

Large Language Model Agents for Evidence Based Genetic Disease Severity Classification

Tohid Ghasemnejad¹, Ahmadrza Argha², Mark Grosser³, John Wang⁴, Min Yang⁵, Thantrira Porntaveetus⁶, Tony Roscioli^{7, 8}, Nigel H. Lovell², Mahmoud Aarabi^{*9,10}, Hamid Alinejad-Rokny^{*1, 11}

¹ UNSW BioMedical Machine Learning Lab (BML), School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia

² School of Biomedical Engineering, UNSW Sydney, NSW 2052, Australia

³ 23Strands, Pyrmont, Australia

⁴ Division of Genetic and Genomic Medicine, Department of Pediatrics, University of Pittsburgh, School of Medicine, Pittsburgh, PA, USA

⁵ Shenzhen Institute of Advanced Technology, Shenzhen, China

⁶ Center of Excellence in Precision Medicine and Digital Health, Department of Physiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand

⁷ New South Wales Health Pathology Genomics, Prince of Wales Hospital, Randwick, NSW, Australia

⁸ Neuroscience Research Australia (NeuRA), University of New South Wales Sydney, NSW, Australia

⁹ Medical Genetics and Genomics Laboratories, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

¹⁰ Departments of Pathology, and Obstetrics, Gynecology, and Reproductive Sciences, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

¹¹ Visiting Scholar (Collaborative Projects), Center of Excellence in Precision Medicine and Digital Health, Chulalongkorn University, Bangkok, Thailand

Correspondence to: H.A.R (h.alinejad@unsw.edu.au) and M.A. (mahmoud.aarabi@pitt.edu)

Abstract

Disease severity classification for genetic conditions is subjective and labor-intensive, creating bottlenecks in genomic screening, where commercial panels vary widely in size and overlap. We developed an autonomous AI agent integrating Reasoning and Acting (ReAct) with Retrieval-Augmented Generation (RAG) to classify 10,211 Human Phenotype Ontology terms. It uses American College of Medical Genetics (ACMG)-endorsed severity guidelines and American College of Obstetricians and Gynecologists (ACOG) quality-of-life criteria to retrieve PubMed literature, generate interpretable reasoning chains, and independently verify claims. At the phenotype level, using expert-curated cohorts, the agent achieved 93.55% accuracy (MCC 0.9237) with 82.6% to 91.4% of claims supported by direct evidence or valid inferences. Gene-level severity was aggregated across 8,738 pairs, identifying 3,283 autosomal recessive pairs with severe or profound presentations. External validation showed 95.2% concordance with Mackenzie's Mission gene list. This system enables standardized panel design by providing reliable, automated classification supported by direct evidence.

Brief non-technical summary

Selecting genes for reproductive screening panels is slow, subjective, and inconsistent across providers. We built an AI system that reads medical literature and automatically classifies the severity of over 10,000 genetic disease features, achieving 94% agreement with expert judgement. This gives clinicians a transparent, reproducible tool to standardise which conditions belong on screening panels.

Introduction

The rapid advancement of artificial intelligence (AI) and computational tools has significantly transformed the landscape of clinical genetics, particularly in the automated classification of sequence variants. Following the establishment of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) variant interpretation guidelines¹, automated tools have been developed to implement these standards, ranging from rule-based pipelines to AI-powered platforms and phenotype-driven prioritization systems²⁻⁵. Recent benchmarking efforts have demonstrated that these tools can achieve considerable accuracy in identifying disease-causing variants, especially when integrating phenotype data alongside genomic features⁶⁻⁸. Despite the impressive progress in use of automated AI-based solutions to classify genetic variants and to interpret genetic testing results, several genetics/genomics processes related to patient care remain to be managed manually. A prominent example is selection of the genes for reproductive screening panels that are commonly used, prenatally or preconception, to determine the risk of couples for having a child affected by recessive genetic conditions. Professional societies such as the American College of Medical Genetics and Genomics (ACMG) have provided guidance on two important requirements for genes to be included in carrier screening¹⁰. The first requirement is for genes to have high frequency of carriers ($>1/200$) for disease causing variants (pathogenic/likely pathogenic) in databases of normal populations such as Genome Aggregation Database (gnomAD). To address this requirement, various lists of genes with high carrier frequency have been generated from different gnomAD versions using well established bioinformatic pipelines¹¹⁻¹⁴. Another major requirement is association of the recessive condition with severe clinical presentations in the affected individuals that justifies screening for heterozygote carriers that are at risk of having an affected child. Both the ACMG and the American College of Obstetricians and Gynecologists (ACOG) state that disease severity and a detrimental effect on quality-of-life (QoL) are critical parts of inclusion criteria for selection of genes in screening panels^{10,15}. However, this requirement complicates the panel design process because determining disease severity is subjective and relies on varying expertise. Consequently, commercially available panels differ significantly in size and gene content, highlighting the need for a more objective, scalable evaluation methodology across all screening panels, from carrier to newborn screening^{16,17}.

Among the efforts to develop a systematic framework for disease severity classification is a study by Lazarin et al.¹⁸. Through surveys of 192 healthcare professionals, the investigators identified key disease characteristics and grouped them into four severity tiers based on perceived impact, ranging from a profound Tier 1 (e.g., shortened lifespan in infancy) to a mild Tier 4 (e.g., reduced fertility). This framework, also utilized by ACMG, enabled the classification of genetic conditions based on their characteristic profiles^{10,18}. In parallel, to operationalize ACOG's "detrimental effect on quality of life" criterion, models such as the Pediatric Quality of Life Inventory™ (PedsQL™) have been proposed^{15,19}. As a validated, multidimensional instrument for measuring health-related quality-of-life (HRQoL), PedsQL assesses the functional impact of health condition across core domains of Physical, Emotional, Social, and School/Cognitive Functioning. These frameworks translate abstract clinical descriptions into their tangible impact on an individual's daily life^{20,21}. However, the study acknowledged limitations, including the labor-intensive nature of manual classification and potential biases with certain diseases remains as an unsolved issue¹⁸. These challenges become particularly pronounced when considering the scale of modern genetic

knowledge. The Human Phenotype Ontology (HPO) provides a standardized vocabulary describing thousands of phenotypic abnormalities associated with human diseases²². Manually classifying the severity of this vast corpus of data is not only time-consuming but also susceptible to inter-rater variability and subjective interpretation¹⁹.

To overcome the limitations of manual curation while preserving the established foundational frameworks, we developed an autonomous AI agent that operationalizes both the ACMG-endorsed severity scale and the ACOG-compliant PedsQLTM functional impact model. Employing the ReAct (Reasoning and Acting) framework for structured logical analysis and a Retrieval-Augmented Generation (RAG) system to ground its conclusions in authoritative medical literature, our agent systematically classifies HPO terms at a scale unattainable through manual curation alone. By automating the application of these established clinical frameworks, this study aimed to establish a more objective, transparent, and scalable process to assist clinical and laboratory experts in making consistent, evidence-based decisions for a medical question of high utility, i.e., severity of genetic conditions.

