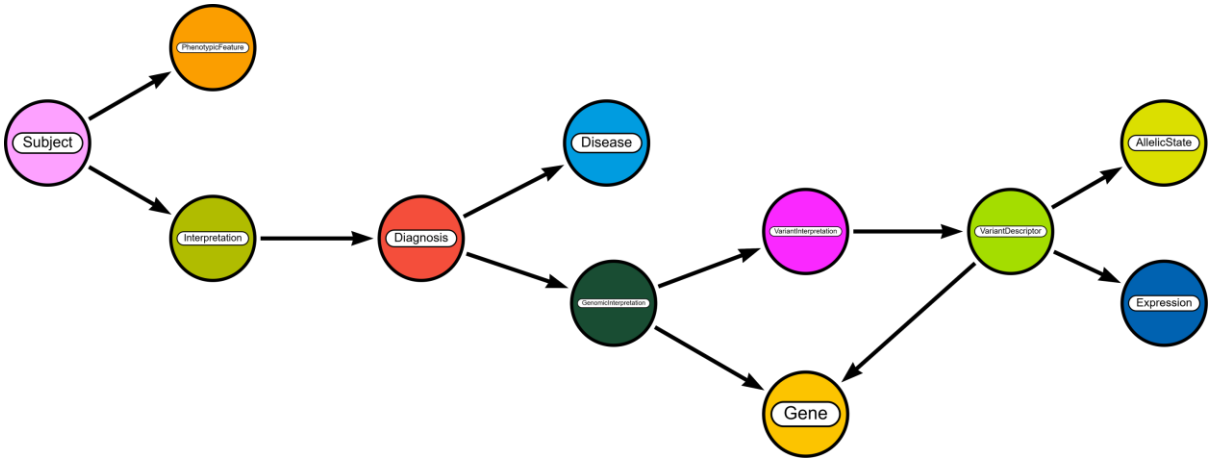

Figure 7. Network design. Shows the different types of nodes or labels that the graph network will have, and how they are related.

The graph model design is characterized by its scalability and its extensibility. It allows the incorporation of new blocks from Phenopacket schema, without requiring a complete redesign of the database. This ensures that the model can evolve and adapt as new data types, ontologies, or clinical descriptors become relevant in the future.

Each patient is represented by the node “Subject”, with basic properties such as identifier, sex or gestational age. This node basically connects with both phenotypical and genomic information. Phenotypes are integrated by the node “PhenotypicFeature”

and they relate to “Subject” through the relation “HAS_PHENOTYPE”. Phenotypes have properties like id (HPO code), label (HPO term) and description.

Even though intuitively could be thought that “Subject” should directly relate to the node which captures the genomic information, if the Phenopcaket schema is followed the genomic information actually represents a lower level of information. That is the reason why “Subject” relates to the node “Interpretation”, which represent a genomic analysis, such as the reports that can be found in the data. So, in case that a patient has several genomic reports, each one of them will refer to one “Interpretation” node. Its properties are progress status and id and it is connected to “Diagnosis”.

“Diagnosis” refers to the disease that is inferred in the inform. It is related, at the same time, with two nodes “Disease” and “Genomic_Interpretation”. “Disease” refers to the proper disease and its properties are id (MONDO code) and label. “Genomic Intepretation” describes the interpretation for an individual variant or gene and its property is “InterpretationStatus” and describes if the variant or gene reported here is related to the diagnosis.

Genomic Intepretation can be connected to one of the following two nodes, “GeneDescriptor” or “VariantInterpretation”. “GeneDescriptor” allow the representation of a gene that is believed to contribute causally to the observed disease phenotypes. VariantInterpretation represents the interpretation of a variant according to the American College of Medical Genetics (ACMG) variant interpretation guidelines. It has properties like acmg_pathogenicity_classification, which must be one of the ACMG pathogenicity categories; and therapeutic_actionability, that establish if there is any therapy for this variant.

“VariantInterpretation” is connected to “VariantDescriptor”, which describe candidate variants or diagnosed causative variants. This will be different depending on which report is been analyzed. If it is an NGS report, the variant will be described by connecting it to the next nodes:

- Allele: it describes in which allele the alteration has taken place, indicating the chromosome (the corresponding NCBI RefSeq) and the start and end position.
- geneContext: which genes have been affected.
- Expressions: the variant is described using HGVS nomenclature.
- Allelic state: the zygosity of a variant is described using the Genotype Ontology (GENO).

If “VariantDescriptor” describes a CNV it will be connected to the allele node previously mentioned and ChromosomeNumber that indicates the number of copies and if it is a gain or loss. This decision was made following [29].

3.5.2 Technical implementation

The integration of clinical and genomic data into a graph database has been made using Neo4J, as has already been said. The goal is to transform semantic complexity from medical and genetic reports to a graph model, where the relationship between

entities (patients, phenotypes, genes and variants) can be explored in a flexible manner.

For this task a Python module was developed to import JSON files with the Phenopacket structure directly into Neo4J through the official driver of the database. The function `phenopacket_to_neo4j` acts as a starter point and it is in charge of opening the JSON file, validating the server connection and executing the writing transaction. Inside this transaction, the function `import_phenopacket` is called, which contains the construction logic of nodes and relationships.

The creation of the graph database using Neo4J requires the use of its programming language *Cypher.* It enables us to define, update and query those nodes, their relationship and their properties. In this project, Cypher is used within the Python implementation to define how clinical and genomic entities are represented and connected. For example, commands such as MERGE ensure that nodes like Subject, Gene, or VariantDescriptor are created only once, while MATCH and MERGE together allow the establishment of relationships such as (:Subject)-[:HAS_PHENOTYPE]->(:PhenotypicFeature) or (:GenomicInterpretation)-[:ASSOCIATED_GENE]->(:Gene).

Moreover, the fact that Cypher can be implemented directly inside the Python module represents a significant advantage for both flexibility and integration. Cypher is interpreted and executed natively by Neo4J, so additional software is not necessary. This means that all operations made to create the database are integrated in the same workflow as the rest of the project.

4 Results

4.1 LLM

During the second phase explained in Methodology, an LLM was applied to the PDF files from echography. The preprocessing of ultrasound reports resulted in simplified texts, saved as .txt files, ready to be processed by the LLM. The goal of this step is to extract HPO terms from raw text in Spanish. A quantitative and qualitative analysis has been made.

4.1.1 Qualitative analysis

In this section, concrete excerpts from text will be commented on with the purpose of illustrating strengths and weaknesses through concrete examples. For example, in this text:

"Amniocentesis genética (HP:0011410) en gestación de 16+4 semanas con feto normodesarrollado con diagnóstico de ventriculomegalia leve (HP:0010952) bilateral y simétrica sin otros hallazgos asociados."

Those words remarked in blue were identified by the LLM and assigned to the correspondent HPO term. On the one hand, the words *"amniocentesis genética"* are identified with the term HP:0011410 (*Caesarean section*), which clearly indicates a total mistake. On the other hand, "*ventriculomegalia leve"* is identified as HP:0010952 (*Mild fetal ventriculomegaly*), which is correctly classified, not only capturing the right phenotype, but also the prenatal context.

However, the analysis should go beyond the simple classification as success / failure, and other aspects should be considered. The next case:

"Gestación de 20+5 semanas con feto con cardiopatía congénita tipo tipo defecto del septo atrio-ventricular (HP:0001674) común con volúmenes ventriculares balanceados con sospecha de hipoplasia de arco aórtico y huesos largos en el límite bajo de la normalidad."

The term assigned is HP:0001674 (Complete atrioventricular canal defect), whose description in the ontology is: "*A congenital heart defect characterized by a specific combination of heart defects with a common atrioventricular valve, primum atrial septal defect and inlet ventricular septal defect.*". In this case the classification could be considered as "partially correct" because, even though it is able to identify the atrioventricular defect it is not complete, as the HP:0001674 says, but the description in the original text does not imply that it also affects the valves. For that reason, a more adjusted term would be HP:0006695 (*Atrioventricular canal defect, "A defect of the atrioventricular septum of the heart."*), which is a "broader term" or an upper-class in the ontology of HP:0001674. However, that granularity failure can be considered a minor mistake.

In the same text, two groups of words were remarked in red. Those are phenotypic characteristics that have not been captured by the LLM model. The first one, "*hipoplasia de arco aórtico" could correspond to the HPO term HP:0012304 (Hypoplastic aortic arch). The second one "huesos largos en el límite bajo de la*

normalidad" could correspond with HP:0003026 (Short long bone). Therefore, it is not only important taking into account those terms that have been wrongly classified, but also those which have not been detected.

