Research Article

Journal of Advances in Artificial Intelligence and Machine Learning

Integrating Genomic and Phenotypic Data for Gene Prioritization: AI Performance Assessment of the InheriNext® Algorithm

Ju-Yuan Chang[1], Kuan-Tsung Li[1], Yu-Shen Tsai[1], Michael Kubal[1], Aaron M Hamby[2], Naomi Thomson[1], Jonathan Sheridan, Shiloh Barfield, Randy Rutz, Frank S Ong, Ramon Felciano, Scott Kahn* and Shao-Min Wu

[1]*Compass Bioinformatics Inc. TX, USA 13400 Briarwick Drive, Unit 103, Austin, Texas, USA.*

[2]*Independent consultant.*

*Correspondence:

Scott D Kahn, Compass Bioinformatics Inc. TX, USA 13400 Briarwick Drive, Unit 103, Austin, Texas, USA.

Received: 11 Oct 2025; **Accepted:** 15 Nov 2025; **Published:** 28 Nov 2025

Citation: Ju-Yuan C, Kuan-Tsung Li, Yu-Shen T, et al. Integrating Genomic and Phenotypic Data for Gene Prioritization: AI Performance Assessment of the InheriNext® Algorithm. J Adv Artif Intell Mach Learn. 2025; 1(1): 1-12.

ABSTRACT

This study presents a comprehensive benchmark analysis of InheriNext®, a domain-specific, AI-powered tool designed for phenotype-driven pathogenic variant prioritization. For this study, 7,244 whole exome test cases were generated using phenotype and genotype data from Phenopackets, along with pools of variants from healthy individuals to serve as genomic backgrounds. Performance was evaluated across diverse testing scenarios and compared against four established tools. The results show InheriNext® achieving a 98.6% sensitivity in identifying pathogenic variants and consistent performance across diverse tests for variant types, phenotype counts, and disease groups—supporting the robustness and adaptability of its methodology. Sharing these benchmarking results and samples is intended to drive progress by assisting clinicians and researchers in evaluating interpretation tools and identifying areas for improvement.

Keywords

InheriNext®, phenotype-driven variant prioritization, Pathogenic variant detection, Whole exome sequencing (WES), AI-powered genomics.

Introduction

Rare diseases, defined as conditions affecting fewer than 200,000 individuals in the United States or fewer than 1 in 2,000 people in the European Union, still collectively impact over 400 million people worldwide, highlighting a significant global health challenge [1]. Many of these diseases, often caused by subtle genetic mutations, go undiagnosed for extended periods. Accurate identification of genetic variants is essential; however, with over 10 million genetic variations present in an average human genome, effective prioritization becomes crucial. Advanced sequencing technologies, such as Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS), provide efficient methods for profiling genetic data. Through these sequencing methods, a patient's specific genetic variants are identified, followed by prioritization based on the pathogenicity of the variants and the phenotypes relevant to the patient [2]. Previous research has highlighted the benefits of enhancing diagnostic accuracy by integrating phenotype data into variant prioritization algorithms [3,4]. However, there remains significant room for improvement. For example, some consequences of genetic variants are difficult to interpret, as their functional impacts may not correspond to their actual pathogenicity.

Several current computational tools utilize clinical phenotype data annotated with HPO terms to rank candidate genes based on established phenotypic and genetic knowledge. Exomiser employs a logistic regression model that integrates variant-based and gene-based scores to generate a final prioritization score [5]. Variant-based scores are influenced by allele frequency, variant type, and pathogenicity predictions from tools. LIRICAL adopts a Bayesian statistical framework to assess candidate diagnoses by computing posterior probabilities based on likelihood ratios (LRs). It integrates *in silico* pathogenicity predictions and phenotype-based LRs, refining genotype-disease associations and improving diagnostic accuracy, particularly in rare disease contexts [6]. Xrare utilizes a machine learning-based approach for prioritizing disease-causing variants by integrating genetic data with phenotypic similarity scores. By leveraging deep learning techniques, Xrare enhances the identification of pathogenic variants and adapts to complex genotype-phenotype relationships, making it a powerful tool for clinical diagnostics [7]. AMELIE sets itself apart by leveraging biomedical literature at scale, analyzing millions of PubMed abstracts and full-text articles to support molecular diagnosis. It employs a logistic regression classifier trained on simulated patient data, allowing it to rank causative variants effectively. By dynamically incorporating new scientific findings, AMELIE improves variant interpretation in the evolving landscape of genomic research [8].

InheriNext® is a domain-specific AI-powered tool for the prioritization of genetic variants through two approaches: one utilizes Human Phenotype Ontology (HPO)-based phenotyping [9] and the other focuses on gene panels to identify causative SNPs and INDELs. It employs three scoring systems: phenotype-correlation score, variant pathogenicity score, and disease-similarity score, which aims to offer a more comprehensive framework for analyzing the complexity and diversity of phenotypes.

To evaluate the performance of gene-ranking tools in identifying causative variants, resources from the Global Alliance for Genomics and Health (GA4GH) were adopted as the benchmarking foundation. The GA4GH Phenopacket dataset is particularly well-suited for benchmarking due to its standardized schema, rich phenotypic detail, and expert-curated variant annotations, which together ensure consistency and clinical relevance in evaluation. By leveraging a widely accepted and biologically diverse dataset, this study enables robust and reproducible comparisons across gene- prioritization methods within realistic diagnostic scenarios.

In 2022, GA4GH introduced the Phenopacket Schema, an ISO-approved standard designed to share detailed clinical and genomic information at the individual level. A Phenopacket links phenotypic features with disease diagnoses, patient data, and genetic variants, thereby enabling the construction of accurate disease models [10]. The GA4GH Phenopacket Store (v0.1.21) provides a comprehensive benchmarking dataset comprising of 7,830 Phenopacket samples covering 489 Mendelian and chromosomal diseases linked to 432 genes and 4,263 unique pathogenic alleles. Previous studies have demonstrated the utility of GA4GH Phenopackets as a benchmark for assessing gene-ranking methods [11].

In this study, the ranking performance of InheriNext® is compared against other established diagnostic tools, with the evaluation focusing on the identification of causative gene variants.

Materials and Methods

To benchmark different variant prioritization approaches, the GA4GH Phenopackets were used to impute a dataset of simulated genetic samples. Genomes from 600 healthy individuals were obtained from the 1,000 Genomes Project, representing genetic diversity across various populations. Considering the population bottlenecks experienced by European, Asian, and American populations—which drastically reduced genetic diversity in these groups [12]—the below method was developed to generate synthetic Whole Exome Sequencing (WES) samples.

Variants from a diverse selection of healthy individuals were pooled to create a rich reservoir from which 50,000 variants were randomly drawn to construct each simulated case of a “healthy” (normal genetic variability) exome. This pooling strategy ensured that the simulated exomes captured a broad spectrum of genetic variability for analysis. Next, data from 7,244 of 7,830 Phenopacket samples that meet the analysis criteria (e.g.: annotated with phenotypic features) were used. Each known disease-causing variant was added into a synthetic healthy exome along with patient’s phenotypic features to simulate a patient with that genetically-driven disease or disorder (Figure 1).

