Genome Medicine

RESEARCH

Open Access

PhenoDP: leveraging deep learning for phenotype-based case reporting, disease ranking, and symptom recommendation

Baole Wen[1], Sheng Shi[1], Yi Long[4], Yanan Dang[1] and Weidong Tian[1,2,3*]

Abstract

Background Current phenotype-based diagnostic tools often struggle with accurate disease prioritization due to incomplete phenotypic data and the complexity of rare disease presentations. Additionally, they lack the ability to generate patient-centered clinical insights or recommend further symptoms for differential diagnosis.

Methods We developed PhenoDP, a deep learning-based toolkit with three modules: Summarizer, Ranker, and Recommender. The Summarizer fine-tuned a distilled large language model to create clinical summaries from a patient's Human Phenotype Ontology (HPO) terms. The Ranker prioritizes diseases by combining information content-based, phi-based, and semantic-based similarity measures. The Recommender employs contrastive learning to recommend additional HPO terms for enhanced diagnostic accuracy.

Results PhenoDP's Summarizer produces more clinically coherent and patient-centered summaries than the general-purpose language model FlanT5. The Ranker achieves state-of-the-art diagnostic performance, consistently outperforming existing phenotype-based methods across both simulated and real-world datasets. The Recommender also outperformed GPT-4o and PhenoTips in improving diagnostic accuracy when its suggested terms were incorporated into different ranking pipelines.

Conclusions PhenoDP enhances Mendelian disease diagnosis through deep learning, offering precise summarization, ranking, and symptom recommendation. Its superior performance and open-source design make it a valuable clinical tool, with potential to accelerate diagnosis and improve patient outcomes. PhenoDP is freely available at https://github.com/TianLab-Bioinfo/PhenoDP.

Keywords Phenotype-driven diagnosis, Mendelian disease, Large language models, Human Phenotype Ontology, Disease ranking, Clinical summarization, Symptom recommendation, Deep learning, Contrastive learning

Background

Mendelian genetic diseases, or monogenic disorders, impact millions of newborns annually, affecting approximately 8% of the global population, with a cumulative incidence ranging from 1.5% to 6.2% [1–3]. Early and accurate diagnosis is critical for predicting disease risks, guiding preventive interventions, and improving patient care [4, 5]. Whole-genome sequencing (WGS) and whole-exome sequencing (WES)—which analyze either the entire genome or just the protein-coding regions—are widely employed in clinical settings to identify

*Correspondence:
Weidong Tian
weidong.tian@fudan.edu.cn
[1] State Key Laboratory of Genetics and Development of Complex Phenotypes, Department of Computational Biology, School of Life Sciences, Fudan University, 2005 Songhu Road, Shanghai 200438, China
[2] Children's Hospital of Fudan University, Shanghai 201102, China
[3] Children's Hospital of Shandong University, Jinan, Shandong 250022, China
[4] School of Medicine, Nankai University, Tianjin 300071, China

pathogenic variants [6–8]. Despite advances in sequencing technology, the diagnostic yield of WES and WGS remains around 40% [5, 7, 9, 10], due in part to the prevalence of non-monogenic conditions with complex genetic or multifactorial etiologies, as well as the limitations of existing diagnostic systems for monogenic diseases. Additionally, stringent filtering often leaves hundreds of variants for manual expert interpretation—a laborious and complex task [11, 12]. The challenges are further compounded by the growing number of Mendelian diseases cataloged in the OMIM database [13], which now includes approximately 7500 conditions associated with 4900 genes, with new disease-gene associations emerging each year. To improve diagnostic efficiency and accuracy, the integration of phenotype data—observable clinical characteristics—into genetic analyses has emerged as a critical approach. This approach is crucial for improving the efficiency of genetic testing, ultimately benefiting patient care and reducing the strain on clinicians tasked with manual variant interpretation.

Incorporating clinical symptoms characteristic of Mendelian disorders into the diagnostic process, alongside gene-driven data, has proven effective in narrowing down potential diseases and accelerating diagnosis [14, 15]. This process begins with clinicians collecting a patient's clinical symptoms, which are then mapped into standardized terms using the Human Phenotype Ontology (HPO) [16]. This conversion can be performed manually or through automated tools, such as PhenoBERT [17]. The HPO not only provides standardized symptom descriptions but also links these symptoms to diseases and genes cataloged in databases like OMIM and Orphanet [18]. Once symptoms are mapped to HPO terms, automated tools can compare the patient's terms to those associated with specific diseases or genes, streamlining the diagnostic process.

Phenotype-driven tools have demonstrated considerable clinical potential. For example, Muffels et al. [19] highlighted the strong clinical utility of the Phrank tool [20], which uses phenotypic overlap to identify potential genetic causes of disease. Existing tools for disease prioritization can be broadly categorized into those that rank diseases based on phenotypic data alone (e.g., GDDP [21]) and those that incorporate both phenotypic and genetic data to prioritize causative genes. These tools vary in their approaches: some prioritize diseases first and then rank genes based on the disease (e.g., PhenoPro [22]), while others directly rank genes without initial disease prioritization (e.g., Phen2Gene [23]). All these methods leverage the directed acyclic graph (DAG) structure of the HPO to relate phenotypes to diseases and genes, though they differ in how they measure the similarity between HPO terms [24].

Despite the success of these tools, significant limitations remain. One key issue is the incomplete use of available data. For instance, Phen2Disease [25] employs bidirectional similarity between patient and disease HPO terms but neglects ancestral terms in the HPO DAG. In contrast, GDDP includes ancestral information but lacks precise weighting of HPO terms. Furthermore, an over-reliance on the HPO DAG for semantic analysis overlooks the richer, real-world context of phenotypic terms—a context that has been successfully incorporated into advanced gene prioritization tools [14]. Another limitation is that patients in clinical settings often present with only a few HPO terms, making it difficult to distinguish among top candidate diseases. Suggesting additional HPO terms to help differentiate closely ranked diseases can reduce diagnostic time and improve accuracy. While PhenoTips [26] provides this functionality, its effectiveness has not been comprehensively validated across diverse scenarios, highlighting an ongoing challenge in the field.

To address these challenges, we introduce PhenoDP, a phenotype-driven, deep-learning-based toolkit for analyzing Mendelian diseases. PhenoDP comprises three core components: Summarizer, Ranker, and Recommender. The Summarizer leverages DeepSeek-R1-671B's advanced reasoning capabilities [27] to fine-tune an open-source large language model (LLM) Bio-Medical-3B-CoT [28], generating high-quality, patient-centered clinical summaries from provided HPO terms. The Ranker integrates multiple similarity measures to prioritize the most likely disease based on the presented HPO terms. It consistently outperforms existing methods in disease ranking across simulated patient datasets and three real-world patient datasets. The Recommender utilizes contrastive learning to identify and suggest missing symptoms from incomplete clinical data, enhancing differential diagnosis. Across various scenarios, incorporating the Recommender's suggested HPO terms more effectively distinguishes target diseases compared to those generated by GPT-4o or PhenoTips. What is more, the Summarizer can also integrate with the Ranker and Recommender to generate structured clinical reports, combining personalized clinical symptoms with probable diagnoses and additional symptoms to support differential diagnosis. By integrating natural language processing (NLP), deep learning, and phenotype-driven analysis, PhenoDP enhances diagnostic accuracy, reduces time to diagnosis, and supports clinical decision-making, positioning itself as a valuable asset for the future of genomic medicine. PhenoDP is implemented in Python 3.7, and

its code is available at https://github.com/TianLab-Bioinfo/PhenoDP.

Methods

Design of PhenoDP

PhenoDP consists of three core modules: the Summarizer, Ranker, and Recommender. Given a set of HPO terms from a patient, the Summarizer generates a patient-centered clinical summary summarizing the patient's clinical symptoms. The Ranker produces a ranked list of Mendelian diseases most likely affecting the patient, and the Recommender suggests additional HPO terms that a physician might examine to refine the diagnosis based on the Ranker's top-ranked diseases.

Below, we describe the construction of each of the three modules separately.

The Summarizer module

The PhenoDP Summarizer is designed to generate concise, patient-centered clinical summaries from Human Phenotype Ontology (HPO) terms, streamlining diagnostic workflows for clinicians. Initially, we began by extracting HPO terms associated with 4731 OMIM entries (April 17, 2024 version) and 3654 Orphanet entries (April 18, 2024 version), focusing on entries with non-empty descriptions or definitions. Using these, we fine-tuned large language models (LLMs), such as FlanT5-Base (250 M). The input consisted of concatenated HPO term definitions (April 26, 2024 HPO version), and the target outputs were expert-written disease descriptions from OMIM ("Description" section) or Orphanet ("Definition" section). This approach aimed to generate expert-level disease definitions. However, it fell short in producing the patient-centered, context-rich clinical summaries needed for practical applications, such as creating abstracts from datasets like SUMPUBMED [29].

To overcome this limitation, we shifted to a more capable model: DeepSeek-R1-671B, a state-of-the-art LLM recognized for its advanced reasoning abilities. We used concatenated HPO term definitions from disease entries, paired with tailored prompt templates, to generate simulated patient-centered clinical summaries. A manual review confirmed that these summaries were of high quality, capturing the detailed, scenario-based insights clinicians depend on. Despite its effectiveness, DeepSeek-R1-671B's large size made it impractical for deployment on standard workstations due to its significant resource requirements.

To create a practical yet powerful solution, we selected Bio-Medical-3B-CoT, a Qwen2.5-3B-Instruct variant released on January 7, 2025, optimized for healthcare tasks. This model, trained on over 600,000 biomedical entries (as exemplified in Additional file 2: Table S1) using chain-of-thought prompting—a technique that improves reasoning by breaking down complex tasks—outperformed FlanT5-Base in both disease definition generation and SUMPUBMED abstract synthesis. We further refined this model using low-rank adaptation (LoRA) technology [30], a method that fine-tunes efficiently by adjusting fewer parameters, reducing computational demands. The training process used concatenated HPO term definitions from all 8385 disease entries as input. The target outputs were the patient-centered clinical summaries and reasoning chains previously generated by DeepSeek-R1-671B. This fine-tuning produced the final PhenoDP Summarizer, a model that balances high performance with practicality for widespread clinical use.

