Science Translational Medicine
AAAS

Diagnosis of genetic diseases in seriously ill children by rapid whole genome sequencing and automated phenotyping and interpretation

Supplementary Materials

Authors: Michelle M. Clark, Amber Hildreth, Sergey Batalov, Yan Ding, Shimul Chowdhury, Kelly Watkins, Katarzyna Ellsworth, Brandon Camp, Cyrielle I. Kint, Calum Yacoubian, Lauge Farnaes, Matthew N. Bainbridge’, Curtis Beebe, Joshua J. A. Braun, Margaret Bray, Jeanne Carroll’, Julie A. Cakici, Sara A. Caylor, Christina Clarke, Mitchell P. Creed, Jennifer Friedman, Alison Frith, Richard Gain, Mary Gaughran, Shauna George, Sheldon Gilmer, Joseph Gleeson, Jeremy Gore, Haiying Grunenwald, Raymond L. Hovey, Marie L. Janes, Kejia Lin, Paul D. McDonagh, Kyle McBride, Patrick Mulrooney, Shareef Nahas, Daeheon Oh, Albert Oriol, Laura Puckett, Zia Rady, Martin G. Reese, Julie Ryu’, Lisa Salz, Erica Sanford, Lawrence Stewart, Nathaly Sweeney’, Mari Tokita, Luca Van Der Kraan, Sarah White, Kristen Wigby’, Brett Williams, Terence Wong, Meredith S. Wright, Catherine Yamada, Peter Schols, John Reynders, Kevin Hall, David Dimmock, Narayanan Veeraraghavan, Thomas Defay, Stephen F. Kingsmore.

Materials and Methods

Natural Language Processing and Phenotype Extraction

Extraction of HPO terms from the EHR entailed four steps:

1) Clinical records were exported from the EHR data warehouse, transformed into a compatible format (JSON) and loaded into CLiX ENRICH.

2) A semi-automated query map was created, using HPO terms (and their synonyms) as the input and CLiX queries as the output. The HPO terms were passed through the CLiX encoding engine, resulting in the creation of CLiX post-coordinated SNOMED expressions for each recognized HPO term or synonym. Where matches were not exact, manual review was used to validate the generated CLiX queries. Where there was no match or incorrect matches, new content was added to the Clinithink SNOMED extension and terminology files to ensure appropriate matches between phenotypes in HPO and those in SNOMED-CT. This was an iterative process that resulted in a CLiX query set that covered 60% (7,706) of 12,786 HPO terms (October 9 2017 HPO build).

3) EHR documents containing unstructured data were passed through the CNLP engine. The natural language processing engine read the unstructured text and encoded it in structured format as post- coordinated SNOMED expressions as shown in the example below which corresponds to HP0007973, retinal dysplasia:

243796009|Situation with explicit context|: {408731000|Temporal context|=410511007|Current or past|, 246090004|Associated finding|=95494009|Retinal dysplasia|, 408732007|Subject relationship context|=410604004|Subject of record|, 408729009|Finding context|=410515003|Known present|}

Each SNOMED expression is made up of several parts, including the associated clinical finding, the temporal context, finding context and subject context all contained within the situational wrapper. Capturing fully post-coordinated SNOMED expressions ensures that the correct context of the clinical note is preserved. Some HPO phenotypes cannot be found in SNOMED and can only be represented using post-coordinated expressions, as shown in the following example which is the encoding of HP0008020, progressive cone dystrophy:

243796009|Situation with explicit context|: {408731000|Temporal context|=410511007|Current or past|, 246090004|Associated finding|=(312917007|Cone dystrophy|:263502005|Clinical course|=255314001|Progressive|), 408732007|Subject relationship context|=410604004|Subject of record|, 408729009|Finding context|=410515003|Known present|}