This process resulted in 7,244 test cases representing disease patients, with each synthesized sample containing annotated phenotypic features, one pathogenic variant from Phenopacket, and alongside 50,000 background variants. These samples were finalized for further analysis and used to evaluate the ranking performance of InheriNext® against four (4) other commonly used variant prioritization tools (A-D).

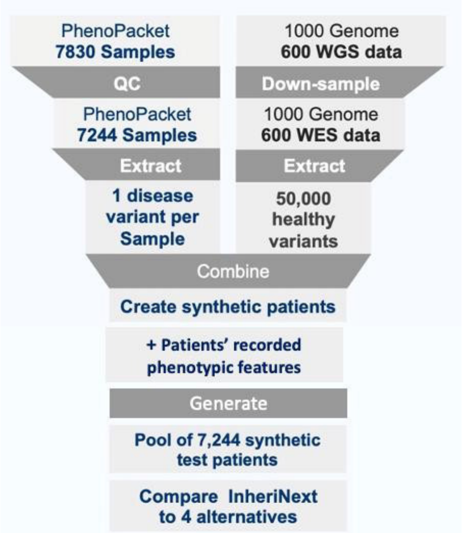

PhenoPacket
7830 Samples
1000 Genome
600 WGS data
QC
Down-sample
PhenoPacket
7244 Samples
1000 Genome
600 WES data
Extract
Extract
1 disease variant per Sample
50,000 healthy variants
Combine
Create synthetic patients
+ Patients' recorded phenotypic features
Generate
Pool of 7,244 synthetic test patients
Compare InheriNext to 4 alternatives

Figure 1: Workflow for generating simulated samples.

The diagram outlines the process for generating synthetic patients. Six-hundred (600) healthy individuals’ background genomes were extracted from the 1000 Genome Project, and the causative variants sourced from Phenopacket. After filtering, 7,244 synthetic samples, along with their respective phenotypic features, are created for the following benchmark study.

The benchmark samples encompass nearly five hundred diseases. Grouping similar diseases helps reveal their distribution for the

Phenopacket, facilitating a better understanding and allowing for performance analysis across different disease groups. Steps taken to group diseases by K-means clustering using Term Frequency-Inverse Document Frequency (TF-IDF) and Principal Component Analysis (PCA) are described in supplementary data (Figure S1., Figure S2., Figure S3., and Table S1.).

Results and Discussions

Ranking Performance in Benchmark Samples

InheriNext® is benchmarked against four (4) different software tools A-D (Exomiser, LIRICAL, Xrare, Amelie) that also utilize phenotype-driven gene prioritization methods to rank candidate pathogenic genes in the 7,244 samples. The "Top-10 Rate" is a practical benchmark for evaluating tool performance, ensuring that the causative gene is captured within the top ranks. It is often visualized using the Cumulative Distribution Function (CDF) plot, which displays the proportion of genes that fall below any given rank. According to previous literature, this method offers an intuitive way to assess and compare the effectiveness of different tools in ranking causative genes [13]. Results show that InheriNext® identified the causative gene within the Top-10 ranks in 98.6% of cases. The corresponding Top-10 rates for tools A, B, C, and D were 95.0%, 86.2%, 87.0%, and 90.9%, respectively (Figure 2). The results indicate that InheriNext® achieved the highest Top-10 capture rate in this benchmark comparison.

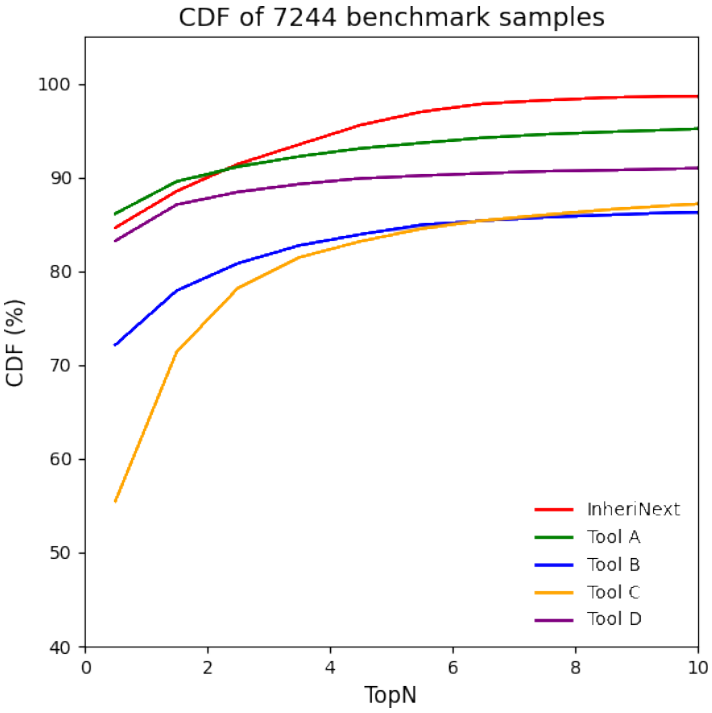


Figure 2: Performance evaluation of InheriNext® with other tools (A-D).

The cumulative distribution function (CDF) shows the ranking distribution of causative genes across all five (5) tools. The CDF plots illustrate the percentage of the samples with causative genes ranked within the top*N* by each tool. *N* could be any integer between 1 and 10. Each tool is represented by a different color.

Ranking Performance in Different Variant Consequences

Variant consequences are pivotal in identifying pathogenic variants, which play a crucial role in diagnosing genetic disorders. The Sequence Ontology (SO) offers standardized annotations to describe the effects of genetic variants on biological sequences, helping categorize their impact on genes, transcripts, and other features to understand their functional implications (Table S2). This knowledge enables clinicians to pinpoint the genetic basis of a disease accurately, facilitating more precise diagnostics. By assessing the functional impact of these variants, one can evaluate the causative genes more precisely in algorithmic calculations. An analysis of the distribution of variant consequences in the benchmarked samples was conducted, revealing 24 distinct types of variant consequences. In some cases, a variant exhibits multiple types of consequences, resulting in duplicated counts. Table 1A indicates that the three (3) most common types of consequences are missense, comprising 49.05% of the total, stop gained at 12.34%, and frameshift truncation accounting for 9.10%. The subsequent evaluation uses the Top-10 rate to assess each tool's performances in identifying causative genes across different variant consequences (refer to Table1B). InheriNext® achieved the highest Top-10 capture rates for the most frequent, therefore relevant, consequence types (missense variant, stop gained, and frameshift truncation). However, InheriNext® falls short in the synonymous variant category, achieving only a 30.43% in the Top-10 rankings.

Table 1: Distribution of Variant Consequences in Benchmark Samples. (A) The descending proportion and counts of 24 variant consequences identified in benchmark samples. (B) The breakdown of variant consequences by the five (5) tools. The percentages reflect the ability of each tool to rank causative genes within their Top- 10s for each consequence.