The Ranker module in PhenoDP

Given a set of HPO terms $Q = \{h_i, i \in \{1, 2, 3, \ldots, n\})\}$ from a patient, the ranker evaluates Q against each disease's HPO terms $D = \{h_j, j \in \{1, 2, 3, \ldots, n_k\}\}$ using three similarity measures, which are then combined to generate an overall similarity score:

(1) IC-based similarity: This measure first computes the similarity between each HPO term from the patient (h_i) and each HPO term from the disease (h_j) using the Jiang and Conrath (JC) method [31], based on the information content (IC) of the terms. The similarity between Q and D is weighted based on each HPO term's specificity:

$$IC(h) = -log_2\left(\frac{|d_h|}{|d_{all}|}\right) \tag{1}$$

$$Sim(h_i, h_j) = \frac{1}{IC(h_i) + IC(h_j) - 2 \times IC(h_{MICA})} \tag{2}$$

where $IC(h)$ denotes the information content of an HPO term, with $|d_h|$ being the number of diseases associated with the term and $|d_{all}|$ the total number of diseases, and h_{MICA} represents the most informative common ancestor between h_i and h_j. Note that when two HPO terms are identical, the *Sim* score is directly set to 1. The similarity between h_i and disease D_c is defined as the maximum similarity between h_i and all HPO terms in D_c. To compute the similarity between Q and D_c, we further weight each HPO term h_i based on its specificity:

$$IC(D) = -log_2\left(\frac{|h_D|}{|h_{all}|}\right) \tag{3}$$

$$weight(h) = IC(h) + mean_{D_r \in T} IC(D_r) \tag{4}$$

where $IC(D)$ denotes the information content of a disease derived from the number of HPO terms associated with

the disease ($|h_D|$) and the total number of HPO terms ($|h_{all}|$), and T represents all diseases associated with the term h. The final similarity score between Q and D_c is then computed as:

$$Sim_{IC-based} = \frac{\sum_{i=1}^{n} Sim(h_i, D_c) \times weight(h_i)}{\sum_{i=1}^{n} weight(h_i)} \quad (5)$$

(2) Phi-based similarity. This method calculates the phi correlation between Q and D_c, by considering both the HPO terms and their ancestors. The correlation is calculated using contingency tables that summarize shared and unique ancestor terms for the patient and disease. The similarity is then computed using the following formula:

$$M = \begin{bmatrix} |A(Q)A(D_c)| & \left|A(Q)\overline{A(D_c)}\right| \\ \left|\overline{A(Q)}A(D_c)\right| & \left|\overline{A(Q)}\,\overline{A(D_c)}\right| \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} = M \quad (7)$$

$$Sim_{phi-based} = \frac{(ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad (8)$$

where $A(\cdot)$ denotes the set of HPO terms and their ancestral HPO terms.

(3) Semantic-based similarity: In this approach, we utilized a graph convolutional network (GCN) [32] with two convolutional layers to capture the semantic relationships among HPO terms within the HPO directed acyclic graph (DAG). We employed the Summarizer to generate semantic embeddings for each HPO term based solely on its name (although incorporating the term's definition yielded similar results). These embeddings were then averaged into 256-dimensional vectors representing the terms in the HPO DAG. To enhance the model's robustness, we randomly masked 20% of the edges and node features and trained the GCN to reconstruct the original graph structure. After 2000 training epochs, we obtained the final model, referred to as the Pre-trained Semantic & DAG-based HPO Encoder (PSD-HPOEncoder).

This method leverages the semantic embeddings generated by the PSD-HPOEncoder to compute similarity between the patient's HPO terms and the disease's HPO terms. The similarity score is derived by calculating the cosine similarity between the embeddings.

$$\boldsymbol{e_h} = PSD - HPOEncdoer(h) \quad (9)$$

$$Sim(\boldsymbol{e_h}, D) = \max_{t \in D} \frac{\boldsymbol{e_h} \cdot \boldsymbol{e_t}}{\|\boldsymbol{e_h}\| \|\boldsymbol{e_t}\|} \quad (10)$$

$$Sim_{semantic-based} = \frac{\sum_{h \in Q} sim(\boldsymbol{e_h}, D_c)}{|Q|} \quad (11)$$

where $\boldsymbol{e_h}$ is the embedding of h from the PSD-HPOEncoder.

The final disease ranking score is obtained by summing the similarity scores:

$$Sim_{disease} = Sim_{IC-based} + Sim_{phi-based} + Sim_{semantic-based} \quad (12)$$

A ranked list of diseases based on these scores is generated, and the coefficient of variation (CV) is computed for the top three candidate diseases to assess the reliability of the rankings.

The Recommender module in PhenoDP

The Recommender module utilizes a Transformer [33] encoder architecture and contrastive learning to suggest missing HPO terms crucial for diagnosis. The input consists of concatenated HPO terms, treated as tokens, and their embeddings are generated by the PSD-HPOEncoder. During each training, 2000 diseases were sampled, creating two sentences for each disease by randomly selecting 70% of its associated HPO terms. Sentences from the same disease were classified as positive pairs of sentences, while those from different diseases served as negative pairs of sentences. The InfoNCE loss [34] maximized cosine similarity for positive pairs and minimized it for negative pairs, leading to the creation of the PCL-HPOEncoder, which refines the semantic vectors from the PSD-HPOEncoder to enhance local semantic representation.

The Recommender suggests additional HPO terms for a patient based on their current term set and contrastive embeddings. First, the term sets of candidate diseases are processed through the PSD-HPOEncoder for semantic embeddings, which are then fed into the PCL-HPOEncoder to generate contrastive embeddings. For each disease, a term not yet seen in the patient or other diseases is added to the patient's term set, and their contrastive embedding is updated. The patient's updated embedding and the disease form a positive pair, while other combinations are treated as negatives. The InfoNCE loss is calculated, and its reciprocal gives the term's score. This process is repeated for all terms of the disease and other diseases, resulting in a ranked list of recommended terms.

Benchmark datasets

Disease definition datasets

Initially, we fine-tuned large LLMs, such as FlanT5-Base [35] for disease definition generation. For this, we sampled 3000 disease entries from each—OMIM and

Orphanet—concatenating the definitions of their associated HPO terms as input and using their respective disease definitions (Definition section for Orphanet and Description section for OMIM) as targets. To avoid data leakage, we trained and tested the models separately within each dataset.

Medical summarization evaluation dataset

The SUMPUBMED dataset, introduced by Gupta et al. in 2021, is a specialized resource for benchmarking biomedical abstract generation models [29]. It comprises 33,772 documents from BioMed Central (BMC) literature, covering various fields such as medicine and nursing. This dataset has been widely used as a benchmark for biomedical abstract generation in multiple studies [36, 37]. Each document in SUMPUBMED consists of two sections: Front section: contains the summary (abstract) of the article; Body section: includes background, results, and conclusions. For our study, we selected the first 1000 documents from the dataset, using the Body section (background, results, and conclusions) as input to generate abstracts. We then compared these generated summaries to the gold-standard abstracts in the Front section.

Synthesized patient-centered clinical summary dataset using DeepSeek-R1

We utilized DeepSeek-R1-671B, a state-of-the-art large language model (LLM) known for its strong reasoning capabilities, to synthesize a patient-centered clinical summary dataset. Using 8385 disease entries from OMIM and Orphanet, we concatenated their HPO term definitions and designed tailored prompts to instruct DeepSeek-R1-671B to generate corresponding patient-centered clinical summaries along with chain-of-thought (CoT) reasoning. To evaluate the fine-tuned Bio-Medical-3B-CoT model, we first fine-tuned it on a randomly selected subset of 3000 OMIM or Orphanet diseases, using concatenated HPO term definitions as input and the synthesized clinical summaries with CoT reasoning from DeepSeek-R1-671B as targets. We then tested the model on the clinical summaries of the remaining OMIM or Orphanet disease entries.

Simulated patient datasets

To assess PhenoDP's disease ranking performance, we generated simulated datasets by randomly selecting three HPO terms for each disease. These datasets were divided into four types based on the relationship between the selected HPO terms and their respective diseases: (1) Precise Only: all three terms are specific to the disease (i.e., terminal nodes), (2) 2 Precise +1 Imprecise: two specific terms and one broader parent term, (3) 2 Precise +1 Noise: two specific terms and one unrelated term, and (4) Mixed: two specific terms, one broader term, and one unrelated term, mimicking real clinical conditions. Each simulation ran 3000 resampling iterations, yielding 3000 patients per dataset.

Real-world patient datasets

PhenoDP was further validated on two real-world datasets. The first dataset comprised 384 cases provided by LIRICAL [24], where the HPO terms and the diagnosed disease of each case were taken from original files. Three cases—"Ge-2019-TJP2-proband," "Javadiyan-2017-MAF-patient_CSA108.01," and "Nevidomskyte-2017-SMAD3-54-year-old_woman"—were excluded due to phenotypes not matching the target disease, leaving 381 cases for analysis.

The second dataset comprised 130 manually curated cases from published articles on genetic and rare diseases. Specifically, we conducted a search on Google Scholar for case reports and selected around 500 relevant publications for manual inspection. From these, we selected 130 reports, each corresponding to a unique OMIM-classified disease, ensuring a one-to-one mapping between disease and case report. For each selected report, we extracted a single patient case description—randomly choosing one if multiple cases were present—to represent the diagnosed disease. Since these reports lacked HPO annotations, we applied three established tools—PhenoBERT [17], PhenoTagger [38], and ClinPhen [39]—with default parameters to infer HPO terms from the text. This process yielded 130 literature-based cases, each comprising a diagnosed disease, a patient description, and a set of annotated HPO terms.

The new LIRICAL dataset now includes 8182 patient cases, with each case including the annotated HPO terms and the diagnosed disease [40]. For robustness, we applied the following filtering criteria: we removed patients with fewer than three HPO terms, those annotated with more than one OMIM disease, and those whose recorded HPOs did not overlap the HPO set of their corresponding OMIM disease. After filtering, 5485 cases remained for testing.

For each case in the three datasets mentioned above, we calculated a score for each OMIM disease based on the provided HPO terms and assessed the ranking of the target disease.

Evaluation of PhenoDP

Evaluation of the Summarizer module

The performance of the Summarizer was assessed using three key metrics: Word Mover's Distance (WMD) [41], BioLinkBERT [42], and PubMedBERT [43]. Each metric captures different aspects of text similarity, ranging from

word-level associations to sentence-level semantics and deep contextual understanding.

(1) Word Mover's Distance (WMD): WMD quantifies text similarity by measuring the "distance" between word meanings in two texts. It leverages pre-trained word embeddings (e.g., Word2Vec) to compute the minimum effort required to align words between the texts, even when they differ lexically (e.g., "sick" vs. "ill"). While WMD is effective for topic-level comparisons, it does not account for sentence structure.
(2) BioLinkBERT: BioLinkBERT extends the LinkBERT architecture by pretraining on PubMed abstracts using citation links between documents. Unlike WMD, which compares individual words, BioLinkBERT captures deep contextual semantics and inter-entity interactions, enabling richer sentence-level representations for biomedical text.
(3) PubMedBERT: PubMedBERT is a domain-specific variant of BERT, pre-trained from scratch on over 14 million PubMed abstracts and full texts, rather than fine-tuned from a general BERT model. It excels at capturing deep contextual relationships, considering both word sense disambiguation (e.g., "cell" as a biological unit vs. a prison term) and syntactic dependencies. This makes PubMedBERT the most sophisticated metric for biomedical text similarity, as it fully models meaning within technical texts. However, it is also the most computationally demanding among the three.

By combining these metrics, we ensure a comprehensive evaluation of the Summarizer's output, balancing lexical, sentence-level, and deep contextual similarity assessments.

Evaluation of the Ranker module

The Ranker was assessed using two metrics: coverage percentage and mean reciprocal rank (MRR). Coverage percentage measures the frequency at which the correct disease is ranked within specified top positions (top 1, top 5, top 10, and top 20), with the average value used as the metric. MRR quantifies the average reciprocal rank of the correct disease across cases. This study focuses on the top 20 diseases, as lower-ranked diseases generally receive less clinical attention and exhibit score convergence. For both metrics, diseases ranked beyond the top 20 are assigned a rank of 21 to ensure consistent evaluation.

