Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# RarePathAI: Predictive Analytics to Identify Rare Diseases in Real-World Data

**Pierantonio Russo**

parusso@eversana.com

EVERSANA Life Sciences

**Ramaa Nathan**

EVERSANA Life Sciences

**Alex Moore**

EVERSANA Life Sciences

**Qian Zhu**

National Institutes of Health

**Eric W.K. Sid**

National Institutes of Health

**Alice Chen Grady**

National Institutes of Health

**Oodaye Shukla**

EVERSANA Life Sciences

**Research Article**

**Keywords:**

**Posted Date:** June 28th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9247937/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

There are millions of individuals affected by a rare disease (RD) in the United States [1]. Yet, many remain undiagnosed or misdiagnosed due to the low prevalence of individual conditions, inconsistent medical coding, highly variable symptoms, and limited clinician familiarity with rare diseases.

We present a new framework that integrates U.S. administrative claims data (~ 320 million records), rare disease ontologies [2, 3], and graph neural networks (GNNs) to support earlier diagnosis of RD through clinician-facing decision tools.

Using 98 curated rare disease cohorts (each with at least one year of pre-diagnosis history and three or more clinical encounters), we built a graph structure comprised of clinical reference data from the ontologies in the form of embedding vectors and the clinical patient journey from claims data to support predictive model development.

We trained a two-step model to provide a likely diagnosis: a binary classifier (whether or not the patient could have any rare disease) (AUC = 0.951), followed by a multiclass model to suggest a set of more specific rare disease diagnoses, achieving a Top 10 weighted recall of 74.8%.

In the future, we envision an AI-Agentic orchestration layer that empowers healthcare providers with real-time, AI-driven support at the point of care. Such a system would recommend diagnosis of potential rare diseases based on clinical observations, while coordinating specialized AI agents to generate explainable insights, prioritized diagnostic suggestions, and interactive guidance. By doing so, it has the potential to shorten diagnostic journeys, direct timely targeted testing (including genetic workups), and facilitate earlier, more effective interventions for patients with rare conditions.


# Introduction

More than 10,000 rare diseases (RDs) collectively impose a significant global health burden, affecting approximately 3.5% to 5.9% of the worldwide population [4, 5]. Despite this substantial collective prevalence, the diagnostic journey for patients is often protracted and a phenomenon widely known as the "diagnostic odyssey" [6, 7, 8]. This delay, frequently spanning multiple years and involving numerous specialist consultations, stems from low individual disease prevalence, nonspecific or heterogeneous symptom presentation, and a general lack of front-line clinician awareness [9]. The delayed diagnosis leads to increased patient morbidity, care giver distress, substantial economic costs [10] and missed opportunities for early therapeutic intervention in the fewer than 5% of RDs that have an FDA-approved treatment [2].

A primary obstacle to timely RD diagnosis is the limitation of existing healthcare data and coding practices [11, 12, 13, 14]. More than half of all known RDs do not have specific International Classification of Diseases (ICD-10) codes, limiting their visibility and recognition within the health care system [10]. This informational gap forces reliance on complex combinations of nonspecific diagnostic

codes to approximate a patient's phenotype, a process that is often challenging. Furthermore, the integration of standardized, granular phenotype terminologies like the Human Phenotype Ontology (HPO) and Orphanet with real-world clinical data captured in administrative claims or electronic health records (EHRs) remains difficult [3].

To address these challenges, several studies have introduced rule-based algorithms, machine learning [15, 16] and computational methods [17] to identify potential RD cases from real-world data (RWD). More recently, advanced learning methodologies have shown the ability to model high-dimensional patient data without extensive manual feature engineering [18, 19]. These approaches integrate diverse data, including diagnostic codes, procedures, and medication to construct a comprehensive, longitudinal representation of a patient's clinical journey [20, 21, 22, 23, 24, 25]. While these models capture a high volume of information, latent temporal features and relationships between symptoms and related disease states may still be missed by these methods.

In this work, we developed proof-of-concept computational frameworks and scalable methods to improve rare disease patient identification. Our machine learning models were designed to flag potential undiagnosed or misdiagnosed cases by analyzing individual diseases, related condition clusters, and key clinical patterns. We designed these models for potential future integration into clinical decision support systems to improve diagnostic precision and reduce time to diagnosis across healthcare settings. We propose a novel framework based on graph neural networks (GNNs), a class of deep learning models well-suited for analyzing complex healthcare data [26, 27]. By representing each patient as a subgraph composed of clinical encounters, diagnoses, and procedures embedded within a knowledge graph, our approach captures relationships that traditional models may miss [26, 27]. This graph-based learning strategy is integrated with domain ontologies and RWD in the form of claims-based patient cohorts, to enable context aware rare disease detection.

# Methodology

1. **Design of RarePathAI**

The overall architecture for RarePathAI as envisioned is shown in Fig. 1. In its present implementation, this architecture incorporates the following components:

**Data Ecosystem**

The data ecosystem comprises RWD, or transactional healthcare data and Reference Data. RWD data is the Symphony Integrated Dataverse offering, and the Reference Data is from both Orphanet [2] and the Human Phenotype Ontology (HPO) [3]. The RWD and Reference Data sources are integrated via their common terms and text descriptions, in a process detailed further in the Graph Schema section.

1. **Symphony Open Claims**: The study was completed using a large United States administrative healthcare database, the Symphony Health Integrated Dataverse. This database aggregates data

from health insurance organizations, hospital systems, and individual providers that covers approximately 320 million unique individuals across pediatric and adult populations. It includes inpatient and outpatient encounters and captures claims from commercial and public health insurance data sources. The data used for this study was from July 2018 through June 2025 with sufficient patient-level longitudinal histories of medical, procedural, and pharmacy claims to support the identification of rare and ultra rare disease populations.

2. **Orphanet**: Orphanet [2] provides high-quality reference information and essential data for rare disease research, as well as a structured vocabulary that is widely integrated into bioinformatics projects and infrastructures worldwide to improve diagnosis and treatment. The Orphanet Scientific Knowledge Base is organized around the Orphanet nomenclature of rare diseases, a dedicated and globally-recognized standard that improves data interoperability across healthcare and research information systems. In this work, the December 2024 Orphanet rare disease alignments and epidemiology datasets were incorporated to support disease exploration and the subsequent mapping of rare diseases into the knowledge graph.
3. **Human Phenotype Ontology (HPO)**: Human Phenotype Ontology (HPO) [3] provides a standardized vocabulary of describing phenotypic abnormalities observed in human disease and is linked to the Orphanet rare disease data, along with associated frequency information. The HPO definitions, identifiers, and frequencies were incorporated as core components of the derived knowledge graph that provides additional insight in detecting a rare disease patient during diagnosis.