(A)

Consequences	Proportion % (count)
Missense variant	49.05 (4345)
Stop gained	12.34 (1093)
Frameshift truncation	9.10 (806)
Coding transcript intron variant	6.48 (574)
Frameshift variant	4.96 (439)
Splice region variant	4.00 (354)
Splice donor variant	3.04 (269)
Frameshift elongation	2.34 (207)
Splice acceptor variant	1.96 (174)
Disruptive inframe deletion	1.51 (134)
Complex substitution	1.12 (99)
Inframe deletion	0.98 (87)
5 prime UTR exon variant	0.72 (64)
Feature truncation	0.58 (51)
Start lost	0.52 (46)
Multi-nucleotide variant	0.33 (29)
Direct tandem duplication	0.26 (23)

Synonymous variant	0.26 (23)
Disruptive inframe insertion	0.25 (22)
3 prime UTR intron variant	0.07 (6)
5 prime UTR intron variant	0.06 (5)
Stop lost	0.03 (3)
Inframe insertion	0.03 (3)
Exon loss variant	0.03 (3)

(B)

Consequences	InheriNext®	Tool A	Tool B	Tool C	Tool D
Missense variant	**98.64**	93.9	87.59	89.07	94.38
Stop gained	**99.27**	97.8	86.55	86.18	91.49
Frameshift truncation	**99.38**	98.39	91.44	89.33	93.05
Coding transcript intron variant	**98.08**	93.55	74.22	81.88	75.44
Frameshift variant	**98.86**	96.58	81.09	83.83	87.24
Splice region variant	**94.63**	94.35	76.27	85.59	66.67
Splice donor variant	**99.63**	98.51	86.99	86.25	95.54
Frameshift elongation	98.55	**100**	92.27	92.27	96.62
Splice acceptor variant	**99.43**	94.25	86.21	83.33	91.38
Disruptive inframe deletion	**99.25**	98.51	94.03	96.27	96.27
Complex substitution	**100**	**100**	93.94	94.95	97.98
Inframe deletion	**98.85**	96.55	83.91	83.91	81.61
5 prime UTR exon variant	**85.94**	82.81	1.56	3.12	3.12
Feature truncation	**100**	**100**	**100**	98.04	98.04
Start lost	**100**	84.78	97.83	91.3	**100**
Multi-nucleotide variant	**100**	86.21	68.97	82.76	82.76
Direct tandem duplication	**100**	86.96	73.91	95.65	86.96
Synonymous variant	30.43	60.87	43.48	**78.26**	21.74
Disruptive inframe insertion	**100**	86.36	68.18	95.45	86.36
3 prime UTR intron variant	**100**	**100**	**100**	**100**	**100**
5 prime UTR intron variant	**100**	60	**100**	**100**	**100**
Stop lost	**100**	**100**	33.33	**100**	**100**
Inframe insertion	**100**	**100**	**100**	**100**	**100**
Exon loss variant	**100**	**100**	**100**	**100**	**100**

Assessment of Diverse Annotated Phenotype Counts

Phenotypic features, as annotated in the Phenopacket schema, describe clinical symptoms in patients and are used to illustrate the similarity between disease- associated features and those present in a patient. InheriNext® integrates the HPO project [9], which provides an ontology of medically relevant phenotypic features and disease-phenotype annotations, to calculate the likelihood of diseases based on the phenotypic features observed in patients. For example, "arachnodactyly" is a relevant feature for "Marfan syndrome" in HPO database, so if a patient exhibits "arachnodactyly," they are more likely to have the disease. InheriNext® and other tools use phenotype-driven methods to rank potential pathogenic genes based on phenotypic features in addition to patient genotype data. In the benchmark samples, diverse annotated phenotype counts are inputted for each sample based on phenotypic features in the Phenopacket record used to create that particular simulated case. However, a large count might decrease the performance in ranking causative genes because it may include more unrelated (noise) features, which are not annotated in the disease. For phenotype counts distribution, the results show that: the range of 6-10 phenotypes is most common, seen in 1821 samples, followed closely by the 21-50 range with 1737 samples (Figure 3A). The number of phenotype inputs across each tool's Top-10s were evaluated to assess performance. The results show that InheriNext® demonstrates consistent performance with percentages ranging from 98% to 100%, indicating strong capability across all phenotype input ranges. Tool A maintains high percentages, reaching 97% in the 6-10 and 11-15 ranges, but drops to 82% in the >50 category. Tool B, Tool C, and Tool D remain steady with over 82% in most ranges; however, they experience a more significant drop in the >50 category (Figure 3B).

Evaluation of Ranking Performance in Four Disease Groups

Each benchmark sample is annotated with a specific disease. However, some diseases do not match a specific MONDO ID and are therefore removed, leaving 7215 samples for analysis. The Phenopacket-annotated diseases were mapped to second-layer categories from MONDO ontologies (Figure S1) to generate a word matrix for TF-IDF similarity calculation (Figure S2 A). The optimal number of clusters for the Phenopacket diseases was determined using k-means clustering with the elbow method applied to the word matrix. This analysis classified the diseases into four major groups, as shown in Figure. S3. To enhance interpretability, these four groups were subsequently renamed with more representative titles, as detailed in Table S1 B.