Following examples:

> *"Evaluación general Actividad cardíaca (HP:0001635) presente."*
>
> *"Ecocardiografia: Situs solitus (HP:0001696)"*

HP:0001635 refers to Congestive heart failure. In this case, the LLM has been able to capture the cardiac context, but there is no presence of any abnormality. Then, *"Situs solitus"* presented in text refers to the normal position of the abdominal and thoracic organs, but HP:0001696 maps to "Situs inversus totalis", where "*A left-right reversal (or mirror reflection) of the anatomical location of the major thoracic and abdominal organs.*". LLM has identified a high coincidence with the HPO term because 2 out 3 words are the same, however that third word completely changes the diagnosis. Those errors can be considered as "semantic errors" where the model relies on partial lexical similarity rather than full semantic understanding, leading to clinically misleading annotations.

Next example:

> *"Si sangra como una regla, pierde líquido amniótico (HP:0001560), dolor abdominal (HP:0002027) y/o fiebre materna acudir a la urgencia."*

Both terms can be considered as well classified, HP:0001560 represents "Abnormality of the amniotic fluid" and HP:0002027 to "Abdominal pain". However, these phenotypes are mentioned as hypothetical scenarios rather than actual findings. Those "contextual errors" can be relevant when analyzing a report, but the LLM was not trained for capturing them.

Last example:

> *"Enfermedades anteriores Cáncer de mama izquierda (HP:0006625)"*

In this case, it was classified as HP:0006625 represents "Multifocal breast carcinoma". Technically, it does not exactly reflect the phenotype because it is not multifocal. However, what should be remarked in this case is that this phenotype does not relate to the fetus, but to the mother. Therefore, in those reports there are also abnormalities that are not necessarily related to pregnancy.

4.1.2 Quantitative analysis

With the aim of conducting a quantitative analysis of the results, the HPO terms were classified according to the patterns identified in the qualitative analysis. Therefore, HPO terms can be grouped into four categories:

- Correct: HPO terms were correctly classified, and they correspond to the phenotype presented in text.
- Granular inconsistency: HPO terms captured the phenotype correctly, but a more specific or more general term in the same ontology hierarchy would have fit better.

- Semantic error: HPO term detected does not correspond to the phenotype described in text. This type of error is typically produced when the model relies on partial lexical similarity rather than full semantic understanding.
- Incorrect: The extracted HPO term has no relation to the phenotype mentioned in the text. These cases represent clear misclassifications.

This classification has been made manually, consulting the terms in the Monarch Initiative webpage, where ontology information can be found, and comparing it with the context given by the original text. It should be remarked that not all HPO terms detected were counted for this evaluation, only those parts of the text which directly refer to phenotype description.

It is important also to note that duplicate HPO terms were not removed from the analysis. This decision was made because the same HPO code can occur multiple times in a report but in different textual contexts, leading to different classifications. For example, the HPO term *HP:0001560 (Abnormality of the amniotic fluid)* may appear once simply introducing the topic "Amniotic fluid" (classified as semantic error), and later in connection with *polyhydramnios* (classified as granular inconsistency). Therefore, each occurrence was considered as an independent observation in quantitative evaluation.

The following tables (Tables 1 and 2) and the graph chart (Figure 8) sum up the classification.

Table 1. Number of terms detected from each class

	Total Number	**Correct**	**Granular inconsistency**	**Semantic error**	**Incorrect**
24L0010	30	6	3	11	10
24L0013	14	6	0	1	7
24L0020	38	12	1	11	14
24L0043	47	11	3	15	18
24L0044	31	9	4	12	6
25V0011	6	3	0	1	2

Table 2. Percentage of terms detected from each class.

	Correct	**Granular inconsistency**	**Semantic error**	**Incorrect**
24L0010	20,00%	10,00%	36,67%	33,33%
24L0013	42,86%	0,00%	7,14%	50,00%
24L0020	31,58%	2,63%	28,95%	36,84%
24L0043	23,40%	6,38%	31,91%	38,30%
24L0044	29,03%	12,90%	38,71%	19,35%
25V0011	50,00%	0,00%	16,67%	33,33%
TOTAL	28,31%	6,63%	30,72%	34,34%

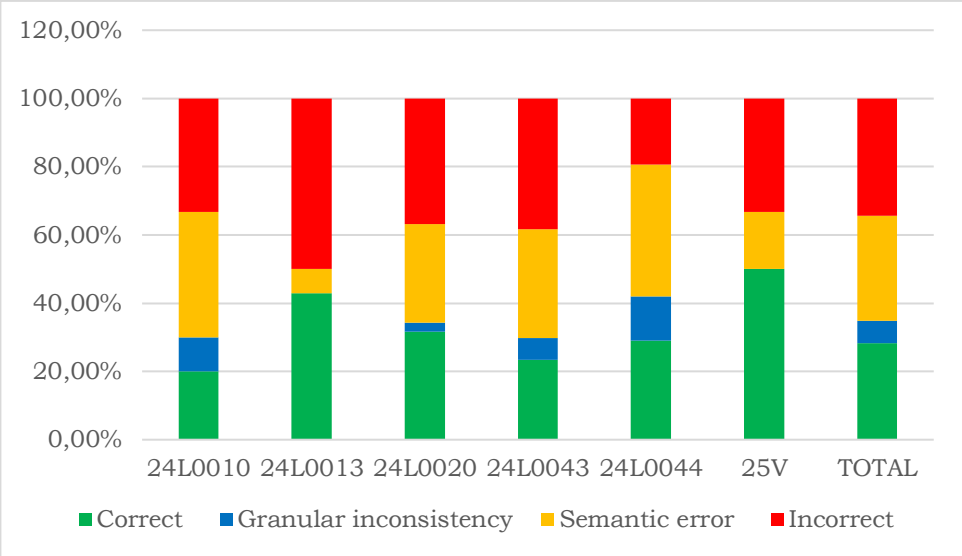

Figure 8. Graph showing the percentage of each type of error detected in each document.

Finally, given the clinical scope of this work, it was considered to also perform an analysis to specifically identify which terms were related to the HPO term studied in [6], *HP:0001197 (Abnormality of prenatal development or birth).* It was done with the objective of evaluating if this term has ability to complete identify the clinical context in real cases. Results are shown in table 3:

Table 3. Number and percentage of terms related to HP:0001197 hierarchy in each document.

	Number of HPO terms related to HP:0001197	**Percentage over total**
24L0010	3	10%
24L0013	5	36%
24L0020	12	32%
24L0043	11	23%
24L0044	11	35%
25V0011	5	83%
TOTAL	47	28%

4.1.3 Clinical relevance of the correct terms

After the review of representative examples that can be found and the general quantitative vision, the next question that should be asked is if those terms correctly classified can capture the phenotype and pathological context of the subject. For that, table 4 shows each report's main clinical features which describe the pathology and the HPO terms that correspond to them.

Table 4. Terms clinically significant and the phenotypes presented in reports.

Report	Phenotypic features	HPO codes related
24L0010	Complete atrioventricular septal defect (CAVC) with ostium primum atrial septal defect (ASD) and a small ventricular septal defect. Competent common valve annulus with mild regurgitation.	HP:0010445 HP:0001671 HP:0001653
24L0013	Mild, bilateral, and symmetrical ventriculomegaly	HP:0010952 HP:0002119
24L0020	Relative microcephaly Hypoplasia of the corpus callosum Cerebral cortex abnormalities Schizencephaly Cerebellar asymmetry	HP:0034206 HP:0009717 HP:0007333 HP:0010636 HP:0006956 HP:0033725 HP:0031933 HP:0002280 HP:0002126
24L0043	Persistent enlarged nuchal fold (11.3 mm). Bilateral cystic hygroma with two laterocervical collections. Cranial and prefrontal subcutaneous edema. Two muscular interventricular septal defects (1.2 mm and one mid-location). Renal/Genitourinary Findings Bilateral hydronephrosis	HP:0011624 HP:0000081 HP:0010880 HP:0003319 HP:0011682 HP:0002280 HP:0000126 HP:0000078 HP:0000070

	Duplication of the right collecting system. Ureterocele measuring 8.2 mm.	
24L0044	Fetal hydrops: mild subcutaneous edema and mild ascites located in the minor pelvis. Polyhydramnios (largest sac: 15 cm). Small subaortic interventricular defect. Narrow aortic valve annulus Venous return abnormality: persistent left superior vena cava. Thoracic/Pleural Findings Bilateral pleural effusion (right-sided predominance), homogeneous and echo-negative in appearance. Right pulmonary lymphangiectasia. Renal/Genitourinary Findings Mild right-sided hydronephrosis.	HP:0002202 HP:0031933 HP:0010945 HP:0001562 HP:0011682 HP:4000138 HP:0001560 HP:0010772 HP:0000969
25V0011	Nuchal translucency (NT)	HP:0010880

4.2 Phenopacket

As it has been already explained before, during the third phase, Phenopacket builder, the clinical and genomic information was transformed into structured instances following the GA4GH Phenopacket standard. The Phenopacket is saved as a JSON file which facilitates long-term storage, sharing, and reuse in future projects. This

structure ensures that genomic variants and prenatal phenotypes are linked, which facilitates posterior genotype-phenotype studies.