Results

Model Selection and System Performance

Across 20 ground-truth HPO terms, GPT-4o at temperature 0.2 achieved the highest accuracy (92%) with a processing time of 7.6 s per term, balancing determinism with flexibility for complex medical reasoning, and was selected as the core model (Figure S1).

Validation on an independent set of 941 samples that matched the population distribution (PSI = 0.0288; Figure 2A) yielded overall accuracy of 93.55% (95% CI: 91.9–95.2%), weighted precision of 94.27%, weighted recall of 93.55%, and weighted F1 of 93.55% (Figure 2B). Multi-class performance remained robust under class imbalance, with Matthews Correlation Coefficient (MCC) of 0.9237, Cohen's kappa of 0.9223, balanced accuracy of 96.85%, and macro-averaged F1 of 0.9446. The confusion matrix concentrated along the diagonal, with the main off-diagonal confusion between phenotypically related categories, particularly "Dysmorphic Features" and "Internal Physical Malformations" (Figure 2C). Per-category precision ranged from 74% to 100% and recall was consistently above 80% (Figure 2D). For ordinal tier prediction, quadratic-weighted kappa was 0.9164, Spearman correlation was 0.9164, Kendall's tau was 0.9003, mean absolute error was 0.0992 tiers, and within ± 1 -tier accuracy was 98.21% (Figure 2E). Per-tier category accuracy ranged from 87.45% (Tier 3) to 100.00% (Tier 1) (Figure 2F).

Clinical Landscape Analysis

The classification of 10,211 phenotypes in HPO terms revealed a structured distribution across severity tiers and functional categories. Among the classified tiers, Tier 2 was the most prevalent (28.8%, $n = 2,939$), followed by Tier 3 (22.6%, $n = 2,306$) and Tier 4 (13.7%, $n = 1,400$). Tier 1 phenotypes, representing the profound conditions, accounted for 2.6% ($n = 269$) of the dataset. A large number of HPO terms were classified under non-tiered designations (32.3%, $n = 3,297$), including non-phenotypic Clinical Sign or Laboratory Abnormality ($n = 2,203$) such as HP:0002315 Headache, HP:0004396 Poor appetite, and HP:0031886 Abnormal LDL cholesterol concentration; Out of Scope: Acquired/Multifactorial conditions ($n = 767$) such as HP:0033235 Difficulty descending stairs, HP:0031446 Erosion of oral mucosa, and

HP:0032448 Achlorhydria; and Not Further Classified (NFC) entries (n = 321) such as HP:0033454 Tube feeding, HP:0000118 Phenotypic abnormality, and HP:0001470 Sex-limited expression. Internal physical malformations (n = 2,196) and dysmorphic features (n = 2,041) represented the most prevalent phenotypic categories. Tier distribution varied substantially across phenotypic categories (Figure 3A). For instance, dysmorphic features showed a bimodal distribution concentrated in Tier 3 (42.4%) and Tier 4 (57.6%), while life-threatening conditions demonstrated clear tier stratification, with shortened lifespan in infancy (n = 129), childhood/adolescence (n = 50), and adulthood (n = 61) exclusively mapping to Tier 1, Tier 1, and Tier 2, respectively.

Early-onset prevalence followed the expected age-of-manifestation pattern, highest in dysmorphic features (83.1%), internal physical malformations (81.1%), and shortened lifespan in infancy (80.6%), and lowest in mental illness (17.8%) and shortened lifespan in adulthood (24.6%) (chi-square test: $\chi^2(12) = 1090.32$, $p < 0.001$; Figure 3B). Management profiles revealed distinct therapeutic approaches across categories and tiers (Figure 3C, D): intellectual disability and mental illness required therapeutic interventions, sensory impairments and impaired mobility relied heavily on assistive devices, and Tier 1 conditions showed the most diverse approach, including therapeutic support, monitoring, palliative care, and pharmacological strategies.

Evidence Quality and Validation

Evidence profiles showed strong grounding in medical literature, with direct evidence plus valid medical inferences covering 82.6% to 91.4% of claims across tiers (Figure 4A). Tier 1 had the highest direct-evidence rate (73.9%) and the lowest unsupported rate (5.7%); Tier 4 was lower on direct evidence (71.8%) and showed the highest unsupported rate (14.5%). Tiers 2 and 3 fell between these bounds, with direct-evidence rates of 73.4% and 71.6% and unsupported rates of 6.7% and 7.2%, respectively. At category level, direct evidence ranged from approximately 65% to 80%, with categories representing severe disease burden (intellectual disability, impaired mobility, internal physical malformations) exceeding 70% (Figure 4B); unsupported claims stayed below 10% for most categories.

Statement support scores concentrated near the maximum across all tiers and categories (Figure 4C, D). Tier 1 and Tier 2 showed the tightest distributions (median ~0.95), while Tier 4 was broader (median ~0.85). Categories such as internal physical malformations, clinical signs/laboratory abnormalities, and dysmorphic features had tight distributions near 1.0, whereas shortened-lifespan categories were wider. Differences in support score across tiers were significant (Kruskal–Wallis $H(3) = 48.92$, $p < 0.001$), as were differences across categories (Kruskal–Wallis $H(12) = 32.27$, $p = 0.001$).

Quality of Life Impact Assessment

PedsQL-based analysis showed measurable functional impact on at least one QoL domain in 63.4% of phenotypes (n = 6,474). Because phenotypes can affect multiple domains, all supported mappings were retained. Domain distributions varied substantially across tiers (chi-square test: $\chi^2(12) = 1001.16$, $p < 0.001$; Figure 5A; Table 1). Tier 1 showed a diverse profile (Physical Functioning 47.8%, Cognitive/School 30.9%, Physical Symptoms 13.1%, Emotional Impact 4.2%, Social Functioning 4.0%); Tier 2 was dominated by Physical Functioning (72.3%) with Physical Symptoms second (14.7%); Tier 3 was led by Physical Functioning (50.2%) but distributed substantially across Cognitive/School (19.5%), Social Functioning

(11.4%), Emotional Impact (10.1%), and Physical Symptoms (8.9%). Tier 4 displayed a markedly distinct profile dominated by Emotional Impact (38.3%) and Physical Functioning (30.9%), with substantial Social Functioning (23.5%), reflecting the predominance of infertility and adult-onset conditions in this tier.

Category-level profiles were highly distinct (Figure 5B; Table S2), with a significant association between category and QoL status (chi-square test: $\chi^2(12) = 2,836.50$, $p < 0.001$). Shortened lifespan in infancy and childhood emphasised Physical Functioning and Physical Symptoms. Mental illness uniquely combined Cognitive/School and Emotional Impact. Sensory impairments split between Cognitive/School and Physical Functioning; intellectual disability was dominated by Cognitive/School; internal physical malformations and dysmorphic features were dominated by Physical Functioning.