**Data Embedding Processes**

The data embedding process was designed to encode information from the reference datasets into vector representations as node features for predictive modeling. BioBERT [28] and ProNE [29] embedding algorithms were applied to ICD-10 code descriptions, Orphanet rare disease entities, and HPO terms. BioBERT is applicable to medical text, while ProNE to fixed network graph structure. They can only be applied to the reference datasets.

**Graph Database**

A graph database was implemented to capture the RWD relevant entities, such as diagnosis codes, procedure code, patient demographics, encounters, and Reference Data entities as nodes. Relationships among these entities (nodes) were represented as graph edges. The Methodology section provides a detailed description of the graph schema and summarizes the number of nodes and edges used to encode data across the 98 rare diseases evaluated in this study.

**Rare Disease Predictive Models**

Two rare disease predictive models were developed within a stacked model approach, a binary classifier and a multi-class classifier (98 classes). The binary classifier first differentiates between rare disease and non-rare disease patients, and a multi-class classification model, then subsequently provides likelihood based diagnostic scores for 98 targeted rare diseases.

## 2. Cohort Selection and Construction

Orphanet and HPO data [2, 3] was systematically explored to identify rare diseases with a prevalence of 1–50 per 100,000 in the United States population that were suitable for cohort construction within the Symphony Open Claims database. In addition, each selected disease was required to have a sufficient number of symptoms commonly observed by physicians, to ensure RarePathAI was trained on the phenotypic features most relevant to frontline clinical practice.

The December 2024 Orphanet, and the December 2024 Human Phenotype Ontology (HPO) data was parsed to extract the following information for all available Orphanet rare diseases:

- OrphaCode
- Disease name
- Prevalence class
- ICD-10 code
- Associated HPO IDs and terms

This compiled data was filtered to retain diseases with prevalence rates between 1 and 50 per 100,000, based on the Orphanet prevalence classification. Among these, priority was given to diseases associated with at least three commonly observed symptoms reported by physicians in the United States, including developmental delay, cognitive impairment, social/behavioral disorders, intellectual disabilities, seizures, growth abnormalities, weakness and fatigue, cardiovascular anomalies, respiratory difficulties, feeding difficulties, vomiting, diarrhea, constipation, hearing impairment, vision problems, musculoskeletal abnormalities, skin abnormalities, difficulty walking, ataxia, and metabolic abnormalities [30, 31]. These symptoms were identified using corresponding HPO terms annotated as “Very frequent” or “Frequent” for each rare disease. The filtered diseases were then manually reviewed to confirm the presence of identifiable and unambiguous ICD-10 codes that would not be associated with any other rare diseases. Additional rare disease cohorts designed in the future may require additional specification for accurate identification in RWD. The overall rare disease selection workflow can be seen in Fig. 2.

This analysis generated a list of 98 rare diseases that formed the patient cohorts used in this study. For model training, only medical encounters occurring prior to the rare disease diagnosis date were included. Accordingly, each patient was required to have at least one year of medical claims history in the Symphony data prior to diagnosis, and a minimum of three clinical encounters during that period. The majority of patient cohorts were defined using unambiguous ICD-10 codes specific to each rare disease, but for some, disease-specific treatments were used. Cohort definitions, including the ICD-10 codes or treatments used to establish diagnosis dates, can be found in Supplemental Table A1. For cohorts defined using ICD-10 codes, a diagnosis required at least two occurrences of the relevant code, at least 60 days apart from each other. The earliest diagnosis date was used as the rare disease diagnosis date for study purposes. Ensuring 60 day interval between diagnosis codes reduced the likelihood of false positive case identification due to provisional or suspected diagnoses without confirmed disease status.

# 3. Graph Schema

A traditional machine learning approach, using relational databases, would represent a patient as a series of clinical features and separate models might be trained to predict different diseases for that patient. This would require extensive training to differentiate among numerous rare diseases using hundreds or thousands of independent clinical features. In contrast, a graph-based machine learning approach using graph neural networks (GNNs) enables the development of a unified model capable of predicting multiple rare diseases simultaneously. This approach takes advantage of graph structures to capture the complex relationships among patients, diseases, symptoms, procedures, and other clinical entities. Through embeddings, these heterogeneous components are represented in a shared latent space, allowing predictive models to gain more information about the similarities between patients and diseases [26, 27]. An initial graph schema was developed to restructure the claims data for each rare disease patient into a graph representation (Fig. 3). Nodes in this schema are organized into two distinct categories, those derived from RWD with their associated properties summarized in Table 1, and those originating from Reference Data with their properties summarized in detail in Table 2.

Table 1

Complete descriptions of all node types derived from Real-World Data, with property and edge details.

| Node Type | Description and Properties |
|---|---|
| Patient | - Containing demographic information such as sex and date of birth year<br>- Connected via edges to Encounter nodes |
| Encounter | - A service date representation of a patient's medical interaction in the time prior to their diagnosis<br>- Temporally linked via "NEXT" edges<br>- Contains a patient's age at the time of the encounter<br>- Includes a temporal weight feature representing the relative time the encounter occurred relative to the index date of the patients' rare disease diagnosis<br>- Temporal weight is a value given to an encounter based on its bucketed time window prior to the rare disease diagnosis (0–3 months, 3–6 months, 6–9 months, 9–12 months, and 12 + months) in increments of -0.25<br>- Connected via edges to Specialty, Diagnosis, Procedure, and Age Range nodes |
| Diagnosis | - Code level representation of an ICD-10 diagnosis<br>- Connected via edges to Encounter nodes |
| Specialty | - Code level representation of a physician specialty<br>- Connected via edges to Encounter nodes |
| Procedure | - Code level representation of a CPT/HCPCS procedure<br>- Connected via edges to Encounter nodes |
| Age Range | - Representation of the age range a patient was in at the time of an encounter<br>- Connected via edges to Encounter and Significant Age nodes |

Table 2
Complete descriptions of all node types derived from Reference Data, with property and edge details.

| Node Type | Description and Properties |
|---|---|
| Rare Disease | - Containing reference OrphaIDs corresponding to rare diseases, to connect to other data sources<br>- The corresponding vector embeddings (BioBERT and ProNE) are added as additional node properties.<br>- Connected to via edges to HPO Term and Significant Age nodes |
| HPO Term | - Containing HPO Terms which are linked to the Rare Disease of interest<br>- The corresponding vector embeddings (BioBERT and ProNE) are added as additional node properties.<br>- Connected via edges to Rare Disease and Diagnosis nodes<br>- Edges to Rare Disease nodes are weighted using the frequency level defined within HPO and the uniqueness of the HPO term amongst the grouping of most similar diseases. |
| Significant Age Triplet<br>(Age-ICD-RD) | - Containing triplet-based information on significant ages at which certain symptoms are observed when associated with certain Rare Diseases<br>- Connected via edges to Diagnosis, Age, and Rare Disease node types |

To create the connections between the RWD and the Reference Data, precise cohort definitions were required to correctly label patients with rare diseases and to examine how HPO terms are linked to the observed ICD-10 signs and symptoms documented in the patients' journeys prior to rare disease diagnosis.