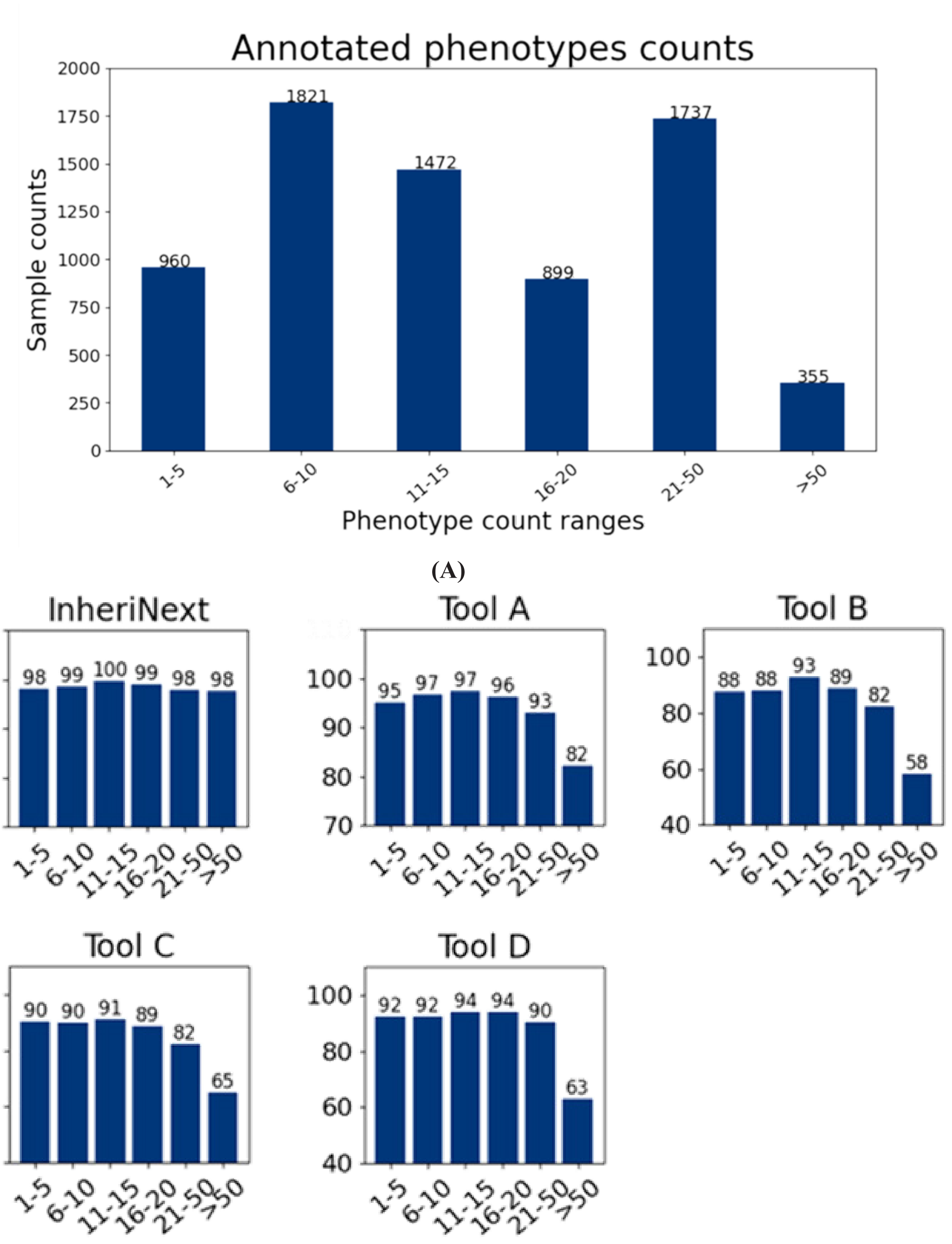
The scatter plot below displays the distribution of different diseases across four major disease groups, with each data point symbolizing a distinct disease. The size of each point reflects the number of samples for that disease. As reflected by the dense cluster of orange points, this plot clearly shows that the Nervous System and Metabolic System Disorder group has the highest number of samples (Figure 4A). A table is presented that lists examples of actual diseases found in the benchmark samples grouped in four major disease categories (Figure 4B). To evaluate ranking performance, the Top-10 rate was used as an indicator across four disease groups. The results show that InheriNext® consistently ranks over 98% of samples' causative genes within the Top-10, demonstrating non-discriminatory performance across major disease groups.

In comparison, other tools show variability based on the disease type. For example, Tool B performs well in the Cancer or Benign Tumor category but less effectively in the Nervous or Metabolic Disorder group (Figure 4C).

(Note: Disease group names were abbreviated slightly while keeping key terms for clarity and consistency:

1. Development or morphogenesis disorder, musculoskeletal system disorder = Development or musculoskeletal disorder
2. Nervous system and metabolic system disorder = Nervous or metabolic disorder
3. Cancer or benign tumor = Cancer or benign tumor
4. Visual system, orbital region disorder = Visual, orbital region disorder)

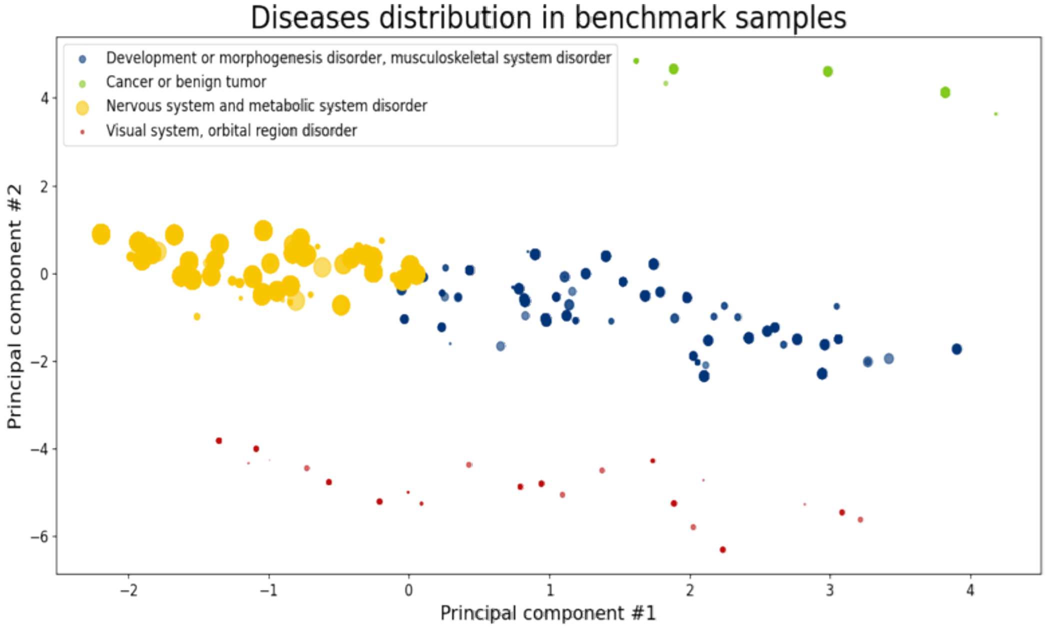
Benchmarking tests for InheriNext® showed some key performance strengths, which could be attributed to the following factors:



(A)

(B)

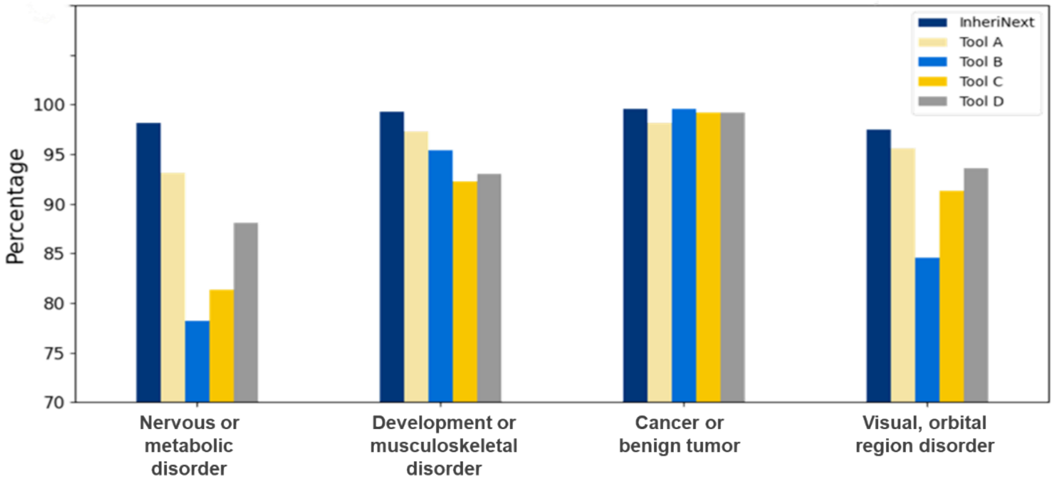
Figure 3: Distribution of Phenotypes Counts in Benchmark Samples. (A) The bar graph illustrates the distribution of samples with different ranges of annotated phenotypic feature counts. Each bar's height represents the sample count within a specific range. (B) The five bar charts represent the performance of different tools— InheriNext®, Tool A, Tool B, Tool C, and Tool D—across various feature count ranges. Each chart shows the percentage of samples' causative genes successfully ranked within the Top-10 for each phenotypic feature count range.