Management Complexity and Early Onset Patterns

Management complexity correlated strongly with the breadth of QoL involvement (Figure 6A). The average number of interventions rose monotonically from 0.93 in phenotypes with no QoL impact to 2.67 in those affecting all five domains ($R^2 = 0.887$, $p = 0.005$). Physical Functioning was the most frequently affected domain ($n = 4,438$), followed by Cognitive/School ($n = 1,142$), Physical Symptoms ($n = 869$), Social Functioning ($n = 504$), and Emotional Impact ($n = 474$) (Figure 6B). QoL-affected phenotypes required significantly more interventions than unaffected ones (Mann–Whitney U test, $p < 0.001$).

Tier-stratified complexity was highest in Tier 1 (mean 1.72 ± 0.91 SD), followed by Tier 3 (1.56 ± 0.80) and Tier 2 (1.48 ± 0.89), with Tier 4 lowest (0.93 ± 0.61); differences across tiers were significant (Kruskal–Wallis $H(3) = 638.28$, $p < 0.001$) (Figure 6C). Life-threatening categories demanded the most intensive management, with shortened-lifespan conditions averaging 1.89–2.00 interventions (Figure 6D). Immunodeficiency/cancer (1.72) and sensory impairments (1.45–1.71) were also elevated, whereas dysmorphic features required the fewest (1.10). Category-level differences were significant (Kruskal–Wallis $H(12) = 550.65$, $p < 0.001$).

Quality of Life Status and Management Profiles

QoL impact varied substantially across categories and tiers (Figure S2A, B; Table S2). Categories lying entirely within Tier 1 (intellectual disability, shortened-lifespan in infancy and childhood/adolescence) or Tier 2 (internal physical malformations, impaired mobility, shortened lifespan in adulthood) showed 100% QoL impact. Mixed-tier categories varied: dysmorphic features (Tiers 3–4) 42.4% affected (865 / 2,041), immunodeficiency/cancer (Tiers 3–4) 85.3% (512 / 600), mental illness (Tiers 3–4) 94.3% (280 / 297), sensory impairment vision 95.1% (521 / 548), sensory impairment hearing 89.7% (61 / 68), and infertility (Tier 4 only) 50.6% (44 / 87). The association between category and QoL status was significant (chi-square test: $\chi^2(12) = 2,836.50$, $p < 0.001$).

Management profiles differed markedly between QoL-affected and unaffected phenotypes (Figure S2C). Affected phenotypes required diverse interventions (surgical, pharmacological, monitoring, therapeutic support, and dietary), whereas unaffected phenotypes were dominated by no-treatment and monitoring strategies. The contrast in management distribution between groups was highly significant (chi-square test: $\chi^2(7) = 2,302.88$, $p < 0.001$), supporting the clinical utility of QoL assessment for predicting therapeutic intensity.

Operational Efficiency

Statement support scores across all 10,211 classifications averaged 0.81 (SD = 0.27), with a median of 0.95; 47% of phenotypes achieved the maximum score of 1.0 and only 11% scored below 0.5, reflecting strong grounding in medical literature (Figure S3A). The ReAct agent required a mean of 6.73 steps (SD = 1.54) per classification, with the majority processed in 6–8 iterations. Processing cost averaged \$0.102 USD per phenotype (range \$0.062–\$0.197; SD = \$0.016), with the middle 50% of classifications costing \$0.091–\$0.113 (Figure S3B). Correlation between agent complexity and cost was modest (Pearson $r = 0.33$, $p < 0.001$), indicating predictable resource use and scalability for clinical deployment. The independent verification system flagged 1,161 of 10,211 classifications (11.4%) for human review.

Gene-Level Severity Classification and Evaluation

Aggregation across 8,738 gene–disease pairs (after exclusion of 17 NFC-only pairs and application of an HPO-Annotation frequency filter retaining only Obligate, Very Frequent, and Frequent annotations; see Methods) revealed a predominance of clinically significant conditions (Figure 7A; Table 2): 38.1% Profound ($n = 3,326$), 41.4% Severe ($n = 3,619$), 16.4% Moderate ($n = 1,435$), and 4.1% Mild ($n = 358$). The combined Profound and Severe categories (79.5%) support inclusion in expanded carrier screening panels under ACMG/ACOG guidelines.

Severity varied across 23 body systems (Figure 7B; Table 2). The highest profound burden was seen in abnormalities of prenatal development or birth (63.9% Profound), respiratory system (63.1%), and nervous system (58.9%); the nervous system was also the largest category by association count ($n = 5,312$), highlighting the burden of severe neurological genetic disease. Body systems with comparatively lower profound burden but high overall severity included blood and blood-forming tissues (26.9% Profound, 58.2% Severe) and constitutional symptoms (26.9% Profound, 68.9% Severe).

Mode of inheritance also shaped severity (Figure 7C). X-linked conditions had the highest profound proportion (57%, $n = 177$ of 311), followed by autosomal recessive (55%, $n = 2,117$ of 3,873) and autosomal dominant (22%, $n = 758$ of 3,495). Among 3,873 autosomal recessive gene–disease pairs, 3,283 (84.8%) were classified as Severe or Profound, and 256 of 311 X-linked pairs (82.3%) fell into the same combined category.

Comparison with established panels validated the framework (Figure 7D). ACMG Secondary Findings genes, associated with severe, medically actionable conditions, showed the highest combined Profound and Severe burden (99.4%; 47.0% Profound and 52.4% Severe). Mackenzie's Mission, comprising severe conditions for carrier screening, showed 75.5% Profound and 19.7% Severe (combined 95.2%). The ACMG-recommended carrier screening panel showed 59.3% Profound and 40.0% Severe (combined 99.3%), with minimal moderate (0.7%) and no mild representation.

For the curated set of 29 genes from published studies, 20 (69.0%) had exact severity matches with the source, 8 (27.6%) received appropriately higher classifications based on broader phenotypic analysis, and only one gene, BCKDHB (3.4%), was underestimated (classified as moderate rather than profound), likely because the ontological databases lacked documentation of the intellectual disability phenotypes described in the wider clinical literature.

Discussion

This study presents the successful development and validation of an autonomous LLM agent that addresses the challenges of subjectivity, scalability, and consistency in analysis of large clinical and published data. Focusing on the challenge of clinical severity for genetic conditions, the agent achieved accurate classification of 10,211 HPO phenotype terms based on severity tiers and quality-of-life metrics, age of onset, and treatment availability. The successful deployment of this model exemplifies a new paradigm for evidence-based medicine that combines rigorous clinical frameworks with computational efficiency.

Various strategies have been recommended for selection of genes for screening panels. Previous attempts at panel standardisation have highlighted significant variability, with an analysis of 16 commercial panels finding that the number of conditions screened ranged from 41 to 1792; only three conditions were screened by all providers and only about 10% of genes appeared on at least half of panels examined, demonstrating the lack of consensus in the field^{16,23}. In the past years, the ACMG recommendation has become the most universally adapted approach, emphasising gene selection based on carrier frequency of under 1/200 and association with moderate, severe, or profound clinical presentations^{10,24}. Carrier frequency has been shown to be a dynamic metric that is subject to change based on the degree of consanguinity, founder effects, bottlenecks, and the availability of genome data from new larger populations^{11,12,25}. Disease severity is less prone to change than carrier frequency as it originates from well-established clinical presentations in the literature. Moreover, the model enables laboratories to design both comprehensive panels for population screening and targeted panels optimised for specific clinical contexts²³. The clinical utility of carrier screening using a carefully-selected gene panel is demonstrated by the recent Mackenzie's Mission study, which identified that 1.9% of 9107 couples had a newly identified increased risk for a severe genetic condition and 76.6% of them acted on the results^{26,27}, validating the trend toward broader screening panels. Our framework also has direct relevance to the rapidly expanding field of genomic newborn screening, where systematic severity assessment is one of the core criteria for gene selection alongside treatability, penetrance, and age of onset. The scalable and reproducible nature of our classification pipeline could support international harmonisation efforts for newborn screening gene lists, which currently show substantial variability across programmes¹⁷.