A total of six Phenopackets were generated, corresponding to the available clinical cases, including one case involving a CNV study. Each section of the JSON file corresponds to a block from the Phenopacket schema. To illustrate this, one representative case based on an NGS study is described in detail below (figure 9):


```
{
  "id": "24L0010_eco",
  "subject": {
    "id": "24L0010_eco",
    "time_at_last_encounter": {
      "gestational_age": {
        "weeks": 20,
        "days": 5
      }
    },
    "alternate_ids": [],
    "sex": "UNKNOWN_SEX",
    "karyotypic_sex": "UNKNOWN_KARYOTYPE"
  },
  "phenotypic_features": [
    {
      "description": "An anomaly of the intra-atrial or intraventricular septum.",
      "type": {
        "id": "HP:0001671",
        "label": "Abnormal cardiac septum morphology"
      },
      "excluded": false,
      "modifiers": [],
      "evidence": []
    },
  "interpretations": [
    {
      "id": "interpretation - 24L0010_ngs",
      "progress_status": "SOLVED",
      "diagnosis": {
        "disease": {
          "id": "MONDO:0015626",
          "label": "Charcot-Marie-Tooth disease"
        },
        "genomic_interpretations": [
          {
            "subject_or_biosample_id": "24L0010_ngs-gi-PRX-chr19:40901570 G>A",
            "variant_interpretation": {
              "acmg_pathogenicity_classification": "LIKELY_PATHOGENIC",
              "variation_descriptor": {
                "id": "var-PRX-NM_181882.3:c.2689C>T(p.Arg897Ter)",
                "variation": {
                  "allele": {
                    "sequence_location": {
                      "sequence_id": "refseq:NC_000019.9",
                      "sequence_interval": {
                        "start_number": {
                          "value": "40901570"
                        },
                        "end_number": {
                          "value": "40901571"
                        }
                      }
                    }
                  }
                }
              },
              "gene_context": {
                "value_id": "HGNC:13797",
                "symbol": "PRX",
                "description": "periaxin",
                "alternate_ids": [
                  "ENSEMBL:ENSG00000105227",
                  "OMIM:605725"
                ]
              "expressions": [
                {
                  "syntax": "hgvs",
                  "value": "NM_181882.3:c.2689C>T(p.Arg897Ter)"
                }
              ],
              "molecule_context": "genomic",
              "allelic_state": {
                "id": "GENO:0000135",
                "label": "heterozygous"
              }
```


Figure 9. Fragment from a JSON file describing a Phenopacket object.

- Subject: the individual is a fetus (unknown sex), whose gestational age is 20+5. Gestational age was used both as a temporal reference for the subject and for clinical and genomic findings.

- Phenotype features: a list of several HPO terms detected in the original documents are grouped here. The information includes, for each one of them, the HPO code and label and the description.

- Interpretation: clinical and genomic information is interpretated here, including which disease the subject suffers, in this case Charcot-Marie-Tooth disease, identified using the MONDO ontology (MONDO:0015626). Furthermore, the genomic information is also included. It is a "likely pathogenic" variant, and its location appears in the allele section, specified with the genomic coordinates from *RefSeq*. The gene involved is "PRX", which HGNC code can be found (HGNC:13797, "ENSEMBL: ENSG00000105227", "OMIM:605725"). The variant is also described according to the HGVS nomenclature. In this case, "NM_181882.3:c.2689C>T(p.Arg897Ter)", indicates that there has been a substitution of a cytosine (C) for one thymine (T) at position 2689 of the coding DNA sequence. At the protein level, this is translated into a replacement of one Arginine (Arg) at position 897 with a termination (STOP) codon, obtaining a truncated protein.

In the following example (figure 10), the patient presents four CNVs. For clarity, only one of these variants will be described in detail, as the remaining ones follow an analogous structure. Each variant is represented by one Genomic Interpretation object. The phenotypic data are represented in the same way as in the previous case.


```
"interpretations": [
  {
    "id": "interpretation - ",
    "progress_status": "IN_PROGRESS",
    "diagnosis": {
      "disease": {
        "id": "MONDO:0700126",
        "label": "trisomy 21"
      },
      "genomic_interpretations": [
        {
          "subject_or_biosample_id": "25V0011_Censurado",
          "variant_interpretation": {
            "variation_descriptor": {
              "id": "25V0011_Censurado-cnv-chr2:73079-2635809",
              "variation": {
                "copy_number": {
                  "allele": {
                    "sequence_location": {
                      "sequence_id": "refseq:NC_000002.11",
                      "sequence_interval": {
                        "start_number": {
                          "value": "73079"
                        },
                        "end_number": {
                          "value": "2635809"
                        }
                      }
                    }
                  },
                  "number": {
                    "value": "1"
                  }
                }
              }
```


Figure 10. JSON file showing the genomic interpretation from a Phenopacket object, based on a CNV

In this case, the patient presents a trisomy 21 (although the interpretation is not definitive as it is marked "*IN_PROGRESS*"). The variation is located in chromosome 2, from the position 73079 to the position 2635809. The chromosome 2 is codified as "*refseq: NC_000002.11"* which refers to reference sequence of human chromosome 2 in the NCBI RefSeq database, specifically to the GRCh38/hg38 human genome

assembly. The variant was classified as a "loss", which means that the number of copies of that genomic region is reduced compared to the normal diploid state. In particular, it corresponds to a heterozygous deletion (State x1), meaning that one copy of this genomic segment is missing while the other remains intact. This reduction in copy number can potentially alter gene dosage, which may impact the expression levels of the genes within this region and influence cellular functions or phenotypic traits, depending on the biological role of these genes.

The Phenopackets were validated using *phenosentry* library [33], a "*Python package for ensuring data quality in Phenopackets and collections of Phenopackets*". This tool allows systematic auditing of both individual Phenopackets and larger collections, detecting inconsistencies, missing metadata, or deviations from the standard. In this project, validation was performed under the STRICT auditor level, which enforces the highest degree of rigor when checking the structure and content of the files. This process guaranteed that the generated Phenopackets were not only syntactically correct JSON documents, but also semantically coherent objects that adhere to the expected ontologies, identifiers, and structural constraints defined by GA4GH. By integrating *phenosentry* into the workflow, it was possible to increase the reliability and reusability of the dataset, minimizing errors and ensuring interoperability with other resources and analytical pipelines.

4.3 Graph-database

The graph-database implemented integrates all standardized clinical cases in Phenopacket format, which includes both phenotypes and genomic variants extracted from the correspondent reports. The result is a graph that permits representing in a structured manner the relationship between patients, phenotypes (HPO), genes, variants and diseases, showing both the phenotypical and genomic information analyzed. From this construction, it is possible to visualize and explore the different connections between nodes, facilitating the identification between relevant associations in the dataset.

The implementation of the graph-database was made using Neo4j, platform which provides an online tool Aura Neo4j, that allows exploring the database and make queries in a visual manner and considerably facilitates the inspection of nodes and relationships. Through this interface it is possible to quickly visualize the connections between patients, phenotypes, genes, variants, and diseases, as well as to execute Cypher queries that retrieve specific subsets of the data. This functionality proved especially useful for validating the correct integration of the different data types and for obtaining an overview of the structure of the dataset.

Each individual Phenopacket is represented as a central node which contains basic patient information, including its id and gestational age. Related to this node validated phenotypes can be found, each one of these codified with its HPO term. In a similar manner, the genomic variants detected in the patient are connected to the corresponding genes and diseases described following proper ontologies. This allows that when observing a unique case, those relations can be clearly visualized. In Figure 11, the graph of document L240010 is shown.

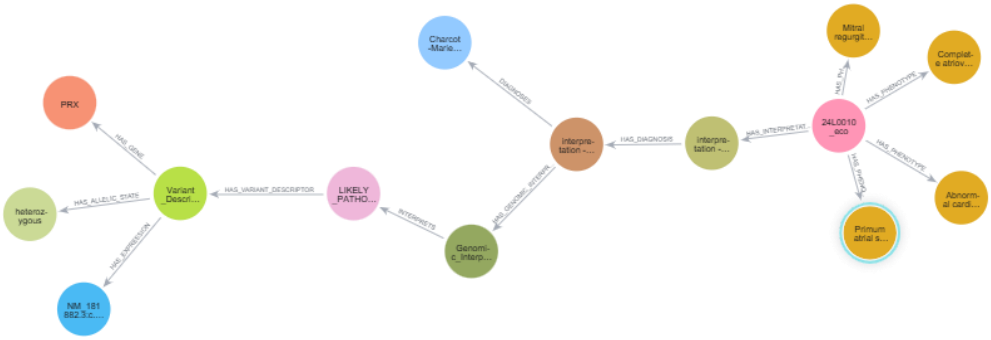

Figure 11. Graph network of the subject L240010_eco shown in Neo4j.

