npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01331-1

Phenotype driven molecular genetic test recommendation for diagnosing pediatric rare disorders

Fangyi Chen[1], Priyanka Ahimaz[2,3], Quan M. Nguyen[4,5], Rachel Lewis[2], Wendy K. Chung[6], Casey N. Ta[1], Katherine M. Szigety[7], Sarah E. Sheppard[7], Ian M. Campbell[7], Kai Wang[4], Chunhua Weng[1,8] ✉ & Cong Liu[6,8] ✉

Patients with rare diseases often experience prolonged diagnostic delays. Ordering appropriate genetic tests is crucial yet challenging, especially for general pediatricians without genetic expertise. Recent American College of Medical Genetics (ACMG) guidelines embrace early use of exome sequencing (ES) or genome sequencing (GS) for conditions like congenital anomalies or developmental delays while still recommend gene panels for patients exhibiting strong manifestations of a specific disease. Recognizing the difficulty in navigating these options, we developed a machine learning model trained on 1005 patient records from Columbia University Irving Medical Center to recommend appropriate genetic tests based on the phenotype information. The model achieved a remarkable performance with an AUROC of 0.823 and AUPRC of 0.918, aligning closely with decisions made by genetic specialists, and demonstrated strong generalizability (AUROC:0.77, AUPRC: 0.816) in an external cohort, indicating its potential value for general pediatricians to expedite rare disease diagnosis by enhancing genetic test ordering.

Individuals with rare diseases often endure a long diagnostic odyssey, filled with an average of 25 examination appointments with specialists[1], and potential misdiagnoses, resulting in anxiety, financial strain, and missed treatment opportunities as their condition progresses[2–4]. A significant portion of rare diseases (50–75%) present initially in children[5], with approximately 80% of rare diseases having a genetic etiology[6]. A genetic diagnosis is essential to understanding the cause and expected natural history of the condition, avoiding unnecessary testing, optimizing management, and facilitating appropriate support systems[7,8].

The standard genetic testing process for suspected genetic disorders follows a stepwise approach[9], depicted in Fig. 1. For example, patients with developmental disabilities undergo Chromosomal Microarray (CMA) and Fragile X syndrome testing as the first-tier tests. If these fail to diagnose, the second-tier tests, which include targeted single-gene tests or gene panels are used based on clinical symptoms to identify underlying genetic variations. If these also prove inconclusive, Whole Exome or Whole Genome Sequencing (ES/EG), known for their higher diagnostic yields but at higher costs, might be used as the third-tier tests.

Alternative testing strategies using ES/GS at different diagnostic stages have been developed. Several cost-effectiveness analyses[9–11] indicate that employing ES/GS directly is more economical for certain conditions. According to the latest American College of Medical Genetics and Genomics (ACMG) guidelines, ES/GS is recommended for conditions like congenital abnormalities and developmental disorders[10,12]. Despite ES/GS having higher yield in detecting pathogenic variants, there are challenges such as interpreting large amounts of data, managing incidental findings unrelated to the primary diagnosis but medically actionable[13,14], inconsistent insurance coverage, and long turnaround times averaging 18 weeks[15,16]. This suggests that well-covered specific disease focused gene panels are more suitable when the phenotypes clearly indicate an underlying genetic disease with a defined set of genes. Recognizing these challenges, the ACMG guidelines[17] recommend that patients with a high likelihood of a specific genetic disorder should first undergo targeted testing or gene panels.

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA. [2]Department of Pediatrics, Columbia University, New York, NY, USA. [3]Institute of Genomic Medicine, Columbia University, New York, NY, USA. [4]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [5]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA. [6]Division of Genetics and Genomics, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [7]Division of Human Genetics, Department of Pediatrics, Children's Hospital of Philadelphia, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [8]These authors contributed equally: Chunhua Weng, Cong Liu. ✉e-mail: cw2384@cumc.columbia.edu; cong.liu@childrens.harvard.edu

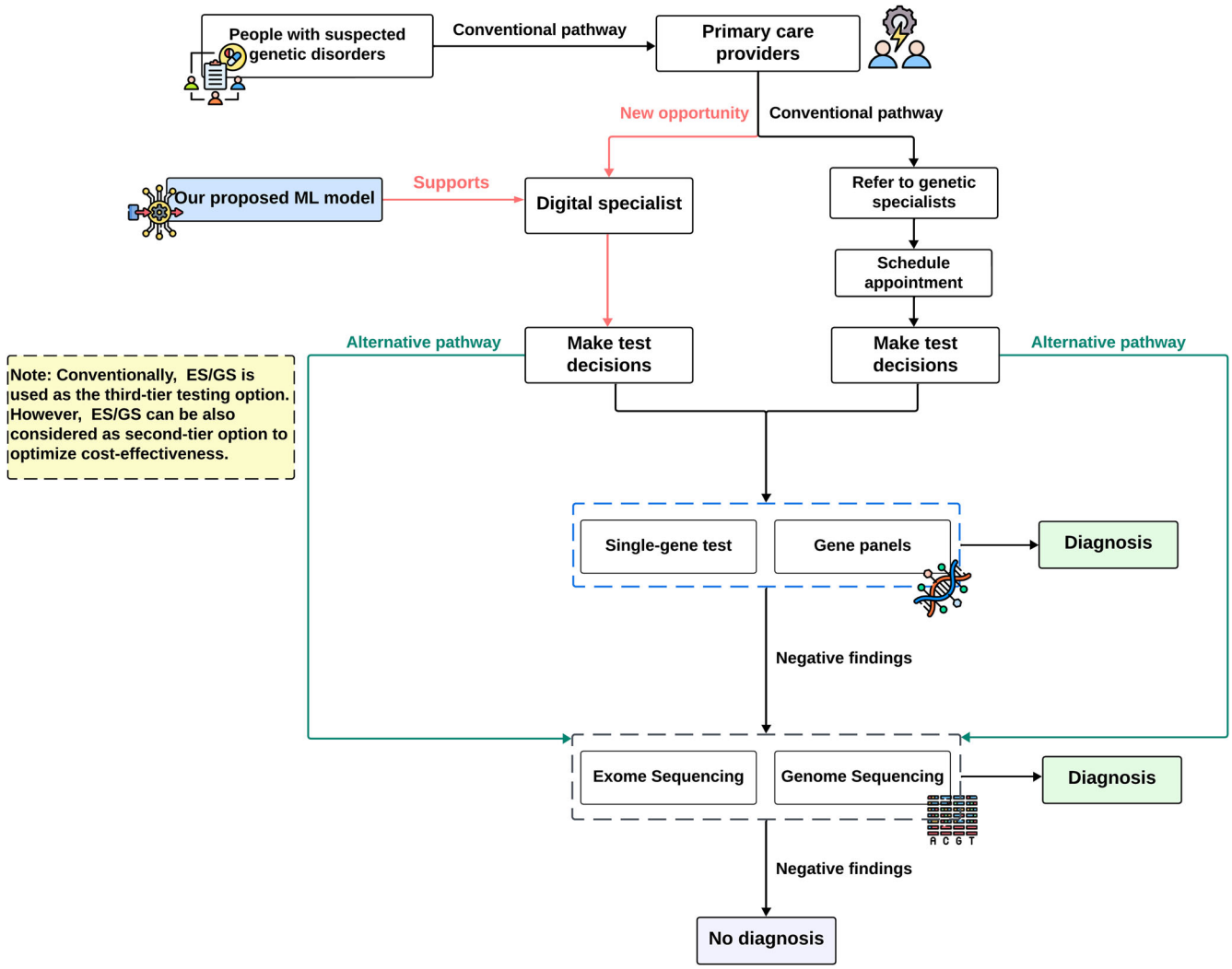


Fig. 1 | Genetic testing pathway for patients with suspected genetic disorders. Conventionally, General pediatricians would refer to genetic specialists to evaluate and order tests, which delays the overall diagnosis. Instead, they could directly make the decisions to expedite the diagnosis process. However, physicians with limited genetic expertise often encounter difficulty in navigating the two options between single gene test/targeted gene panels and exome sequencing (ES)/genome sequencing (GS). Our model aims to support this decision by predicting a geneticist's decision.