The scalability challenges inherent in genetic disease classification have historically imposed severe constraints on comprehensive carrier screening implementation. Lazarin and colleagues pioneered systematic classification through surveys of 192 healthcare professionals but explicitly acknowledged the prohibitive labour requirements and potential biases inherent in manual approaches¹⁸. These limitations were starkly illustrated when Arjunan and colleagues employed twelve genetics professionals to manually classify 176 genes, achieving only 60.8% initial inter-rater concordance¹⁹. In Australia's Mackenzie's Mission project, the gene selection process reported an overall discordance rate of 25% between clinical geneticists²⁶. Our agent processes phenotypes at very low cost (\$0.10 per classification) with high accuracy (93.55%), transforming what was previously an insurmountable manual task into a feasible, reproducible process. The integration of severity tier classifications with quality-of-life assessments, management profiles, and age of onset data are among the unique features of this model that create a multidimensional resource that enhances clinical decision-making in genetic counselling settings.

The RAG–ReAct architecture proposed here represents a major step forward in creating transparent medical AI. Many conventional AI systems operate as black boxes, which has raised concerns about their interpretability in clinical settings. In contrast, the system generates an explicit reasoning trace for each classification, averaging 6.73 steps, creating an auditable record that clinicians can review and validate. The hierarchical evidence retrieval strategy successfully grounds the majority of phenotypes in high-quality sources, ensuring that classifications derive from authoritative medical literature rather than statistical patterns alone. Our MCC of 0.9237 and Cohen's kappa of 0.9223 demonstrate that integration of structured clinical frameworks with evidence-based reasoning can achieve superior performance in medical classification tasks. Testing the classification of severity against various established lists of severe conditions, along with our expert-curated lists, demonstrated the robustness of this model as a tool used by both clinicians and laboratorians.

An important limitation of this study is the reliance on structured HPO terms that may not capture all clinical nuances present in narrative descriptions. Despite containing over 19,672 terms as of 2025²², some phenotypic features may still lack appropriate HPO representation. The system's inability to dynamically incorporate emerging literature represents a structural limitation that future iterations must address through continuous learning frameworks. Furthermore, approximately one-third of HPO terms could not be definitively assigned to existing severity tiers, which may reflect the non-genetic nature of these clinical presentations but could also reveal limitations in current categorical frameworks. Tier 4 evidence rates were also lower (71.8% direct evidence; 14.5% unsupported) than for Tiers 1–3, reflecting both the heterogeneity of adult-onset and infertility phenotypes and the comparatively sparser indexed literature on their long-term functional impact.

In conclusion, this work establishes a new standard for the intersection of artificial intelligence and clinical genetics, demonstrating that autonomous systems can successfully mimic complex medical frameworks at population scale. By providing transparent, validated, and scalable tools that amplify rather than replace clinical expertise, this approach enables the systematic application of professional guidelines across diverse healthcare data resources. As genetic knowledge expands and population-wide screening programmes become standard care¹⁷, the integration of AI-assisted classification systems offers a pathway to ensure that advances in genetic medicine translate into improved outcomes for all populations. The modular design and evidence-grounded RAG–ReAct reasoning allow the system to process novel phenotypes even in the absence of established gene or disease associations, transforming a previously static catalogue of clinical terms into an actionable, evidence-based landscape.

Methods

Framework for Systematic Phenotype Severity Classification

The analytical foundation of this study is a synthesised framework designed to systematically evaluate clinical phenotypes (Figure 1). The integrated workflow combines a hierarchical severity assessment system used by the ACMG¹⁰⁻¹⁸ with a QoL evaluation based on PedsQL™ domains, age of onset, and treatment availability, all within a broader framework aligned with ACOG recommendations¹⁵. These classification criteria were provided to a ReAct agent that executed four sequential phases: contextualisation, query generation, evidence retrieval, and adaptive expansion through iterative reasoning

loops. Phenotype term classifications were then mapped to gene and disease associations and aggregated through a hierarchical algorithm to produce gene-level severity designations: Profound, Severe, Moderate, or Mild.

Adapting the severity framework developed by Lizarin et al.¹⁸, we assigned each of the 10,211 HPO terms with a phenotype–gene association to one of four severity tiers. Tier 1 (most severe) covered conditions that are life-limiting in infancy or childhood/adolescence or that cause intellectual disability. Tier 2 covered conditions that are life-limiting in adulthood, impaired mobility, and internal physical malformations. Tier 3 encompassed sensory impairments, immunodeficiency, cancer, mental illness, and dysmorphic features that have demonstrable functional consequences. Tier 4 covered adult-onset diseases, infertility, and cosmetic-only or non-functional variants of categories that overlap with Tier 3 (specifically dysmorphic features without functional impact, mild sensory variants, and mental illness traits without major functional consequences). This within-category subdivision, operationalised through the PedsQL-based QoL assessment described below, allowed phenotypes within categories such as dysmorphic features (Tier 3 = 42.4%; Tier 4 = 57.6%), immunodeficiency/cancer (Tier 3 = 85.7%; Tier 4 = 14.3%), mental illness (Tier 3 = 94.3%; Tier 4 = 5.7%), and sensory impairments (Tier 3 = 89.7 to 95.1%; Tier 4 = 4.9 to 10.3%) to be distinguished by their actual functional impact rather than category membership alone. Full category definitions are provided in Supplementary Table 1.

Three non-tiered designations handled terms outside this framework. "Clinical Sign or Laboratory Abnormality" captured objective findings such as transient symptoms and laboratory results rather than persistent functional disability. "Out of Scope: Acquired/Multifactorial" covered phenotypes that are primarily acquired, such as hypertension. "Not Further Classified" (NFC) was applied to terms that did not fit any Lizarin category. Terms in these categories did not receive a numerical tier.

Quality of Life Assessment, Management, and Age of Onset

The ACOG's criterion of a "detrimental effect on quality of life" lacks a standard evaluation method. Therefore, to systematically quantify this effect, we adopted a framework based on the Pediatric Quality of Life Inventory (PedsQL 4.0 Generic Core Scales)^{20,21}, an approach previously validated by Arjunan et al. for the same purpose¹⁹. The agent mapped evidence to five domains: Physical Functioning, Emotional Functioning, Social Functioning, School/Cognitive Functioning, and Physical Symptoms (the last added to capture gastrointestinal, respiratory, and dermatological manifestations)^{20,21}. Full domain definitions and items are listed in Supplementary Table 2.