A key aspect of constructing the graph was establishing connections between HPO terms and ICD-10 codes. To support this mapping, we adopted prior work from groups such as McArthur et al. [32], who used the Unified Medical Language System (UMLS), WikiMedMap, and string and substring-matching techniques to establish connections between HPO terms and groups of ICD-9 and ICD-10 codes. Those links between HPO terms and ICD-10 codes were directly integrated into the graph database for usage in this research framework.

Symptoms of rare diseases manifest within specific age range. A key feature in identifying a rare disease is the age at which symptoms present, an aspect that is sparsely documented in available reference datasets. To address this gap, we introduced "age triplet" nodes connecting symptoms (HPO terms), age ranges, and Orphanet labeled rare diseases present in the graph network here. Programmatic prompts were designed and sent to the large language model (LLM) Claude 3.5 to generate tables of data labeling the typical age ranges at which specific HPO symptoms manifest in association with corresponding rare diseases. These tables of generated data were then used to augment the reference data nodes as seen in Fig. 3.

Prompts were initialized with role specification, namely, “You are a medical informatics and epidemiology expert with comprehensive knowledge of medical classification ontologies such as ICD-10 codes, HPO (human phenotype ontology) terms, ORPHA rare-disease codes, and HCPCS procedure codes.” Responses were required to be returned in structured JSON format, and instructed to be grounded as “base responses on current medical knowledge and guidelines”, as well as to “return factual, clinically relevant information”. The prompt for the age ranges associated with given pairs of rare diseases and HPO terms then continues “For each ORPHA and HPO term combination given, determine the typical age ranges when the HPO feature would manifest in patients who have been diagnosed or eventually will be diagnosed with that disease.” Predefined age categories included: 0–1 years, 2–4 years, 5–12 years, 13–17 years, 18–39 years, 40–64 years, and 65 + years. Multiple age ranges could be selected for each pair of rare disease and HPO term.

## 4. Node Features

Reference data nodes also include additional embeddings, not directly related to the disease classification models, as features of the nodes themselves. Two types of embedding were incorporated into the graph schema:

- Text Embeddings, generated using the BioBERT [28] model from descriptions of a given Diagnosis, Procedure, HPO Term, or Rare Disease node
- Reference Graph Embeddings, generated using the PRoNE [29] algorithm from a fixed knowledge graph composed of Diagnosis, HPO Term, and Rare Disease node types and their interconnecting edges

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a language representation model pre-trained on large-scale biomedical corpora, an extension of the general BERT model to better account for the word distribution shift in biomedical texts [28]. In this study, expanded descriptive text for each Diagnosis, Procedure, HPO Term, and Rare Disease node was generated using Claude 3.5. Prompts were designed similarly to those used for the age range extraction, specifying medical expertise and factual information. Instead of requesting the age ranges of manifestation, the model was instructed to generate a 5–10 sentence description of the Diagnosis, Procedure, HPO Term, or Rare Disease. The expanded description was then processed by BioBERT to generate 768-dimensional floating point embeddings. This approach enables nodes with semantically similar descriptions to occupy nearby positions in embedding space, capturing latent description similarities.

ProNE, a propagation-based network embedding algorithm, is one example of a technique to pre-train embeddings on large-scale pre-existing networks, along with the likes of LINE, DeepWalk, and node2vec [29]. ProNE achieves high efficiency in this task by initializing network embeddings as sparse matrix factorizations, then enhancing the embeddings by propagating them in spectrally modulated space. In this work, the ProNE algorithm was used to generate embeddings of length 512 floating point values

from the graph network consisting of Diagnosis, HPOTerm, and RareDisease nodes and their inter-connecting edges (see “From reference data” in Fig. 3). These embeddings capture relationships among signs, symptoms, and rare diseases within the curated Reference Data network.

Comparative evaluation against several alternative network embedding algorithms, such as DeepWalk and node2vec, demonstrated that ProNE performed the best in terms of capturing Orphanet-listed differential diagnoses in close proximity under L2 distance between RareDisease node embeddings. Consequently, ProNE embeddings were then used as the reference graph embedding features within the final graphs used for predictive model training.

## 5. Graph Neural Network Model Design

The model presented here utilizes graph neural networks (GNNs) trained on a combination of RWD and Reference Data to generate rare disease diagnostic predictions [26, 27]. Similar approaches trained on biomedical reference data and simulated patient records have shown strong performance across diagnostic tasks, such as causal gene discovery and retrieval of “patients-like-me based on phenotypic and genotypic data knowledge [33]. The framework presented incorporates RWD, reflecting actual delivery of healthcare services rendered by front-line physicians.

Data corresponding to each patient was fed into the model by extracting their relevant subgraph including encounters, diagnosis codes, procedures, and HPO terms derived from their medical history. These patient subgraphs were then processed by a graph convolutional neural network built in PyTorch using GraphConv layers [34]. The network is composed of nine message passing layers, each with 256 channels. Within each layer, messages are transformed through a hyperbolic tangent (Tanh) activation function and aggregated via the mean. The output of the heterogeneous graph neural network is in the form of a vector embedding, representing the patient and their relationships based on their encounters, diagnoses, procedures, and subsequent HPO terms and rare diseases from the Reference Data. These neural network embeddings were then ultimately used as inputs to the downstream predictive classifiers, which generate the likelihood scores for rare disease diagnoses. The number of layers and channels, as well as the activation and aggregation function types were optimized via iterative grid search during model development.

A stacked predictive modeling approach, seen in Fig. 4, was implemented to generate the final rare disease prediction. The framework consists of two sequential models, a binary classification model to predict if a patient is likely to have a rare disease and its output was subsequently fed into the second model, a multi-class classification model to estimate the likelihood of specific rare diseases. The two models share the same model configuration detailed previously, but were trained using different sets of parameters and patient cohorts.

The first stage binary classification model predicts if a patient has one of the 98 rare diseases or not. The model is trained on a target cohort including all patients diagnosed with any of the 98 rare diseases

and the control cohort consisting of a general population demographically matched on age and sex, with no clinical codes in their claims history associated with any of the ICD codes that are used to identify the 98 rare diseases. Additionally, patients were excluded from the control cohort if they had ICD codes identified in Orphanet as associated with rare diseases. The reference date for patients in the target cohort were defined as the earliest rare disease diagnosis date, while the reference date for the patients in the control cohort were determined randomly based on an encounter that ensured at least one year of prior claims history.