Consequently, deciding whether to use ES/GS or gene panels as the first test is becoming a crucial and complex, case-by-case decision for clinicians, particularly when there is strong suspicion of a genetic basis for the condition. This is especially challenging for general pediatricians, who usually do not have specialized training in genetics and do not routinely order genetic testing. As clinical guidelines increasingly lean towards recommending ES/GS as a diagnostic test for conditions with long and ill-defined genetic differentials, pediatricians face growing challenges in navigating these options in everyday practice[18]. As a result, most general pediatricians refer patients to geneticists, delaying the timely ordering of crucial tests (Fig. 1).

In light of this, the study aims to develop a machine learning model (Phen2Test) that uses patient phenotypes documented in electronic health records (EHRs) to predict whether an experienced specialist will directly order ES/GS, bypassing the need for a gene panel test. Trained with data from specialists knowledgeable in genetic test ordering, this model is designed to assist general pediatricians and other physicians who have limited experience in genetics in choosing the appropriate testing strategy.

Results

Data Characteristics

Table 1 showed the demographics of the three cohorts. In the curated cohort ($n = 1005$), 37.1% were white. Initially, there were 570 individuals recommended for ES/GS and 435 for gene panels. The subcategories of gene panels were described in Supplementary Data 1. 139 individuals were shifted to the ES/GS group according to guideline-adjusted outcomes. The mean ages of the patients recommended for ES/GS were 4.88, while the mean ages of the gene panel group were older (5.88). The age difference was also observed in the phenotyping-based cohort ($n = 6458$), with a mean age of around 5.56 years in ES/GS group and 10.05 years in the panel group. It is important to note that in the phenotyping-based cohort, based on the extraction results, the number of panel cases ($n = 3427$) exceeded the number of ES/GS cases ($n = 3031$). In the CHOP cohort ($n = 997$), 537 patients were recommended for ES/GS directly, and 460 patients were suggested for gene panels at first, but 55 were later escalated for ES/GS. The mean ages of the CHOP cohort were older (mean age: 10.05 for ES/GS and 10 for Panel) than the Columbia cohort. The other demographics distributions, such as gender and race, were similar in both CUIMC and CHOP data.

Model optimization

After experimenting with different feature sets and sampling strategies, the model achieving the highest average AUPRC (0.918, std: 0.023) was Random Forest built on features including phecodes aggregated by frequency derived from structured data, demographics characteristics and the number

Table 1 | Demographics characteristics of initial manually curated and larger phenotyping-based cohorts

Demographics Characteristics	Clinicians–Curated Genetic Cohort (n = 1005)		EHR Phenotyping-based Cohort (n = 6458)		Children's Hospital of Philadelphia Cohort (n = 997)		Total
	ES/GS[a]	Panel[a]	ES/GS[b]	Panel[b]	ES/GS[c]	Panel[c]	
Self-reported Race							
White	271 (227)	102 (146)	1447	1642	384	244	4090
Black or African American	79 (58)	34 (55)	344	363	69	63	952
Asian	30 (25)	9 (14)	116	78	26	27	286
Other (e.g., American Indians or Alaska nation)	2 (2)	4 (4)	504	42	-	-	552
Not described	176 (139)	79 (116)	42	592	88	46	1023
Decline or not specified	151 (119)	68 (100)	578	710	25	25	1577
Sex							
Male	452 (362)	169 (259)	1687	1502	347	219	4376
Female	257 (208)	127 (176)	1342	1924	245	186	4081
Other	-	-	2	1	-	-	3
Age at the index date							
Mean age ± Std	4.88 ± 4.87 (4.92 ± 5.00)	5.88 ± 5.91 (5.50 ± 5.46)	5.56 ± 5.53	10.05 ± 6.09	10.52 ± 6.16	10 ± 6.23	-
Median age	3.48 (3.46)	3.65 (3.61)	5.27	10.87	8.74	7.79	-
0–5	443 (359)	165 (249)	1470	914	83	103	3178
5–10	140 (103)	52 (89)	722	675	261	143	1993
10–15	93 (79)	47 (61)	505	837	120	71	1673
>= 15	33 (29)	32 (36)	334	1001	128	88	1616
Total cases	709 (570)	296 (435)	3,031	3,427	592	405	8460

ES/GS whole-exome/whole-genome sequencing, *Panel* gene panel.
[a]Data collected from Columbia University Irving Medical Center (CUIMC). Genetic testing outcomes curated by clinicians, followed by the outcome counts after adjustment (statistics before adjustment).
[b]Data collected from Columbia University Irving Medical Center (CUIMC). Genetic testing outcomes extracted from clinical narrative notes based on phenotyping approach.
[c]Data collected from Children's Hospital of Philadelphia (CHOP), served as the external testing dataset for the algorithm. Genetic testing outcomes were curated by clinicians.

Table 2 | Average performances across different feature sets based on the held-out testing sets in the initial cohort after outcome adjustment

Feature Sets	Feature Dimension	Average AUROC	Average AUPRC
Phecodes [structured] + Demographics	1228	0.831 ± 0.040	0.917 ± 0.022
HPO + Demographics	26	0.737 ± 0.051	0.852 ± 0.042
Phecodes [unstructured] + Demographics + note counts	422	0.803 ± 0.049	0.90 ± 0.025
Phecodes [structure] + Demographics + note counts	1229	0.823 ± 0.045	0.918 ± 0.023
HPO + Demographics + note counts	27	0.736 ± 0.054	0.864 ± 0.035
Phecodes [structure and unstructured] + Demographics + note counts	1297	0.838 ± 0.052	0.917 ± 0.028

of notes. It applied class weight adjustments to address class imbalance and did not perform feature reduction. Phecode-based feature sets yielded better performance compared with features derived from HPO ontology, as shown in Table 2. Additionally, we did not perceive a meaningful gain in the performance while leveraging phenotypes extracted from both structured data and narrative notes. The ROC curve (the highest average AUROC 0.823, std: 0.045) and precision-recall curve for three classifiers trained using the optimal feature set were depicted in Fig. 2a. The performance of all trained models across various features, strategies and classifiers were summarized in Supplementary Data 2.

The top 10 important phenotypes and HPO phenotype categories in our optimal model were visualized in Fig. 2b, c. Systems of phenotypic abnormalities significantly (p-value < 0.05) and positively correlated with ES/GS ordering included nervous system, genitourinary system abnormality, and birth abnormality, whereas abnormality of the digestive system showed a negative correlation (Supplementary Data 3). Furthermore, phenotypes significantly enriched in phecode sets included neurological, mental disorders, metabolic, sense organs, and dermatologic (see Supplementary Data 4 for the complete list). Overlap was observed between top important HPO phenotypic abnormalities (abnormality of nervous system, integument, ear, and metabolic) and identified significant phecode sets (neurological, dermatologic, sense organ, endocrine/metabolic).