In this image the relation between disease-phenotype-genotype can be seen. The subject presents several phenotypes (Primum atrial septal defect, Mitral regurgitation…) and is connected to the disease (MONDO:0015626, Charcot-Marie-Tooth disease). It can also be viewed that it is interpreted as likely pathogenic and the heterozygous variant is related with the gen PRX and is defined as *"NM_181882.3:c.2689C>T(p.Arg897Ter"*.

The Aura Neo4j interface allows getting all properties which describes one node by simply clicking on it. Figure 12 shows all the information collected by several nodes.

PhenotypicFeature

Key	Value
<id>	4:6a5fcd4f-e2e4-4053-b978-15260497ad30:3
description	"An ostium primum atrial septal defect is located in the most anterior and inferior aspect of the atrial septum. The ostium primum refers to an anteri... Show all
id	"HP:0010445"
label	"Primum atrial septal defect"

Subject

Key	Value
<id>	4:6a5fcd4f-e2e4-4053-b978-15260497ad30:0
gestational_age_days	5
gestational_age_weeks	20
id	"24L0010_eco"
sex	"UNKNOWN_SEX"

Disease

Key	Value
<id>	4:b97175d0-a6ff-4dad-9690-3e99a9cc04c5:7
id	"MONDO:0015626"
label	"Charcot-Marie-Tooth disease"

Gene

Key	Value
<id>	4:6a5fcd4f-e2e4-4053-b978-15260497ad30:11
description	"periaxin"
id	"HGNC:13797"
symbol	"PRX"

Figure 12. Properties from different nodes shown in Aura Neo4j.

Another more complex example is presented in document 24L0013. Here, the subject presents three genomic variations, two of them related to the same gene "GJB2" and all of them classified as heterozygous. Furthermore, they are related with the disease *"sensorineural hearing loss disorder"* (MONDO:0020678), which presents the phenotypes *"Mild fetal ventriculomegaly"* (HP:0010952) and *"Ventriculomegaly"* (HP:0002119). These results are shown in the image.

The document which contains the CNV analysis is represented in Figure 13.

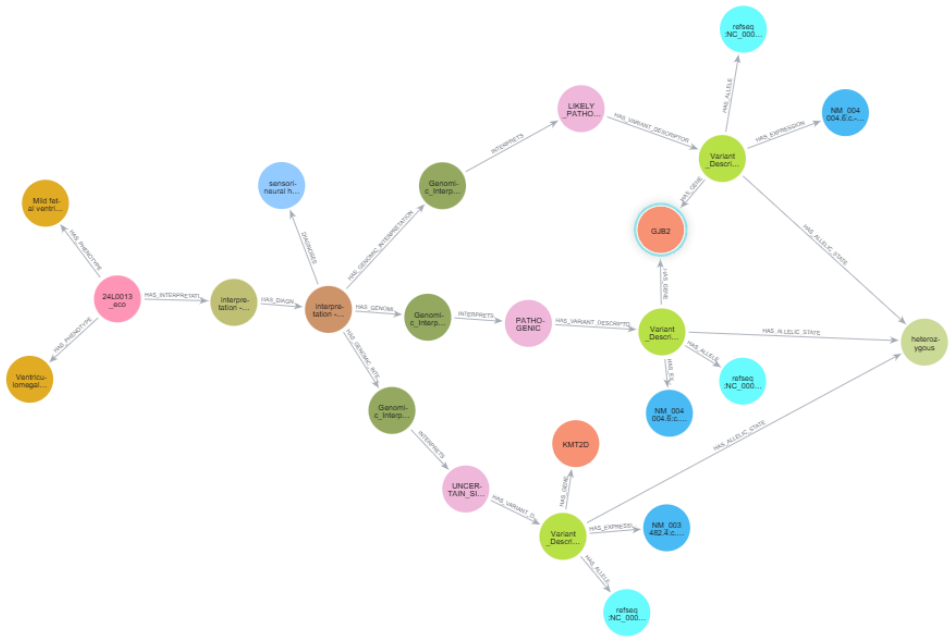

Figure 13. Graph network of the subject 24l0013 shown in Aura Neo4j.

Here, in Figure 14 the four CNVs from document 25V0011 can be viewed, alongside the phenotype which describes the subject. In the last node the location of those variations can be consulted.

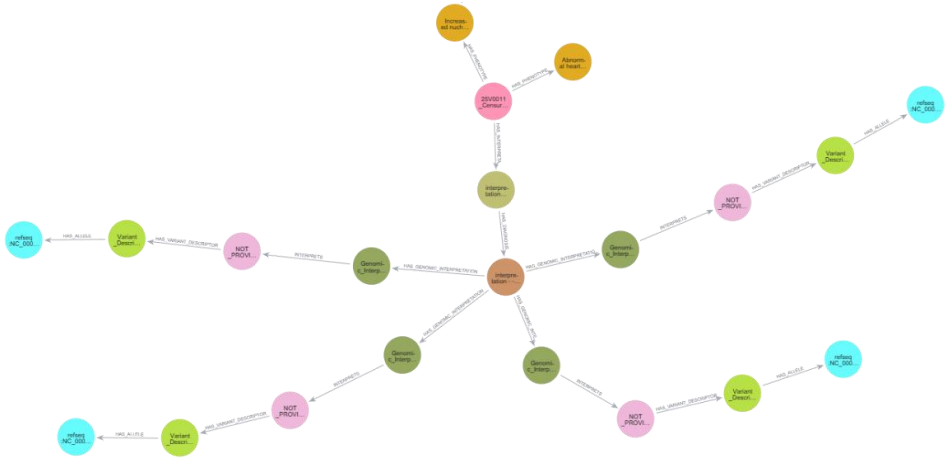

Figure 14. Graph network of the subject 25V011 shown in Aura Neo4j.

Finally, the hole net is constituted by 117 nodes and 123 relationships between them. A general view is shown in Figure 15.

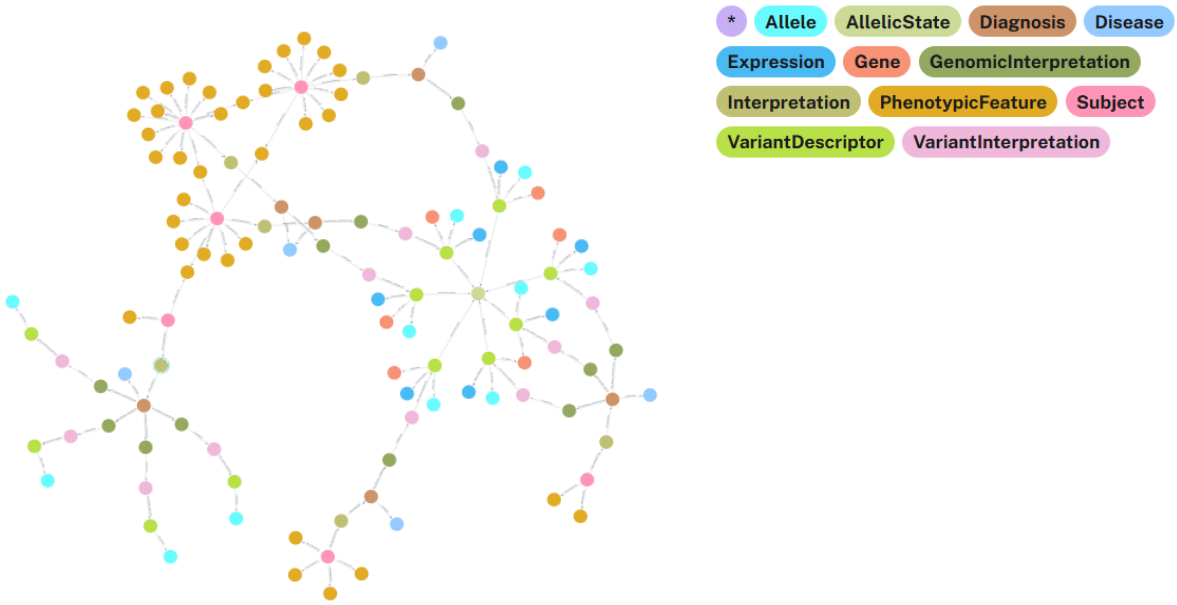

Figure 15. Complete graph network database.

In this image two characteristics can be highlighted. First, most variants (green nodes) are related to the allelic state “heterozygous” (grey node). Secondly, some phenotypes (orange nodes) are shared by some subjects (pink nodes). Subjects 24L0043_eco and 24L0044_eco share the same disease (Figure 16).