Evaluation of the Recommender module

The Recommender was assessed by comparing its suggested HPO terms with those generated by GPT-4o and PhenoTips through the following three experiments:

(1) Impact on disease ranking confidence: For cases where the Ranker initially placed the target disease in the first position, we evaluated how adding the recommended HPO terms influenced ranking confidence. Specifically, we measured the difference in ranking scores between the top-ranked disease and the second-ranked disease before and after incorporating the recommended terms.
(2) Improvement in disease ranking position: For cases where the Ranker initially ranked the target disease in the second or third position, we assessed the effectiveness of the recommended HPO terms by calculating the percentage of cases where the target disease moved to the first position after their inclusion.
(3) Impact on overall disease ranking performance: For cases where the Ranker initially placed the target disease within the top three, we analyzed how adding the recommended HPO terms affected overall ranking performance. This was quantified using the mean reciprocal rank (MRR) metric across multiple disease-ranking models, including PhenoDP, PhenoPro, and GDDP.

These experiments provide a comprehensive evaluation of the Recommender's ability to enhance disease ranking accuracy and confidence.

Competing methods of PhenoDP's Ranker

PhenoPro

PhenoPro [22] is a pathogenic gene ranking algorithm that prioritizes genes by integrating phenotype and variant data. It ranks diseases based on the probability of each phenotype being associated with a disease, using Bayesian principles and a one-sided KS test to assess the likelihood of a patient's phenotypes. Diseases are then ranked according to their p values. The phenotype-only disease ranking function of PhenoPro is implemented via the Ranked_Score_Disease_Pheno function, which takes a set of HPO terms as input and outputs a ranked list of candidate diseases. PhenoPro is available at https://github.com/TianLab-Bioinfo/PhenoPro.

GDDP GDDP [21] is an online disease ranking tool. It introduces the ancestor nodes of the patient's and disease's HPO terms, weights these nodes based on IC values, constructs contingency tables, and applies Fisher's exact test to measure the similarity between the patient's

phenotypes and the disease. The GDDP website is https://gddp.research.cchmc.org/, and the benchmark uses default algorithm settings.

Phen2Disease

Phen2Disease [25] is a recently proposed algorithm for pathogenic gene and disease ranking. Its design also follows the approach of ranking diseases first, then matching genes. The algorithm weights HPO terms based on IC values and introduces a bidirectional set similarity measure, both patient-centric and disease-centric. The phenotype-only disease ranking function of Phen2Disease is executed using the scripts Phen2Disease-patient.py and score2disease_patient.py. Phen2Disease is available at https://github.com/ZhuLab-Fudan/Phen2Disease.

Phrank

Phrank [20] is a classic algorithm for pathogenic gene and disease ranking. Its core disease ranking method expands the phenotypes of a disease by using gene-associated phenotypes, and it introduces a set similarity measure based on Bayesian networks and information theory. The phenotype-only disease ranking function of Phrank is implemented via the rank_diseases function. Phrank can be accessed at https://github.com/meng-ma-biomedical-AI/F29_Phrank.

LIRICAL LIRICAL [24] is a Java-based application for pathogenic gene and disease ranking. It evaluates the likelihood of each disease by calculating the interpretable clinical genomics likelihood ratio (LR) between the patient's phenotypes and the target disease. The tool generates an LR based on the probability of observing specific phenotypes with and without the target disease, supporting diagnosis. The phenotype-only disease ranking function of LIRICAL is executed using the command-line argument –observed-phenotypes. LIRICAL is available at https://github.com/TheJacksonLaboratory/LIRICAL.

Phenomizer

Phenomizer [44] is an online disease ranking tool. It is the first tool to use HPO terms for phenotype-disease association profiling and is the recommended analysis tool by the HPO project. Its algorithm is primarily based on traditional information theory. Phenomizer is available at https://compbio.charite.de/phenomizer.

Exomiser

Exomiser [11] is a computational tool designed to prioritize genes and variants in next-generation sequencing (NGS) data. However, Exomiser also features an "only phenotype" mode, enabling phenotype-driven prioritization, which we have incorporated into our evaluation. Exomiser is available at https://github.com/exomiser/Exomiser.

Competing methods of PhenoDP's Recommender

GPT-4o

The prompt template designed for GPT-4o to generate recommended HPO terms is detailed in Additional file 2: Table S2. GPT-4o was accessed via its official website (https://chatgpt.com/), and the evaluation was conducted using the version released on September 27, 2024. All cases were manually input into the chat interface to obtain the recommended terms.

PhenoTips

We utilized the latest open-source version of PhenoTips (https://github.com/phenotips/phenotips/). Given a case with a known target disease, we manually entered its associated HPO terms into PhenoTips and extracted the top recommended HPO terms.

Results

Overview of PhenoDP and its core modules

PhenoDP is an innovative tool designed to enhance disease diagnosis through comprehensive phenotype data analysis, comprising three core modules: the Summarizer, the Ranker, and the Recommender. The Summarizer harnesses DeepSeek-R1-671B's advanced reasoning ability to fine-tune Bio-Medical-3B-CoT, producing high-quality, patient-centered clinical summaries from provided HPO terms. The Ranker compares a patient's HPO term set against each OMIM disease using multiple similarity measures, producing a ranked list of possible diagnoses. Building on this, the Recommender suggests additional HPO terms that may refine the diagnosis, employing a contrastive learning framework to identify the most discriminative phenotype terms. Collectively, these modules empower PhenoDP to deliver precise disease rankings enhanced with contextual insights, thereby facilitating improved clinical decision-making. The workflow of PhenoDP is illustrated in Fig. 1, with further details available in the Methods section.

PhenoDP's Summarizer: utilizing DeepSeek-R1 to generate high-quality patient-centered clinical summaries

To enhance the ability of the PhenoDP Summarizer to generate clinically relevant summaries, we undertook a systematic refinement process, involving multiple iterations of model training and evaluation. Initially, we fine-tuned the FlanT5-Base model using disease definitions sourced from expert-curated databases: Orphanet and OMIM. Specifically, we used the "Definition" section from Orphanet and the "Description" section from OMIM, as both provided concise and structured disease

information. We sampled 3000 disease entries from each dataset, concatenating the definitions of their associated HPO terms as input, with the disease definitions as targets. To prevent data leakage, we ensured that training and testing were conducted separately for each dataset. The fine-tuned models successfully outperformed the baseline FlanT5-Base in generating disease definitions (Fig. 2).

Despite these promising results, we realized that relying solely on disease definitions—whether from OMIM or Orphanet—posed significant limitations for the Summarizer's ultimate goal of producing real-world clinical summaries. Disease definitions are abstract and generalized, offering broad descriptions that lack the nuance and patient-centered detail required for clinical practice. Without a dataset of paired clinical symptoms and summaries, we turned to the SUMPUBMED dataset—a collection of 33,772 biomedical documents commonly used for abstract generation in previous research. We selected the first 1000 documents, using the "Body" sections (background, results, and conclusions) as input to generate abstracts, and compared them to the gold-standard "Front" sections (summary). Surprisingly, the fine-tuned models underperformed the baseline FlanT5-Base, suggesting that training on disease definitions did not translate effectively to more complex summarization tasks like biomedical text abstraction (Fig. 2).

To overcome these limitations, we explored large-scale biomedical language models (LLMs) that are specifically optimized for healthcare applications. We adopted Bio-Medical-3B-CoT, a Qwen2.5-3B-Instruct variant released on January 7, 2025, which had been trained on over 600,000 biomedical samples with chain-of-thought prompting. Without additional fine-tuning, this model outperformed FlanT5-Base in both disease definition generation and SUMPUBMED abstract synthesis. We then fine-tuned Bio-Medical-3B-CoT separately on the OMIM and Orphanet datasets. While the fine-tuned model performed well in generating disease definitions, it struggled to synthesize coherent clinical narratives from sets of HPO terms—highlighting the challenge of adapting disease-focused training to more dynamic, patient-oriented summarization tasks (Fig. 2).

To address the challenge of generating high-quality, patient-centered summaries, we turned to DeepSeek-R1-671B, an advanced LLM renowned for its reasoning capabilities. We leveraged 8385 disease entries from OMIM and Orphanet, concatenating their HPO term definitions and crafting tailored prompts to instruct DeepSeek-R1-671B to generate detailed, patient-centered clinical summaries. A manual review confirmed that these outputs were of high quality, capturing the nuanced insights that clinicians rely on for decision-making. However, the computational demands of DeepSeek-R1-671B made it impractical for widespread deployment in a clinical setting.

To create a more efficient solution, we applied knowledge distillation to fine-tune Bio-Medical-3B-CoT using the patient-centered summaries and reasoning chains generated by DeepSeek-R1-671B. Figure 3a illustrates an example, showing the input text, prompt, undistilled model output, clinical summary with reasoning chains from DeepSeek-R1-671B, and the distilled model output. The resulting model, Bio-Medical-3B-CoT-R1-Distilled, retained the advanced capabilities of the larger model while being lightweight enough to run on standard hardware. This distilled model demonstrated strong performance in both SUMPUBMED abstract generation and a

(See figure on next page.)

Fig. 1 The PhenoDP framework. PhenoDP consists of three core modules: the Summarizer, the Ranker, and the Recommender. **a** Deep learning architecture of PhenoDP. PhenoDP fine-tunes the pre-trained Bio-Medical-3B-CoT model using synthetic cases generated by DeepSeek-R1-671B to develop the Summarizer. The Summarizer's encoder captures real-world semantic nuances by processing Human Phenotype Ontology (HPO) terms. Each HPO directed acyclic graph (DAG) node represents a term's semantics, while edges encode term-to-term relationships, including anatomical and conceptual connections (e.g., between HP:0040279 [Frequency] and HP:0040283 [Occasional]). PhenoDP applies random masking and trains a graph convolutional network (GCN) via unsupervised learning, creating the PSD-HPOEncoder to generate hidden embeddings for all nodes. These embeddings treat phenotype terms as tokens and diseases as sentences, enabling the PCL-HPOEncoder to produce representations for patient or disease phenotype sets using the CLS token. **b** Workflow of the HPO Recommender. Given a patient's phenotype set and candidate diseases, the Recommender suggests additional symptoms. For each candidate disease, an unobserved phenotype term is added to the patient's set. This augmented set, along with the term sets of all candidate diseases, is processed by the PCL-HPOEncoder to generate embeddings. Positive pairs are formed from disease-patient embedding pairs, while others are treated as negatives. The InfoNCE score is calculated for each symptom, generating relevance scores. **c** Functional overview of PhenoDP. The Summarizer generates a patient-centered clinical summary based on the patient's phenotype. The Ranker evaluates candidate diseases using three similarity metrics: (1) IC-based similarity, derived from information theory and phenotype-disease relationships; (2) phi-based similarity, calculated using the phi coefficient after considering shared ancestor terms between the patient and disease; and (3) semantic-based similarity, obtained by computing the cosine similarity between hidden embeddings generated by the PSD-HPOEncoder. The Ranker outputs a list of candidate diseases with corresponding scores, and the coefficient of variation (CV) is calculated—higher CV values indicate higher confidence in the ranking, while lower values suggest insufficient symptom data. Finally, the Recommender suggests additional symptoms based on the ranked diseases