Internal evaluation and baseline assessment

The performance (AUROC, AUPRC) distribution of Phen2Test, evaluated under two reference standards, was depicted in Fig. 3b. The average AUROC and AUPRC were 0.916 (95% CI: 0.911–0.921) and 0.962 (95% CI: 0.959–0.964), respectively, when the standard was based on a genetic expert's decisions. The model performance was similar

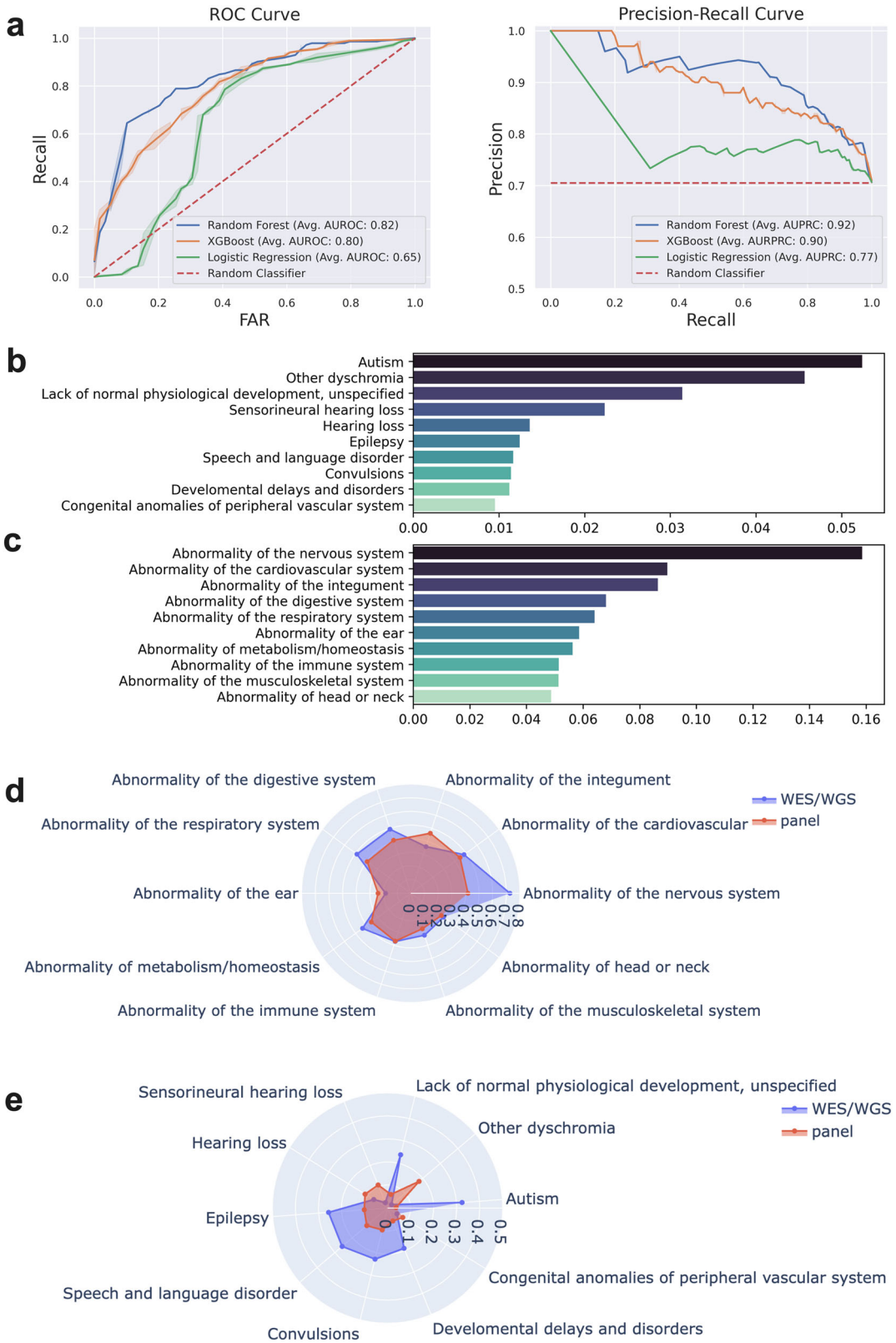
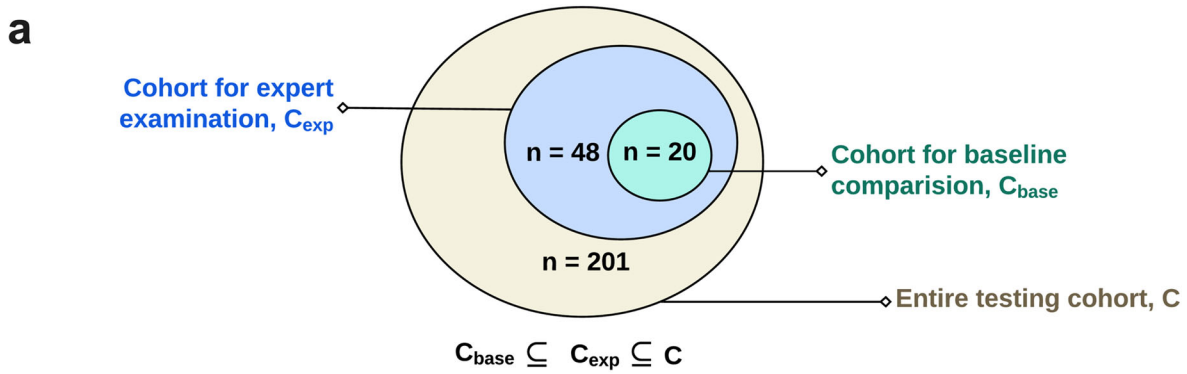


Fig. 2 | Model optimization and feature importance visualization. a Receiver operating characteristic (ROC) Curve (left) and Precision-recall Curve (right) derived from the Initial Curated Testing Cohort. Receiver operating characteristic (ROC) curves and Precision-Recall curves were used to illustrate the performances of the optimal feature sets across three classifiers on the held-out testing dataset from the initial cohort. The shaded area represents the confidence interval calculated from 1000 iterative results. The top 10 feature importance scores calculated by Gini impurity for both (**b**) Phecodes ($n = 1225$), and (**c**) HPO phenotypic abnormalities ($n = 23$). Two radar charts illustrate the proportion of (**d**) the top 10 HPO phenotypic abnormalities, and (**e**) Phecodes within two different groups: whole-exome sequencing/whole-genome sequencing (WES/WGS) and panel.