The second stage model is a muti-class classification model, trained on the 98 rare disease cohorts. This model was trained using patients in the “rare disease” population, and outputs a likelihood score for each of the targeted 98 rare disease.

## 6. Feature Importance

A key metric of the model interpretation and validation is the assessment of feature importance, identifying the features contribute most strongly to a model’s prediction. In the case of a graph-based classification model, a commonly used method for measuring node importance is perturbation analysis. Perturbation is the removal of a node from the graph, and re-calculation of the classification scores, and subsequent measurement of the average change in the score from the baseline with that node removed. The importance of each node is then quantified by measuring the average change in score from the baseline obtained from the original graph.

## Results

## 1) Real-World Data & Reference Data

The final counts for each node type and edge type including groupings for Rare Disease patients and the Non-Rare Disease control patients where applicable, are summarized in Table 3. The final multi-class classification model was trained on a graph containing over 40 million nodes and more than 350 million edges.

Table 3
Counts of each Node Type and Edge Type present in the final graph, broken down by Rare Disease and Non-rare Disease where applicable.

| Node Type | Node Count | |
|---|---|---|
| | Node Count (Rare Disease) | Node Count (Non-Rare Disease) |
| Patient | 309,940 | 1,053,224 |
| Encounter | 17,250,385 | 24,895,099 |
| Diagnosis | 103,241 | |
| Age Range | 7 | |
| Specialty | 123 | |
| Procedure | 28,622 | |
| Age-ICD-RD Triplet | 119,463 | |
| HPO Term | 19,435 | |
| Rare Disease | 3,879 | |

| Edge Type | Edge Description | Edge Count | |
|---|---|---|---|
| | | Edge Count (Rare Disease) | Edge Count (Non-Rare Disease) |
| HAS_ENCOUNTER | Connects Patients to their Encounters | 17,250,385 | 24,895,099 |
| NEXT | Connects an Encounter to that Patient's next Encounter | 16,940,445 | 23,841,875 |
| HAS_DIAGNOSIS | Connects an Encounter to an included Diagnosis | 61,573,996 | 40,445,626 |
| HAS_AGE_RANGE | Connects an Encounter to the relevant Age Range | 17,250,385 | 24,895,099 |
| PROVIDED_ENCOUNTER | Connects an Encounter to the relevant Specialty | 17,250,385 | 24,895,099 |
| HAS_PROCEDURE | Connects an Encounter to an included Procedure | 61,811,728 | 45,212,383 |
| IN_AR | Connects a Diagnosis to the relevant Age-ICD-RD Triplet | 119,463 | |
| IS_SIGNIF_AGE | Connects an Age Range to the relevant Age-ICD-RD Triplet | 119,463 | |
| IS_SIGNIF_RD | Connects a Rare Disease to the relevant Age-ICD-RD Triplet | 119,463 | |
| MATCHED_TO_ICD | Connects HPO Term to any matched Diagnosis | 31,349 | |
| HAS_HPOTERM | Connects Rare Disease to any prevalent HPO Term | 100,921 | |
| HAS_PARENT_ICD | Connects Diagnosis to any parent Diagnosis based on ICD hierarchy | 101,959 | |
| HAS_PARENT_HPO | Connects HPO Term to any parent HPO Term based on HPO hierarchy | 23,409 | |

An important outcome of this work is to examine the overlap between what is observed in real-world data and what is reported in the reference data for rare disease patients. Table 4 summarizes the proportion of the 98 targeted rare disease patient cohorts who exhibit an HPO term associated with their labeled rare disease, grouped by the HPO frequency categories. Because the predictive models were trained using only the pre-diagnosis data, the average proportion is provided for both the pre-diagnosis period as well as the patients' entire claims history.

From these results, we see that the expected trend of decreasing proportion for the "Very frequent", "Frequent", and "Occasional" labels, holds for the rare disease patients, while the actual percentage

values are not necessarily the same as the HPO labeled percentages. This highlights limitations of the claims based matching approach, such as imperfect matching between HPO term and ICD-10 codes, and some symptoms and signs being more likely to be reported in administrative claims compared to others.

Table 4
Table of the % of patients to exhibit HPO Terms of a given frequency type, averaged over the patients from one of the 98 rare diseases targeted for this study, for both the pre-diagnosis encounters used in the predictive models, as well as the all-time claims history for each rare disease patient.

| HPO Frequency Categories | Avg % of Patients from Claims Data (Pre-Diag.) | Avg % of Patients from Claims Data (All Time) | # of Disease-Term Matches from HPO/ Orphanet Reference Data |
|---|---|---|---|
| Very frequent (99 – 80%) | 6.2% | 12.9% | 376 |
| Frequent (79 – 30%) | 4.7% | 9.1% | 590 |
| Occasional (29 – 5%) | 3.3% | 6.3% | 800 |

# 2) Predictive Models

The first stage binary prediction model achieved a receiver operating characteristic area under the curve (ROC-AUC) of 0.951 on held-out test set, as shown in Fig. 5. The figure also shows the corresponding curves for true positive rate, false positive rate, and precision, reflecting the approximate 4:1 ratio of negative to positive patients in the dataset. All models were developed using train-test split of 80%/20%. Here, rare disease patients were labeled as the positive class, while general population patients were labeled as the negative label. Two decision thresholds highlighted the trade-offs between sensitivity and specificity. At the lower threshold, the model achieved a true positive rate of 99%, but a false positive rate of 63%. At the higher threshold, the false positive rate decreased to 5% while the true positive rate remained relatively high at 84%. These thresholds represent a practical range for deployment, with the final implementation threshold decided by the impact of true positives vs. false positives within a given the health system.

The results of the second stage multiclass classifier are summarized in Table 5, which reports average recall values for variable length lists of the most likely rare diseases predicted for each patient. A correct prediction is the patient's true rare disease label appears anywhere within the top "X" ranked rare diseases generated by the model. Two summary metrics are reported, macro average recall, which represents the average recall across all diseases regardless of cohort size, and weighted average recall, which accounts for the number of patients in each disease cohort. The results indicate that to achieve a per-patient accuracy level of 60–80%, the model is most effectively used to generate lists of the 3–15 most likely rare diseases for supporting healthcare providers by narrowing the diagnostic search space and guiding further clinical evaluation.

Table 5

Macro and weighted average recall results for variable length lists of the most likely rare diseases predicted for each test set patient. In each case, a patient is considered positively labeled if their true label appears anywhere in the top “X” most likely rare diseases, based on model score.

| Rare Disease | Top 1 RD | Top 2 RD | Top 3 RD | Top 5 RD | Top 10 RD | Top 15 RD | Top 20 RD |
|---|---|---|---|---|---|---|---|
| **Macro Average Recall** | 20.0% | 28.0% | 33.7% | 40.2% | 50.8% | 58.0% | 63.5% |
| **Weighted Average Recall** | 41.7% | 52.3% | 58.7% | 65.6% | 74.8% | 80.3% | 84.2% |

Another key result is the measurement of feature importance towards model decision making. In this model, nodes associated with key symptoms and signs were measured for their contribution to predictions using the perturbation approach outlined in the Feature Importance section. Examples of top-ranked diagnostic features identified by this metric for three rare diseases in the multiclass classification model are shown in Table 6.