Here, an additional attribute for ‘Clinical Course’ and an appropriate value, ‘Progressive’, are used to further qualify the expression. Clinithink used references to these SNOMED expressions, linked with Boolean logic, to create the queries corresponding to HPO terms. *S*hown below is an example query for HP0008866, failure to thrive secondary to recurrent infections:

c*hp0008866_Failure_to_thrive_secondary_to_recurrent_infections (hp0008866_1_1_Failure_to_thrive_q AND hp0002719_1_1_Infection_Recurrent_q)

q-hp0008866_1_1_Failure_to_thrive_q 243796009|Situation with explicit context|:{408731000|Temporal context|=410511007|Current or past|,246090004|Associated finding|=54840006|Failure to thrive|,408732007|Subject relationship context|=410604004|Subject of record|,408729009|Finding context|=410515003|Known present|}

q-hp0002719_1_1_Infection_Recurrent_q 243796009|Situation with explicit context|:{408731000|Temporal context|=410511007|Current or past|,246090004|Associated finding|=(40733004|Infection|:263502005|Clinical course|=255227004|Recurrent|),408732007|Subject relationship context|=410604004|Subject of record|,408729009|Finding context|=410515003|Known present|}

For an encoding created from the unstructured data to trigger one of these queries, all of the components must be matched. Therefore, the encoding of a clinical note describing an affected sibling will not trigger the query since the encoding is that of family history whilst the query looks for the term in the subject of the record (i.e. the patient). Furthermore, it should be noted that some individual HPO synonyms generate more than one SNOMED expression. Therefore, each query used in the query set is a compound of often more than 2 SNOMED expressions. If we strip out the above constants from each expression (the associated clinical finding, the temporal context, finding context and subject context all contained within the situational wrapper) from each expression in the query set (along with all of the associated SNOMED codes), we can create a more readable format to show linguistically what is included in each query created by Clinithink (data file S1).