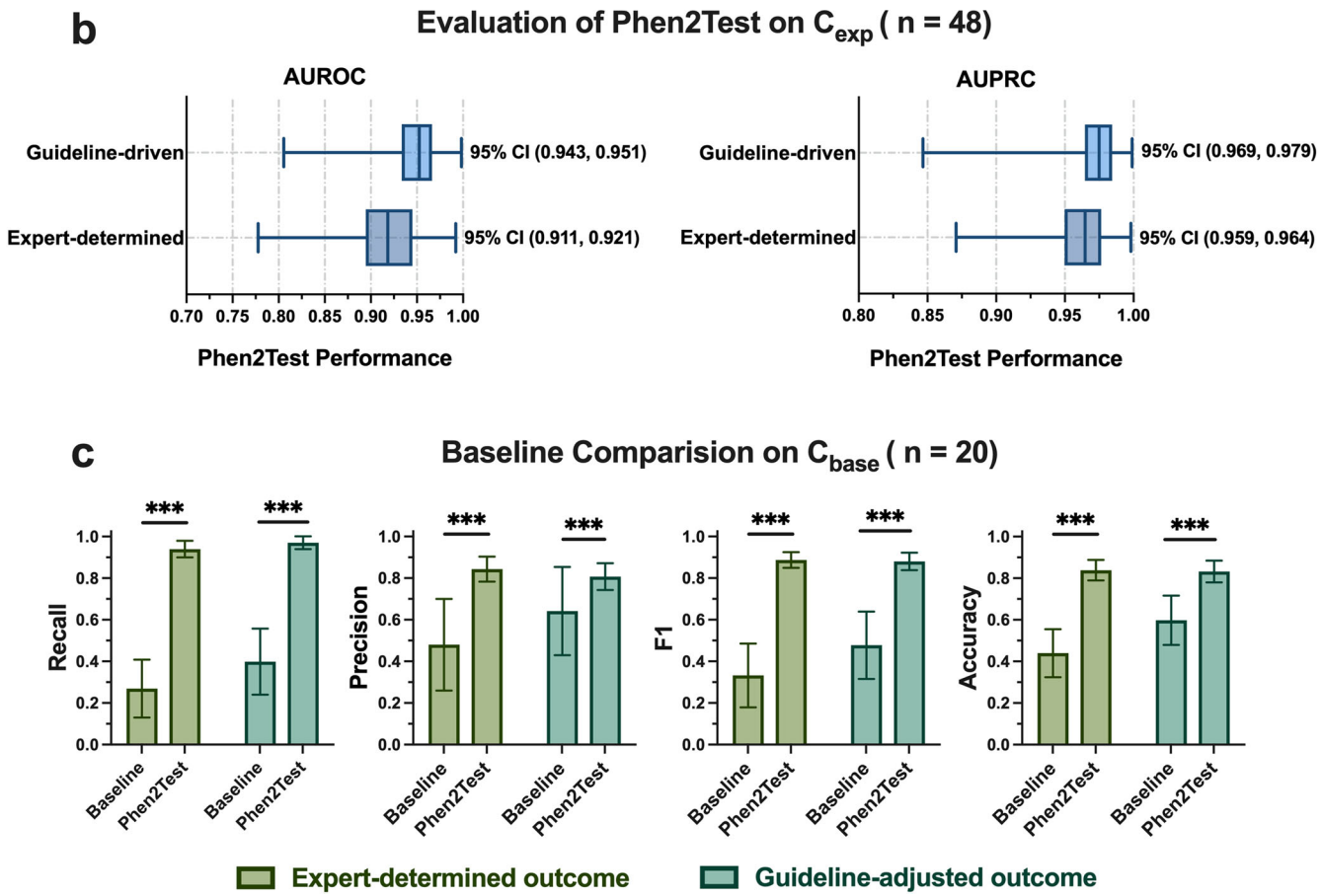


Fig. 3 | Results of internal evaluation and baseline assessment. a In addition to evaluating the algorithm on the entire testing cohort, we curated two specific subsets for further validation: C_{exp} ($n = 48$) and C_{base} ($n = 20$). C_{exp} was reassessed by a genetic specialist, and C_{base} was randomly selected from C_{exp} cohort. Two evaluation standards were applied in these two cohorts. **b** The bootstrap performance (AUROC, AUPRC) of Phen2Test model evaluated on cohort C_{exp} under two standards: guideline-adjusted and expert-determined outcomes. **c** The performance comparison between the model and a general pediatrician (assuming the 2 unknowns as ES/GS) in determining the genomic tests based on phenotypic manifestation. The bars in (**c**) represented the average performance of the bootstrap samples in C_{base}, with the error bar indicating standard deviation of the mean. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

against the guideline-adjusted evaluation standard, where the average AUROC was 0.947 (95% CI: 0.943–0.951) and AUPRC was 0.972 (95% CI: 0.969–0.979). Furthermore, we compared the model with a general pediatrician who does not order genetic tests routinely. When we switch the 'not sure' decisions ($n = 2$) to 'ES/GS' (or 'gene panels'), the average accuracy for a general pediatrician becomes, 0.44 (or 0.56) and 0.598 (or 0.588) in the expert-determined and outcome-adjusted standard, respectively. In contrast, the average accuracy of Phen2Test was 0.838 and 0.795, significantly higher ($p < 0.001$) than a general pediatrician's performance (Fig. 3c).

External Validation

Phen2Test was applied to an external cohort ($n = 997$) with data characteristics documented in Table 1. An overall AUROC of 0.77 and AUPRC of 0.816 were achieved on the CHOP cohort (Fig. 4a), with around a 5% decline in AUROC and an 10% reduction in AUPRC when compared with the CUIMC initial testing cohort. However, Phen2Test evaluated on CHOP data outperformed the CUIMC EHR Phenotyping-based cohort (Fig. 4b). Figure 4a showed improvement in both AUROC and AUPRC when comparing across two time periods: 2016-2019 and 2020-2021, with the AUPRC increasing from 0.807 in 2016–2019 to 0.820 in 2020–2021.

Sensitivity analysis

The average accuracy is lower (0.718 AUROC; 0.671 AUPRC) when evaluated on the EHR-based cohort. As our model was trained based on the outcomes assigned according to the current guideline, we observed a discernible trend of performance improvement as the time approached more recent dates, with the AUROC increasing from 0.68 in 2012 to 0.76 in 2020 (Fig. 4b). The drop in performance post-2021 was likely caused by shifts in healthcare behavior due to the pandemic which significantly disrupted traditional clinical practices[19].

We examined the performance disparity in different subgroups categorized under self-reported race, sex, and age groups (Fig. 4c, detailed in Supplementary Data 5). The model performance remained stable when assessed across different demographic subsets, with low variance of AUROC and AUPRC within each of the three testing sets (Fig. 4d) among the subgroups.

Cost-effectiveness analysis

A brief cost-benefit analysis was conducted to better understand and showcase the clinical utility of Phen2Test. We compared the potential costs of the test recommendation made by Phen2Test and two clinical choices: 1) always select targeted gene panels first and transit to ES/GS when negative results are found (tiered approach), 2) directly opt for ES/GS. The diagnostic yield of target gene panels in non-consanguineous population ranges between 25–45%[11,20,21]. In our analysis, we used a diagnostic yield rate of 40% to estimate costs. The prices for genetic tests were referenced from Blueprint Genetics (GTR Lab ID: 500188). The average price of gene panels was estimated to be $1600, and the average price of ES/GS was estimated to be $3250. Under such diagnostic yield and average prices, at the present time evaluated under the CHOP institution with a prediction threshold set as 0.5 and assuming geneticists always select the optimal testing sequence, Phen2Test yields an expected savings of $236 in testing cost per patient compared to the ES/GS-only approach and $536 relative to the tiered approach (Table 3). An interactive web demo was published online, which allows users to manipulate parameters such as input prices, diagnostic yield rates, and test preferences. It dynamically reflects the total costs for each scenario.

Discussion

Clinical manifestation has been consistently leveraged to facilitate rare disease diagnosis. One of the applications is to better assist interpretation of candidate variants yielded from the sequencing results. For example, bioinformatics tools such as Exomiser[22], VarElect[23], Phenolyzer[24], and Phen2Gene[25] have been developed to provide more comprehensive variants interpretation. Also, those tools can be used to design virtual gene panels (with ES/GS as the backbone) by considering the phenotype-genotype relationships[26]. Subsequently, researchers have explored additional aspects of utilizing phenotypes to support clinical decisions at various stages of the rare disease diagnosis process. For example, a previous study has shown that shared phenotypes among patient groups can be leveraged to identify individuals likely to benefit from CMA testing with relatively high accuracy[27]. As ES/GS becomes more prevalent, our study further supports the genetic test ordering process by providing personalized test ordering recommendations based on patients' phenotypes. Our findings contribute to a future workflow for systematically using EHR-based phenotypic features for rare disease diagnosis.

Phen2Test achieved the highest performance using features from structured billing codes (converted to Phecodes) and demographics. While detailed HPO terms offer refined phenotypic descriptions, beneficial for variant interpretation and identifying causal variants, we did not pursue this due to the resource-intensive nature of manual extraction and the limited accuracy of automated NLP approaches[28–30]. Our analysis showed no performance improvement when extracting HPO phenotypes from relevant notes using a context-aware query approach[31]. Although our keyword-based approach is basic, it serves as a proof of concept for using phenotypic information to guide genetic test selection. Future studies might consider transformer-based algorithms[30,32,33] to enhance recognition of HPO or non-HPO concepts from narrative notes[34].

The lack of improvement using narratives might be due to the simplicity of our phenotype extraction algorithm, or essential information for predicting genetic test ordering is already effectively summarized in structured data (as conditions). Individuals who benefit most from ES/GS are likely to show abnormalities across multiple systems[35], which can be inferred from the ICD codes. Furthermore, our recommendation aligned with established clinical guidelines, such as recommending ES/GS for conditions like multiple congenital anomalies, autism, and developmental delay, all of which are likely to be summarized within the structure data.