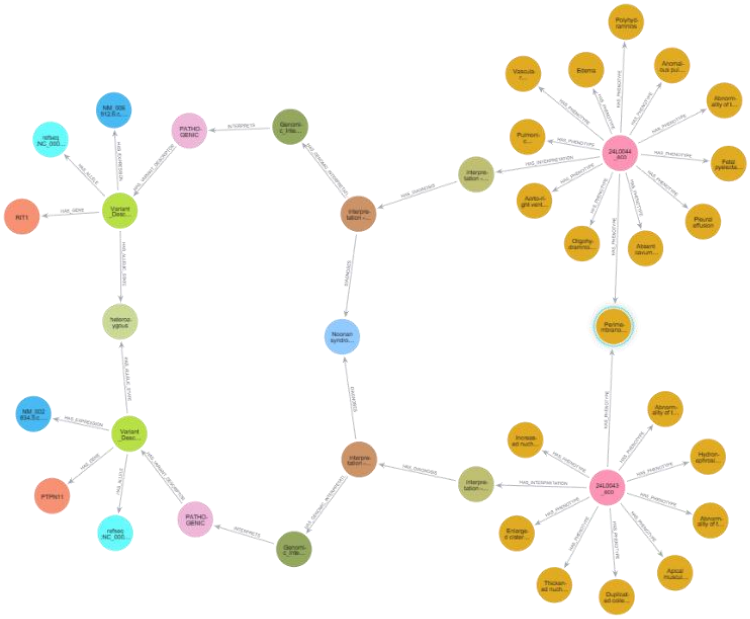

Figure 16. Graph network of the subject 24L0043_eco and 24L0044_eco.

5 Discussion

5.1 LLM

The results of this project show both the current potential and limitations in the application of LLMs for HPO term extraction from prenatal clinical reports. The model has shown its capacity to identify a considerable number of fetal phenotypes, with several terms correctly classified and assigned to their respective HPO codes. However, the analysis also reveals a significant proportion of errors. These findings point out that although LLMs can be really helpful to automate phenotype extraction, their performance heavily depends on the understanding of the context, especially in complex prenatal narratives.

First thing that must be remarked is that a traditional AI analysis could not be carried out. The main reason is the lack of a golden standard. This means that the reports should have been revised by experts and indicate which words can be associated with HPO terms. Then, once the LLM has been applied, a comparison between the LLM results and the expert curation could have been made. If this methodology had been followed, important metrics for the evaluation of LLM performance could have been obtained.

This situation represents a clear limitation in this work. It would be interesting to know the precision of the model (the percentage of terms which were correctly detected by the LLM over the experts) or the recall (percentage of terms which the LLM was not able to capture). The precision can be approximated with the following analysis. However, the recall cannot be approximated because this would have been made by comparing the terms annotated by experts which were not detected by the model.

In this project, as an alternative to expert curation, terms were manually reviewed and classified into four categories (correct, granular inconsistency, semantic error, and incorrect), which allowed an intrinsic evaluation of the model.

The results (Tables 1 and 2) indicate that most of terms detected are not correctly assigned to its HPO code. Only the documents 25V0011, reaching 50% of correct and 24L0013 with 42.86%, show a better performance. The rest of them barely reach the third of correct terms. Semantic errors play a fundamental role representing 30.72% of the total reports. This indicates that the model can detect where important information can be relied, but it is not able to assign them to the correct feature. Lastly, 34,34% of the terms in all reports are incorrectly detected.

However, despite the amount of the terms incorrectly classified, an exhaustive review report by report demonstrates that the terms which were correctly classified consistently capture the core pathology and phenotypic characteristics of the subject. Table 4 shows that the main clinical findings are well described by the HPO codes. For example, the first case which indicates *"Complete atrioventricular septal defect (CAVC) with ostium primum atrial septal defect (ASD) and a small ventricular septal defect"* and *"Competent common valve annulus with mild regurgitation"*, the terms are *HP:0010445 (Primum atrial septal defect), HP:0001671(Abnormal cardiac septum morphology),* and *HP:0001653 (Mitral regurgitation).* The same happens with the rest of the cases. Even though only around 30% of terms were correctly classified in the best reports, these represent the most clinically relevant phenotypes.

Moreover, these results highlight the practical utility of the LLM in supporting prenatal clinical annotation. The fact that essential fetal abnormalities are detected demonstrates the potential of this approach to aid clinicians when identifying relevant features quickly. By concentrating on the terms that are correctly mapped, it becomes evident that the model can serve as a valuable tool for highlighting key clinical features, reducing manual annotation workload, and providing a reliable starting point for expert curation.

Therefore, if the main clinical information is collected by LLM, it can be said that there might be an overestimation of the number of terms needed made by the model. This is partially explained the number of Semantic errors. One example is the one shown in results: when the text indicates that it is going to talk about some topic (heart rate) or describes a normal situation (*situs solitus*), the model that relies in the lexical similarity, identifies them as abnormality, thus a HPO term.

Hence, for the application of LLMs model in real reports, it is essential for them to understand the context in which the terminology is fitted. If the model only focuses on lexical similarity, the results can be altered. Something similar happens in sentences indicating negative or hypothetical situations. It is normal in this kind of documents sentences like "*If the patient suffers…*" or "*The subject does not present…*" and then naming some symptoms. In parental reports clinical characteristics of the mother can also be described and they do not correspond with the fetus. Those, although identified by the model, do not represent an actual feature. As it was said in the Results section, if the model correctly identifies a term in one of those cases, it was decided to count them as correct because the model was not trained for those situations.

Sometimes, an important phenotypic feature is not captured by the model. However, in these types of reports (EHR) the more important a clinical characteristic is, the more times is repeated in the text. Thus, the model can detect it later in the text. This can put in perspective the number of errors committed by the model, because they can be corrected.

The last topic that is going to be discussed is the HPO term HP:0001197 (Abnormality of prenatal development or birth) also discussed in [6]. Table 3 shows how many terms detected are part of this hierarchy. Once again, these results on their own do not say anything, so they have to be shown in their clinical context. In some cases, those terms perfectly capture the pathology, for example in *L24*0013 when using HP:0010952 (Mild fetal ventriculomegaly), or when describing specific phenotypes like HP:0010880 (Increased nuchal translucency) in documents L240043 and 25V0011 and HP:0001562 (Oligohydramnios) in L240044. In other cases, like document L240020, HP:0034206 (Abnormal fetal central nervous system morphology) is a general term and others might provide more information. In document L240010, any of those terms refers to HP:0001197.

The fact that several correctly identified terms in this project are descendants of this parent node suggests that the ontology provides an appropriate framework to classify key prenatal abnormalities. However, the presence of very general terms, such as HP:0034206 (*Abnormal fetal central nervous system morphology*), highlights the limitation that some annotations may lack specificity. Furthermore, there are terms that can describe the fetus like *HP:0010445 (Primum atrial septal defect)* and they do not belong to that hierarchy. Expanding this branch with concepts that are often encountered in fetal medicine could ensure that both specific and broadly relevant

findings are adequately captured, increasing the practical utility of HPO in prenatal contexts.

5.2 Phenopacket

5.2.1 Interoperability and structured objects

A really common problem nowadays when treating biomedical data is their standardization, especially if this data came from unstructured formats. This project is one example of it, where the clinical and genomic data were presented in reports or EHR, written down by the proper medical team. This situation presents several problems when a computational analysis is pretended to be done. First, it limits the sharing data between departments or hospitals due to each one of them following their own rules, restricting the amount of data available. Moreover, they are not machine-readable, so it is needed to preprocess the information.

In this context, the GA4GH Phenopacket constitutes an essential advancement for the computational representation of biomedical data. A common schema is presented where most of the information that a EHR could contain is collected here. This is especially relevant in this project, whose main goal is to relate genomic and phenotypical data. Both types of information can be described in a standardized way using the Phenopacket schema. The capacity of containing several levels of information in a unique object addresses the need to understand disease mechanisms, where the genotype-phenotype correlation is essential both for advancing their study and for driving progress in precision medicine.

This structural flexibility permits to represent not only genomic or phenotypic data, as already said, but also other elements that can be vital for understanding the clinical context in which the case is developed. Information about measurements that can be reflected on the EHR or the possible medical actions that can be taken are taken into account in the Phenopacket schema. However, it was decided that this information would not be included for this project and only information about the subject, disease, phenotype and genotype would be represented. There are several reasons for this choice. Firstly, it does not necessarily align with the project objectives. Furthermore, to do this there were two possible ways. It could have been manual annotation which does not really add significant scientific value while considerably increasing the workload. The other option would have been employing automated pipelines and the use of other LLMs models, but the lack of data for training them was the main reason to discard this alternative. For these reasons, it was decided to prioritize the core elements.

Therefore, three main elements can be found in the created Phenopackets: Subject, which includes the id, sex and gestational age; Phenotype features referring to the echography descriptions; and Interpretation, which at the same time includes the disease and genomic interpretation. This structured representation ensures that phenotypic manifestations can be directly linked with their underlying genomic variants, providing a consistent framework for exploring genotype–phenotype correlations.

5.2.2 Ontologies

The representation of heterogeneous data within Phenopackets is based on the extensive use of ontologies and other nomenclatures, which provides several key advantages. Firstly, each component is defined with unique identifiers derived from these resources, avoiding the ambiguity inherent in natural language, where terms may vary depending on the author, language, or clinical context. Secondly, standardization guarantees reuse and traceability, as the same term can be localizable and comparable between different studies and repositories. Thirdly, the hierarchical structure of ontologies adds a semantic layer that enables computational analyses beyond exact matches; related terms can be grouped or inferred through parent–child relationships, which is particularly valuable for genotype–phenotype correlation studies and for uncovering broader disease mechanisms. For that reason, Phenopacket can be considered as a strategic tool for scientific collaboration.