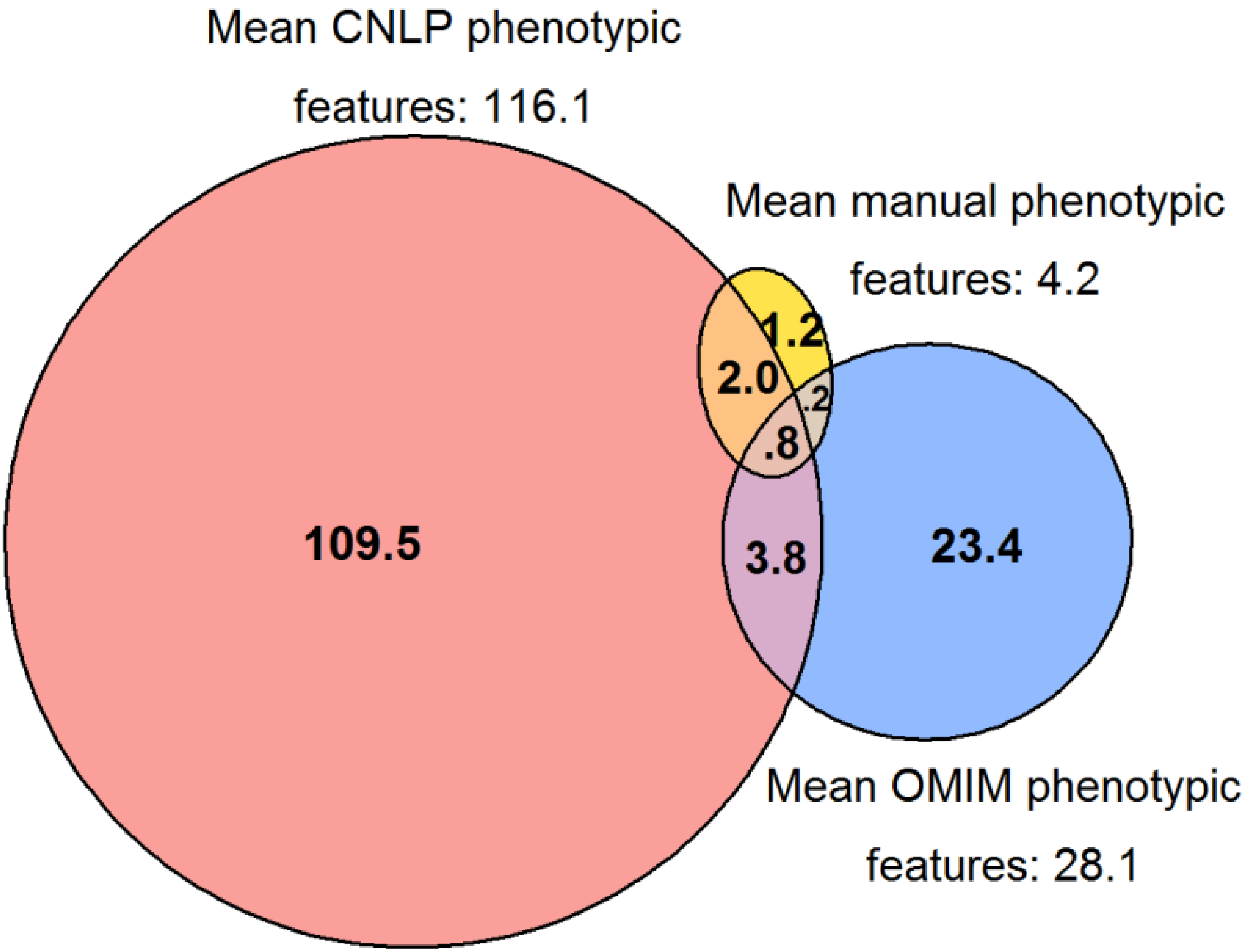

fig. S1. Venn diagram showing overlap of observed and expected patient phenotypic features in 95 children diagnosed with 97 genetic diseases. Phenotypic features identified by expert manual EHR review during interpretation are shown in yellow. Phenotypic features identified by CNLP are shown in red. The expected phenotypic features are derived from the OMIM Clinical Synopsis and are shown in blue. We excluded eight diagnoses that were considered to be incidental findings. Phenotypes extracted by CNLP overlapped expected OMIM phenotypes (mean 4.55, SD 4.62, range 0-32) more than phenotypes that were manually extracted (mean 0.97, SD 1.03, range 0-4).