ES/GS was initially introduced into the clinical settings in 2012[36], a time when both clinical application and supporting evidence were limited[37]. As clinical guidelines evolved, more physicians began to incorporate ES/GS into their diagnostic workup. Our model, adjusted for the latest guidelines, effectively reflected this temporal trend – exhibiting improved performance in more recent cohorts. This underscores its potential as a component of a learning health system[38], capable of direct training using the EHR data, facilitating the dynamic updating of clinical guidelines, and consequently offering decision support to general pediatricians. Moreover, while ES/GS has been used for newborn screening[39], the limited phenotypes present at birth challenge the data analysis at prenatal or neonatal stages[40]. Re-analyzing sequencing data as knowledge evolves is beneficial[26,41], though the appropriate timing is unclear[42]. Phen2Test is primarily designed for the selection of genomic/genetic tests based on the latest EHR data, which can also identify the optimal timing for re-analysis by examining the latest phenotypes, aiding in the selection of virtual panels and ES/GS.

Several concerns must be addressed before deploying this system into a routine clinical workflow. First, we observed a higher ratio of ES/GS in our cohorts, likely due to the high-resource settings of the study. Hence, it is essential to acknowledge that the system's generalizability may be limited when applied to lower-resource healthcare systems, as the training data might not accurately represent the true distribution within the intended cohort. Unfortunately, this challenge is a pervasive limitation encountered in many machine learning or AI-based approaches, where models are often trained within high-resource academic centers but are increasingly sought after in low-resource clinical settings, such as rural pediatric clinics[43]. While our simple billing-code-based feature engineering may help adaptability, addressing this imbalance remains crucial. Additionally, Phen2Test lacks the capacity to incorporate facial photos and videos, which could enhance multi-modal decision support. Furthermore, financial costs and resource availability are also relevant factors in the real-world decision-making process[18,44] but are not accounted for currently. While Phen2Test's performance aligns well with genetic specialists, general pediatricians may still lack the confidence to use it for ordering tests. In practice, they may prefer to refer patients to geneticists, causing inevitable delays in diagnosis due to long waiting times for appointments. Integrating retrospectively validated models into frontline clinical practice remains a sociotechnical challenge. On the other hand, payer barriers have limited the clinical use of ES/GS for patients with suspected genetic diseases[45,46], with reimbursement sometimes being denied[47]. Although Phen2Test can provide more objective decisions to justify the need for specific genetic tests, it remains uncertain how willing payers are to accept AI-supported orders.

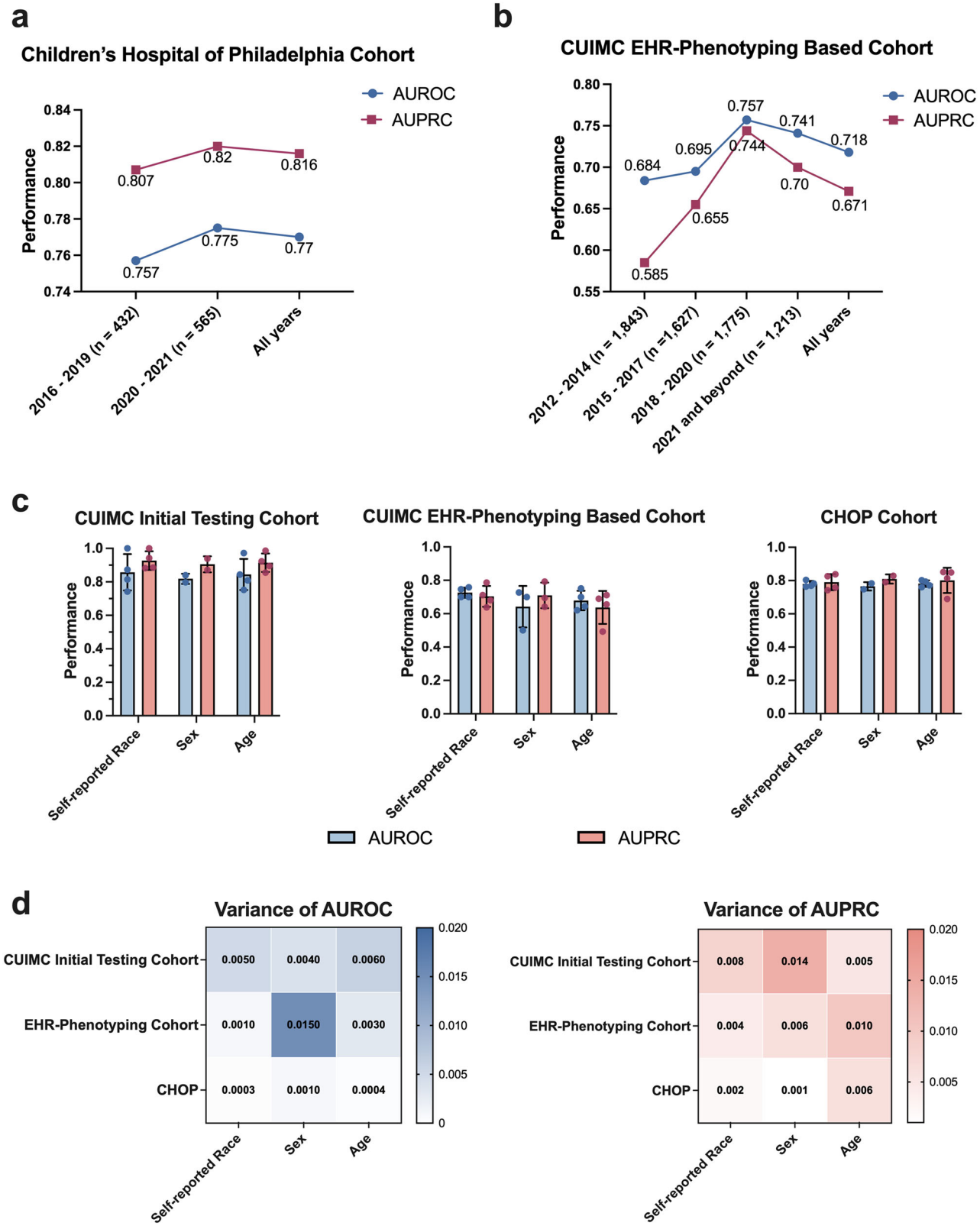


Fig. 4 | Phen2Test external evaluation and sensitivity analysis. The area under receiver operator characteristics curve (AUROC) and precision-recall curve (AUPRC) divided by calendar years were shown for (**a**) Children's Hospital of Philadelphia, and (**b**) CUIMC EHR-phenotyping based cohort. **c** The AUROC and AUPRC of each subgroup divided by self-reported race, sex and age categories were represented in dotted points, with the bar represented as the mean, error bar representing standard deviation of the mean. **d** Heatmap visualization of the performance variability for demographic subgroups.

Table 3 | Comparative cost analysis

	Targeted Gene Panels	ES/GS-only	Our Model Prediction
Average price	$1600	$3250	Not applicable
Total cost for 100 patients	$355,000	$325,000	$301,400
Average cost per patient	$3550	$3250	$3014
Expected savings per patient using our model	$536	$236	Not applicable

The average prices for gene panels and ES/GS were estimated based on Blueprint Genetics (https://blueprintgenetics.com/). Blueprint Genetics offers target gene panels priced between $1450 to $1750, and whole exome/genomic sequencing services range from $2500 to $4000. The gene panel diagnostic yield rate was set to 40% for the calculations. Follow-up and secondary finding management costs were not considered in this analysis. Phen2Test's prediction accuracy was evaluated against the CHOP cohort, with a threshold set at 0.5 (assign to the class with higher predictive probability). The interactive cost analysis demo was available at https://comparative-cost-demo-63ivhjau4u4r8436pyblbe.streamlit.app/.