The QoL framework served two roles. First, it determined whether each phenotype met the ACOG QoL criterion. Second, it refined severity within Tier 3 by distinguishing phenotypes with different functional consequences. For example, orofacial cleft (HP:0000202) and thick eyebrow (HP:0000574) are both dysmorphic features, but orofacial cleft affects Physical Functioning (e.g., feeding difficulties)²⁸ and Social Functioning (e.g., speech impairment)²⁹, and is classified as QoL-affected (Tier 3), whereas thick eyebrow is an isolated cosmetic trait with no functional impact and is classified as not affected (Tier 4).

Two additional ACOG-inspired criteria were incorporated. Age of onset was captured through evidence of congenital, neonatal, or early childhood presentation, since earlier onset correlates with greater disease burden. Treatment availability served as a proxy for clinical severity, with management strategies extracted

from the literature and grouped into nine management profile categories (No Treatment, Monitoring, Pharmacological, Surgical, Therapeutic Support, Dietary, Assistive Devices, Palliative Care, and Radiation Therapy).

Model Selection and Optimization

We benchmarked six LLMs (GPT-5, Grok-4, Gemini-2.5-pro, Claude-3.5-sonnet, GPT-4o, and GPT-4o-mini) at three temperatures (0.0, 0.2, 0.5) against 20 ground-truth HPO terms spanning all tiers and categories. Models were scored on classification accuracy, response consistency, and processing time.

Prompt Engineering

A structured prompt was designed to guide an LLM through a defined clinical reasoning process, aiming to classify an HPO term into predefined severity categories based on Lazarin et al.¹⁸ and to assess its impact on QoL domains^{20,21}. All inputs (HPO context and curated literature) were supplied in a standardised JSON structure with a unique identifier per source, which the LLM cited for every factual claim. The model first determined whether the phenotype was congenital or acquired; congenital phenotypes were then traversed through a hierarchical decision tree and assigned to the first applicable category, followed by a PedsQL-based QoL assessment. Mandatory citation and a prohibition on internal knowledge kept reasoning grounded in the supplied evidence. The output was a single JSON object containing the reasoning trace, classification, management profile, and cited references, enabling automated parsing and validation.

Agent Workflow

Each HPO term was processed through five sequential phases. In Phase 1 (Term Contextualisation), the system retrieved the full definition, synonyms, and hierarchical relationships from the HPO database (January 2025 release) and mapped the term to its corresponding Medical Subject Heading (MeSH) via NCBI to improve search precision. In Phase 2 (Query Generation), the LLM expanded the vocabulary with common clinical diagnoses and synonyms, which were combined with HPO synonyms and the MeSH term. Five predefined search templates (congenital and genetic associations, hereditary and familial patterns, paediatric and neonatal onset, childhood prognosis, and management of inherited disorders) were then applied to produce up to 15 deduplicated queries per term. In Phase 3 (Evidence Retrieval), PubMed was searched hierarchically, first combining the MeSH term with genetic-disease keywords and progressively broadening to include all synonyms, then removing context filters, retrieving up to 20 abstracts per search. In Phase 4 (Adaptive Expansion), triggered when fewer than five relevant articles were found, the LLM identified commonly associated genetic syndromes and repeated the search, ensuring adequate evidence for rare conditions. In Phase 5 (Classification), the system synthesised retrieved evidence, applied the hierarchical rules, assigned a category based on the primary manifestation, and determined QoL status.

Source Verification System

A separate verification agent assessed the factual accuracy of each classification using a modified SourceCheckup framework³⁰. In the extraction stage, a parser model identified all verifiable medical claims in the reasoning trace, preserving each reference tag. In the verification stage, an independent LLM instance acting as a fact-checker matched each claim to its cited source and assigned a weighted support score: direct statement from the source (1.0), valid medical inference (0.85), weak inferential support (0.60), or absence

of support (0.0). Scores below 0.5 triggered review flags while preserving the original classification for consistency. Across the full 10,211-phenotype dataset, 11.4% of classifications ($n = 1,161$) were flagged for human review, and the remaining 88.6% ($n = 9,050$) passed verification with no flags.

Phenotype level Performance Evaluation

We created a manually curated ground truth dataset through a two-part evaluation strategy to ensure representativeness and validity. The first part focused on practical representativeness by comparing the category distribution in the curated ground truth against the full predicted population. To quantify distribution, the Population Stability Index was calculated as:

$$\text{PSI} = \sum_{i=1}^k (p_{\text{actual},i} - p_{\text{expected},i}) \times \ln \left(\frac{p_{\text{actual},i}}{p_{\text{expected},i}} \right) \quad (\text{Equation 1})$$

where $p_{\text{actual},i}$ represents the proportion of category i in the ground truth sample, and $p_{\text{expected},i}$ represents the proportion in the full population. The ideal sample size for each category was then calculated using:

$$n_{\text{ideal},i} = \text{round}(P_{\text{population},i} \times N_{\text{target}}) \quad (\text{Equation 2})$$

Over- and under-represented categories are identified using the representativeness gap:

$$\Delta_i = n_{\text{ideal},i} - n_{\text{current},i} \quad (\text{Equation 3})$$

Expert annotations were established through independent review by two clinical and laboratory geneticists per term, with discrepancies resolved by structured consensus. Distributional differences were tested by chi-square.

Performance was evaluated using multiple metrics. Matthews Correlation Coefficient (MCC), robust for imbalanced multi-class problems, served as the primary metric:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (\text{Equation 4})$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Cohen's kappa assessed chance-corrected agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (\text{Equation 5})$$

where p_o is the observed agreement and p_e is the expected agreement by chance. For ordinal tier evaluation, quadratic weighting was applied with $w_{ij} = \frac{(i-j)^2}{(k-1)^2}$. Balanced accuracy ensured equal weight across all phenotypic categories:

$$\text{Balanced Accuracy} = \frac{1}{n_{\text{classes}}} \sum_{i=1}^{n_{\text{classes}}} \frac{TP_i}{TP_i + FN_i} \quad (\text{Equation 6})$$

Both macro-averaged and weighted F1-scores (Equations 7a and 7b) were calculated to evaluate performance across rare and common phenotypes:

$$F1_{\text{macro}} = \frac{1}{n_{\text{classes}}} \sum_{i=1}^{n_{\text{classes}}} F1_i \quad (\text{Equation 7a})$$

$$F1_{\text{weighted}} = \sum_{i=1}^{n_{\text{classes}}} \frac{n_i}{N} F1_i \quad (\text{Equation 7b})$$

where n_i is the number of samples in class i and N is the total sample size. Bootstrap resampling (1,000 iterations) was employed to generate 95% confidence interval for all metrics, providing robust estimates of performance stability. The reasoning for each classification was manually reviewed by clinicians to ensure that the assigned labels were not only statistically consistent but also clinically meaningful.