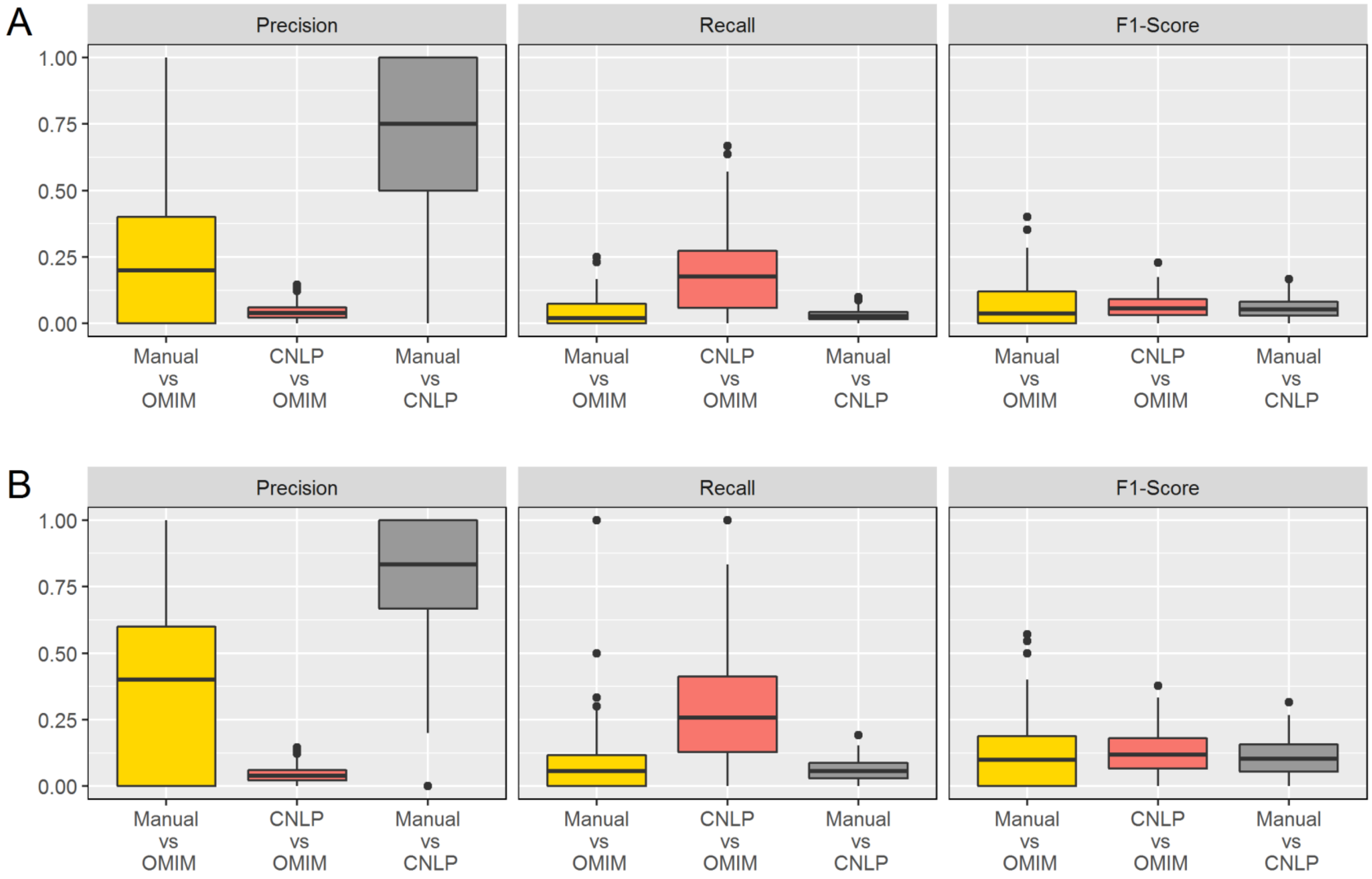

fig. S2. Precision, recall, and F1-score of phenotypic features identified manually, by CNLP, and OMIM. Data are from 101 children with 105 genetic diseases. Precision (PPV) was given by tp/tp+fp, where tp were true positives and fp were false positives. Recall (sensitivity) was given by tp/tp+fn, where fn were false negatives. **A.** Precision and recall calculated based on exact phenotypic feature matches. Manual vs OMIM - Precision: mean 0.25, SD 0.30, range 0-1; Recall: mean 0.04, SD 0.06, range 0-0.25; F_1: mean 0.07, SD 0.09, range 0-0.40. cNLP vs OMIM - Precision: mean 0.04, SD 0.03, range 0-0.15; Recall: mean 0.20, SD 0.16, range 0-0.67; F_1: mean 0.06, SD 0.05, range 0-0.23. Manual vs cNLP - Precision: mean 0.71, SD 0.28, range 0-1; Recall: mean 0.03, SD 0.02, range 0-0.1; F_1: mean 0.06, SD 0.04, range 0-0.17. **B.** Precision and recall calculated allowing for inexact phenotype matches (terms with one degree of hierarchical separation). Manual vs OMIM - Precision: mean 0.4, SD 0.34, range 0-1; Recall: mean 0.09, SD 0.13, range 0-1; F_1: mean 0.13, SD 0.13, range 0-0.57. cNLP vs OMIM - Precision: mean 0.09, SD 0.07, range 0-0.38; Recall: mean 0.29, SD 0.22, range 0-1; F_1: mean 0.12, SD 0.08, range 0-0.38. Manual vs cNLP - Precision: mean 0.79, SD 0.24, range 0-1; Recall: mean 0.06, SD 0.04, range 0-0.19; F_1: mean 0.11, SD 0.07, range 0-0.32.