This retrospective study develops and validates an accurate phenotype-driven approach to identify suitable molecular genetic tests needed to diagnose rare pediatric disorders for clinicians with minimal knowledge about genetic tests. The model's performance is comparable to that of genetic specialists. More studies are warranted to test the effectiveness of this approach prospectively in clinical settings and the generalizability to different EHR systems.

Methods

The study developed binary classification models that use phenotypic features extracted from the EHR as inputs. The outcomes determine whether ES/GS (*positive*) should be ordered directly or gene panels (*negative*) should be considered first. Figure 5 provides an overview of the study design, with further details elaborated in subsequent sections.

Ethical statement

The study received ethical approval from Columbia University Irving Medical Center Institutional Review Board (protocol number: AAAR3954) and Children's Hospital of Philadelphia Institutional Review Board (protocol number: 18-015712). This EHR-based research was determined by Columbia University Irving Medical Center Institutional Review Board and Children's Hospital of Philadelphia Institutional Review Board to qualify for a waiver of consent as per 45CFR46.116(d) as the following criteria are met in this study:

The research involves no more than minimal risk to the subjects. The research involves analysis of existing data only, and the risk from this research is minimal. There is a risk that participants could be harmed in the unlikely event that information was disclosed outside the study in an identifiable way. We will take multiple measures to protect the privacy and confidentiality of all involved participants, and to minimize the risks associated with possible distress and burden. All data analysis will be performed in secure and CUIMC-approved servers, and we will safeguard the data sets to ensure the privacy of the patients, and to eliminate the risks of data leak.

The waiver or alteration will not adversely affect the rights and welfare of the subjects. The waiver will not adversely affect the rights and welfare of the study participants, because the analysis is based on existing data without any new data collection. Additionally, the data will not be released, and strict confidentiality will be maintained.

The research could not practicably be carried out without the waiver or alteration. The research involves retrospective analysis of existing phenotype data of a large number of patients. Without the waiver, this analysis cannot be carried out.

Whenever appropriate, the subjects will be provided with additional pertinent information after participation. The subjects will be provided with relevant information on refined phenotype analysis of the raw clinical phenotype information. When requested, publications resulting from the proposed study will also be shared to study participants.

Data collection and outcome preprocessing

To train a model that helps general pediatricians select appropriate testing strategies without referring to genetic specialists, the intuition here is to align the model's prediction outcomes with genetic specialists' decisions. We therefore collected genetic test orders on patients seen by geneticists at Columbia University Irving Medical Center's (CUIMC) Department of Pediatrics between 2012 and 2023. We excluded genetic tests with non-diagnostic purposes, CMA-only cases, and patients aged 19 or older when tests were ordered. The genetic test orders were categorized as (1) direct ES/GS (positive), (2) gene panels only (negative), or (3) gene panels followed by ES/GS. The third category was considered 'positive' in this study because ES/GS was ultimately required to make the diagnosis. As the ACMG guidelines evolve with new evidence, past test decisions may no longer conform with current standards, potentially leading to suboptimal outcomes, such as low diagnostic yields and reduced cost-effectiveness. We, therefore, adapted guideline-adjusted outcomes by adjusting the test orders according to the ACMG recommended practice. We leveraged the clinical summaries (Supplementary Table 1) as the basis to adjust the labels. The summaries were manually crafted by clinicians and was not used for model training. For instance, ES/GS was recommended as the first-tier test for patients with congenital anomalies, developmental delay, intellectual disability, neurological developmental disability (e.g., autism spectrum disorder, attention-deficit/hyperactivity disorder)[10,12], or seizures[48,49]. If any of these phenotypic manifestations were found in the patient clinical summaries, we would update the test order to ES/GS if the initial order was a gene panel. Only those with well-established guidelines were adjusted for the recommendation labels. Accordingly, there were 139 patients with test order adjusted to ES/GS (Supplementary Table 2). Finally, the guideline-adjusted outcomes were leveraged for model training.

Feature engineering

Phenotypic features were extracted from both structured and unstructured EHR data, focusing on features available before the index date. The index date was defined as the earliest test order date or, if missing, the genetic appointment visit date, assuming minimal delay between the two. The feature extraction process is detailed below; refer to Supplementary Table 3 for a list of the extracted features.

Structured data

Condition concepts collected based on the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) were mapped to corresponding phecodes (version 1.2)[27,50–52]. We utilized two methods to aggregate phecodes per individual: (1) frequency of each phecode (Freq_phecodes) and (2) sum of unique phecodes (Sum_phecodes), counting identical phecodes recorded on the same date only once. Alternatively, OMOP condition concepts were mapped to Human Phenotype Ontology (HPO)[53] terms following our previously developed approach[31], aggregating them by phenotypic abnormality (e.g., musculoskeletal abnormalities). We then counted each HPO-based organ system of phenotypic abnormality (23 systems) as input features (Freq_HPO). Demographic characteristics (sex, race, age) at the index date were also extracted.