Gene-Level Severity Classification

Following completion of phenotype-based classifications, we aggregated HPO terms for each gene through their disease associations, creating comprehensive gene-to-phenotype mappings that preserve tier classifications. For each gene-disease pair (g, d) , tier counts were calculated as:

$$n_{\text{Tier}i} = |\{p \in \text{Phenotypes}(g, d) \mid \text{tier}(p) = i\}| \quad (\text{Equation 8})$$

where p represents an individual phenotype and $i \in \{1, 2, 3\}$. Gene-disease pairs consisting exclusively of phenotypes labeled as *NFC* were excluded from subsequent severity assessment. Gene-level severity was determined using a hierarchical classification algorithm, whereby genes were evaluated sequentially and assigned to the first severity category whose criteria were met.

- Profound: $n_{\text{Tier}1} > 1$, indicating multiple severe, life-limiting, or intellectually disabling phenotypes.
- Severe: $n_{\text{Tier}1} = 1$ or $(n_{\text{Tier}2} \geq 1 \wedge n_{\text{Tier}2} + n_{\text{Tier}3} \geq 4)$, representing either a single Tier 1 phenotype or a higher cumulative number of Tier 2 and Tier 3 phenotypes.
- Moderate: $(n_{\text{Tier}2} \geq 1 \wedge n_{\text{Tier}2} + n_{\text{Tier}3} < 4)$ or $n_{\text{Tier}3} \geq 1$, representing intermediate severity driven by fewer high-tier phenotypes or exclusively Tier 3 phenotypes alone.
- Mild: reserved for genes with no associated Tier 1, Tier 2, or Tier 3 phenotypes.

An illustrative example of this gene-level aggregation process is shown in Figure S4.

Gene level Severity Evaluation

To validate the gene-level classification model, we initially tested it on a small dataset of curated genes. Gene lists with known severity from two published studies by Goldberg et al.³¹ and Taber et al.³², which contained 65 and 75 genes, respectively. Of the 43 genes common to both, 32 had consistent severity ratings and 11 were discordant. Retaining the 32 aligned genes and excluding three without disease-type information left 29 genes with 700 distinct HPO terms. Validation was expanded with three established expert-curated panels: ACMG carrier screening (113 genes)³¹, Mackenzie's Mission (1,300 genes)²⁶, and ACMG Secondary Findings v3.0 (73 genes)³³. Although these panels do not assign explicit severity, they represent consensus lists of clinically actionable genes, enabling implicit validation of the automated classifications against expert judgement of clinical actionability.

Statistics and reproducibility

No statistical method was used to predetermine sample size. The 10,211 HPO terms evaluated represent the complete set of HPO terms with a documented gene-phenotype association in the 2024 HPO release. The validation subset ($n = 941$) was sized using Population Stability Index calibration against the full population distribution, with category-level target sizes calculated as described in Equations 1 to 3. No data

were excluded from analyses other than gene-disease pairs composed exclusively of Not Further Classified terms, which were excluded from gene-level severity assessment because they contain no tiered phenotype on which the hierarchical aggregation rules operate. Expert annotators for the ground-truth set were blinded to the agent's predictions during independent review, and discrepancies between the two annotators were resolved by structured consensus.

All statistical tests were two-sided, and a significance threshold of $\alpha = 0.05$ was applied throughout. Comparisons of continuous or ordinal outcomes across more than two independent groups (for example, management complexity across severity tiers or across phenotypic categories) were performed using the Kruskal-Wallis H test, as the underlying distributions were non-normal by Shapiro-Wilk testing. Pairwise comparisons between two independent groups (for example, QoL-affected versus unaffected phenotypes) were performed using the Mann-Whitney U test. Associations between categorical variables were assessed using Pearson's chi-square test of independence, with expected-frequency assumptions verified prior to testing. The relationship between quality-of-life domain breadth and mean intervention count was modelled by ordinary least squares linear regression, with strength of association reported as the coefficient of determination and Pearson correlation coefficient. Classification performance was quantified using Matthews Correlation Coefficient as the primary metric, together with Cohen's kappa (quadratic-weighted for ordinal tier evaluation), balanced accuracy, and macro- and weighted-averaged F1 scores, as defined in Equations 4 to 7b. 95% confidence intervals for all performance metrics were estimated by non-parametric bootstrap resampling with 1,000 iterations. Exact P values are reported for all tests; values below 0.001 are reported as $P < 0.001$. Test statistics, degrees of freedom, and exact n values are reported in the relevant Results paragraphs and figure legends. Error bars in figures represent standard deviation unless stated otherwise in the legend. All analyses were performed in Python 3.11 using SciPy 1.11, statsmodels 0.14, scikit-learn 1.3, and pandas 2.1; figures were generated with Matplotlib 3.8 and seaborn 0.13.

Data Availability

The Human Phenotype Ontology (hp.obo, January 2024 release) is publicly available from the HPO consortium (<https://hpo.jax.org/>). PubMed abstracts and MeSH terms used for evidence retrieval were obtained via NCBI E-utilities. The reference panels used for external validation are available from the original publications: the ACMG carrier screening panel¹⁰, the ACMG Secondary Findings v3.0 list³³, and Mackenzie's Mission²⁶. The input HPO list, ground-truth annotations, hyperparameter results, gene-level severity table, and formatted comparison panels are deposited at Zenodo (<https://doi.org/10.5281/zenodo.19904041>) and mirrored at <https://github.com/T0hid/hpo-classification-agent>. The full per-term classification output can be regenerated end-to-end by running main.py and is also available to editors and peer reviewers during manuscript assessment, and from the corresponding author.

Code Availability

All source code is openly available on GitHub (<https://github.com/T0hid/hpo-classification-agent>), with a fully executable reproducibility capsule on Code Ocean (<https://doi.org/10.24433/CO.8502400.v1>) and a permanent archival snapshot on Zenodo (<https://doi.org/10.5281/zenodo.19904041>). The repository contains the ReAct + RAG classification pipeline, the source verification module, the gene level severity aggregation, the hyperparameter analysis, and the figure generation scripts. The prompt distributed publicly

is a structural template; the full production prompt is available to editors and peer reviewers during manuscript assessment, and to academic researchers under a Material Transfer Agreement with UNSW Sydney, by contacting the corresponding author.

Author Contributions

TG conceptualised the study, designed and developed the ReAct–RAG agent framework, implemented the classification pipeline, performed all computational analyses, and wrote the first draft of the manuscript. AA provided input on methodology and reviewed the manuscript. MG contributed to manuscript editing and revision. NHL contributed to the study design, provided oversight on the biomedical engineering aspects, and reviewed the manuscript. JW contributed to the clinical validation of gene-level severity classifications and reviewed the manuscript. MY contributed to methodology review. TP contributed to clinical review. TR provided clinical genetics expertise and reviewed the manuscript. MA contributed to the conceptualisation of the clinical framework, provided expert curation of gene and phenotype datasets for validation, supervised the clinical genetics aspects of the study, and critically revised the manuscript. HAR conceptualised and supervised the study, contributed to the study design and analytical framework, and revised the manuscript. TG and MA directly accessed and verified the underlying data reported in this manuscript. All authors had full access to all the data in the study and accept responsibility to submit for publication.