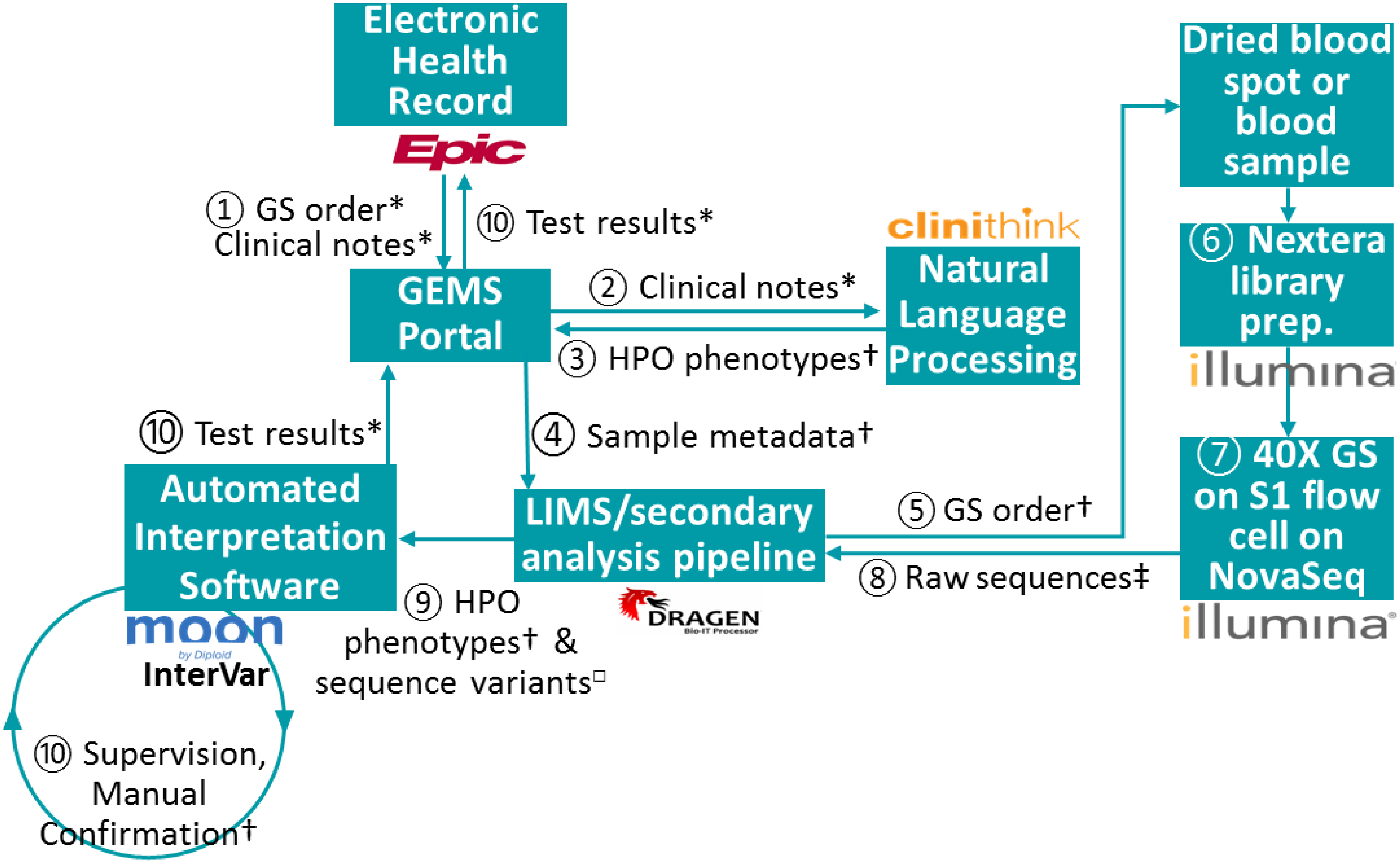

fig. S3. Flow diagram of the software components of the autonomous system for provisional diagnosis of genetic diseases by rapid genome sequencing. Abbreviations: GS: rapid whole genome sequencing; GEMS: Genome management system; HPO: Human Phenotype Ontology; LIMS: Clarity laboratory information management system. Data types were as follows: *: HL7/FHIR; †: JSON; ‡: bcl; □: vcf.

table S1. Comparison of the analytic performance of standard and new library preparation, and standard and rapid genome sequencing in retrospective samples. The standard library preparation and genome sequencing methods were TruSeq PCR-free library preparation and 2 x 100 nt sequencing on a NovaSeq 6000 with S2 flow cell, respectively. The new library preparation and genome sequencing methods were Nextera Flex library preparation and 2 x 100 nt sequencing on a NovaSeq 6000 with S1 flow cell, respectively. The "Median" column is the median of runs R17AA978, R17AA978, R17AA059, and R17AA119. Controls 1 and 2 are mean values for five and fifty-two samples, respectively. Analytic performance of variant calls was assessed in sample NA12878, with comparison to the NIST Genome-in-a-bottle results *(76)*. Note: The NA12878 control run with the S1 flowcell and TruSeq PCR free library (far right) was 2 x 151 nt.