(A)

Disease Groups	Examples	
Development or morphogenesis disorder, musculoskeletal system disorder	KBG syndrome	Holt-Oram syndrome
Nervous system & metabolic system disorder	Cockayne syndrome	Developmental and epileptic encephalopathy 4
Cancer or benign tumor	Multiple endocrine neoplasia type 2A	Neurofibromatosis type 1
Visual system, orbital region disorder	Leber congenital amaurosis 6	Ectopia lentis, familial’

(B)



(C)

Figure 4: Disease Distribution in Benchmark. (A) The scatter plot illustrates the distribution of diseases by condensing complex data into two principal components (Principal Component #1 and Principal Component #2), highlighting similarities and clustering in four disease groups among the samples. Each point represents a unique disease, and the size indicates the number of samples for that disease; larger points signify a greater number of samples. (B) Example of actual diseases observed in the benchmark samples across four disease groups. (C) The performance comparison across the five tools shows the percentage of samples' causative genes in Top-10 distributed among the four groups of diseases (abbreviated).

(1) Integration of Disease Similarity Score Enhances InheriNext® Ranking Performance:

InheriNext® incorporates clinician feedback to refine the prioritization of pathogenic genes. One such enhancement was prompted by a request to improve the assessment of similarity between disease symptoms and patient phenotypes, leading to the integration of a disease similarity score. This feature contributes to stable ranking performance across a variety of disease conditions (Figure 4C). By aligning more closely with clinical realities, this integration supports more accurate and relevant predictions.

(2) Delay Filtering and Removal of Candidates Until After Prioritization:

InheriNext® evaluates all potential causative variants and genes, including those with a lower initial likelihood of pathogenicity, to account for their possible clinical relevance. In contrast, other tools may apply early variant consequence filters—for example, excluding 5′ UTR exon variants—which may result in the removal of relevant pathogenic candidates (Table 1B: 5′ UTR exon variant, Tools B–D). InheriNext® retains all candidates through the prioritization stage, enabling more consistent ranking performance across variant types and supporting the identification of atypical or rare variants.

Despite overall strong performance, several challenges remain. These include limitations in the classification of variants of uncertain significance (VUS) in ClinVar, the underestimation of certain high-allele-frequency variants' pathogenicity, and ambiguous functional consequences of specific variants. Benchmarking also identified reduced performance in a subset of samples. Further analysis identified two

(2) primary contributors, offering clear targets for future refinement:

(1) Variants Annotated with Synonymous Consequence Did Not Rank Well (Table 1B):

In its initial implementation, InheriNext® excluded synonymous variants from ranking due to their typically neutral effect relative to non-synonymous variants. This exclusion was intended to streamline prioritization by focusing on more likely pathogenic candidates. However, this approach led to the omission of clinically relevant variants, counter to the broader goal of comprehensive evaluation (see "Delay Filtering and Removal of Candidates Until After Prioritization"). Recent studies have shown that some synonymous variants may have functional consequences, such as impacting splicing motifs or cryptic splice sites, altering microRNA binding [14], and affecting mRNA structure or protein expression via reduced codon optimality [15-17]. In response, InheriNext® has begun to incorporate synonymous variants into the ranking process to improve sensitivity and better reflect their potential clinical relevance.

(2) High Allele Frequency of the Variants Affect Ranking:

The ClinGen Sequence Variant Interpretation Working Group has refined the ACMG/AMP variant pathogenicity guidelines for rule BA1, which designates variants with a minor allele frequency (MAF) > 0.05 as benign. However, they identified nine (9) variants with MAF > 0.05 that may exhibit pathogenicity. These findings suggest that MAF and pathogenicity do not always correlate. As a result, InheriNext® algorithm may inadvertently penalize genes based on the high allele frequency and affect the ranking accuracy. While high-frequency variants are typically deemed benign, some may still possess pathogenic potential, emphasizing the need to evaluate additional factors when assessing variant pathogenicity. Future optimizations will incorporate additional criteria for adjusting gene scores. The intention is to retain the variant frequency filter as recommended by ACMG/AMP guidelines [18]. However, specific exceptions are systematically reviewed and curated as they appear in the clinical genetics literature, allowing potentially pathogenic variants to be considered despite their relatively high frequency.

Conclusion

The benchmarking analysis demonstrates that InheriNext® reliably prioritizes candidate pathogenic genes under varied testing scenarios among the tools assessed. In evaluations of phenotype-driven gene prioritization methods, InheriNext® achieved the highest Top-10 capture rate at 98.6% and the lowest missed rate, reflecting high sensitivity. This result is supported by consistent performance across various variant consequences, particularly the most common types—missense variants, stop gains, and frameshift truncations—although some limitations were noted (and addressed) with synonymous variants. Additional tests demonstrate that it maintained ranking accuracy as phenotype complexity increased and across different disease groups, indicating its adaptability across different clinical contexts.

Taken together, these findings suggest that InheriNext® is a reliable tool for gene prioritization, with effective integration of genomic and phenotypic data. Its performance across multiple testing dimensions supports its potential utility to assist in genetic diagnostics and research applications.

Acknowledgements

The authors acknowledge Melissa Haendel and the Monarch Initiative (https://monarchinitiative.org/) in developing a standardized framework for integrating genomic, environmental, and phenotypic data [10]. Phenopackets provide a critical foundation for ensuring that phenotype profiles are FAIR++ for disease diagnosis and discovery [19]. Their efforts in standardizing data formats, including gene data (GFFs, VCFs, BED files), environmental data, and the novel PXF format, have been instrumental in advancing interoperability and data sharing in rare disease research.

Availability of Data

The benchmark simulated samples from the GA4GH Phenopacket are available from the corresponding author upon request.