Acknowledgements

This study was supported by the UNSW Scientia Program Fellowship and the Australian Research Council Discovery Early Career Researcher Award (DECRA), under grant DE220101210 to HAR. The project was also supported by 23Strands Pty Ltd. This collaborative project was also supported by the Second Century Fund (C2F), Chulalongkorn University. Additionally, the project has support from the Ratchadaphiseksomphot Endowment Fund, Top-up Grant of C2F, Chulalongkorn University.

Competing Interests

UNSW Sydney holds intellectual property rights (trade secret and copyright) related to the ReAct–RAG autonomous agent framework and classification pipeline, the ACMG severity classification framework, the ACOG quality-of-life assessment module (PedsQL™ domain mapping and functional impact evaluation), the management and therapeutic strategies for each HPO term, the source verification system, and the gene-level severity aggregation algorithm described in this work. The ACMG-endorsed severity classification framework is licensed to 23Strands Pty Ltd. MG is CEO of 23Strands Pty Ltd. TG reports a top-up scholarship from 23Strands. All other authors declare no competing interests.

Materials & Correspondence

Correspondence and requests for materials should be addressed to Hamid Alinejad-Rokny (h.alinejad@unsw.edu.au) and Mahmoud Aarabi (mahmoud.aarabi@pitt.edu).

References:

1. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
2. Nicora, G., Zucca, S., Limongelli, I. & Bellazzi, R. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Sci. Rep.* **12**, 2517 (2022).
3. Stawiński, P. & Płoski, R. Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. *Clin. Genet.* **106**, 119–126 (2024).
4. Nicora, G. *et al.* CardioVAI: An automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases. *Hum. Mutat.* **39**, 1835–1846 (2018).
5. Kopanos, C. *et al.* VarSome: the human genomic variant search engine. *Bioinformatics* **35**, 1978 (2019).
6. Ghasemnejad, T. *et al.* Comprehensive Evaluation of ACMG/AMP-based Variant Classification Tools. *Bioinformatics* btaf623 (2026) doi:10.1093/bioinformatics/btaf623.
7. Kelly, C. *et al.* Phenotype-aware prioritisation of rare Mendelian disease variants. *Trends Genet.* <https://doi.org/10.1016/j.tig.2022.07.002> (2022) doi:10.1016/j.tig.2022.07.002.
8. Wang, J. *et al.* Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief. Bioinform.* **23**, bbac019 (2022).
9. Kraft, S. A., Duenas, D., Wilfond, B. S. & Goddard, K. A. B. The evolving landscape of expanded carrier screening: challenges and opportunities. *Genet. Med.* **21**, 790–797 (2019).
10. Gregg, A. R. *et al.* Screening for autosomal recessive and X-linked conditions during pregnancy and preconception: a practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1793–1806 (2021).
11. Schmitz, M. J. *et al.* Carrier frequency of autosomal recessive genetic conditions in diverse populations: Lessons learned from the genome aggregation database. *Clin. Genet.* **102**, 87–97 (2022).
12. Schmitz, M. J. *et al.* Leveraging diverse genomic data to guide equitable carrier screening: Insights from gnomAD v4.1.0. *Am. J. Hum. Genet.* **112**, 181–195 (2025).
13. Guo, M. H. & Gregg, A. R. Estimating yields of prenatal carrier screening and implications for design of expanded carrier screening panels. *Genet. Med.* **21**, 1940–1947 (2019).
14. Hotakainen, R., Järvinen, T., Kettunen, K., Anttonen, A.-K. & Jakkula, E. Estimation of carrier frequencies of autosomal and X-linked recessive genetic conditions based on gnomAD v4.0 data in different ancestries. *Genet. Med.* **27**, 101304 (2025).
15. Committee Opinion No. 690: Carrier Screening in the Age of Genomic Medicine. *Obstet. Gynecol.* **129**, e35–e40 (2017).
16. Chokoshvili, D., Vears, D. & Borry, P. Expanded carrier screening for monogenic disorders: where are we now? *Prenat. Diagn.* **38**, 59–66 (2018).
17. Saier, C. *et al.* TREAT: systematic and inclusive selection process of genes for genomic newborn screening as part of the Screen4Care project. *Orphanet J. Rare Dis.* **20**, 231 (2025).
18. Lazarin, G. A. *et al.* Systematic classification of disease severity for evaluation of expanded carrier screening panels. *PLoS One.* **9**, e114391 (2014).
19. Arjunan, A. *et al.* Evaluation and classification of severity for 176 genes on an expanded carrier screening panel. *Prenat. Diagn.* **40**, 1246–1257 (2020).
20. Varni, J. W., Seid, M. & Kurtin, P. S. PedsQL™ 4.0: Reliability and Validity of the Pediatric Quality of Life Inventory™ Version 4.0 Generic Core Scales in Healthy and Patient Populations: *Med. Care* **39**, 800–812 (2001).
21. Varni, J. W. *et al.* The PedsQL™ Infant Scales: feasibility, internal consistency reliability, and validity in healthy and ill infants. *Qual. Life Res.* **20**, 45–55 (2011).
22. Gargano, M. A. *et al.* The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
23. Wang, T., Scuffham, P., Byrnes, J., Delatycki, M. B. & Downes, M. An overview of reproductive carrier screening panels for autosomal recessive and/or X-linked conditions: How much do we know? *Prenat. Diagn.* **43**, 1416–1424 (2023).

24. Guha, S. *et al.* Laboratory testing for preconception/prenatal carrier screening: A technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **26**, 101137 (2024).
25. Gruzin, M. J. *et al.* Optimizing gene panels for equitable reproductive carrier screening: The Goldilocks approach. *Genet. Med.* **27**, 101387 (2025).
26. Kirk, E. P. *et al.* Gene selection for the Australian reproductive genetic carrier screening project (“Mackenzie’s Mission”). *Genet. Med.* **29**, 79–87 (2021).
27. Kirk, E. P. *et al.* Nationwide, Couple-Based Genetic Carrier Screening. *N. Engl. J. Med.* **391**, 1877–1889 (2024).
28. De Vries, I. A. C. *et al.* Prevalence of feeding disorders in children with cleft palate only: a retrospective study. *Clin. Oral Investig.* **18**, 1507–15 (2014).
29. Hocevar-Boltezar, I., Jarc, A. & Kozelj, V. Ear, nose and voice problems in children with orofacial clefts. *J. Laryngol. Otol.* **120**, 276–281 (2006).
30. Wu, K. *et al.* An automated framework for assessing how well LLMs cite relevant medical references. *Nat. Commun.* **16**, 3615 (2025).
31. Goldberg, J. D., Pierson, S. & Johansen Taber, K. Expanded carrier screening: What conditions should we screen for? *Prenat. Diagn.* **43**, 496–505 (2023).
32. Taber, K. J. *et al.* A guidelines-consistent carrier screening panel that supports equity across diverse populations. *Genet. Med.* **24**, 201–213 (2022).
33. Lee, K. *et al.* ACMG SF v3.3 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **27**, 101454 (2025).

Table 1. Quality of Life (PedsQL) Domain Distribution and Impact Across Severity Tiers.