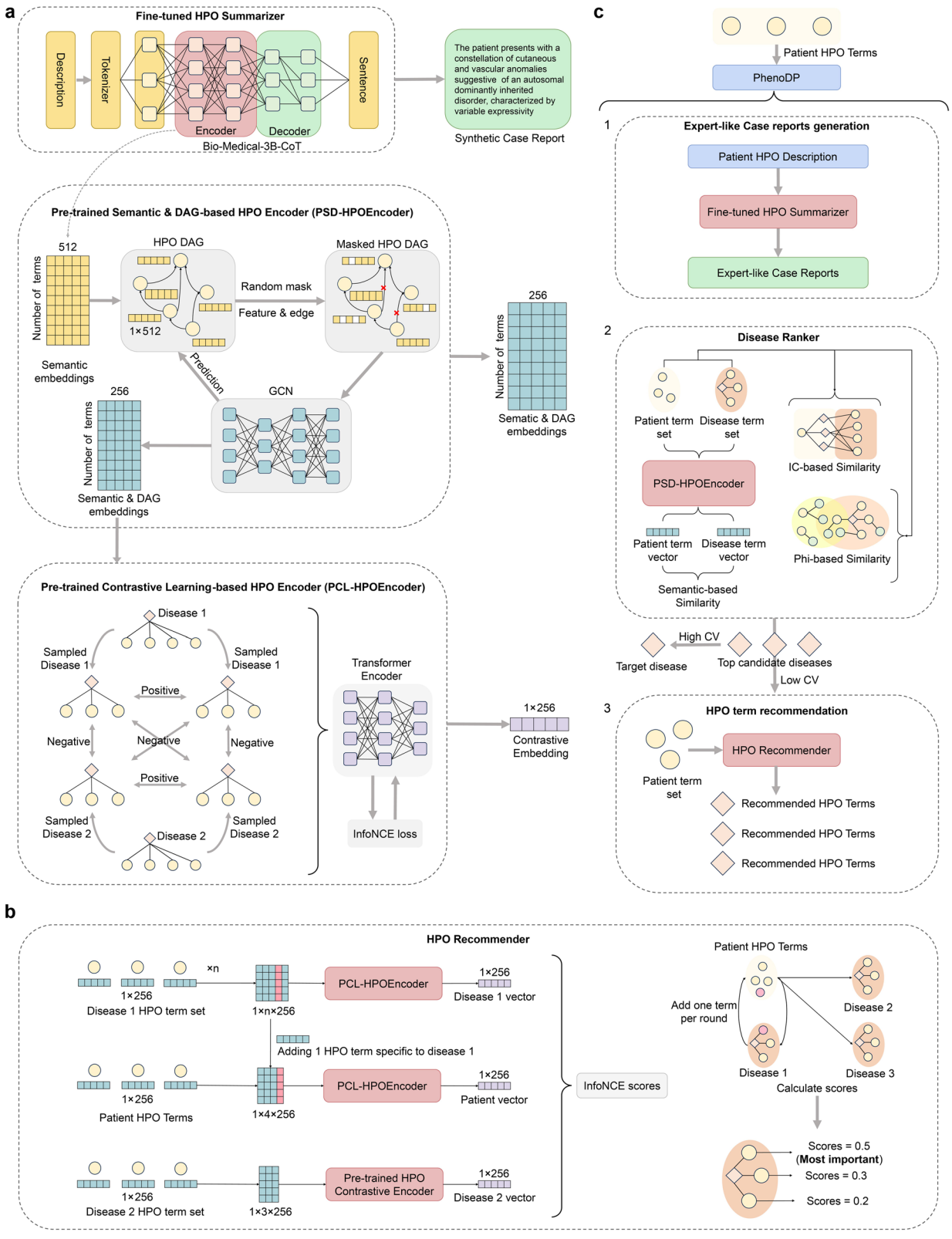


Fig. 1 (See legend on previous page.)

Disease Definition (Disease Description)								
	FlanT5 (OMIM)	FlanT5-lora (OMIM)	FlanT5 (Orphanet)	FlanT5-lora (Orphanet)	Bio-Medical (OMIM)	Bio-Medical-lora (OMIM)	Bio-Medical (Orphanet)	Bio-Medical-lora (Orphanet)
WMD	3.09	2.51	3.00	2.37	1.09	1.02	1.20	1.15
BioLinkBERT	0.65	0.71	0.67	0.74	0.91	0.94	0.87	0.91
PubMedBERT	0.19	0.23	0.23	0.26	0.55	0.60	0.46	0.56

SUMPUBMED						
	FlanT5	FlanT5-lora (OMIM)	FlanT5-lora (Orphanet)	Bio-Medical	Bio-Medical-lora (OMIM)	Bio-Medical-lora (Orphanet)
WMD	0.96	1.13	1.21	**0.69**	1.35	1.33
BioLinkBERT	0.90	0.88	0.86	**0.96**	0.91	0.92
PubMedBERT	0.71	0.64	0.59	**0.86**	0.66	0.63

Fig. 2 Evaluation results of disease definition generation and medical abstract summarization performance. Top. Performance of FlanT5 and Bio-Medical-3B-CoT in disease definition generation before and after fine-tuning. "FlanT5" and "Bio-Medical" denote the original models, while " + lora" indicates fine-tuning with specific datasets. "(OMIM)" and "(Orphanet)" represent training on 3000 randomly sampled OMIM and Orphanet disease entries, respectively, with testing conducted on the remaining entries (1731 for OMIM, 654 for Orphanet). Bottom. Performance of FlanT5 and Bio-Medical-3B-CoT in medical abstract summarization before and after fine-tuning. Models were trained using the described methods, with the prompt "Summarize the following text: {Text Input}" applied consistently during training and testing. Identical parameters were used for both training and evaluation

test set of unseen clinical summaries produced by DeepSeek-R1-671B (Fig. 3b).

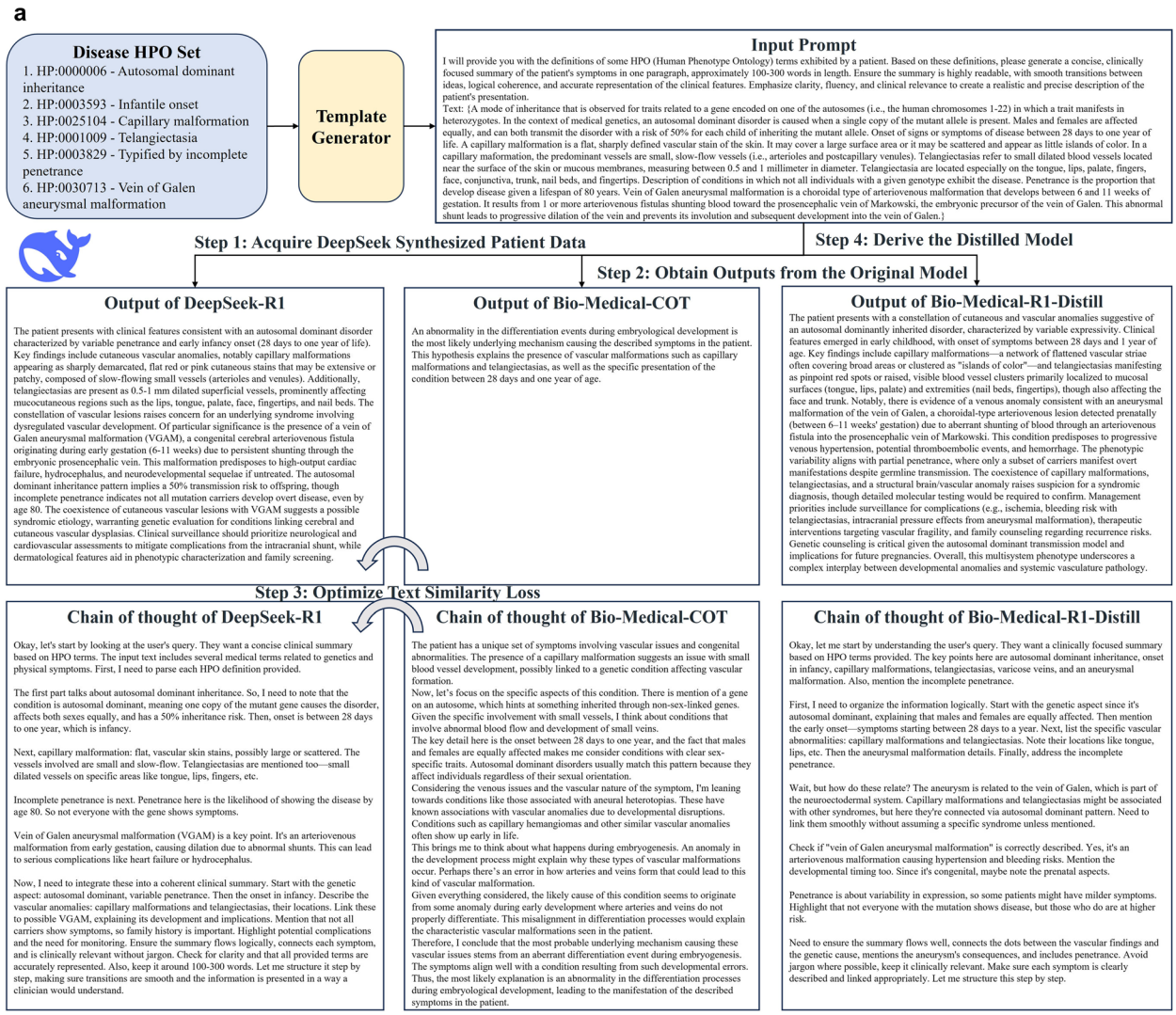
This enhancement ensures that PhenoDP's Summarizer not only achieves its original objective of structured clinical summary but also adapts to the complexities of real-world medical decision-making, making it a useful tool for clinicians.

PhenoDP's Ranker: strong disease ranking performance across simulated and real-world datasets

We evaluated PhenoDP's ability to prioritize target diseases based on a patient's HPO terms using both simulated and real-world datasets. PhenoDP was compared to seven competing methods: PhenoPro, GDDP, Phen2Disease, Phrank, LIRICAL, Phenomizer, Exomiser. It is important to note that Phenomizer was only assessed in two of the three real-world datasets due to its nature as an online tool, making it impractical to run on the simulated datasets and one of the real-world datasets, which contain thousands of patients.

We constructed four types of simulated datasets representing patient phenotypic data: Precise Only, 2 Precise + 1 Imprecise, 2 Precise + 1 Noise, and Mixed, ordered by increasing difficulty (refer to the Methods section for simulation details). As illustrated in Fig. 4a–d, all methods exhibited a decline in coverage percentage with stricter criteria; however, PhenoDP consistently attained higher coverage percentages across all datasets. Notably, PhenoDP more frequently ranked the target disease within the top 20 than other methods, demonstrating its superior performance in these simulated scenarios. When compared to PhenoPro, which ranked second in coverage, PhenoDP achieved an average improvement of 11.7% across all datasets, with a significant 10.4% improvement in the more complex Mixed dataset. Additionally, PhenoDP outperformed all other methods in terms of mean reciprocal rank (MRR), achieving the highest overall MRR, while PhenoPro ranked second. PhenoDP's average MRR improvement over PhenoPro was 12.6%, highlighting its effectiveness in prioritizing the correct diagnosis at the top of the ranked list (Fig. 4e).