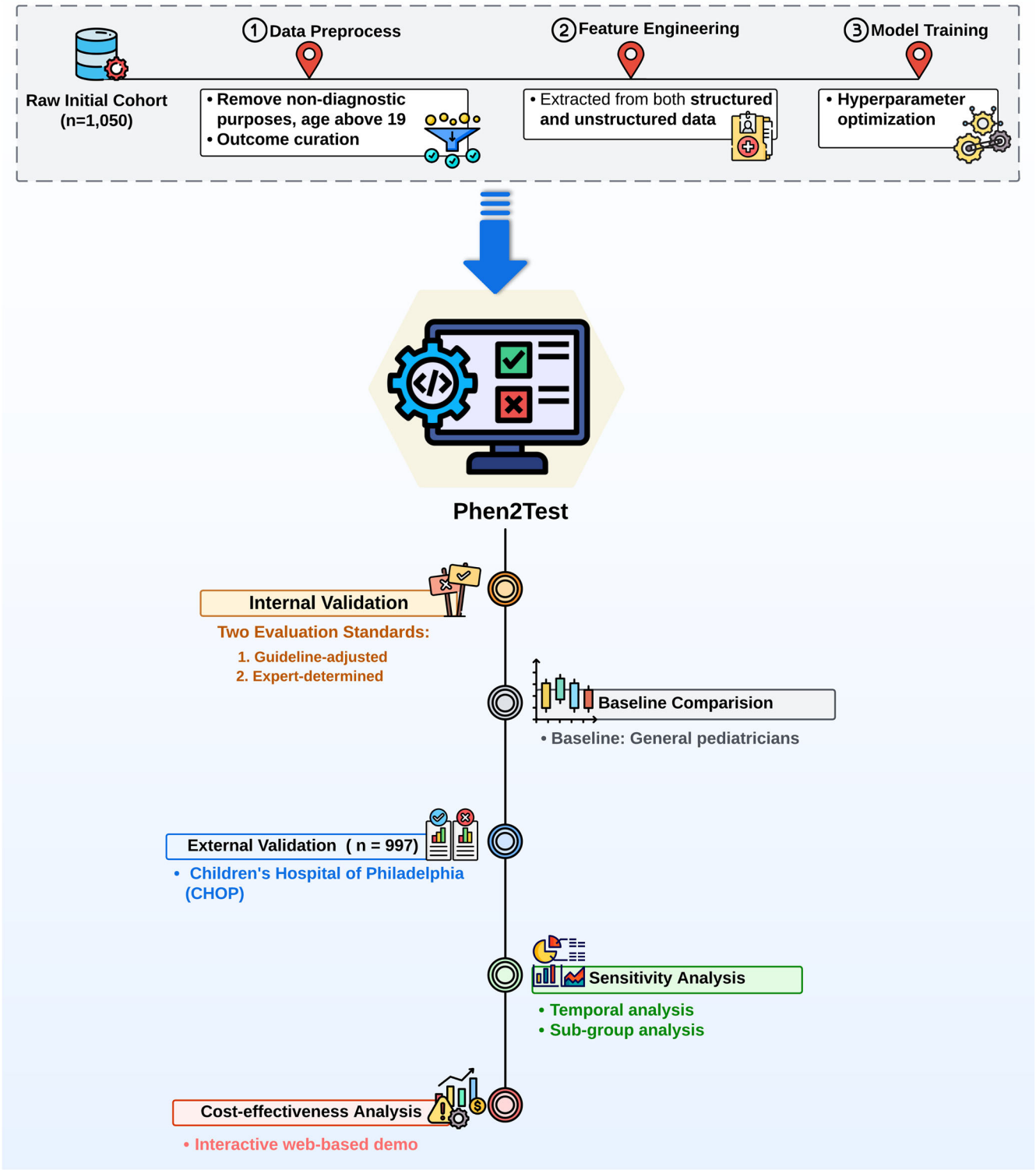


Fig. 5 | Study overview. There are three main steps in constructing and optimizing the models, which includes data preprocessing, feature engineering and model training. During the data preprocessing stage, non-diagnostic purposes and individuals aged 19 and older at the index date were excluded. We adopted a guideline-adjusted approach to systematically adjust test orders to align with latest ACMG recommended practice. In step 2, we extracted features from both structured data and narrative notes. In the end, 6 different feature sets were passed into 3 classifiers. Each classifier was trained with optimized hyperparameters. The model with the highest average AUPRC would be regarded as the best performing model (Phen2Test), which would be used for the proceeding analysis. Four parts of analysis were conducted, including both internal evaluation and external validation of Phen2Test, sensitivity and cost-effectiveness analysis.

Unstructured data

We applied regular expressions to extract additional phenotypic features from clinical narratives that matched with the 1755 phecode terms. Additionally, a trained model[54] for negation and scope detection was utilized to only include observed phenotypes. Features were aggregated using the same aggregation methods previously described (Freq_phecodes_notes and Sum_phecodes_notes). Previous findings[55,56] suggested that the higher frequency of clinical notes correlated with worse health, we therefore counted

all clinical notes documented before the index date (Num_notes) to reflect indidivuals' healthcare utilization.

Model training and optimization

We developed multiple feature sets and trained them using Logistic Regression, Random Forest, and XGBoost classifiers. We conducted nested cross-validation for hyperparameter tuning and model evaluation, using a three-fold inner cross-validation on 80% of the data to select models with best F1-scores, and further evaluated the model using the remaining 20%. To address class imbalance, we used class weight adjustment, SMOTE[57], and random duplication of minority class data points. Furthermore, we performed principal component analysis (PCA) to assess the impact of feature reduction on model performance. The experiments were repeated in five iterations and averaged performance metrics—precision, recall, F1-measure, area under the ROC curve (AUROC) and precision-recall curve (AUPRC) were calculated. The model with the highest averaged AUPRC (optimal) was used for further analysis.

We assessed feature importance in our optimal model using the Gini impurity criterion. We also explored the relationship between phenotypic abnormalities and test recommendations (ES/GS, panel) using ordinary least squares (OLS) regression. Phenotypes were grouped into 17 categories to examine system-level impacts on test choices using a Chi-Square test. The *p*-values were adjusted using Benjamin/Hochberg approach. We compared these findings with the HPO ontology to identify discrepancies introduced due to different vocabularies.

Internal validation and baseline assessment

We assessed the optimal model in a subset C_{exp} ($n = 48$), randomly sampled from the entire testing cohort (Fig. 3a). Two ground-truth evaluation standards applied, (1) guideline-adjusted outcomes (as described above), and (2) expert-determined outcomes, where a genetic counselor (PA) conducted chart reviews for the C_{exp} cohort and provided genetic test decisions (ES/GS, gene panel) according to current practices. The model's predictions were evaluated on the C_{exp} against both standards, with 95% confidence intervals estimated using 200 bootstrap iterations.

Given the primary objective to support general pediatricians with limited genetic expertise, we further conducted a comparison study by establishing a baseline based on a general pediatrician's decision. A pediatrician (RL) reviewed a cohort C_{base} ($n = 20$), randomly sampled from the C_{exp} cohort, and was asked to select the appropriate genetic test (not sure, ES/GS, gene panel) based on the patient's prior-to-index conditions and demographics. The decisions were also evaluated against the two aforementioned standards.

External validation

We used an independent cohort from the Children's Hospital of Philadelphia (CHOP) to assess the portability and generalizability of our model. This dataset included 997 patients referred to CHOP clinical geneticists, containing ICD-10 codes summarizing their conditions before encounters and their genetic test orders. The same feature engineering procedures were applied to prepare model inputs for validation. The trained model was available on GitHub.

Sensitivity analysis: temporal effect and sub-group analysis

We derived an EHR-based 'genetic' cohort from 2012-2023, which included patients with visits to genetic clinics, tests, or measurements (Supplementary Data 6), excluding patients aged 19 or older at their appointment. Testing outcomes were extracted by first identifying clinical notes containing keywords like "genetic", "letter", "visit", and "progress note" in their titles; and then pinpointing notes with "exome", "genomic", "WES", or "WGS as positives (Supplementary Table 4). To refine the negatives, we excluded non-genetic testing (e.g. biochemical panels). The outcome extraction algorithm achieved an 89% accuracy rate in identifying labels on a pre-tested cohort. Index dates were set as the dates when keywords were identified in notes. Patients with both ES/GS and gene panel test orders were considered positives, with the index at the earliest test order dates. The same feature extraction pipeline of the optimal feature set was applied in this cohort. We divided the cohort of 6458 individuals into four periods (2012–2014, 2015–2018, 2019–2021, and 2021–2023) to analyze temporal trends in model performance. Similarly, we also investigated the temporal effects in the external cohort. Model performances across different sub-populations were also examined in all three testing cohorts.

Data availability

The clinical data used in this study contains Protected Health Information (PHI) and, as such, cannot be made readily available for distribution. Requests for access to the data will undergo review by the institutional IRB (Institutional Review Board) for consideration. We provided synthetic data (https://github.com/stormliucong/RARE-GOrder/tree/master/data_preprocessing/demo_data) for users to better understand and execute the training pipelines.

Code availability

The code and the final trained model utilized in this study are available on GitHub at https://github.com/stormliucong/RARE-GOrder. The repository provides a configuration file for users to setup the virtual environment with the required Python packages installed to run the scripts. To avoid potential errors caused by incompatibility, Python version above 3.9 is preferred.

Received: 10 November 2023; Accepted: 7 November 2024;
Published online: 21 November 2024

References

1. Willmen, T. et al. Rare diseases: why is a rapid referral to an expert center so important? *BMC Health Serv. Res.* **23**, 904 (2023).
2. Dawkins, H. J. et al. Progress in rare diseases research 2010–2016: an IRDiRC perspective. *Clin. Transl. Sci.* **11**, 11 (2018).
3. Zurynski, Y. et al. Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. *Orphanet J. Rare Dis.* **12**, 1–9 (2017).
4. Kwon, J. M. & Steiner, R. D. "I'm fine; I'm just waiting for my disease" The new and growing class of presymptomatic patients. *Neurology* **77**, 522–523 (2011).
5. Rode, J. Rare diseases: understanding this public health priority. *EURORDIS Paris Fr.* **5**, 3 (2005).
6. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's online Mendelian inheritance in man (OMIM®). *Nucleic Acids Res* **37**, D793–D796 (2009).
7. Iglesias, A. et al. The usefulness of whole-exome sequencing in routine clinical practice. *Genet. Med.* **16**, 922–931 (2014).
8. ACMG Board of Directors Clinical utility of genetic and genomic services: a position statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **17**, 505–507 (2015).
9. Li, C. et al. Cost-effectiveness of genome-wide sequencing for unexplained developmental disabilities and multiple congenital anomalies. *Genet. Med.* **23**, 451–460 (2021).
10. Manickam, K. et al. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 2029–2037 (2021).
11. Platt, C. D. et al. Efficacy and economics of targeted panel versus whole-exome sequencing in 878 patients with suspected primary immunodeficiency. *J. Allergy Clin. Immunol.* **147**, 723–726 (2021).
12. Arteche-López, A. et al. Towards a change in the diagnostic algorithm of autism spectrum disorders: evidence supporting whole exome sequencing as a first-tier test. *Genes* **12**, 560 (2021).
13. Westerfield, L., Darilek, S. & Van den Veyver, I. B. Counseling challenges with variants of uncertain significance and incidental findings in prenatal genetic screening and diagnosis. *J. Clin. Med.* **3**, 1018–1032 (2014).