Declaration of Interests

All authors are either employees or scientific advisors that work or have worked for Compass Bioinformatics, developers of InheriNext®.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve language and readability, with caution. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Groft SC, Posada M, Taruscio D. Progress, challenges and global approaches to rare diseases. Acta Paediatr. 2021; 110: 2711-2716.
2. Basel-Salmon L. Phenotypic compatibility and specificity in genomicvariant classification. Eur J Hum Genet. 2024; 32: 471-473.
3. Robinson PN, Köhler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014; 24: 340-348.
4. Jacobsen JOB, Kelly C, Cipriani V, et al. Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. Hum Mutat. 2022; 43:1071-1081.
5. Smedley D, Jacobsen JOB, Jager M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015; 10: 2004-2015.
6. Robinson P N, Ravanmehr V, Jacobsen JOB, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Am J of Hum Genet. 2020; 107: 403-417.
7. Li Q, Zhao K, Bustamante C D, et al. Xrare: A machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med. 2019; 21: 2126-2134.
8. Birgmeier J, Haeussler M, Deisseroth CA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020; 12.
9. Robinson PN, Köhler S, Bauer S, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008; 83: 610-615.
10. Danis D, Bamshad MJ, Bridges Y, et al. A corpus of GA4GH Phenopackets: Case-level phenotyping for genomic diagnostics and discovery. Human Genetics and Genomics Advances. 2024; 6: 100371.
11. Bridges Y, de Souza V, Cortes KG, et al. Towards a standard benchmark for variant and gene prioritisation algorithms: PhEval - Phenotypic inference Evaluation framework. bioRxiv. 2025.
12. Zheng-Bradley X, Flicek P. Applications of the 1000 Genomes Project resources. Brief Funct Genomics. 2017; 16: 163-170.
13. Masino AJ, Dechene ET, Dulik MC, et al. Clinical phenotype-based gene prioritization: An initial study using semantic similarity and the Human Phenotype Ontology. BMC Bioinformatics. 2014; 15: 248.
14. Supek F, Miñana B, Valcárcel J, et al. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell. 2014; 156: 1324-1335.
15. Brest P, Lapaquette P, Souidi M, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn’s disease. Nat Genet. 2011; 43: 242-245.
16. Bartoszewski RA, Jablonsky M, BartoszewskaS, et al. A Synonymous Single Nucleotide Polymorphism in ΔF508 CFTR Alters the Secondary Structure of the mRNA and the Expression of the Mutant Protein. J Biol Chem. 2010; 285: 28741-28748.
17. Kim A, Douce JL, Diab F, et al. Synonymous variants in holoprosencephaly alter codon usage and impact the Sonic Hedgehog protein. Brain. 2020; 143: 2027-2038.
18. Kim SY, Kim BJ, Oh DY, et al. Improving genetic diagnosis by disease- specific, ACMG/AMP variant interpretation guidelines for hearing loss. Sci Rep. 2022; 12: 12457.
19. Haendel M. Phenopackets: Making phenotype profiles FAIR++ for disease diagnosis and discovery. figshare Presentation. 2016.

Supplementary Materials

K-means Clustering Diseases Using Term Frequency-inverse Document Frequency (TF-IDF) and Principal Component Analysis (PCA)

(1) Building Ontology Structure:

MONDO ontologies are used to standardize disease terminology within a unified framework to establish initial relationships and processes specific MONDO for analysis [1]. The diagram below shows a simplified version of the MONDO ontology (Figure S1). At the root is "Disease" highlighted in the red box. Following it is "human disease", which branches into several second-layer categories, such as "cardiovascular disorder", "acute disease", "cancer or benign tumor", and "disorder of development or morphogenesis". In MONDO system, these second-layer categories composition terms are broken down into individual words for calculating similarity between diseases.

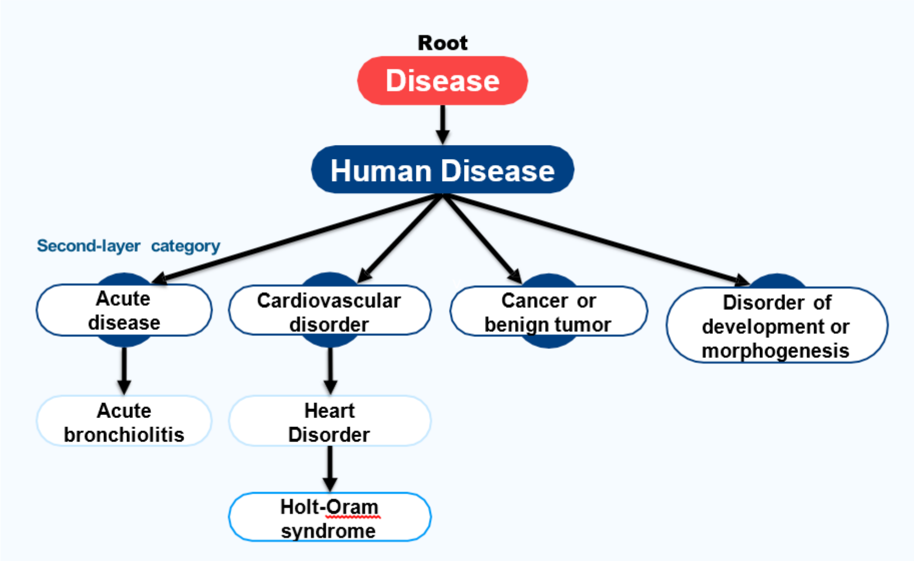
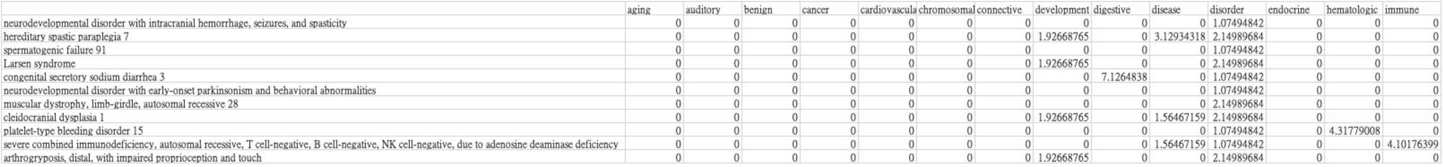


Figure S1: The simplified ontology hierarchical structure of diseases in the MONDO System. This diagram represents a portion of the MONDO disease ontology, illustrating the hierarchical relationships among various diseases. The "Root" node at the center is labeled "Disease," from which several branches extend to different disease categories and sub-categories.