We further assessed PhenoDP's real-world applicability using three patient phenotype datasets derived from actual clinical cases, with detailed data acquisition information provided in the Methods section. The first dataset contained 381 cases provided by LIRICAL [24]. The second dataset included 130 manually curated cases from published articles on genetic and rare diseases, compiled by our team. The third dataset consisted of 5485 cases, recently updated by LIRICAL [40]. Both the first and third datasets included annotated HPO terms and the diagnosed disease for each case. For the second dataset, diagnosed diseases were sourced from the articles, while HPO terms were extracted from patient symptom descriptions using automated tools, such as PhenoBERT, ClinPhen, and PhenoTagger. Notably, all disease ranking methods performed optimally with HPO terms extracted by PhenoBERT (Additional file 2: Table S3), which we used for the second dataset.



b

	OMIM - Synthesized Case			Orphanet - Synthesized Case			SUMPUBMED		
	WMD	BioLinkBERT	PubMedBERT	WMD	BioLinkBERT	PubMedBERT	WMD	BioLinkBERT	PubMedBERT
Bio-Medical	0.64	0.93	0.70	0.80	0.90	0.59	0.69	0.96	0.86
R1-Distill	**0.55**	**0.96**	**0.88**	**0.64**	**0.94**	**0.73**	**0.66**	**0.96**	**0.87**

Fig. 3 Model distillation process and evaluation of text summarization performance before and after distillation. **a** Example of the distillation process for the disease capillary malformation-arteriovenous malformation 2 (OMIM:618196), annotated with six HPO terms. Definitions of these HPO terms were concatenated and input into a fixed template to generate model inputs. The leftmost boxes represent the output text (synthesized patient-centered summary) and chain-of-thought (CoT) from DeepSeek-R1-671B. The middle boxes show the output and CoT from the unmodified Bio-Medical-3B-CoT model, while the rightmost boxes display the output and CoT from the distilled model. **b** Evaluation of text summarization performance before and after distillation. "OMIM-Synthesized Case" and "Orphanet-Synthesized Case" were generated using synthesized patient-centered summaries and CoTs from 3000 randomly selected disease entries (from OMIM or Orphanet) as training templates, with synthesized summaries from the remaining entries used for testing. In both evaluations, "R1-Distill" represents the fine-tuned model, while Bio-Medical-3B-CoT remained unmodified. "SUMPUBMED" is an independent medical literature summarization dataset, with the first 1000 articles and their summaries used for evaluation

PhenoDP's Ranker consistently outperformed all other methods across the three datasets, achieving an improvement of 8.1%, 8.6%, and 2.6% over the second-best approach in the first, second, and third datasets, respectively (Fig. 5a–c). GDDP ranked second in the first dataset, while PhenoPro ranked second in the second and

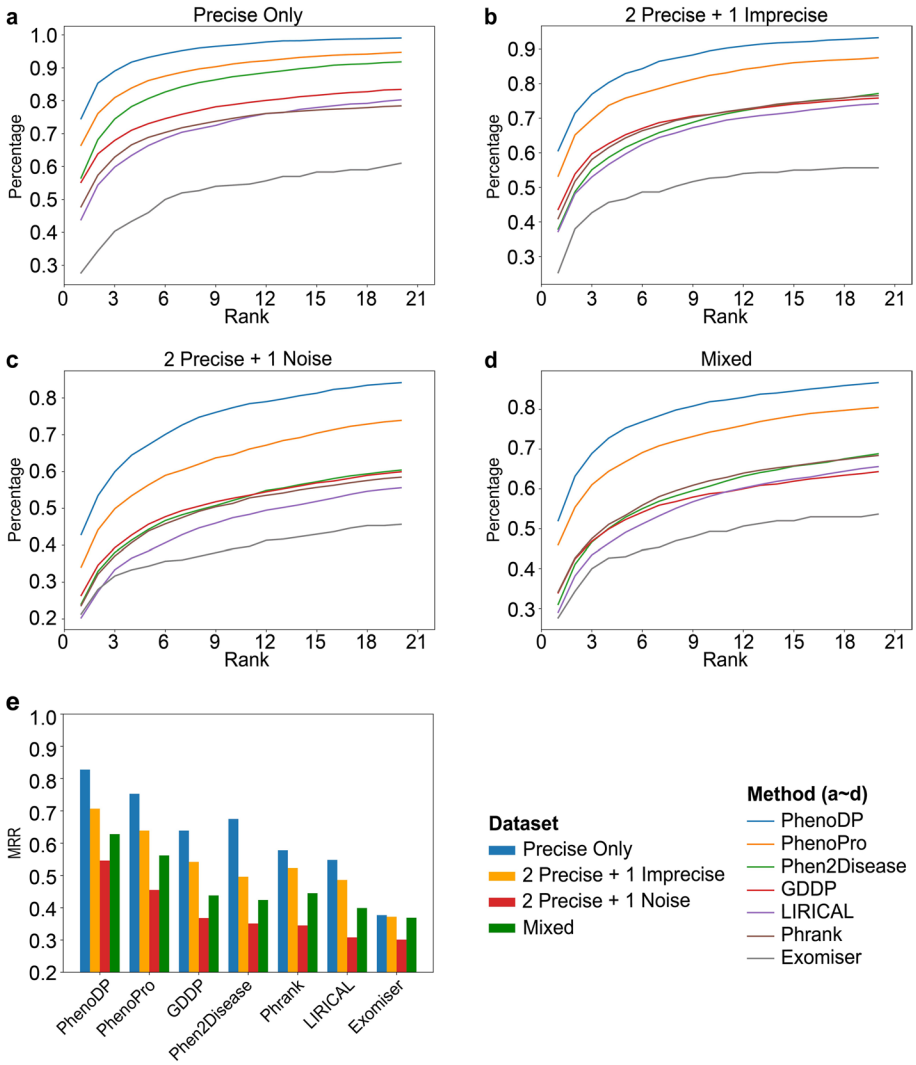


Fig. 4 Performance of PhenoDP's Ranker on simulated datasets. **a–d** depict the coverage percentage of various disease ranking methods across four simulated datasets: "Precise Only", "2 Precise + 1 Imprecise", "2 Precise + 1 Noise" and "Mixed". The first three datasets each contain three terms, while the "Mixed" dataset includes two precise terms, one imprecise term, and one noise term. Each dataset consists of 3000 samples. Coverage percentage indicates the proportion of cases where the ranking tool places the target disease (i.e., the true disease) within a specific rank. **e** shows a bar chart comparing the mean reciprocal rank (MRR) for each method across the four simulated datasets. MRR measures the average reciprocal rank of the target disease, providing an overall assessment of ranking accuracy

third datasets. In the second dataset, PhenoDP's superior performance was maintained across HPO terms extracted by the other tools as well (Additional file 2: Table S3).

Following PhenoDP's ranking, we calculated the coefficient of variation (CV) for the scores of the top three diseases. A higher CV indicates greater variability among the top disease scores, which we hypothesized could signal the presence of the true target disease among the top-ranked options. Our findings supported this hypothesis: in real-world dataset 1, the high CV group (CV > 2) exhibited an MRR of approximately 0.7, significantly higher than the low CV group (MRR: 0.256, $p = 1.18 \times 10^{-26}$). Similarly, in real-world dataset 2, the high CV group had an MRR of 0.555, compared to 0.378 in the low CV group ($p = 0.014$). In real-world dataset 3, the high CV group exhibited an MRR of 0.570, compared to 0.367 in the low CV group ($p < 1 \times 10^{-100}$) (Fig. 5d).

The Ranker combines three similarity-based metrics—IC-based, phi-based, and semantic-based similarity—to rank diseases. Further analysis revealed that IC-based similarity played the most crucial role (Additional file 1: Fig. S1a), while phi-based similarity also contributed significantly. In contrast, the contribution of semantic-based

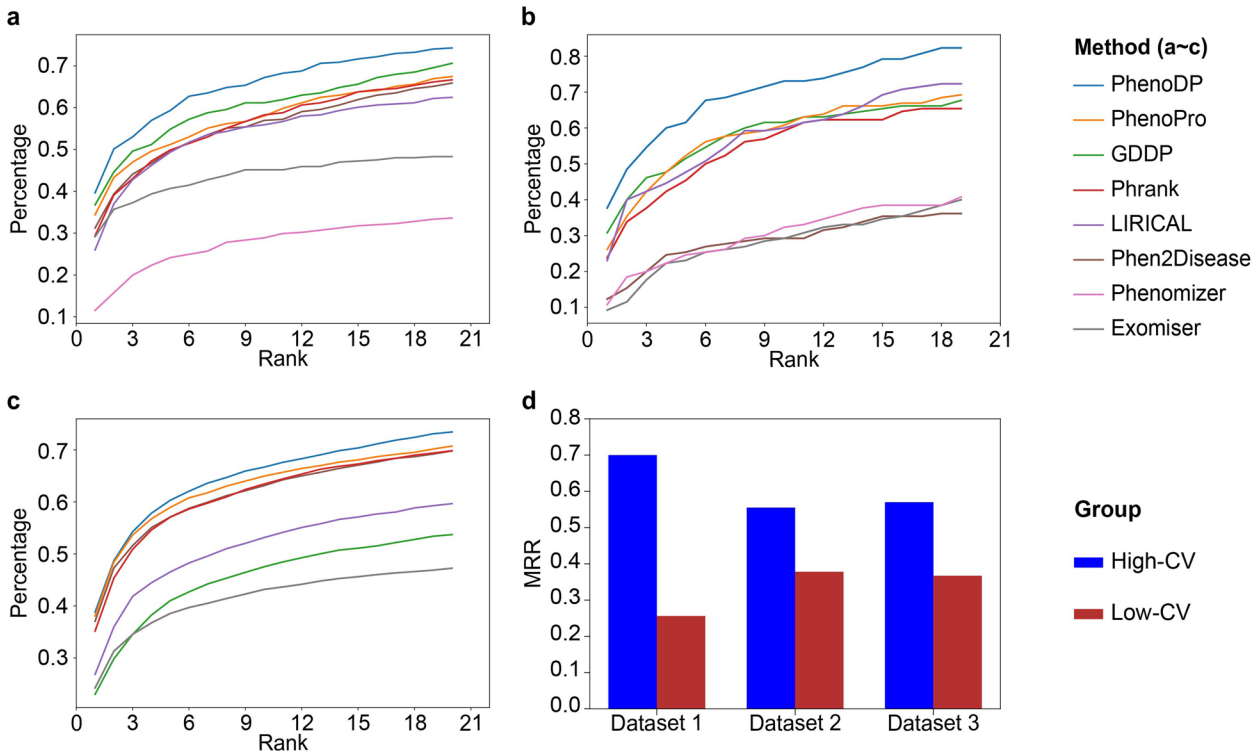


Fig. 5 Performance of PhenoDP's Ranker on real-world datasets. **a**–**c** Line plot showing the coverage percentage of different disease ranking tools across three real-world datasets respectively ($N = 381$; $N = 130$; $N = 5485$). **d** Bar plot of the mean reciprocal rank (MRR) of disease ranking results generated by PhenoDP, with cases grouped according to their coefficient of variation (CV) values. These groups are based on PhenoDP's ranking results, considering three candidate diseases for each case. A CV threshold of 2 is used, with cases categorized as high CV (above 2) and low CV (2 or below)

similarity was minimal, possibly due to real semantic data being derived from a limited dataset.