The choice of ontologies within the Phenopacket framework is not arbitrary but rather strategic, as each resource adds a complementary dimension to the structured representation of clinical and genomic data.

MONDO provides a unified language for diseases through integration of terms coming from different resources such as OMIM, Orphanet, NCIt and DOID. Its hierarchical organization permits precise describing annotations at different levels, from broader categories to more specific subtypes, without losing consistency. Thanks to its adoption, the Phenopackets in this project can be easily aligned with other repositories and facilitate joint analyses across heterogeneous datasets.

HPO plays a fundamental role in the description of phenotypical abnormalities. Its detailed vocabulary allows the systematic codification of clinical findings, such as the ultrasound findings in prenatal cases. Moreover, the hierarchical and relational structure supports computational processing, enabling, for example, the grouping of phenotypes that share common parent terms.

RefSeq (NCBI Reference Sequence) is used for specifying reference sequences and the genomic coordinates, which describe the variants. The genome assembly (GRCh37 or GRCh38) is referenced using RefSeq ids, which ensures that the reported variants are unequivocal and reproducible in different databases and genomic pipelines. This can be critical in both CNV and NGS cases, because the genomic positions could be interpreted differently depending on the reference genome used. Thus, Phenopackets achieve consistency with genomic standards, guaranteeing compatibility with other resources.

For the description of the implicated genes the HGNC ontology was used, giving (as the rest of ontologies) a unique identity for each gene. Furthermore, alternative ids coming from other databases, like Ensembl and OMIM, were included, ensuring interoperability and enriching the genomic context.

Variant expressions were made using the HGVS annotation, which has been explained the significance in the Results section. Lastly, the allelic state was defined using GENO ontology, which classifies conditions like heterozygosity or homozygosity.

In conclusion, the coordinated use of these ontologies within the Phenopacket framework guarantees phenotypic and genomic information to be represented in a consistent and interoperable manner. Each resource adds a complementary layer of precision. Together, they enable the integration of heterogenous datasets, support

computational analyses at different levels and facilitate genotype–phenotype correlation studies. This structured and ontology-driven approach not only strengthens the reliability of the data but also enhances its reusability and impact in collaborative research and clinical applications.

5.2.3 Validation

Another important step for obtaining high quality Phenopackets is validation process. A critical level of assurance is added to the data generated by the *phenosentry* library's validation of Phenopackets, particularly at the STRICT auditor level. Through systematic inspections, the tool ensures that each Phenopacket is syntactically correct not only as a JSON document but also it is semantically coherent according to the GA4GH standard. Because a syntactically valid JSON does not necessarily imply that the clinical and genomic data it contains is internally consistent, suitably formatted, or compliant with the pertinent ontologies. Validation contributes to the overall robustness and reproducibility of the research. Furthermore, it also enables achieving the same goals previously mentioned: interoperability with other systems and reuse by other investigators.

5.2.4 Limitations

Notwithstanding, the advantages that Phenopacket schema has provided to this project, there are still some limitations or difficulties that can arise during its implementation.

- Dependency on the original data. As it was explained in the methodology chapter, the initial data is preprocessed for the posterior Phenopacket creation. However, this preprocess stage was designed adapting to that data. If other format was presented, due to any reason, this preprocess stage would need to be redefined. Furthermore, if the original data have incomplete information, the Phenopacket objects would inherit those limitations.

- Manual curation. For the creation of Phenopacket, characteristics like sex, gestational age or disease was needed to be annotated by hand. In a similar manner that was explained with the other structures (measurements and medical action), its integration into an automated pipeline would have needed the training of several LLMs, for which more data and workload would have been necessary.

- Partial ontology coverage: Although MONDO, HPO, HGNC, RefSeq, etc. are very complete, they don't always cover very specific terms, especially in poorly documented contexts such as prenatal care. This may require the use of more general descriptions, losing clinical detail.

- Inherent complexity to the standard. The several blocks and different levels in which the schema is divided permit its flexibility, but at the same time

increase its complexity. Initially, it was difficult to properly understand the differences between blocks like Interpretations, “GenomicInterpretations” and where each variant described should be included.

- Ontology and standard evolution. Databases and ontologies are frequently updated. A valid Phenopacket could turn obsolete or inconsistent with future versions of the ontologies or the proper standard. That is the reason why automatically updating systems should be worked on.

- Adoption in real clinical environments. Although Phenopacket is very promising, it is not yet implemented in most hospitals or electronic health systems, which may limit practical interoperability in the short term.

Despite these challenges, the adoption of the GA4GH Phenopacket standard in this project demonstrates its potential as a robust framework for structuring biomedical data. Addressing the mentioned limitations will be crucial to ensure its scalability, clinical applicability, and integration into routine workflows.

5.3 Graph-database

The discussion of the presented results will analyze the relevance and scope of integrating standardized clinical cases into a database, as well as the possibilities offered by this representation in biomedical research. From the database created, it is examined how the organization of clinical and genomic information not only allows validating the correct data structure but also exploring in a more efficient manner the relationships between patients, phenotypes, genomic variants and diseases. Moreover, the advantages and limitations found during its implementation are highlighted. The goal is to place the results in a broader context and discuss its potential for further applications in phenotype-genotype correlation analysis.

The use of a structured database allows integrating clinical and genomic patient information in a systematic way, which facilitates establishing relationships between phenotypes and genomic variants. When centralizing data coming from different sources and codified them with standardized ontologies, it is possible to link each observable characteristic in the patient with genes or associated variants and diseases. This organization not only ensures information tradability, but also permits exploring patrons, identifying correlations and generating hypothesis about possible phenotype-genotype associations, contributing to a deeper and rigorous analysis of clinical and genetic data.

The graph-based database building is especially convenient in the phenotype-genotype analysis context. It allows representing in a flexible and natural manner the complex relationship between entities such as patients, phenotypes, genes, variants and diseases. Unlike traditional relational models, graphs do not require rigid table structures, which makes it easier to integrate heterogeneous and constantly growing data. Furthermore, its designed based on nodes and edges enables intuitive and efficient exploration of connections, allowing the detection of

indirect associations, the identification of recurring patterns, and a clear visualization of how different elements in the dataset are interrelated. This makes them a powerful tool in phenotype-genotype correlation analysis and when generating new hypothesis form available information.

5.3.1 Neo4j

The implementation of graph database in this project was made through Neo4j. It is a platform for creating graph-based databases widely used in research and complex data analysis. Its design is optimized for storing and managing information, which is saved naturally as nodes and relationships. This makes it an appropriate tool for the analysis of biomedical data, where multiple interconnected entities such as patients, phenotypes, genes, variants, and diseases are involved. Neo4j allows representing the network in an explicit and flexible manner, avoiding traditional relational models and facilitating the exploration of connections that can be relevant in the research.

One of the main advantages of Neo4j is its integration with querying language Cypher, especially designed for interacting with graphs. This functionality not only enables highly specific searches within the database but also allows for the identification of complex patterns and indirect associations. Furthermore, its compatibility with Python modules makes it possible to connect the database with the workflow of the rest of the project. This enables the construction of a continuous path from the data extraction from reports to the database creation.

Another remarkable characteristic is the platform Aura Neo4j, which offers an intuitive and user-friendly online interface. It allows users to visualize the nodes and relationships graphically, as well as to explore their properties with a simple click. Thus, it is useful for data validation and an initial approach to analysis, since it offers a clear and accessible visual representation even for researchers without prior experience in graph databases. Moreover, Neo4j Aura allows researchers to execute Cypher queries directly from the interface, which provides a balance between visualization and detailed analytical capability.

However, the use of Neo4j also presents certain limitations. Although Cypher is relatively easy to learn compared to other programming languages, it still requires time to become familiar with its syntax. At the same time, although Aura Neo4j interface is useful for initial exploration, some advanced functionalities needed for a deeper network analysis (such as calculating centrality measures, modularity, or community detection) are only available in enterprise versions or require additional libraries. Moreover, being in a different environment to traditional databases, researchers should adapt their analysis practices to this platform.

5.3.2 Individual case analysis

Individual case analysis within the database offers a clear advantage, as it allows both phenotypes and genomic variants presented in a subject to be observed in just an image. Each Phenopacket block is represented as a node permitting phenotype-genotype-disease relationships to be intuitively explored and a fast understanding of clinical and genomic characteristics of a particular case.