Run	R17AA978	R17AA978	R17AA059	R17AA119	Median	NA12878	Control 1	Control 2	NA12878
NovaSeq 6000 Flowcell	S1				S1		S2	S2	S1
Library Preparation Method	Nextera Flex				Nextera Flex		Nextera Flex	TruSeq PCR-free	
Sample	263	263	6124	3003	263 x 2, 6124, 3003	1 sample	5 samples	52 samples	1 sample
Raw Yield Per Flowcell (Gb)	416	419	404	432	418	435	933	897	537
% Reads Q>30	92.00%	92.07%	92.11%	94.84%	92.09%	90.69%	91.50%	91.70%	91.96%
Trimmed Yield (Gb)	153.9	158.9	165.0	160.7	159.8	148.9	183.3	152.8	164.5
% Reads Mapped	97.9%	97.9%	98.1%	96.9%	97.9%	98.9%	98.6%	98.7%	98.8%
% Duplicate Reads	9.3%	10.4%	7.6%	19.1%	9.8%	8.50%	11.4%	6.3%	17.2%
Mean Insert Size (nt)	386.0	348.0	336.0	274.0	342.0	345.1	315.1	423.4	514.6
Average genome coverage	42.0	43.0	44.4	39.0	42.5	47.5	49.4	43.6	32.9
% OMIM genes with 100% coverage at ≥10X	96.0%	95.7%	94.9%	65.1%	95.3%	95.8%	96.8%	97.7%	98.00%
Variants	4,910,055	4,915,843	4,847,506	4,655,831	4,878,781	4,733,000	4,976,974	4,922,188	4,747,231
Variants passing QC	96.0%	96.1%	96.6%	96.8%	96.3%	96.8%	98.1%	98.4%	98.5%
CD Variants	0.53%	0.53%	0.55%	0.54%	0.53%	0.58%	0.53%	0.53%	0.58%
Indels	17.8%	17.9%	18.0%	17.5%	17.8%	17.5%	18.6%	18.8%	19.4%
CD Homozygous/ Heterozygous Variant Ratio	0.59	0.59	0.57	0.60	0.59	0.60	0.56	0.59	0.60
Ti/Tv ratio	2.02	2.02	2.02	2.03	2.02	2.02	2.02	2.02	2.01
CD Ti/Tv ratio	2.85	2.87	2.88	2.94	2.88	2.81	2.85	2.85	2.82
Analytic Performance									
PPV (SNV)	n.a.	n.a.	n.a.	n.a.	n.a.	99.8%	99.8%	99.9%	99.9%
PPV (indels)	n.a.	n.a.	n.a.	n.a.	n.a.	99.0%	97.0%	99.3%	99.7%
Sensitivity (SNV)	n.a.	n.a.	n.a.	n.a.	n.a.	99.7%	99.6%	99.7%	99.8%
Sensitivity (indels)	n.a.	n.a.	n.a.	n.a.	n.a.	95.5%	96.3%	99.0%	99.4%

Abbreviations: nt: Nucleotides; FC: flowcell; Gb: gigabase; Q: Quality score; OMIM: Online Mendelian Inheritance in Man; QC: Quality Control; CD: Coding Domain; Ti/Tv ratio: ratio of the number of nucleotide transitions to the number of nucleotide transversions; PPV: Positive predictive value; SNV: single nucleotide variants; indels: nucleotide insertion-deletion variants.

table S2. Comparison of the analytic performance of standard and new library preparation and genome sequencing methods in seven matched prospective samples. The standard library preparation and genome sequencing methods were TruSeq PCR-free library preparation and NovaSeq 6000 with S2 flow cell, respectively, with the exception of subjects 7052 and 412, where the library preparation was done with the KAPA Hyper kit. The new library preparation and genome sequencing methods were Nextera Flex library preparation and NovaSeq 6000 with S1 flow cell, respectively.