In conclusion, PhenoDP's Ranker consistently outperforms other methods across both simulated and real-world datasets, providing significant advantages in coverage percentage and MRR. These results underscore its robustness and effectiveness in accurately prioritizing diagnoses, particularly in complex clinical scenarios.

PhenoDP's Recommender: suggesting relevant symptoms for clinical assessment

In numerous clinical cases, the observed symptoms (HPO terms) may prove insufficient for accurately differentiating the target disease from other conditions with similar presentations. To address this challenge, we developed a Recommender system that suggests additional symptoms for clinicians to consider based on the patient's observed symptoms. The effectiveness of this Recommender was evaluated through three benchmark scenarios.

In the first scenario, for cases where the Ranker initially placed the target disease in the first position, we evaluated the impact of incorporating the recommended HPO terms on ranking confidence. Specifically, we measured the difference in ranking scores between the top-ranked disease and the second-ranked disease before and after adding the suggested terms, assessing how the inclusion of these terms influenced the disease ranking confidence.

In the second scenario, for cases where the Ranker initially ranked the target disease in the second or third position, we assessed the effectiveness of the recommended HPO terms by calculating the percentage of cases in which the target disease moved to the first position after their inclusion, thereby measuring the improvement in ranking position.

In the third scenario, for cases where the Ranker initially placed the target disease within the top three, we analyzed how adding the recommended HPO terms affected overall ranking performance. This was quantified using the mean reciprocal rank (MRR) metric, which was applied across multiple disease-ranking models, including PhenoDP, PhenoPro, GDDP, Phen2Disease, and LIRICAL.

For comparison, we included the latest version of GPT-4o and PhenoTips in the benchmark. A specific prompt was designed for GPT-4o, wherein it was

provided with the symptoms (HPO terms) for both the target and competing diseases, alongside the observed patient symptoms. GPT-4o was then instructed to recommend an additional symptom that would aid in diagnosing the patient with the target disease. The prompt templates utilized in this comparison are outlined in Additional file 2: Table S2. For PhenoTips, given a case with a known target disease, we manually entered its associated HPO terms into PhenoTips online webpage and extracted the top recommended HPO terms.

In the first scenario, the Recommender consistently outperformed both GPT-4o and PhenoTips in increasing the score difference between the top-ranked target disease and the second-ranked disease (Fig. 6a). Among 195 cases where the Ranker placed the target disease in the top position ($N = 150$ in real-world dataset 1, $N = 45$ in real-world dataset 2), the Recommender achieved an average score improvement of 0.086, significantly surpassing GPT-4o's improvement of 0.071 ($p = 5.63 \times 10^{-4}$, paired t-test) and PhenoTips' improvement of 0.0581 ($p = 5.04 \times 10^{-9}$) (Fig. 6b).

In the second scenario, the Recommender successfully elevated the target disease to the top-ranked position in 78.1% of cases, markedly outperforming GPT-4o (53.4%) and PhenoTips (23.3%) (Fig. 6c). In the third scenario, incorporating HPO terms suggested by the Recommender resulted in superior performance across all methods, as evidenced by higher MRR scores compared to those suggested by GPT-4o or PhenoTips (Fig. 6d). This finding underscores the superior quality of the Recommender's suggestions in more effectively distinguishing the target disease.

Illustrating the Recommender: case study 1

In this scenario, a patient was diagnosed with Immunodeficiency 103 (OMIM:212050, IMD103). The patient presented with symptoms including increased circulating IgE concentration, eosinophilia, increased circulating IgG concentration, lymphopenia, abnormal proportions of CD8-positive T cells, nail dystrophy, chronic oral candidiasis, and abnormal proportions of CD4-positive T cells. While PhenoDP's Ranker correctly identified IMD103

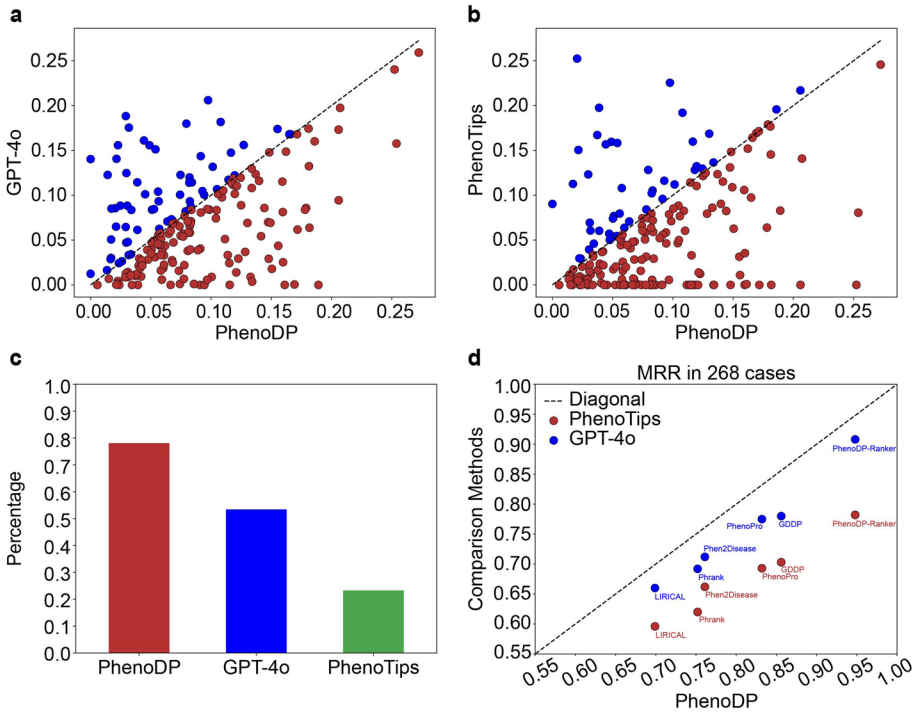


Fig. 6 Performance of PhenoDP's Recommender on real-world datasets. **a**–**b** Scatter plots comparing the score differences between the top-ranked and second-ranked diseases after term recommendations and re-ranking by PhenoDP's Recommender versus GPT-4o/PhenoTips, evaluated on 195 cases (150 cases from real-world dataset 1 and 45 cases from real-world dataset 2, where PhenoDP's Ranker initially ranked the target disease as top 1). The y-axis represents score differences for GPT-4o/PhenoTips, while the x-axis shows those for PhenoDP. The black dashed line indicates equality, with red points denoting cases where PhenoDP performed as well or better than GPT-4o/PhenoTips, and blue points where GPT-4o/PhenoTips outperformed PhenoDP. **c** Histogram showing the percentage of cases where the target disease moved to the top 1 position after term recommendation by PhenoDP, GPT-4o, and PhenoTips, evaluated on 73 cases (51 cases from real-world dataset 1 and 22 cases from real-world dataset 2, where PhenoDP's Ranker initially ranked the target disease as top 2 or top 3). **d** Comparison of mean reciprocal rank (MRR) between PhenoDP and GPT-4o/PhenoTips after term recommendation. Results are based on re-ranking by PhenoDP, PhenoPro, GDDP, Phrank, Phen2Disease, and LIRICAL after adding the first recommended term to the HPO set for a total of 268 cases (195 cases + 73 cases)

as the top-ranked disease, the scores for the second and third-ranked diseases—Immunodeficiency 25 (IMD25) and Immunodeficiency 23 (IMD23)—were close, with scores of 0.800 for IMD103, 0.742 for IMD25, and 0.704 for IMD23.

To better distinguish IMD103 from IMD25 and IMD23, the Recommender suggested "Lymphadenopathy" (HP:0002716), while GPT-4o proposed "Hypereosinophilia" (HP:0032061), and PhenoTips recommended "Protracted diarrhea" (HP:0032061). Incorporating the Recommender's suggestion raised IMD103's score to 0.832, further separating it from IMD25 (0.708) and IMD23 (0.670). In contrast, adding GPT-4o's suggested term increased IMD103's score to 0.817, with IMD25 at 0.729. However, using PhenoTips' recommendation caused IMD103 to drop to second place (0.764), while IMD25 rose to the top (0.774), leading to an incorrect ranking.

Supporting this finding, a related study documented lymph node enlargement in 9 out of 15 patients with fungal infections and CARD9 mutations [45], the causative gene of IMD103. Meanwhile, increased eosinophils, diarrhea, or protracted diarrhea have been reported in cases of IMD103, IMD25, and IMD23 [46–48], suggesting that the latter two suggested terms lacked specificity in distinguishing IMD103.

Case study 2

In this scenario, a patient was diagnosed with Catecholaminergic Polymorphic Ventricular Tachycardia-3 (OMIM:614021, CPVT3), presenting with symptoms such as atrial arrhythmia, abnormal QT interval, ventricular arrhythmia, cardiac arrest, syncope, prolonged QT interval, and ventricular fibrillation. Initially, PhenoDP's Ranker placed Short QT syndrome 2 (SQT2) in the top position with a score of 0.809, followed by Long QT syndrome 6 (LQT6) at 0.805, while CPVT3 ranked third with a score of 0.795. Incorporating the Recommender's suggested term—"Polymorphic ventricular tachycardia" (HP:0031677)—elevated CPVT3 to the top position with a score of 0.835, while SQT2 dropped to third place (0.756) and LQT6 fell to fifth (0.751). The new second-ranked disease was Cardiac arrhythmia syndrome, with or without skeletal muscle weakness (OMIM: 615441, CARDAR), with a score of 0.792.

GPT-4o recommended "Juvenile onset" (HP:0003621), which, when included, increased CPVT3's score to 0.857 at the top position, but ranked Long QT syndrome 9 (OMIM: 611818, LQT9) second with a close score of 0.830. Notably, both SQT2 and LQT6 have also been associated with juvenile onset [49, 50], reducing the specificity of this feature for CPVT3. Despite CPVT3 moving to the top position after adding the terms suggested by the Recommender and GPT-4o, the narrow score differences between CPVT3 and other top-ranked diseases suggested the need for additional distinguishing features.

In contrast, PhenoTips recommended "Abnormality of the ear" (HP:0000598), after which CPVT3 remained in third place (0.782), while SQT2 retained the top position (0.799). This result underscores the importance of selecting highly relevant phenotypic terms to enhance disease differentiation.

In conclusion, the Recommender module of PhenoDP has demonstrated strong performance in enhancing disease differentiation and improving diagnostic accuracy by effectively suggesting additional symptoms. It outperformed both GPT-4o, a state-of-the-art generative language model, and PhenoTips, a tool specifically designed for phenotype-based disease prioritization. While GPT-4o is not specifically optimized for HPO term recommendation, making its lower performance unsurprising, the Recommender's superior results compared to PhenoTips highlight its effectiveness in refining differential diagnoses. With its innovative design and consistent performance across diverse datasets, PhenoDP's Recommender proves to be a valuable tool for clinicians navigating complex diagnostic challenges.