Table 6

Top ranked ICD-10 features by the importance metric measured by average change in score from baseline when removing the listed feature node from the graph used to evaluate the test set patients.

| Diagnosis Feature Rank | Friedreich ataxia | Prader-Willi syndrome | Phenylketonuria |
|---|---|---|---|
| **1** | Ataxia, unspecified | Unsp lack of expected normal physiol dev in childhood | Disorders of sulfur-bearing amino-acid metabolism, unsp |
| **2** | Early-onset cerebellar ataxia, unspecified | Other nonspecific abnormal finding of lung field | Unspecified convulsions |
| **3** | Dysphagia, unspecified | Pectus excavatum | Attention-deficit hyperactivity disorder, unspecified type |
| **4** | Muscle weakness (generalized) | Developmental disorder of speech and language, unspecified | Difficulty in walking, not elsewhere classified |
| **5** | Other hereditary ataxias | Unspecified asthma, uncomplicated | Failure to thrive (child) |

# Discussion

The proposed system addresses a critical gap in rare disease care: frontline clinicians often lack effective tools to support timely triage of rare diseases. By integrating structured ontologies like Orphanet and HPO with unstructured data embeddings derived from BioBERT, the model produces clinically relevant outputs. The inclusion of real-world claims data further provides population-level

context, helping move rare disease diagnosis out of isolated research settings and closer to routine clinical practice.

Graph neural networks (GNNs) are particularly advantageous in characterizing and detecting rare diseases. Traditional machine learning techniques often fail to capture the interdependent and temporal nature of patient journeys, especially when symptom patterns are heterogeneous. In contrast, GNNs capture complex relationships between data points by passing information through connected nodes and continuously learning from the data. This enables the discovery of patterns without the need for manually engineered feature. A crucial innovation in our study is the introduction of an age-weighted encounter feature, which stratifies clinical encounters by patient age and allows the model to simulate clinical reasoning about when symptoms appeared during disease progression. In addition, age-specific symptom nodes derived from large language model (LLM) outputs also bridges a gap in reference ontologies, which typically lack age-specific manifestations.

To facilitate the application of the predictive model outputs into clinical practice, in the future, we envision the use of an AI agent-based orchestration layer. The agentic framework would create an interactive interface for clinicians not trained in the diagnosis and management of rare diseases to recommend the clinical differential diagnosis of the most prevalent rare diseases.

The AI agent architecture proposed for deployment marks an improvement in user interaction. Most clinical decision support tools are either static dashboards or rule-based alerts embedded in EHRs. The agent-based orchestration proposed here introduces continuous learning, making the system responsive to the individual clinician needs.

The framework also holds promise for operational deployment. The modular graph schema supports adaptation to new diseases or new data. The AI agent system can be expanded with disease-specific reasoning engines or integrated into multi-modal platforms incorporating genomics, lab data, or patient-reported outcomes. The binary-multiclass model also supports a triage where clinicians receive a ranked differential list an intuitive aid aligned with real clinical reasoning aligned with the differential diagnosis process.

By integrating AI-driven decision support within clinical workflows, the framework enables earlier identification of rare disease cases, facilitates more targeted diagnostic evaluations, and supports timely additional diagnostic tests like genetic tests, likely reducing diagnostic delays and low value-based care

## Limitations

Claims data lacks granularity and therefore the model may miss phenotypes manifesting only in unstructured formats. While HPO mappings to ICD-10 were strengthened using prior literature and Claude-generated associations, some semantic mismatches are inevitable. Therefore, clinical validation is essential to ensure the model's reliability, safety, and effectiveness in real-world settings. Clinical

validation is currently planned to assess performance across diverse populations and clinical environments and to ensure integration into EMR systems that remains critical for provider adoption.

The current model structure is trained on a graph of more than 40 million nodes and more than 400 million edges to make predictions based on the identification of 98 rare diseases. If this were to be scaled up at the same rate to 1000 diseases for example, the graph would then be approximately 400 million nodes and 4 billion edges, a model-training process, that, while large, would be manageable with sufficient compute resources available today. An additional consideration to scaling the model up would be the design of the additional patient cohorts which may not always have unambiguous ICD-10 codes associated, where phenotypic matching algorithms and clinical confirmation would need to be properly utilized.

## Conclusion

Graph-based modeling of rare disease patient trajectories combined with AI agentic orchestration offers a scalable, interpretable, and impactful method to reduce diagnostic odysseys. With ongoing validation and clinical integration, such tools can transform rare disease detection in routine care.

## Declarations

**Research funding**

P.R., R.N., A.M., and O.S. received funding via contract with the National Institutes of Health, requisition #6949555.

**Competing interests:**

The authors declare no competing interests that might be perceived to influence the results and/or discussion reported in this paper.

# Ethics

Not applicable

# Author Contribution

P.R., R.N., A.M., and O.S. performed the research, wrote the main manuscript text and prepared manuscript figures. A.G.C., E.W.K.S., and Q.Z. made substantive contributions to research design and manuscript text. All authors reviewed the manuscript.

# Acknowledgement

Acknowledgement for Ainslie Tisdale, MPH

# Data Availability

The administrative claims data underlying this study are not publicly available due to contractual agreements governing the use of the claims database sourced from Symphony Health. Access may be granted upon reasonable request, subject to approval by the data provider. Disease classification resources were obtained from Orphanet, which is publicly available.
https://github.com/Orphanet/Orphadata_aggregated

