Supplementary Information for "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 Ta[1], Katherine M. Szigety[7], Sarah E. Sheppard[7], Ian M. Campbell[7], Kai Wang[4], Chunhua Weng[1,*], Cong Liu[1,*]

[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]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;

(*: equal-contribution senior corresponding authors; emails: Dr. Cong Liu - cl3720@cumc.columbia.edu; Dr. Chunhua Weng - cw2384@cumc.columbia.edu)

Table of Contents

Supplementary Table 1. Example of patient clinical summaries. The clinical summaries entail key phenotype indicators that genetic experts believe are relevant to the genetic disorders. It is important to note that the phenotype summary was manually recorded in the research database, separate from the original EHR data (used for model training), and was not directly utilized as features for model training.

MRN	Primary Indication	Test	Test Date
xxxxxxx	ACL tear	Gene panel (Ehlers Danlos Syndrome panel)	xx/xx/xxxx
xxxxxxx	Marinesco-Sjögren syndrome and a history of catatonia	Whole Exome	xx/xx/xxxx
xxxxxxx	pulmonary hypertension, severe hydronephrosis, and posterior urethral valves, status post repair	Whole Genome	xx/xx/xxxx

Supplementary Table 2. Clinical summary of patients with test order adjusted to ES/GS.

Primary Indication	Counts
Seizures	59
Autism spectrum disorder	52
Congenital heart defect	13
Developmental delay	10
Multiple birth defects	5

Supplementary Table 3. Features used to train the classification models.

Categories	Feature Names & Description	Input Feature Dimension
Clinical Features (Structured Data)	- *Freq_phecodes*: frequency of each phecodes/phenotype	1,225
	- *Sum_phecodes*: total number of unique phecodes/phenotypes	1
	- *Freq_HPO*: frequency of each HPO-based organ systems of phenotypic abnormality	23
Clinical Features (Unstructured Data)	- *Freq_phecodes_notes*: frequency of each phecodes/phenotypes derived from clinical narratives	418
	- *Sum_phecodes_notes*: total number of unique phecodes/phenotypes derived from clinical narratives	1
	- *Num_notes*: cumulative sum of notes	1
Demographics Characteristics	- *Age*	1
	- Sex assigned at birth time	1
	- *Race self-reported by patients*	1

Supplementary Table 4. Lists of keywords used to filter genetic-related notes, and search for molecular tests ordered by clinicians, as well as exceptions words for panel group were being noted.

Keywords in Note Titles	Keywords in label determination	Exceptions in Panel Group*
"genetic", "letter", "progress", "visit", "progress"	"WES", "WGS", "exome", "genomic", "panel"	blood', 'screen', 'screening','viral','virus', 'pcr','metabolic', 'hepatic','lipid', 'tcell', 't cell', 't-cell', 'iron', 'respiratory', "pathogen", "feeding", "liver", "thyroid", "immunoglobulin", "allergy", "allergen", "celiac", 'antigen', "hepatitis",'vitamin', "chemistry"

Supplementary Data

Please refer Supplementary Data to excel sheet.

Supplementary Data 1. Categories of gene panels in Columbia University Irving Medical Center (CUIMC) initial genetic cohort.

Supplementary Data 2. Experimental results of various combinations of models, sorted by average areas under precision-recall curve (AUPRC) in descending order. Performance metrics were averaged and displayed below.

Supplementary Data 3. Feature Importance of HPO phenotypic abnormalities calculated by Gini impurity, along with OLS estimated coefficients.

Supplementary Data 4. Feature importance pf phecodes calculated by Gini impurity, along with phecode sets and total sum of feature importances by corresponding phecodes.

Supplementary Data 5. Model Performance within subgroups characterized by demographic characteristics.

Supplementary Data 6. Lists of OMOP concept IDs used to identify the "genetic" cohort from the structured OMOP database.