A representative example is the document L240010 (Figure 11), in which the phenotype-genotype-disease relation is clearly visualized. The graph can be summarized in few sentences: *"The patient, who has symptoms like mitral regurgitation, abnormal cardiac septum morphology, atrioventricular canal defect and primum atrial septal defect; presents a likely pathogenic genomic variant associated to the disease Charcot-Marie-Tooth. This variant is identified as heterozygous, affects the gene PRX and is described as NM_181882.3:c.2689C>T(p.Arg897Ter)".* If more information of each node is needed, it can be seen just by clicking any of those nodes (Figure 12). The graph visualization integrates all that information in a compact representation, being especially useful for complex clinical cases.

Another example is 24L0013 (Figure 13). This subject presents three heterozygous genomic variants: one likely pathogenic, another pathogenic and another with uncertain significance. It can be seen that two of those variants (the pathogenic and likely pathogenic ones) are related to the same gene GJB2. This can be interpreted as the affectation of that gene might can provoke the presence of the disease sensorineural hearing loss disorder.

Finally, the document containing the CNV analysis (Figure 14) demonstrates how the graph can be adapted to represent other types of structural genomic variations. In this case, four CNVs are visualized alongside the subject's phenotype and the genomic location of each variant. Here, pathological information of the variants is not provided, as it does not appear in the reports offered. However, it could be obtained following the guidelines of ACMG [8], [34]. The classification of pathology variants will be discussed in future lines.

5.3.3 General network analysis

When visualizing the hole network, it can be defined as modular, that is, it presents well-defined modules, each one of them representing an individual case. However, all those modules are interconnected, which is quite surprising because apparently those cases are not related. The first important node, where most cases converge, is the AllelicState, as all the variants identified in the NGS reports are heterozygous. This node explains the connection between two subjects with the rest of them.

The subjects 24L0043_eco and 24L0044_eco (Figure 16) are joined by the same disease Noonan syndrome. The description given by the NGS reports of this disease is:

"Noonan syndrome is an autosomal dominant disorder characterized by short stature, facial dysmorphism, and a wide spectrum of congenital heart defects. Cardiac involvement is present in up to 90% of patients. Pulmonary stenosis and hypertrophic cardiomyopathy are the most common forms of heart disease, but other abnormalities are also observed. Other relatively frequent features include multiple skeletal defects (thoracic and spinal deformities), pterygium colli, intellectual disability, cryptorchidism, and bleeding diathesis."

This description is consistent with the phenotypes found in both cases. For example, 24L0044_eco presents Pulmonic stenosis, Anomalous pulmonary venous return and Aorto-right ventricular tunnel. 24L0043_eco presents Apical muscular ventricular septal defect, Abnormality of the cervical spine and Abnormality of the genital system. They have one characteristic in common, Perimembranous ventricular septal defect,

which aligns with the description of cardiac abnormalities presented in Noonan syndrome. However, the variants affect two different genes in each case, which may contribute to explaining this variability (one subject is more affected by pulmonic stenosis and the other symptoms like the affectation of the cervical spine and the genital system), but other factors such as genetic modifiers, epigenetic influences, or environmental conditions may also play a role. Anyway, this is an example where the integration of genotype and phenotype can be vital for clinical interpretation.

There are other examples where different subjects share the same phenotype. For instance, 24L0043_eco and 25V0011 presents, increased nuchal translucency. However, the completely different genomic information that the graph provides might suggest that one single phenotype can be fitted in different pathological processes. Something similar might happen with the phenotypes shared by 24L0020_eco with 24L0043_eco and 24L0044_eco.

A relevant aspect that emerges from this analysis is the phenomena of both genetic and phenotypic heterogeneity present in the network. On the one hand, a same observed phenotype can be associated with different genes or diseases, showing genetic heterogeneity and highlighting the necessity of no considering a clinical characteristic as exclusive of a single pathology. On the other hand, a single variant or gene is connected to different phenotypes, which correspond to the phenotypic heterogeneity. The possibility of visualizing at the same time those relationships within the network adds significant value, as it provides a deeper understanding of the complexity of genotype–phenotype interactions and facilitates the formulation of more precise clinical hypotheses.

Nevertheless, the network analysis also presents some limitations that should be taken into account. Firstly, the number of cases integrated in the database is still limited, which conditions the emergence of well-defined and sparsely interconnected modules. As more cases are incorporated, more complexity is expected to be gained by the network, emerging subnetworks with high connection degree. Moreover, the over-representation of certain properties, such as the fact that most of the identified variants are heterozygous, may introduce a bias in the overall connectivity, limiting the diversity of observable patterns. These limitations do not reduce the value of the approach, but they indicate that the full potential can only be achieved with larger and more heterogeneous cohorts.

Finally, the network analysis highlights its clinical value and potential as a tool to support the interpretation of genomic and phenotypic data. The modular structure and the visualization of relationships allow easily recognizing patrons shared by cases (for instance in the Noonan syndrome) enabling the identification of clinical characteristics and their relationship with specific genomic variants. In addition, the integration of phenotype and genotype in the same environment permits the formulation of hypotheses regarding the clinical variability between patients with the same disease, while considering potential influences from genetic, epigenetic, or environmental factors. Lastly, this approach has potential diagnostic value, as it can help prioritize relevant genetic variants and support clinical decision-making, consolidating the network as a useful and versatile tool in both research and clinical practice.

6 Conclusions

6.1 Conclusions

This TFM is a transversal project that, in order to achieve its main goal, (studying relationships between phenotype and genotype in the prenatal clinical context) has addressed several challenges presented in current biomedical research, such as the use of LLMs, standardized ontologies, data interoperability or data representation in knowledge graphs. Several conclusions can be extracted from the project as a whole and from all these problems studied in particular.

Regarding the use of LLMs for detecting ontology terms in real clinical reports, specifically HPO ontology, it has been demonstrated its potential. The model used in this project has been able to identify a large number of phenotypes and assign the corresponding HPO identifiers. Therefore, the model is able to extract relevant clinical information from prenatal reports written in Spanish and associate it with HPO terms describing the patient phenotype, highlighting its practical utility.

However, LLMs still present several challenges. A proper evaluation of the model would have required a gold standard, where clinicians would have determined which terms can be classified as HPO terms and then compare it with the results of the LLM. Furthermore, the detection of relevant terms cannot only be based on the lexical similarity, but also on the context where the words are placed. Despite these limitations, LLMs represent a promising tool to support clinical annotation.

In addition to the findings related to LLMs, this project has demonstrated the importance of structured and standardized data representation through the use of GA4GH Phenopackets. Clinical and genomic information from prenatal reports, which is typically unstructured and heterogeneous, was successfully organized into Phenopacket objects. This structured approach ensures that phenotypic characteristics can be related to their underlying genomic variants, providing a consistent framework for exploring phenotype-genotype correlations.

The use of ontologies and standardized nomenclatures within Phenopackets was vital for achieving unequivocal, interoperable and reused data. The main limitation that can be found is that those ontologies do not cover specific terms in the prenatal context. Nevertheless, Phenopackets constitute a robust framework which, along with LLMs, allow extracting and organizing key information for prenatal investigations and decision-making.

This project has also proved the utility of integrating clinical and genomic information into a graph-based database, allowing data to be organized in a structured manner and efficiently explore relationships between patients, phenotypes, genomic variants and diseases. The representation through nodes and relations facilitates patron identification, indirect associations and genotype-phenotype correlation.

The global network analysis showed genetic and phenotypic heterogeneity phenomena, showing that a same phenotype can be associated with different genes or diseases and vice versa. This allows formulating hypothesis about clinical variability between patients and considering possible genotypic influences. Despite the limited number of cases in this database, the approach shows a high clinical and

research value, facilitating genomic and phenotypic data interpretation, prioritizing relevant variants and support decision-making.

In conclusion, this project demonstrates that the combination of LLMs, structured data frameworks like Phenopackets, and graph-based databases provides a powerful and complementary approach for the analysis of prenatal clinical and genomic information. Together, these tools facilitate the identification of clinically relevant patterns, support hypothesis generation, and enhance decision-making in prenatal research. While limitations remain, this integrated approach lays a robust foundation for future studies, enabling more precise, reproducible, and collaborative investigations in the prenatal biomedical field.

6.2 Future lines

Several avenues can be explored to extend the findings of this project and enhance the study of prenatal genotype–phenotype relationships:

1. Creation of a gold standard for LLM evaluation: designing and validating a set of prenatal reports annotated by experts for adequately evaluating the LLM model in HPO term extraction. This would allow objectively quantifying the reliability of the model.

2. LLM improvement: training or adjusting LLM models for understanding better the clinical context, negative or hypothetical sentences and fetus and mother characteristics distinction.

3. Phenopacket expansion: including other elements from EHR, such as measurements, treatments or medical actions, using automatized pipelines.

4. Integration of greater and more diverse cohorts: increasing the number of cases in the database for exploring more complex networks, detecting emerging submodules and improving genetic and phenotypic heterogeneity.

5. Clinical integration and assisted diagnosis: exploring the use of these tools as a support for clinical decision making, prioritizing relevant variants and detecting critical phenotypes, with potential application in personalized prenatal medicine.