Finally, we integrated the Summarizer, Ranker, and Recommender to generate structured clinical reports using a tailored Summarizer prompt (see Additional file 2: Table S4). The report includes patient-centered symptoms derived from HPO terms, probable diagnoses ranked by the Ranker, and additional symptoms for differential diagnosis suggested by the Recommender. Figure 7 presents an example for case 1. Notably, in this example, the language model also suggested *FGFR2* and *TNXB* as candidate genes. While our prompt did not request gene prioritization, this output reflects the model's tendency to incorporate inferred biological associations when not explicitly restricted. The suggestion of *FGFR2* is biologically plausible, as the epithelial isoform FGFR2b is critical for normal morphogenesis of the skin and related structures, and its disruption has been associated with impaired skin elasticity and cutis laxa-like features (Katoh [51]). The suggestion of *TNXB* is even more consistent with current biological knowledge, as *TNXB* has been identified as a causative gene for Ehlers-Danlos syndrome, according to the report by Vanakker et al. [52].

Discussion

PhenoDP introduces a new approach in the field of genetic disease diagnostics, leveraging large language models (LLMs) and deep learning methodologies to address ongoing challenges in clinical practice. By comprising three core modules—the Summarizer, Ranker, and Recommender—PhenoDP introduces a

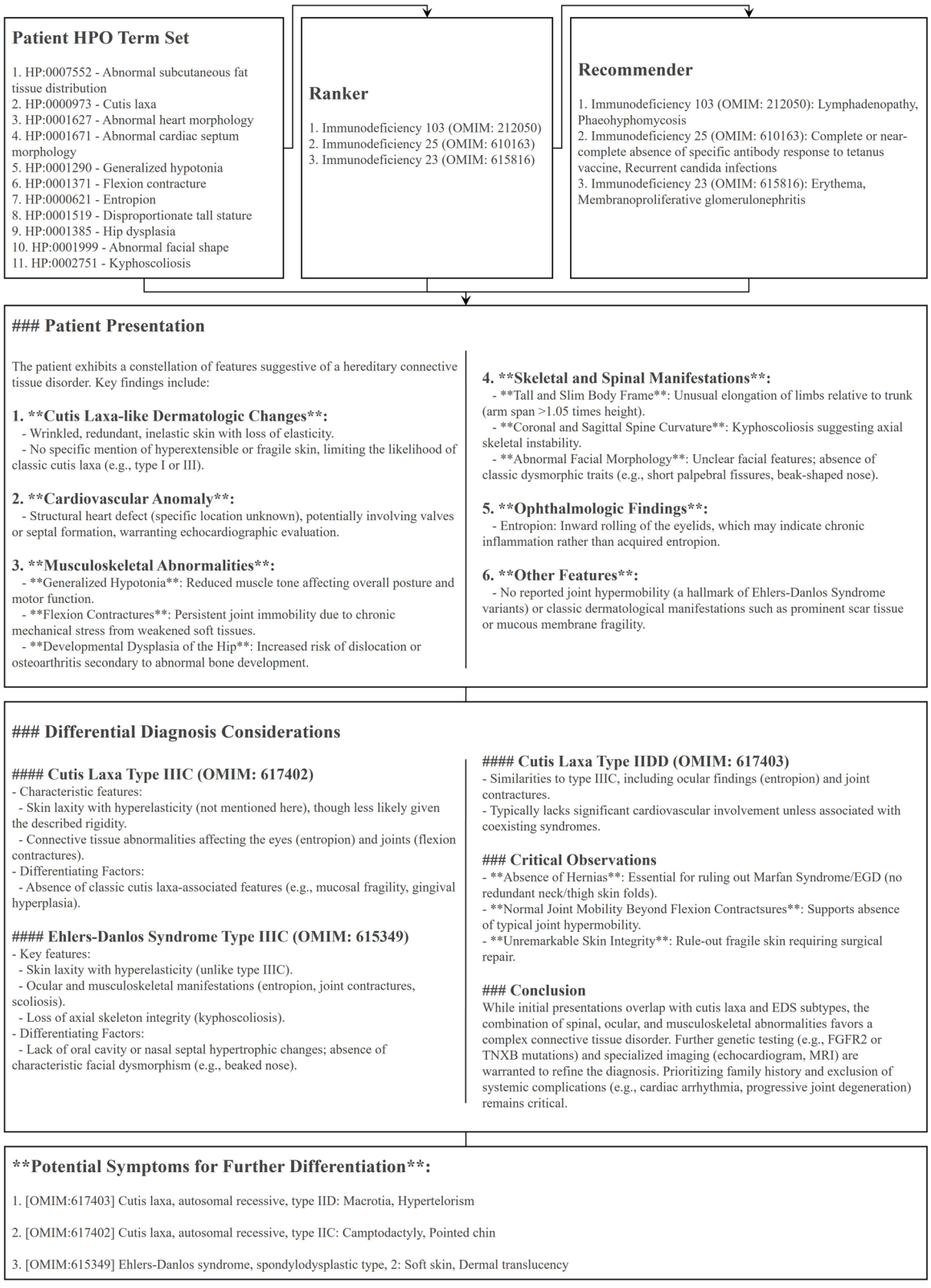


Fig. 7 Example of a structured clinical report generated by the summarization model. The case from case study 1 was used, with the prompt provided in Additional file 2: Table S4

comprehensive toolset that enhances diagnostic workflows through automated case report generation, robust disease ranking, and precise symptom recommendation. These innovations contribute to a more accurate, efficient, and scalable diagnostic process, positioning PhenoDP as a valuable tool in modern healthcare.

The Summarizer module represents a significant advancement in generating patient-centered clinical summaries, a critical component for clinical decision-making. Our results demonstrate that while fine-tuning language models like FlanT5-Base on static disease definitions from Orphanet and OMIM improved disease description generation, this approach did not translate effectively to the dynamic, nuanced task of clinical summarization. The performance gap observed using the SUMPUBMED benchmark underscored the need for models that can capture the rich contextual detail of real-world clinical scenarios. By leveraging DeepSeek-R1-671B's advanced reasoning capabilities and subsequently distilling its performance into Bio-Medical-3B-CoT, we developed Bio-Medical-3B-CoT-R1-Distilled—a lightweight yet powerful model that excels in generating high-quality, patient-centered summaries. This refined Summarizer now seamlessly integrates with PhenoDP's Ranker and Recommender to deliver structured clinical reports featuring tailored case summaries, probable diagnoses, and additional differential symptoms. Moving forward, potential improvements include further fine-tuning on larger, more diverse clinical datasets and incorporating real-time feedback from clinical experts to enhance the model's ability to capture evolving clinical language and context-specific nuances.

The Ranker module integrates three similarity-based measures—IC-based, phi-based, and semantic-based similarity—to prioritize diseases based on patient phenotypes. Among these, the IC-based similarity measure is particularly impactful, leveraging the Jiang and Conrath method, which has consistently outperformed other similarity metrics [25]. This method's effectiveness is augmented by a bidirectional weighting system, which mitigates noise commonly associated with disease-centric data. Phi-based similarity complements this by incorporating ancestral information and indirectly factoring in disease-specific similarities. Semantic-based similarity, though limited by the size of its training dataset, still contributes to improved disease ranking. Future iterations of PhenoDP should focus on expanding the dataset used for semantic training [53] and refining its integration, which could further enhance ranking performance. Notably, the Ranker also exhibits impressive computational efficiency, maintaining minimal time overhead when processing input sets with variable numbers of HPO terms (Additional file 1: Fig. S1c). This scalability underscores its suitability for real-world clinical environments where rapid and accurate diagnoses are imperative [12].

The relationship between the Summarizer and Ranker modules is critical to understanding the overall performance of PhenoDP. Importantly, the Summarizer was fine-tuned to generate coherent and clinically relevant summaries from concatenated HPO definitions rather than to differentiate between diseases. Although this process may capture some phenotype co-occurrence patterns, our experiments demonstrate that these patterns do not inadvertently enhance disease ranking. For example, when we compared cosine similarities of specific HPO term groups using both the base FlanT5 model and its fine-tuned counterpart or the Bio-Medical-3B-CoT model and the distilled model, we observed no significant improvement in ranking the correct disease (Additional file 1: Fig. S2a, b). Moreover, the node vectors derived from the Summarizer, when used in the graph model, yielded consistent Ranker performance regardless of the source model (Additional file 1: Fig. S2c). This suggests that the advantage in disease differentiation is more dependent on the underlying model architecture and pretraining data—evidenced by the superior performance of the Bio-Medical model over FlanT5—than on fine-tuning aimed at improving summary quality. Thus, the design of the Summarizer module, focused on enhancing clinical summary generation, does not compromise the evaluation of the Ranker's performance in distinguishing diseases.

The Recommender addresses the challenge of identifying unobserved but relevant symptoms by using contrastive learning. By utilizing a contrastive learning approach, the Recommender predicts missing symptoms based on observed data, thus aiding clinicians in refining the diagnostic process. Case studies demonstrate the module's capacity to suggest overlooked symptoms that enhance diagnostic accuracy. Nonetheless, further development, including the incorporation of real-world HPO term frequencies and gene-phenotype relationships [54], would improve the accuracy and relevance of recommendations. Additionally, expanding the Recommender's ability to account for more complex and variable phenotypic presentations would increase its clinical utility and adaptability [55]. However, the phenotype annotations from resources such as OMIM and Orphanet may be incomplete or biased toward well-documented features. Consequently, the recommended HPO terms may not always provide effective discriminatory power in real-world scenarios, especially when the phenotype spectrum of diseases is broader or underreported. We also observed that the model may occasionally generate candidate gene suggestions, even without explicit prompting. While such outputs can reflect biologically meaningful

associations—as in the case of *FGFR2* and *TNXB*—they should be interpreted cautiously. These inferences are shaped by the model's training data and are not validated clinical recommendations. Users should evaluate such suggestions within the appropriate clinical or experimental context.

Looking ahead, PhenoDP's modular design provides a robust foundation for future enhancements. Key areas of development include the integration of gene-phenotype associations, expansion of the semantic data corpus, and leveraging advancements in long-token processing technologies. These improvements will enhance diagnostic accuracy while ensuring adaptability to evolving clinical needs.

In conclusion, PhenoDP represents a transformative platform for genetic disease diagnostics through its integration of advanced large language models (LLMs) and deep learning techniques. This integration streamlines clinical workflows and enhances diagnostic processes. With continued refinement, it has the potential to further revolutionize clinical workflows, offering healthcare professionals more personalized, data-driven, and efficient diagnostic support.

Conclusions

PhenoDP offers a transformative approach to Mendelian disease diagnosis by integrating deep learning into a modular, phenotype-centric framework. Its three core modules—Summarizer, Ranker, and Recommender—collectively address key diagnostic challenges, delivering structured clinical summaries, accurate disease prioritization, and context-aware symptom recommendations.

The Summarizer, leveraging a distilled large language model, generates patient-centered clinical summaries that provide actionable insights for clinicians. The Ranker, combining multiple similarity measures, achieves superior disease prioritization, as demonstrated by its consistent outperformance of existing tools across simulated and real-world datasets. The Recommender, powered by contrastive learning, enhances differential diagnosis by suggesting discriminative HPO terms, outperforming both traditional systems and advanced generative models like GPT-4o.