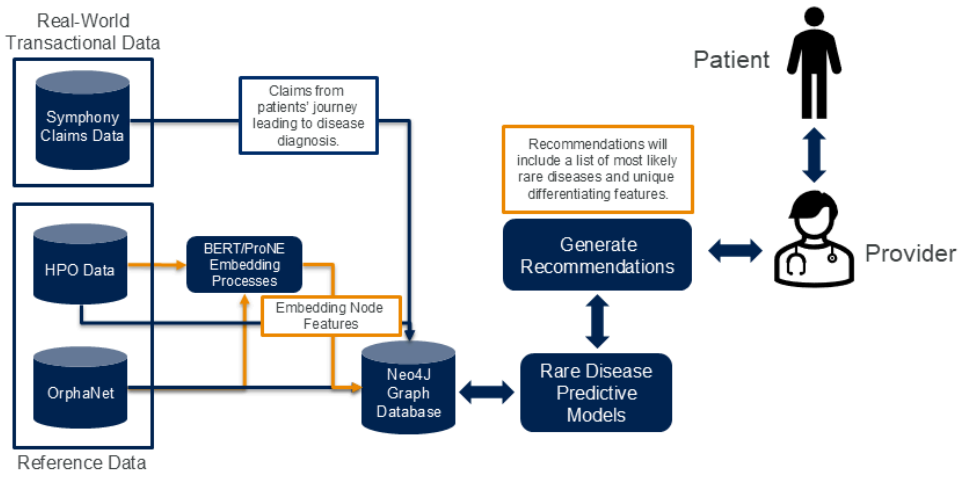
## References

1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28:165–73. 10.1038/s41431-019-0508-0.
2. Orphanet: an online rare disease and orphan drug data base, INSERM. 1999. [Online]. Available: https://www.orpha.net. [Accessed 2025.].
3. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, Avillach P, Bagley AM, Bakštein E, Balhoff JP, Baynam G. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res. 2024;52(D1):D1333–46. 10.1093/nar/gkad1005.
4. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, Hamosh A, Baynam G, Groza T, McMurry J, Dawkins J, Rath A, Thaxon C, Bocci G, Joachimiak MP, Kohler S, Robinson PN, Mungall C, Oprea RI. How many rare diseases are there? Nat Rev Drug Discov. 2020;19(2):77–8. 10.1038/d41573-019-00180-y.
5. Hartley T, Balci TB, Rojas SK, Eaton A, Care4Rare Canada, Dyment DA, Boycott KM. The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM®. Am J Med Genet. 2018;178C:458 – 63. 10.1002/ajmg.c.31662
6. Klimova B, Storek M, Valis M, Kuca K. Global view on rare diseases: A mini review. Curr Med Chem. 2017;24:3153–8. 10.2174/0929867324666170511111803.
7. Angelis A, Tordrup D, Kanavos P. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. Health Policy. 2015;119:964–79. 10.1016/j.healthpol.2014.12.016.
8. Ryder S, Leadley RM, Armstrong N, Westwood M, de Kock S, Butt T, Jain M, Kleijnen J. The burden, epidemiology, costs and treatment for Duchenne muscular dystrophy: an evidence review. Orphanet J Rare Dis. 2017;12:79. 10.1186/s13023-017-0631-3.
9. Molster C, Urwin D, Di Pietro L, Fookes M, Petrie D, van der Laan S, Dawkins H. Survey of healthcare experiences of Australian adults living with rare diseases. Orphanet J Rare Dis. 2016;11:30. 10.1186/s13023-016-0409-z.

10. Tisdale A, Cutillo CM, Nathan R, Russo P, Laraway B, Haendel M, Nowak D, Hasche C, Chan CH, Griese E, Dawkins H. The IDeaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. Orphanet J Rare Dis. 2021;16(1):429. 10.1186/s13023-021-02061-3.
11. Zhu Q, Nguyen DT, Grishagin I, Southall N, Sid E, Pariser A. An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). J biomedical Semant. 2020;11(1):13. 10.1186/s13326-020-00232-y.
12. Boycott KM, Hartley T, Biesecker LG, Gibbs RA, Innes AM, Riess O, Belmont J, Dunwoodie SL, Jojic N, Lassmann T, Mackay D, Temple IK, Visel Z, Baynam G. A diagnosis for all rare genetic diseases: The horizon and the next frontiers. Cell.2019;1:32 – 7. 10.1016/j.cell.2019.02.040
13. Cohen AM, Chamberlin S, Deloughery T, Nguyen M, Bedrick S, Meninger S, Ko JJ, Amin JJ, Wei AJ, Hersh W. Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. PLoS ONE. 2020;15:e0235574. 10.1371/journal/pone.0235574.
14. Gunne E, McGarvey C, Hamilton K, Treacy E, Lambert DM, Lynch SA. A retrospective review of the contribution of rare diseases to paediatric mortality in Ireland. Orphanet J Rare Dis. 2020;15:311. 10.1186/s13023-020-01574-7.
15. Roman-Naranjo P, Parra-Perez AM, Lopez-Escamez JA. A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases. J Biomed Inform. 2023;143:104429. 10.1016/j.jbi.2023.104429.
16. Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. The use of machine learning in rare diseases: a scoping review. Orphanet J Rare Dis. 2020;15(1):145. 10.1186/s13023-020-01424-6.
17. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annual Rev biomedical data Sci. 2018;1(1):53–68. 10.1146/annurev-biodatasci-080917-013315.
18. Lee J, Liu C, Kim J, Chen Z, Sun Y, Rogers JR, Chung WK, Weng C. Deep learning for rare disease: A scoping review. J Biomed Inform. 2022;135:104227. 10.1016/j.jbi.2022.104227.
19. Banerjee J, Taroni JN, Allaway RJ, Prasad DV, Guinney J, Greene C. Machine learning in rare disease. Nat Methods. 2023;20(6):803–14. 10.1038/s41592-023-01886-z.
20. Carrasco-Ribelles LA, Llanes-Jurado J, Gallego-Moll C, Cabrera-Bean M, Monteagudo-Zaragoza M, Violán C, Zabaleta-del-Olmo E. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. J Am Med Inform Assoc. 2023;30(12):2072–82. 10.1093/jamia/ocad168.
21. Wojtara M, Rana E, Rahman T, Khanna P, Singh H. Artificial intelligence in rare disease diagnosis and treatment. Clin Transl Sci. 2023;16(11):2106–11. 10.1111/cts.13619.
22. Visibelli A, Roncaglia B, Spiga O, Santucci A. The impact of artificial intelligence in the odyssey of rare diseases. Biomedicines. 2023;11(3):887. 10.3390/biomedicines11030887.