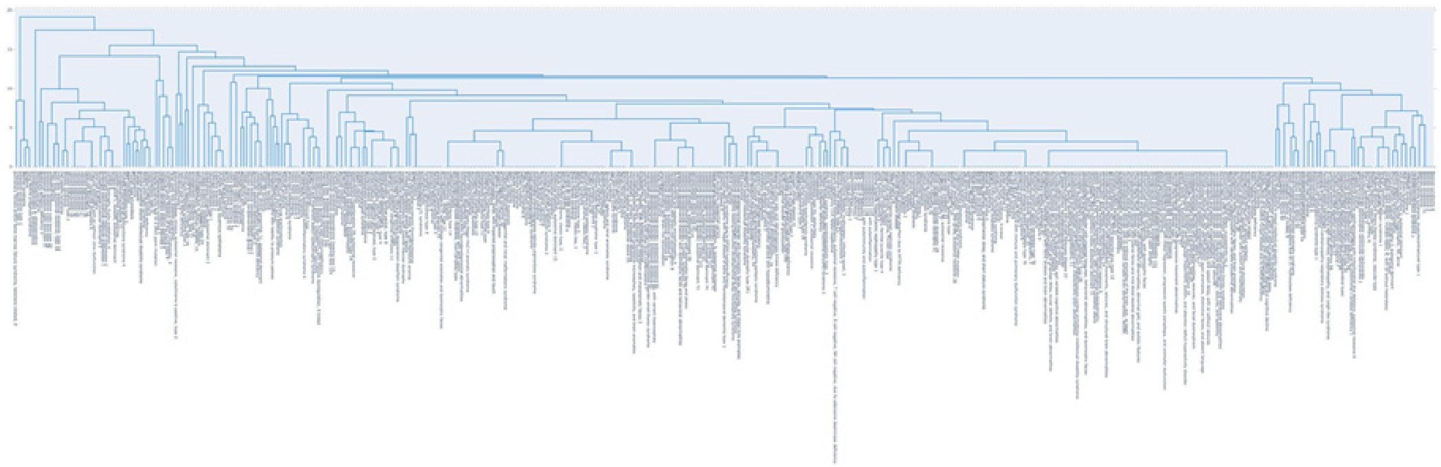
(2) Text Processing and Term Frequency-Inverse Document Frequency (TF-IDF) Calculation:

In the Phenopacket-annotated diseases, there are 28 unique second-layer categories from MONDO ontologies. These second-layer categories composition terms are broken down into individual words, with stopwords such as "the", "is", "that" removed, and assembled into a word matrix (Figure S2 A). Each Phenopacket disease's corresponding second-layer categories are then represented by a set of word vectors, which are converted into TF-IDF vectors for subsequent similarity analysis [2]. The dendrogram was used to interpret how disease terms cluster together based on their TF-IDF similarity scores. In F Figure. S2 B, a lower linkage height indicates greater similarity between terms. For example, if two disease terms are joined at a low height, they share a high degree of textual similarity based on their TF-IDF scores. In this analysis, the dendrogram visually represents the hierarchical relationships between disease terms, with the height reflecting their TF-IDF-based distances. It effectively highlights these distances, helping to distinguish closely related disease terms from more distinct ones. Furthermore, dendrograms rely on the ultrametric tree assumption, which rarely holds in real-world text analysis, and may lead to misleading representations of term similarities. To overcome these challenges, Principal component analysis (PCA) and k-means clustering were adopted for disease term classification. PCA reduces the dimensionality of TF-IDF vectors while preserving key features, improving visualization and interpretability. K-means, on the other hand, provides a scalable and flexible clustering approach that determines a specific number of groups, allowing for further evaluation and analysis of these disease clusters.



	aging	auditory	benign	cancer	cardiovascula	chromosomal	connective	development	digestive	disease	disorder	endocrine	hematologic	immune
neurodevelopmental disorder with intracranial hemorrhage, seizures, and spasticity	0	0	0	0	0	0	0	0	0	0	1.07494842	0	0	0
hereditary spastic paraplegia 7	0	0	0	0	0	0	0	1.92668765	0	3.12934318	2.14989684	0	0	0
spermatogenic failure 91	0	0	0	0	0	0	0	0	0	0	1.07494842	0	0	0
Larsen syndrome	0	0	0	0	0	0	0	1.92668765	0	0	2.14989684	0	0	0
congenital secretory sodium diarrhea 3	0	0	0	0	0	0	0	0	7.1264838	0	1.07494842	0	0	0
neurodevelopmental disorder with early-onset parkinsonism and behavioral abnormalities	0	0	0	0	0	0	0	0	0	0	1.07494842	0	0	0
muscular dystrophy, limb-girdle, autosomal recessive 28	0	0	0	0	0	0	0	0	0	0	2.14989684	0	0	0
cleidocranial dysplasia 1	0	0	0	0	0	0	0	1.92668765	0	1.56467159	2.14989684	0	0	0
platelet-type bleeding disorder 15	0	0	0	0	0	0	0	0	0	0	1.07494842	0	4.31779008	0
severe combined immunodeficiency, autosomal recessive, T cell-negative, B cell-negative, NK cell-negative, due to adenosine deaminase deficiency	0	0	0	0	0	0	0	0	0	1.56467159	1.07494842	0	0	4.10176399
arthrogryposis, distal, with impaired proprioception and touch	0	0	0	0	0	0	0	1.92668765	0	0	2.14989684	0	0	0

A



B

Figure S2: The ontology hierarchical relationship among diseases in the dendrogram based on TF-IDF similarity from a word matrix. (A) A word matrix generated from 28 unique second-layer disease categories, with category names tokenized into individual terms after removing common stop words. (B) The dendrogram illustrates the hierarchical clustering of disease terms based on TF-IDF similarity, where lower linkage heights indicate greater textual similarity. As the hierarchy builds upward, the structure may become less precise in reflecting actual semantic distances.

(3) Principal Component Analysis (PCA) and K-means Clustering:

To present the results more intuitively, PCA was applied to reduce dimensionality and capture the most significant differences in the TF-IDF vectors of Phenopacket diseases [3]. This reduction process enabled visualization in a 2D plot. After applying PCA, k-means clustering was used to group the diseases based on these reduced dimensions. K-means clustering is an unsupervised learning algorithm used for grouping unlabeled data points into distinct clusters [4]. To effectively apply k-means, the elbow method [5] was used to determine the optimal number of clusters. The optimal K was identified at the point where Within-Cluster Sum of Squares (WCSS) stops decreasing sharply, indicating a balance between cluster compactness and flexibility. The results showed that when K=4, SSE exhibited a steep decline before leveling off, suggesting that four clusters provided the most effective grouping without overfitting or unnecessary complexity (Figure S3).

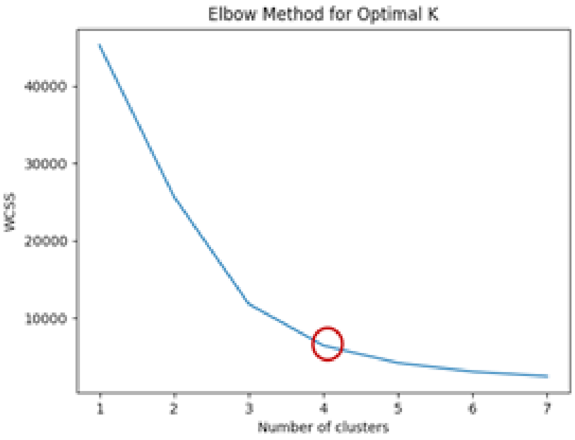


Figure S3: The Elbow Method for determining the optimal number of clusters. K values from one to seven were tested, using WCSS as a performance metric. The plot showed a sharp decline in WCSS when K = 4; beyond this point, adding more clusters did not significantly improve clustering performance, as the rate of WCSS decrease slowed down.