6. Pathological classification: the pathological classification of CNVs is not provided. Following the [8], [34] and the calculator available in webpage, an automated software can help the clinician for this task. The software would be able of consulting in databases such as DECIPHIER if the genomic content contains protein-coding segments, if the region overlaps with established Triplosensitive (TS), Haploinsufficient (HI), or Benign Genes or Genomic Regions or how many genes are affected by the variation.

7 Bibliography

[1] J. O. B. Jacobsen *et al.*, «Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease», *Hum. Mutat.*, vol. 43, n.º 8, pp. 1071-1081, ago. 2022, doi: 10.1002/humu.24380.

[2] A. Schmidt *et al.*, «Next-generation phenotyping integrated in a national framework for patients with ultrarare disorders improves genetic diagnostics and yields new molecular findings», *Nat. Genet.*, vol. 56, n.º 8, pp. 1644-1653, ago. 2024, doi: 10.1038/s41588-024-01836-1.

[3] K. J. Gray, L. E. Wilkins-Haug, N. J. Herrig, y N. L. Vora, «Fetal phenotypes emerge as genetic technologies become robust», *Prenat. Diagn.*, vol. 39, n.º 9, pp. 811-817, ago. 2019, doi: 10.1002/pd.5532.

[4] M. A. Shear, P. N. Robinson, y T. N. Sparks, «Fetal imaging, phenotyping, and genomic testing in modern prenatal diagnosis», *Best Pract. Res. Clin. Obstet. Gynaecol.*, vol. 98, p. 102575, feb. 2025, doi: 10.1016/j.bpobgyn.2024.102575.

[5] F. Mone *et al.*, «Evolving fetal phenotypes and clinical impact of progressive prenatal exome sequencing pathways: cohort study», *Ultrasound Obstet. Gynecol.*, vol. 59, n.º 6, pp. 723-730, jun. 2022, doi: 10.1002/uog.24842.

[6] F. Dhombres *et al.*, «Prenatal phenotyping: A community effort to enhance the Human Phenotype Ontology», *Am. J. Med. Genet. C Semin. Med. Genet.*, vol. 190, n.º 2, pp. 231-242, jun. 2022, doi: 10.1002/ajmg.c.31989.

[7] L.-X. Huang *et al.*, «Case report and a brief review: Analysis and challenges of prenatal imaging phenotypes and genotypes in Joubert syndrome», *Front. Genet.*, vol. 13, p. 1038274, nov. 2022, doi: 10.3389/fgene.2022.1038274.

[8] E. R. Riggs *et al.*, «Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen)», *Genet. Med.*, vol. 22, n.º 2, pp. 245-257, feb. 2020, doi: 10.1038/s41436-019-0686-8.

[9] «View of Comparative Efficacy of CNV-Sequencing and Karyotyping in Prenatal Genetic Diagnosis».

[10] D. Smedley y P. N. Robinson, «Phenotype-driven strategies for exome prioritization of human Mendelian disease genes», *Genome Med.*, vol. 7, n.º 1, p. 81, jul. 2015, doi: 10.1186/s13073-015-0199-2.

[11] D. L. Rubin, N. H. Shah, y N. F. Noy, «Biomedical ontologies: a functional perspective», *Brief. Bioinform.*, vol. 9, n.º 1, pp. 75-90, oct. 2007, doi: 10.1093/bib/bbm059.

[12] M. Amith, Z. He, J. Bian, J. A. Lossio-Ventura, y C. Tao, «Assessing the practice of biomedical ontology evaluation: Gaps and opportunities», *J. Biomed. Inform.*, vol. 80, pp. 1-13, abr. 2018, doi: 10.1016/j.jbi.2018.02.010.

[13] N. A. Vasilevsky *et al.*, «Mondo: Unifying diseases for the world, by the world», 16 de abril de 2022, *Health Informatics*. doi: 10.1101/2022.04.13.22273750.

[14] S. Köhler *et al.*, «The Human Phenotype Ontology in 2021», *Nucleic Acids Res.*, vol. 49, n.º D1, pp. D1207-D1217, ene. 2021, doi: 10.1093/nar/gkaa1043.

[15] M. A. Gargano *et al.*, «The Human Phenotype Ontology in 2024: phenotypes around the world», *Nucleic Acids Res.*, vol. 52, n.º D1, pp. D1333-D1346, ene. 2024, doi: 10.1093/nar/gkad1005.

[16] M. Mohtashamian, R. Abeysinghe, X. Hao, y L. Cui, «Identifying Missing IS-A Relations in Orphanet Rare Disease Ontology», en *2022 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA: IEEE, dic. 2022, pp. 3274-3279. doi: 10.1109/BIBM55620.2022.9995614.
[17] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, y S. Ayme, «Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users», *Hum. Mutat.*, vol. 33, n.º 5, pp. 803-808, may 2012, doi: 10.1002/humu.22078.
[18] K. A. Shefchek *et al.*, «The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species», *Nucleic Acids Res.*, vol. 48, n.º D1, pp. D704-D715, ene. 2020, doi: 10.1093/nar/gkz997.
[19] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, y A. Hamosh, «OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders», *Nucleic Acids Res.*, vol. 43, n.º D1, pp. D789-D798, ene. 2015, doi: 10.1093/nar/gku1205.
[20] J. Castell-Díaz, J. A. Miñarro-Giménez, y C. Martínez-Costa, «Supporting SNOMED CT postcoordination with knowledge graph embeddings», *J. Biomed. Inform.*, vol. 139, p. 104297, mar. 2023, doi: 10.1016/j.jbi.2023.104297.
[21] J. Foreman *et al.*, «DECIPHER: Supporting the interpretation and sharing of rare disease phenotype-linked variant data to advance diagnosis and research», *Hum. Mutat.*, p. humu.24340, feb. 2022, doi: 10.1002/humu.24340.
[22] D. Trujillano, G. Oprea, Y. Schmitz, A. M. Bertoli-Avella, R. Abou Jamra, y A. Rolfs, «A comprehensive global genotype–phenotype database for rare diseases», *Mol. Genet. Genomic Med.*, vol. 5, n.º 1, pp. 66-75, ene. 2017, doi: 10.1002/mgg3.262.
[23] B. Li *et al.*, «GPCards: An integrated database of genotype–phenotype correlations in human genetic diseases», *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1603-1611, 2021, doi: 10.1016/j.csbj.2021.03.011.
[24] J. Kim, K. Wang, C. Weng, y C. Liu, «Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease», *Am. J. Hum. Genet.*, vol. 111, n.º 10, pp. 2190-2202, oct. 2024, doi: 10.1016/j.ajhg.2024.08.010.
[25] L. Liang *et al.*, «Genetic Transformer: An Innovative Large Language Model Driven Approach for Rapid and Accurate Identification of Causative Variants in Rare Genetic Diseases», 19 de julio de 2024, *Genetic and Genomic Medicine*. doi: 10.1101/2024.07.18.24310666.
[26] F. J. Moreno-Barea *et al.*, «Named entity recognition for de-identifying Spanish electronic health records», *Comput. Biol. Med.*, vol. 185, p. 109576, feb. 2025, doi: 10.1016/j.compbiomed.2024.109576.
[27] M. Baddour, S. Paquelet, P. Rollier, M. De Tayrac, O. Dameron, y T. Labbe, «Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era», en *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, Varna, Bulgaria: IEEE, ago. 2024, pp. 1-8. doi: 10.1109/IS61756.2024.10705235.
[28] L. C. Seller, «Named Entity Recognition of Human Phenotype Ontology Concepts in Spanish Clinical Texts».
[29] M. S. Ladewig *et al.*, «GA4GH Phenopackets: A Practical Introduction», *Adv. Genet.*, vol. 4, n.º 1, p. 2200016, mar. 2023, doi: 10.1002/ggn2.202200016.
[30] D. Danis *et al.*, «A corpus of GA4GH phenopackets: Case-level phenotyping for genomic diagnostics and discovery», *Hum. Genet. Genomics Adv.*, vol. 6, n.º 1, p. 100371, ene. 2025, doi: 10.1016/j.xhgg.2024.100371.
[31] D. Danis *et al.*, «Phenopacket-tools: Building and validating GA4GH Phenopackets», *PLOS ONE*, vol. 18, n.º 5, p. e0285433, may 2023, doi: 10.1371/journal.pone.0285433.

[32] «What is a Phenopacket? — phenopacket-schema 2.0 documentation». Accedido: 12 de septiembre de 2025. [En línea]. Disponible en: https://phenopacket-schema.readthedocs.io/en/latest/basics.html
[33] *phenosentry: A tool and library for quality control and validation of GA4GH Phenopackets*. Python.
[34] H. M. Kearney, E. C. Thorland, K. K. Brown, F. Quintero-Rivera, y S. T. South, «American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants», *Genet. Med.*, vol. 13, n.º 7, pp. 680-685, jul. 2011, doi: 10.1097/GIM.0b013e3182217a3a.