With its open-source availability and modular design, PhenoDP seamlessly integrates into clinical workflows, offering scalability and adaptability. Future enhancements, such as incorporating gene-phenotype associations and expanding semantic data, will further strengthen its diagnostic capabilities. By improving accuracy, efficiency, and transparency, PhenoDP empowers clinicians to make faster, more confident decisions, marking a significant step forward in rare disease diagnostics.

Abbreviations

HPO	Human Phenotype Ontology
LLM	Large language model
MRR	Mean reciprocal rank
IC	Information content
GCN	Graph convolutional network
DAG	Directed acyclic graph
LoRA	Low-rank adaptation
CoT	Chain-of-thought
PSD-HPOEncoder	Pre-trained Semantic & DAG-based HPO Encoder
PCL-HPOEncoder	Phenotype Contrastive Learning HPO Encoder
WGS	Whole-genome sequencing
WES	Whole-exome sequencing
OMIM	Online Mendelian Inheritance in Man
WMD	Word Mover's Distance

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01496-8.

Additional file 1. Includes two figures of additional results. Fig. S1 Ablation study and performance evaluation of PhenoDP's Ranker; Fig. S2 Supplementary results for evaluating model performance before and after fine-tuning.

Additional file 2. Includes three tables related to language model examples and one table presenting additional results. Table S1 An example of a biomedical entry used for training Bio-Medical-3B-CoT; Table S2 Example of prompt templates for term recommendation using GPT-4o, with "{}" representing a replaceable entity; Table S3 The performance of difference disease ranking methods using patent terms extracted by various automated tools in real-world dataset 2; Table S4 Example of prompt templates for clinical case report generating.

Additional file 3. Includes one table for real-world dataset 2. Table S5 Sources of real-world dataset 2 and the corresponding sets of HPO terms extracted using three different tools.

Acknowledgements

We sincerely thank our research team members—Qiming Liu, Tian Yao, Lejin Tian, Yuqing Su, Yucan Zhu, Minghan Li, Xiang Zeng, Qing Wen, Liu Liu, and Yizhou Tang—for their valuable suggestions throughout this study.

Authors' contributions

W.T. designed the overall project and methodology. B.W. developed and implemented the method and performed data analysis. B.W., S.S., Y.L., and Y.D. carried out data visualization. W.T. and B.W. drafted and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32370719, No. 32170667).

Data availability

The datasets analysed during the current study are available in public repositories. The disease definitions and phenotype annotations were obtained from the Human Phenotype Ontology [16] (HPO, https://hpo.jax.org/), OMIM [13] (https://www.omim.org/), and Orphanet [18] (https://www.orpha.net/). The benchmark dataset for clinical summarization was sourced from SUMPUBMED [29], which is available at https://github.com/vgupta123/sumpubmed. The first and third real-world patient phenotype datasets were derived from the LIRICAL benchmark [24, 40], which is publicly available at https://github.com/TheJacksonLaboratory/LIRICAL. The second dataset is provided in Additional file 3: Table S5, and details of its preprocessing are described in the Methods section. The DeepSeek-R1 model [27] used to generate patient-centered summaries is available at https://www.deepseek.com/, and the Bio-Medical-3B-CoT model [28] is publicly accessible through Hugging Face at https://huggingface.co/ContactDoctor/Bio-Medical-3B-CoT-012025. The source code of PhenoDP, including the Summarizer, Ranker, and Recommender modules,

as well as the fine-tuned models and necessary preprocessing files, is available at https://github.com/TianLab-Bioinfo/PhenoDP.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 November 2024 Accepted: 27 May 2025
Published online: 06 June 2025

References

1. McCandless SE, Brunger JW, Cassidy SB. The burden of genetic disease on inpatient care in a children's hospital. Am J Hum Genet. 2004;74(1):121–7.
2. Baird PA, Anderson TW, Newcombe HB, Lowry RB. Genetic disorders in children and young adults: a population study. Am J Hum Genet. 1988;42(5):677–93.
3. Ferreira CR. The burden of rare diseases. Am J Med Genet A. 2019;179(6):885–92.
4. Zemojtel, T., et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014 Sep 3;6(252):252ra123.
5. Tan TY, Dillon OJ, Stark Z, et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions[J]. JAMA Pediatr. 2017;171(9):855–62.
6. Ewans LJ, Minoche AE, Schofield D, et al. Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. Eur J Hum Genet. 2022;30:1121–31.
7. Ewans LJ, Schofield D, Shrestha R, et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genet Med. 2018;20:1564–74.
8. Dillon OJ, Lunke S, Stark Z, et al. Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders[J]. Eur J Hum Genet. 2018;26(5):644–51.
9. Iglesias A, Anyane-Yeboa K, Wynn J, et al. The usefulness of whole-exome sequencing in routine clinical practice[J]. Genet Med. 2014;16(12):922–31.
10. Dragojlovic N, Elliott AM, Adam S, et al. The cost and diagnostic yield of exome sequencing for children with suspected genetic disorders: a benchmarking study[J]. Genet Med. 2018;20(9):1–9.
11. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser[J]. Nat Protoc. 2015;10(12):2004–15.
12. Saunders C J, Miller N A, Soden S E, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units[J]. Science translational medicine, 2012, 4(154): 154ra135–154ra135.
13. Amberger J S, Bocchini C A, Schiettecatte F, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders[J]. Nucleic acids research, 2015, 43(D1): D789-D798.
14. Yuan X, Wang J, Dai B, et al. Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases[J]. Briefings in Bioinformatics, 2022, 23(2): bbac019.
15. Jagadeesh KA, Paggi JM, Ye JS, et al. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing[J]. Nat Genet. 2019;51(4):755–63.
16. Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021[J]. Nucleic Acids Res. 2021;49(D1):D1207–17.
17. Feng Y, Qi L, Tian W. PhenoBERT: a combined deep learning method for automated recognition of human phenotype ontology[J]. IEEE/ACM Trans Comput Biol Bioinf. 2022;20(2):1269–77.
18. Weinreich SS, Mangon R, Sikkens JJ, et al. Orphanet: a European database for rare diseases[J]. Ned Tijdschr Geneeskd. 2008;152(9):518–9.
19. Muffels I J J, Wiame E, Fuchs S A, et al. NAA80 bi-allelic missense variants result in high-frequency hearing loss, muscle weakness and developmental delay[J]. Brain communications, 2021, 3(4): fcab256.
20. Jagadeesh KA, Birgmeier J, Guturu H, et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization[J]. Genet Med. 2019;21(2):464–70.
21. Chen J, Xu H, Jegga A, et al. Novel phenotype–disease matching tool for rare genetic diseases[J]. Genet Med. 2019;21(2):339–46.
22. Li Z, Zhang F, Wang Y, et al. PhenoPro: a novel toolkit for assisting in the diagnosis of Mendelian disease[J]. Bioinformatics. 2019;35(19):3559–66.
23. Zhao M, Havrilla J M, Fang L, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases[J]. NAR genomics and Bioinformatics, 2020, 2(2): lqaa032.
24. Robinson PN, Ravanmehr V, Jacobsen JOB, et al. Interpretable clinical genomics with a likelihood ratio paradigm[J]. The American Journal of Human Genetics. 2020;107(3):403–17.
25. Zhai W, Huang X, Shen N, et al. Phen2Disease: a phenotype-driven model for disease and gene prioritization by bidirectional maximum matching semantic similarities[J]. Briefings in Bioinformatics, 2023, 24(4): bbad172.
26. Girdea, Marta, et al. "PhenoTips: patient phenotyping software for clinical and research use." Human mutation 34.8 (2013): 1057–1065.
27. Guo, Daya, et al. "DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning." arXiv preprint arXiv:2501.12948 (2025).
28. ContactDoctor. Bio-Medical-3B-CoT: a high-performance biomedical language model with reasoning capabilities. 2025. Available at: https://huggingface.co/ContactDoctor/Bio-Medical-3B-CoT.
29. Gupta, Vivek, et al. "SumPubMed: summarization dataset of PubMed scientific articles." Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing: student research workshop. 2021.
30. Devalal, Shilpa, and A. Karthikeyan. "LoRa technology-an overview." 2018 second international conference on electronics, communication and aerospace technology (ICECA). IEEE, 2018.
31. Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy[J]. arXiv preprint cmp-lg/9709008, 1997.
32. Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.
33. Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.
34. Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.
35. Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." Journal of Machine Learning Research 25.70 (2024): 1–53.
36. Li, Jianning, et al. "ChatGPT in healthcare: a taxonomy and systematic review." Computer Methods and Programs in Biomedicine 245 (2024): 108013.
37. He, Kai, et al. "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics." Information Fusion (2025): 102963.
38. Luo L, Yan S, Lai PT, et al. PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology[J]. Bioinformatics. 2021;37(13):1884–90.
39. Deisseroth CA, Birgmeier J, Bodle EE, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis[J]. Genet Med. 2019;21(7):1585–93.
40. Danis, Daniel, et al. "A corpus of GA4GH Phenopackets: case-level phenotyping for genomic diagnostics and discovery." Human Genetics and Genomics Advances 6.1 (2025).
41. Kusner M, Sun Y, Kolkin N, et al. From word embeddings to document distances[C]//International conference on machine learning. PMLR, 2015: 957–966.
42. Yasunaga M, Leskovec J, Liang P. Linkbert: Pretraining language models with document links[J]. arXiv preprint arXiv:2203.15827, 2022.
43. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3(1):1–23.

44. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies[J]. The American Journal of Human Genetics. 2009;85(4):457–64.
45. Lanternier, Fanny, et al. "Deep dermatophytosis and inherited CARD9 deficiency." New England Journal of Medicine 369.18 (2013): 1704–1714.
46. Dinauer, Mary C. "Primary immune deficiencies with defects in neutrophil function." Hematology 2014, the American Society of Hematology education program book 2016.1 (2016): 43–50.
47. Navabi B, Upton JEM. Primary immunodeficiencies associated with eosinophilia. Allergy Asthma Clin Immunol. 2016;12:1–12.
48. Justiz-Vaillant, Angel A., et al. "Severe combined immunodeficiency—classification, microbiology association and treatment." Microorganisms 11.6 (2023): 1589.
49. Harrell, Daniel Toshio, et al. "Genotype-dependent differences in age of manifestation and arrhythmia complications in short QT syndrome." International journal of cardiology 190 (2015): 393–402.
50. Roberts, Jason D., et al. "Loss-of-function KCNE2 variants: true monogenic culprits of long-QT syndrome or proarrhythmic variants requiring secondary provocation?." Circulation: Arrhythmia and Electrophysiology 10.8 (2017): e005282.
51. Katoh M. Cancer genomics and genetics of FGFR2[J]. Int J Oncol. 2008;33(2):233–7.
52. Vanakker O, Callewaert B, Malfait F, et al. The genetics of soft connective tissue disorders[J]. Annu Rev Genomics Hum Genet. 2015;16(1):229–55.
53. Shorten C, Khoshgoftaar TM, Furht B. Text data augmentation for deep learning[J]. Journal of big Data. 2021;8(1):101.
54. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes[J]. Genome medicine. 2015;7:1–11.
55. Ely JW, Graber ML. Preventing diagnostic errors in primary care[J]. Am Fam Physician. 2016;94(6):426–32.

Publisher's Note