(4) Defining Representative Disease Terms Across Four Clusters:

After categorizing the diseases into four groups, the top two highest-frequency terms in each group were selected as representative names by Mondo ontology second-layer disease terms, providing a clearer and more direct overview of diseases in benchmark. When choosing a representative name, the International Classification of Diseases (ICD) was considered, as it offers a comprehensive disease classification system that typically includes standardized disease names [6]. Although syndromic disease is a recognized classification in the MONDO ontology system—referring to disorders that affect multiple organ systems—it cannot be classified under ICD-11 (Table S1 A). Due to its lack of specificity and limited representative power for precise disease grouping, syndromic disease was excluded to enable a clearer understanding of the disease categories. Note that only the terms within each cluster observed in the benchmark are presented in Table S1 A. Finally, representative names were selected from the top two most frequent terms within each cluster to define a more informative and meaningful disease label (Table S1 B).

Table S1: Top Two High-Frequency Terms in Each Cluster Defining the Representative Disease Group Name. (A) ICD-11 is considered for selecting the representative disease term. (B) In the MONDO ontology system, 28 unique second- layer terms are used to classify disease groups, providing a structured hierarchy for disease representation. The table presents the top two highest-frequency terms from each of the four clusters, which will be considered in defining a representative disease term.

A

ICD-11 for Disease Classifications System	Mondo disease terms (second layer)
01 Certain infectious or parasitic diseases	
02 Neoplasms	cancer or benign tumor
03 Diseases of the blood or blood-forming organs	
04 Diseases of the immune system	
05 Endocrine, nutritional or metabolic diseases	
06 Mental, behavioural or neurodevelopmental disorders	disorder of development or morphogenesis
07 Sleep-wake disorders	
08 Diseases of the nervous system	nervous system disorder
09 Diseases of the visual system	disorder of orbital region disorder of visual system
10 Diseases of the ear or mastoid process	
11 Diseases of the circulatory system	
12 Diseases of the respiratory system	
13 Diseases of the digestive system	
14 Diseases of the skin	
15 Diseases of the musculoskeletal system or connective tissue	musculoskeletal system disorder
16 Diseases of the genitourinary system	
17 Conditions related to sexual health	
18 Pregnancy, childbirth or the puerperium	
19 Certain conditions originating in the perinatal period	
20 Developmental anomalies	
21 Symptoms, signs or clinical findings, not elsewhere classified	
22 Injury, poisoning or certain other consequences of external causes	
23 External causes of morbidity or mortality	
24 Factors influencing health status or contact with health services	
25 Codes for special purposes	
26 Supplementary Chapter Traditional Medicine Conditions	

B

Clusters	Mondo disease terms	Term	Representative Disease
	(second layer)	frequency	Classification
Class 0	disorder of orbital region	357	Visual system, orbital
	disorder of visual system	357	region disorder
Class 1	cancer or benign tumor	474	Cancer or benign tumor
	syndromic disease	474	
Class 2	nervous system disorder	2254	Nervous system and metabolic system disorder
	metabolic disease	696	
Class 3	disorder of development or morphogenesis	2035	Development or morphogenesis disorder, musculoskeletal system disorder
	syndromic disease	1925	
	musculoskeletal system disorder	1185	

Table S2: Variant Consequence Annotations in the Sequence Ontology. This table illustrates the standardized annotations provided by the Sequence Ontology to describe the consequences of genetic variants on biological sequences. The columns include the SO ID, the corresponding SO term, and the description of each term. Only the consequences observed in the benchmark samples are presented in this figure.

SO Term	SO ID	Description
Missense variant	SO:0001583	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved.
Stop gained	SO:0001587	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened polypeptide.
Frameshift truncation	SO:0001910	A frameshift variant that causes the translational reading frame to be shortened relative to the reference feature.
Coding transcript intron variant	SO:0001969	A transcript variant occurring within an intron of a coding transcript.
Frameshift variant	SO:0001589	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three.
Splice region variant	SO:0001630	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.
Splice donor variant	SO:0001575	A splice variant that changes the 2 base pair region at the 5' end of an intron.
Frameshift elongation	SO:0001909	A frameshift variant that causes the translational reading frame to be shortened relative to the reference feature.
Splice acceptor variant	SO:0001574	A splice variant that changes the 2 base pair region at the 5' end of an intron.
Disruptive inframe deletion	SO:0001826	An inframe decrease in cds length that deletes bases from the coding sequence starting within an existing codon.
Complex substitution	SO:1000005	When no simple or well defined DNA mutation event describes the observed DNA change, the keyword \"complex\" should be used. Usually there are multiple equally plausible explanations for the change.
Inframe deletion	SO:0001822	An inframe non synonymous variant that deletes bases from the coding sequence.
5 prime UTR exon variant	SO:0002092	A UTR variant of exonic sequence of the 5'UTR.
Feature truncation	SO:0001906	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence.
Start lost	SO:0002012	A codon variant that changes at least one base of the canonical start codon.
MNV	SO:0002007	An MNV is a multiple nucleotide variant (substitution) in which the inserted sequence is the same length as the replaced sequence.
Direct tandem duplication	SO:1000039	A tandem duplication where the individual regions are in the same orientation.
Synonymous variant	SO:0001819	A sequence variant where there is no resulting change to the encoded amino acid.
Disruptive inframe insertion	SO:0001824	An inframe increase in cds length that inserts one or more codons into the coding sequence within an existing codon.
3 prime UTR intron variant	SO:0002090	A UTR variant of intronic sequence of the 3' UTR.
5 prime UTR intron variant	SO:0002091	A UTR variant of intronic sequence of the 5' UTR.
Stop lost	SO:0001578	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript.
Inframe insertion	SO:0001821	An inframe non synonymous variant that inserts bases into in the coding sequence.
Exon loss variant	SO:0001572	A sequence variant whereby an exon is lost from the transcript.

References

1. Vasilevsky NA, Matentzoglu NA, Toro S, et al. Mondo: Unifying diseases for the world, by the world. medRxiv. 2022.
2. Arivarasan A, M Karthikeyan. Data Mining K-Means Document Clustering using TFIDF and Word Frequency Count. International Journal of Recent Technology and Engineering (IJRTE). 2019; 8: 2542-2548.
3. Jolliffe IT, Cadima J. Principal Component Analysis: A Review and Recent Developments. Philos Trans A Math Phys Eng Sci. 2016; 374: 20150202.
4. Ding C, He X. K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery. New York NY USA. 2004; 29.
5. Syakur MA, Khotimah BK, Rochman EMS, et al. Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. IOP Conf Ser Mater Sci Eng. 2018: 336 012017.
6. World Health Organization. QE84 Acute stress reaction. In International statistical classification of diseases and related health problems (11th ed.). 2019. https://icd.who.int/browse11/l-m/en#http%3a%2f%2fid.who.int%2ficd%2fentity%2f505909942
7. Eilbeck K, Lewis SE, Mungall CJ, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6: R44.