Tier/ PedsQL domains	N Phenotypes	Physical Funct. %	Cognitive / School %	Physical Symptoms %	Emotional %	Social %
Tier 1	269	47.8	30.9	13.1	4.2	4.0
Tier 2	2,939	72.3	9.3	14.7	1.9	1.8
Tier 3	2,306	50.2	19.5	8.9	10.1	11.4
Tier 4	1,400	30.9	2.5	4.9	38.3	23.5

Table 2. Gene-Level Severity Distribution Across Body Systems. Severity distribution of gene–disease associations grouped by HPO parent term (body system). Severity was assigned by the hierarchical aggregation algorithm operating on tier counts of constituent phenotypes (see Methods, Equation 2). Associations were filtered using HPOA frequency annotations to retain only Obligate, Very Frequent, and Frequent ($\geq 30\%$) phenotypes, removing Occasional, Very rare, Excluded, and NOT-qualified annotations. A total of 8,738 gene–disease pairs were retained. Body systems are sorted by descending Profound burden. Each row sums 100% across the four severity categories.

Body System (HPO Parent Term)	N Associated genes	Profound %	Severe %	Moderate %	Mild %
Abnormality of prenatal development or birth	679	63.9	32.3	2.7	1.2
Abnormality of the respiratory system	2,424	63.1	31.8	4.9	0.1
Abnormality of the voice	260	59.6	37.7	2.7	0.0
Abnormality of the nervous system	5,312	58.9	38.5	2.4	0.2
Abnormality of the breast	271	57.9	33.9	8.1	0.0
Abnormality of the musculoskeletal system	4,854	55.2	42.7	2.0	0.1
Growth abnormality	3,748	54.5	43.4	1.8	0.3
Abnormality of the ear	2,164	54.2	34.9	10.8	0.1
Abnormality of head or neck	4,917	53.3	40.9	5.0	0.9
Abnormality of limbs	3,730	51.3	47.4	1.2	0.2
Abnormal cellular phenotype	652	49.5	39.4	10.4	0.6
Abnormality of the eye	4,916	48.7	44.5	6.8	0.0
Abnormality of the digestive system	3,273	46.7	48.0	4.7	0.6
Abnormality of the cardiovascular system	2,774	42.0	53.4	4.0	0.6
Abnormality of the immune system	829	39.8	45.6	13.5	1.1
Abnormality of metabolism / homeostasis	2,628	38.6	51.4	7.1	2.9
Abnormality of the genitourinary system	2,879	37.8	50.7	6.2	5.3
Abnormality of the integument	2,785	34.8	53.9	9.8	1.6
Abnormality of the endocrine system	1,616	30.8	60.0	6.1	3.1
Neoplasm	760	27.5	44.3	27.5	0.7
Constitutional symptom	1,501	26.9	68.9	3.5	0.7
Abnormality of blood and blood-forming tissues	1,873	26.9	58.2	10.3	4.6
Abnormality of the thoracic cavity	13	23.1	76.9	0.0	0.0
Overall (all 8,738 gene–disease pairs)	8,738	38.1	41.4	16.4	4.1

Figure legends

Figure 1. AI-based Workflow for Automated Phenotype and Gene Severity Classification. Schematic overview of the complete classification pipeline demonstrating: (1) Phenotype classification criteria integrating ACMG severity tiers and ACOG criteria including PedsQL domains, age of onset, and treatment availability; (2) ReAct agent framework with iterative thought-action-observation loops executing contextualization, query generation, evidence retrieval, and adaptive expansion; (3) Classification synthesis with source verification; and (4) Gene-level severity aggregation algorithm producing final classifications of Profound, Severe, Moderate, or Mild based on associated phenotype profiles.

Figure 2. HPO Classification Model Performance Metrics. (A) Distribution comparison between validation sample and full population across phenotypic categories, with Population Stability Index (PSI) reported above the panel. (B) Overall model performance metrics including accuracy, weighted F1-score, balanced accuracy, and Matthews Correlation Coefficient. (C) Normalized confusion matrix displaying classification patterns across 14 phenotypic categories, with minimal off-diagonal misclassifications concentrated between phenotypically related categories. (D) Per-category performance metrics showing precision, recall, and F1-scores across all phenotypic classes. (E) Jitter scatter plot of predicted versus true severity tiers, with quadratic-weighted kappa and mean absolute error reported. (F) Severity tier-stratified performance metrics showing precision, recall, and F1-scores across all four ordinal severity levels.

Figure 3. HPO Terms Classifications based on Clinical Landscape, Management Complexity and Early Onset. (A) Distribution of severity tiers within each of 14 phenotypic categories, revealing clear stratification of life-limiting conditions. (B) Prevalence of early onset designation by phenotypic category. (C) Management profiles showing distribution of intervention types across categories. (D) Management complexity across severity tiers, demonstrating increased interventions with higher severity.

Figure 4. Evidence Quality and Classification Confidence Metrics. (A) Tier-stratified evidence profiles showing proportions of direct evidence, valid inferences, weak inferences, and unsupported claims across severity classifications. (B) Category-specific evidence profiles demonstrating variation in evidence quality across all 14 phenotypic classification groups. (C) Distribution of statement support scores by severity tier, with violin plots revealing confidence patterns from Tier 1 (most severe) to Tier 4 (least severe). (D) Category-stratified statement support score distributions demonstrating evidence strength and classification confidence across all phenotypic categories.

Figure 5. Association of HPO Classifications with Quality-of-Life (QoL) Domain Impact Profiles. (A) Tier-stratified distribution of PedsQL™ quality of life domains, demonstrating domain-specific functional impact patterns across severity classifications, with Tier 1 showing the most diverse multi-domain involvement and Tier 4 exhibiting concentrated single-domain effects. (B) Category-specific quality of life domain distributions across all 14 phenotypic classifications, revealing domain-specific impact signatures, with Physical Functioning predominating in structural abnormalities, Cognitive/School functioning in neurodevelopmental conditions, and balanced multi-domain profiles in life-threatening phenotypes.

Figure 6. Management Complexity and Quality of Life Domain Analysis of Predicted Severity Tiers.

(A) Correlation between QoL domain involvement and therapeutic intensity, with the fitted linear relationship overlaid on the data. (B) Overall frequency distribution of PedsQL™ domains across all classified phenotypes. (C) Tier-stratified management complexity showing mean number of interventions with standard deviation error bars. (D) Category-specific management complexity with 95% confidence intervals across the 14 phenotypic categories.

Figure 7. Gene-Level Severity Classification and Validation.

(A) Overall severity distribution across gene-disease pairs (NFC-only pairs excluded), showing the proportions of Profound, Severe, Moderate, and Mild classifications. (B) Body system stratified severity burden across 23 physiological systems, ranked by proportion of Profound classifications. (C) Mode of inheritance specific severity distributions across gene-disease pairs, stratified by autosomal dominant, autosomal recessive, X-linked, and complex inheritance patterns. (D) External validation through comparative analysis across established clinical gene panels (ACMG Secondary Findings, Mackenzie's Mission, and ACMG carrier screening), with combined Profound/Severe burden reported for each panel.