14. Crawford, G., Foulds, N., Fenwick, A., Hallowell, N. & Lucassen, A. Genetic medicine and incidental findings: it is more complicated than deciding whether to disclose or not. *Genet. Med.* **15**, 896–899 (2013).
15. Bourchany, A. et al. Reducing diagnostic turnaround times of exome sequencing for families requiring timely diagnoses. *Eur. J. Med. Genet.* **60**, 595–604 (2017).
16. Atwal, P. S. et al. Clinical whole-exome sequencing: are we there yet? *Genet. Med.* **16**, 717–719 (2014).
17. 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**, 405–423 (2015).
18. Stoller, J. K. The challenge of rare diseases. *Chest* **153**, 1309–1314 (2018).
19. Beshyah, S. A. et al. Impact of the COVID-19 Pandemic on Clinical Practice, Medical Education, and Research: An International Survey Impact de la pandémie de COVID-19 sur la pratique clinique, la formation médicale et la recherche: une enquête internationale. *Tunis. Médicale* **98**, 610–618 (2020).
20. Dillon, O. J. et al. Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders. *Eur. J. Hum. Genet.* **26**, 644–651 (2018).
21. Tan, T. Y. et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. *JAMA Pediatr.* **171**, 855–862 (2017).
22. Robinson, P. N. et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. **24**, 340–348 (2014).
23. Stelzer, G. et al. VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics* **17**, 195–206 (2016).
24. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–843 (2015).
25. Zhao, M. et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics Bioinforma.* **2**, lqaa032 (2020).
26. Molina-Ramírez, L. P. et al. Personalised virtual gene panels reduce interpretation workload and maintain diagnostic rates of proband-only clinical exome sequencing for rare disorders. *J. Med. Genet.* **59**, 393–398 (2022).
27. Morley, T. J. et al. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat. Med.* **27**, 1097–1104 (2021).
28. Zhang, Y. et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat. Protoc.* **14**, 3426–3444 (2019).
29. Liu, C. et al. Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Res*. **47**, W566–W570 (2019).
30. Luo, L. et al. PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics* **37**, 1884–1890 (2021).
31. Liu, C. et al. OARD: Open annotations for rare diseases and their phenotypes based on real-world data. *Am. J. Hum. Genet.* **109**, 1591–1604 (2022).
32. Dong, H. et al. Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Med. Inform. Decis. Mak.* **23**, 1–17 (2023).
33. Kim, D. et al. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* **7**, 73729–73740 (2019).
34. Yang, J. et al. Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT. *ArXiv Prepr. ArXiv230806294* (2023).
35. Wapner, R. et al. 8: Whole exome sequencing in the evaluation of fetal structural anomalies: A prospective study of sequential patients. *Am. J. Obstet. Gynecol.* **216**, S5–S6 (2017).
36. Lee, H., Martinez-Agosto, J. A., Rexach, J. & Fogel, B. L. Next generation sequencing in clinical diagnosis. *Lancet Neurol.* **18**, 426 (2019).
37. Halbisen, A. L. & Lu, C. Y. Trends in Availability of Genetic Tests in the United States, 2012–2022. *J. Pers. Med.* **13**, 638 (2023).
38. Greene, S. M., Reid, R. J. & Larson, E. B. Implementing the learning health system: from concept to action. *Ann. Intern. Med.* **157**, 207–210 (2012).
39. Adhikari, A. N. et al. The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nat. Med.* **26**, 1392–1397 (2020).
40. Best, S. et al. Promises, pitfalls and practicalities of prenatal whole exome sequencing. *Prenat. Diagn.* **38**, 10–19 (2018).
41. Dai, P. et al. Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected Mendelian disorders: A systematic review and meta-analysis. *Genet. Med.* **24**, 1618–1629 (2022).
42. Robertson, A. J. et al. Evolution of virtual gene panels over time and implications for genomic data re-analysis. *Genet. Med. Open* 100820 (2023).
43. Wahl, B., Cossy-Gantner, A., Germann, S. & Schwalbe, N. R. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob. Health* **3**, e000798 (2018).
44. Paul, J. L., Leslie, H., Trainer, A. H. & Gaff, C. A theory-informed systematic review of clinicians' genetic testing practices. *Eur. J. Hum. Genet.* **26**, 1401–1416 (2018).
45. Bertier, G., Hétu, M. & Joly, Y. Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. *BMC Med. Genomics* **9**, 1–12 (2016).
46. Lennerz, J. K. et al. Health care infrastructure for financially sustainable clinical genomics. *J. Mol. Diagn.* **18**, 697–706 (2016).
47. Reuter, C. M. et al. Yield of whole exome sequencing in undiagnosed patients facing insurance coverage barriers to genetic testing. *J. Genet. Couns.* **28**, 1107–1118 (2019).
48. Zhang, L. et al. Pathogenic variants identified by whole-exome sequencing in 43 patients with epilepsy. *Hum. Genomics* **14**, 1–8 (2020).
49. Fernández, I. S., Loddenkemper, T., Gaínza-Lein, M., Sheidley, B. R. & Poduri, A. Diagnostic yield of genetic tests in epilepsy: a meta-analysis and cost-effectiveness study. *Neurology* **92**, e418–e428 (2019).
50. Hripcsak, G. et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. in *MEDINFO 2015: eHealth-enabled Health* 574–578 (IOS Press, 2015).
51. Wu, P. et al. Developing and Evaluating Mappings of ICD-10 and ICD-10-CM codes to Phecodes. *BioRxiv* 462077 (2018).
52. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111 (2013).
53. Robinson, P. N. et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
54. Van Aken, B. et al. Assertion detection in clinical notes: Medical language models to the rescue? in 35–40 (2021).
55. Collins, S. A. & Vawdrey, D. K. "Reading Between the Lines" of Flowsheet Data: Nurses' Optional Documentation Associated with Cardiac Arrest Outcomes. *Appl. Nurs. Res. ANR* **25**, 251 (2012).
56. Collins, S. A. et al. Relationship between nursing documentation and patients' mortality. *Am. J. Crit. Care* **22**, 306–313 (2013).
57. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

Acknowledgements

This study was supported by National Human Genome Research Institute grant R01HG012655, and National Center for Advancing Translational Sciences, National Institutes of Health (NIH), through Grant Number UL1TR001873. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We thank Dr.

Natarajan Karthik and the CUIMC DBMI operational team for providing access to the CUIMC clinical data warehouse and clinical narratives.

Author contributions

The author contributions in this study are outlined as follows: F.C. conceptualized project idea and research aims, preprocessed and analyzed data, developed and evaluated models, investigated outcomes, performed data visualization and drafted the initial manuscript. P.A. collected and curated data used for this study, performed label annotation and validation, reviewed and edited manuscript. R.L. provided baseline. K.W., W.C. and C.T. assisted in methodology development, reviewed and edited manuscript. Q.N., K.W., K.S., S.S. and I.C. contributed to the validation on the CHOP dataset. C.W. and C.L. co-supervised research project, conceptualized research idea and scope, managed project administration, oversaw the development and validation of algorithms, funding and grants supported for the study, reviewed and edited manuscript. All authors have read and approved the manuscript.

Competing interests

W.C. is on the Board of Directors of both Prime Medicine and Rallybio. Other authors declare no financial competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01331-1.

Correspondence and requests for materials should be addressed to Chunhua Weng or Cong Liu.