23. Morley TJ, Han L, Castro VM, Morra J, Perlis RH, Cox NJ, Bastarache L, Ruderfer DM. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. Nat Med. 2021;27(6):1097–104. 10.1038/s41591-021-01356-z.
24. Faviez C, Chen X, Garcelon N, Zaidan M, Billot K, Petzold F, Faour H, Douillet M, Rozet JM, Cormier-Daire V, Attié-Bitach T. Objectivizing issues in the diagnosis of complex rare diseases: lessons learned from testing existing diagnosis support systems on ciliopathies. BMC Med Inf Decis Mak. 2024;24(1):134. 10.1186/s12911-024-02538-8.
25. Yahia R, Rafi I, Evans W. Implementing a digital rare disease case-finding tool in UK primary care: a qualitative study of health professionals' experience. Br J Gen Practice: J Royal Coll Gen Practitioners. 2024;74(suppl 1). 10.3399/bjgp24x738105.
26. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. Nat biomedical Eng. 2022;6(12):1353–69. 10.1038/s41551-022-00942-x.
27. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans neural networks Learn Syst. 2020;32(1):4–24. 10.1109/TNNLS.2020.2978386.
28. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40. 10.1093/bioinformatics/btz682.
29. Zhang J, Dong Y, Wang Y, Tang J, Ding M, Prone. Fast and scalable network representation learning. InIJCAI 2019 Aug 10 (19, pp. 4278–84). 10.24963/ijcai.2019/594
30. Hirtz D, Ashwal S, Berg A, Bettis D, Camfield C, Camfield P, Crumrine P, Elterman R, Schneider S, Shinnar S. Practice parameter: evaluating a first nonfebrile seizure in children: report of the quality standards subcommittee of the American Academy of Neurology, The Child Neurology Society, and The American Epilepsy Society. Neurology. 2000;55(5):616–23. 10.1212/WNL.55.5.616.
31. Rodan LH, Stoler J, Chen E, Geleske T. Genetic evaluation of the child with intellectual disability or global developmental delay: clinical report. Pediatrics. 2025;156(1):e2025072219. 10.1542/peds.2025-072219.
32. McArthur E, Bastarache L, Capra JA. Linking rare and common disease vocabularies by mapping between the human phenotype ontology and phecodes. JAMIA open. 2023;6(1):ooad007. 10.1093/jamiaopen/ooad007.
33. Alsentzer E, Li MM, Kobren SN, Noori A, Undiagnosed Diseases Network, Kohane IS, Zitnik M. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. npj Digit Med. 2025;8(1):380. 10.1038/s41746-025-01749-1.
34. Morris C, Ritzert M, Fey M, Hamilton WL, Lenssen JE, Rattan G, Grohe M. Weisfeiler and leman go neural: Higher-order graph neural networks. InProceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 4602–4609). 10.1609/aaai.v33i01.33014602

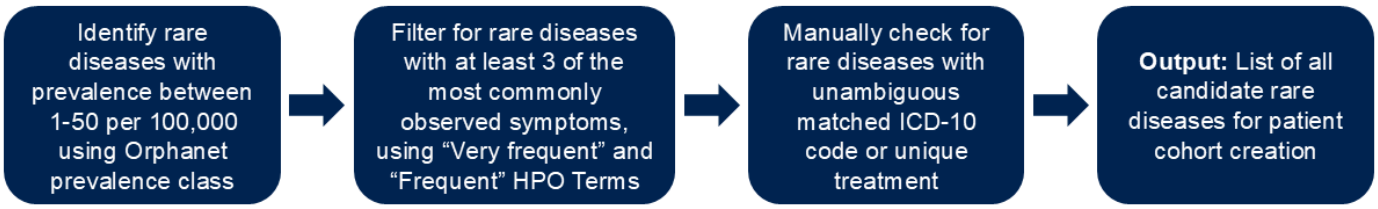
# Figures




1 © 2024 EVERSANA. All Rights Reserved. CONFIDENTIAL.

**Figure 1**

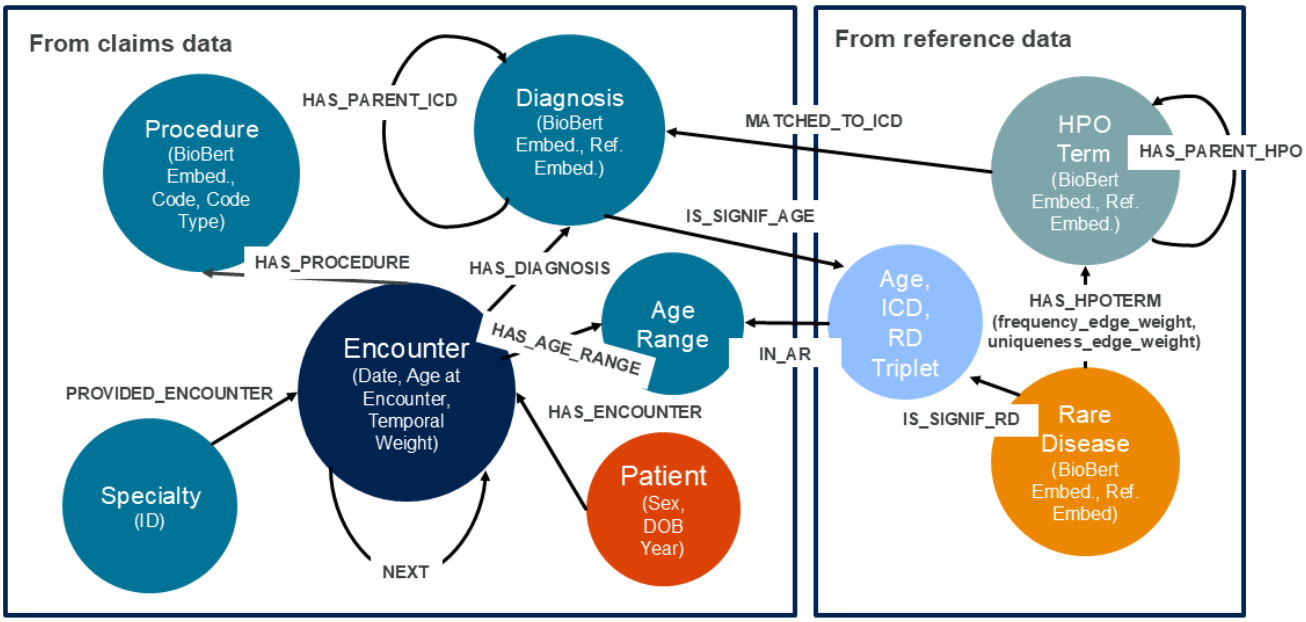
**Overall architecture representing the flow of information in the clinical decision support tool**




2 © 2024 EVERSANA. All Rights Reserved. CONFIDENTIAL.

## Figure 2

**Flowchart of the process used to identify the rare diseases targeted for this study**




3 © 2024 EVERSANA. All Rights Reserved. CONFIDENTIAL.

## Figure 3

**Outline of the overall graph structure used to train predictive graph neural network (GNN) models, including nodes originating from Real-World administrative claims Data, and nodes originating from Reference Data**

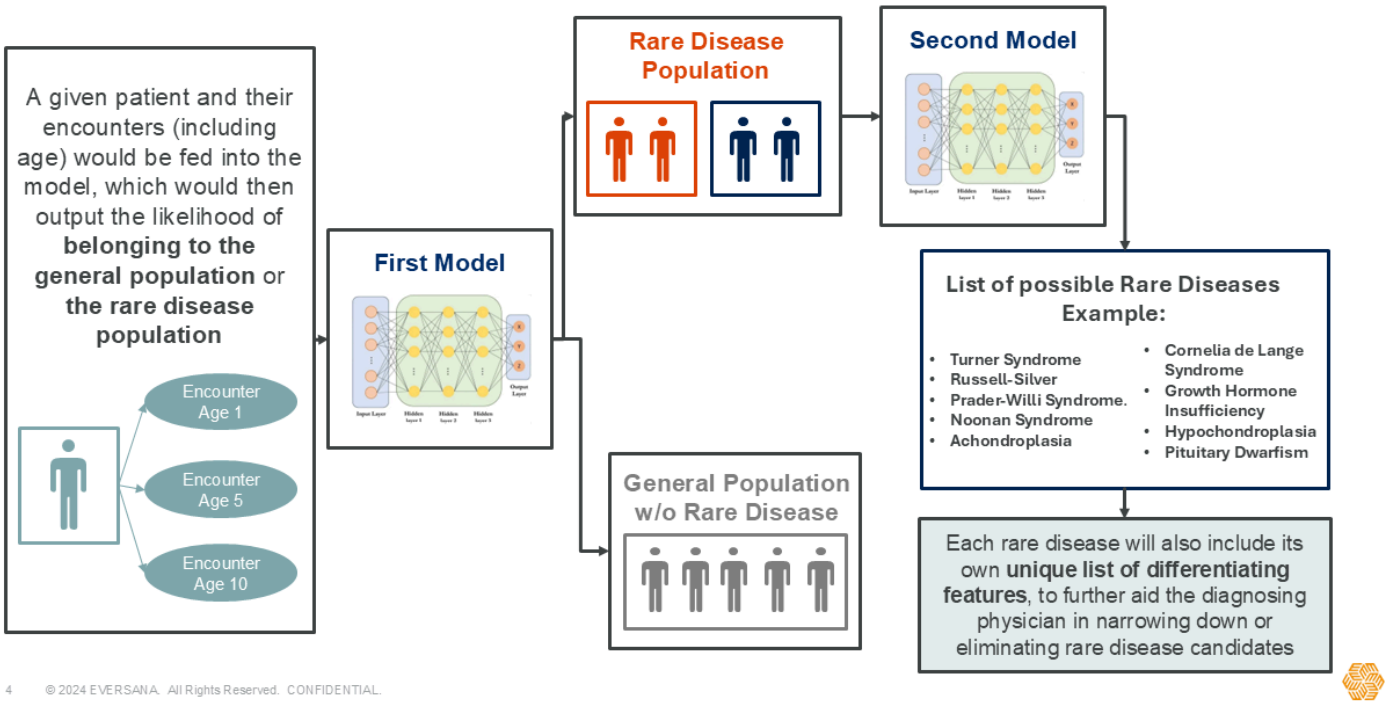



**Figure 4**

**Representation of the stacked predictive model approach to predicting a patient's potential for rare disease diagnosis**

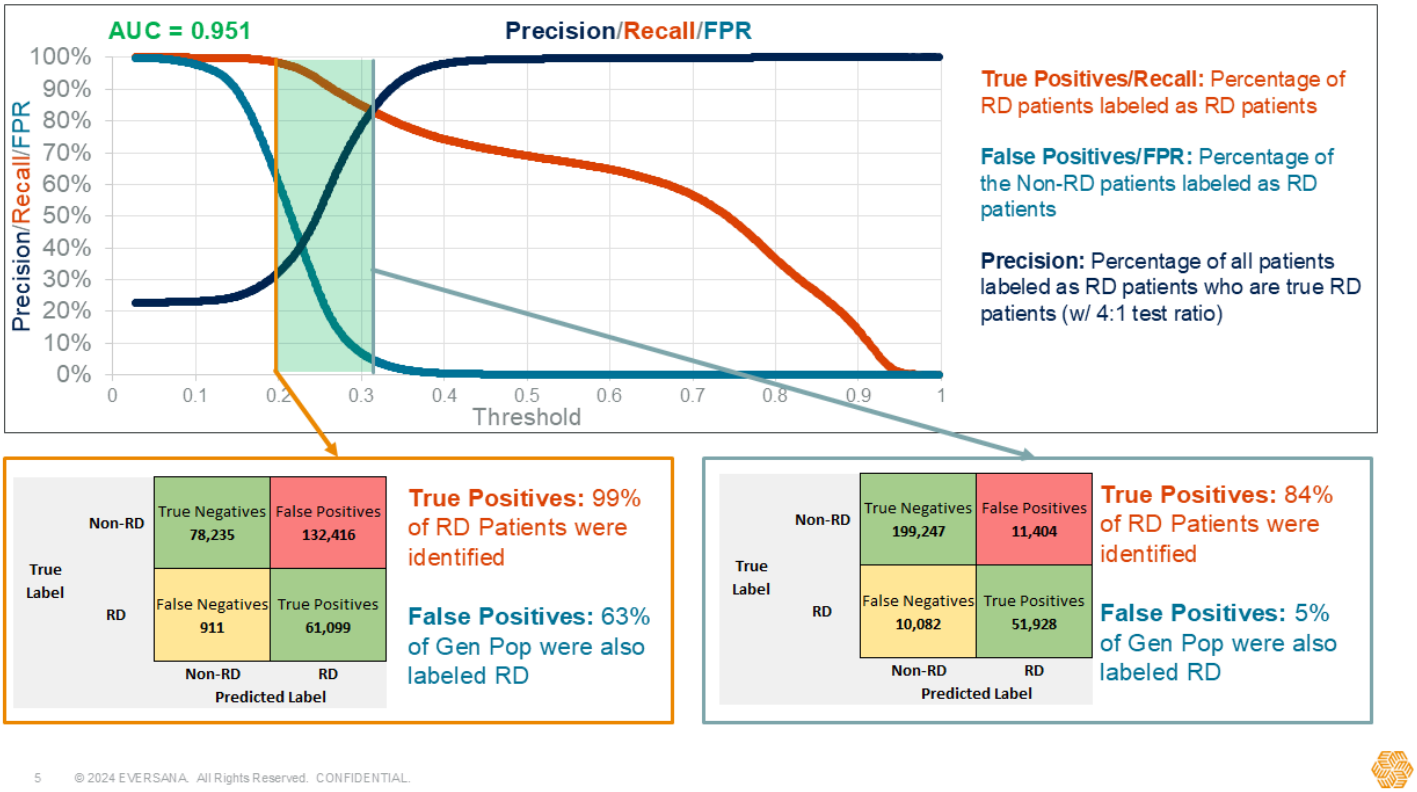



**Figure 5**

**Results of the binary classification model designed to identify rare disease patients from the general, non-rare disease population. A test set receiver operating characteristic area under the curve of 0.951 was observed in the final model**

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OrphanetSupplementaryInfoDraft20260325.docx