Run	**R18AA202**	**Std.**	**R18AA218**	**Std.**	**R18AA922**	**Std**	**R18AB113**	**Std**	**R18AB229**	**Std**	**R18AB352**	**Std**	**R18AB672**	**Std**
Subject	6194 (Prospective)		290 (Prospective)		352 (Prospective)		362 (Prospective)		374 (Prospective)		7052 (Prospective)		412 (Prospective)	
Library Prep Method	Nextera	TruSeq	Nextera	TruSeq	Nextera	TruSeq	Nextera	TruSeq	Nextera	KAPA Hyper	Nextera	KAPA Hyper	Nextera	KAPA Hyper
Flow cell	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Raw Yield Per Flow cell (Gb)	389.9	945.4	381.8	946	365.3	869.9	398.3	440.7	420.8	899.1	383.4	860.2	422.1	908.2
Reads Q>=30	90.90%	93.70%	91.30%	93.10%	89.80%	90.70%	92.20%	90.00%	93.30%	91.60%	90.10%	90.10%	92.90%	91.60%
% Cluster passing filter, L1/L2	69.8/82.9	82.1/82.0	73.9/75.6	82.2/82.0	73.8/69.3	75.5/75.5	78.9/77.1	36.7/39.9	83.0/81.8	78.3/77.8	75.49/74.7	75.2/74.1	83.1/82.3	78.9/78.8
% Error rate (ΦX174), R1/R2	0.19/0.42	0.27/0.47	0.25/0.65	0.27/0.37	0.25/0.45	0.31/0.37	0.20/0.36	0.33/0.41	0.20/0.40	0.25/0.35	0.26/0.50	0.31/0.36	0.22/0.32	0.28/0.29
Trimmed Yield (Gb)	174.1	172.3	168.6	218.2	141	144.2	164.3	148.4	185.5	267.8	156.4	138	183.4	203
Reads Mapped	97.70%	98.60%	97.30%	98.30%	97.20%	98.60%	97.40%	98.50%	98.00%	98.50%	97.30%	98.30%	98.60%	98.60%
Duplicate Reads	11.50%	6.50%	11.60%	7.30%	8.90%	9.20%	9.90%	3.90%	11.70%	14.60%	8.30%	9.40%	14.00%	13.40%
Mean Insert Size (nt)	361.2	405.8	223.7	430	373.4	419.8	369	410	266.9	423.8	371.4	428.4	338.1	416.2
Average genome coverage	44.8	48.4	54	60.4	39.1	39.3	43.1	42.8	48	68.4	41.6	37.3	47.6	50.9
%OMIM genes w. >10X x 100% nt	95.80%	97.90%	93.30%	98.20%	95.80%	97.80%	95.70%	96.60%	96.00%	98.40%	95.20%	97.80%	96.90%	98.20%
Variants	4,687,590	4,881,456	4,776,648	5,016,422	4,765,467	4,934,554	4,719,091	4,917,044	4,758,713	5,001,708	4,821,433	4,981,748	4,958,194	4,965,915
Variants passing QC	96.90%	98.30%	97.00%	98.20%	97.00%	98.60%	97.00%	98.20%	98.10%	98.00%	98.10%	98.60%	98.10%	98.20%
CD Variants	0.57%	0.52%	0.57%	0.53%	0.54%	0.56%	0.55%	0.54%	0.55%	0.53%	0.56%	0.53%	0.56%	0.53%
Indels	18.20%	18.90%	18.00%	18.90%	18.00%	18.60%	17.70%	18.50%	19.60%	18.80%	17.60%	18.50%	18.70%	18.90%
Ti/Tv ratio	2.02	2.02	2.03	2.03	2.02	2.03	2.02	2.01	2.01	2.01	2.03	2.02	2.01	2.02

Abbreviations: L: lane; R: read; nt: Nucleotides; Gb: gigabase; Q: Quality score; OMIM: Online Mendelian Inheritance in Man; QC: Quality Control; CD: Coding Domain; Ti/Tv ratio: ratio of the number of nucleotide transitions to the number of nucleotide transversions.

table S3: Characteristics of sixteen children with genetic diseases used to train CNLP.

Family	S, D, T	rWES or rWGS	Disease	Affected Gene	OMIM ID	Inheritance	de novo or inherited	Variant 1 (V1)	Variant 2 (V2)	V1 P/ LP	V2 P/ LP	Age at enrollment (days)	Sex	Consanguinity
6007	T	rWGS	EIEE9	*PCDH19*	300088	AD	DN	Xq22del				423	F	No
6008	S	rWGS	Glioblastoma	*BRCA1*	604370	AD	n.d.	c.5159G>A, p.Arg1720Gln				4563	F	No
6012	S	rWGS	Coffin-Siris syndrome 1	*ARID1B*	135900	AD	DN	c.3096_3100delCAAAG; p.Lys1033ArgfsTer32				231	F	No
6014	S	rWGS	Nemaline myopathy 2	*NEB*	256030	AR	n.d.	c.19262+1G>A	c.2416-1G>C			35	M	No
6024	T	rWGS	Hypophosphatemic rickets, X-linked dominant	*PHEX*	307800	XLD	I	c.1604C>T,p.Thr535Met				137	M	No
6026	T	rWGS	Alagille syndrome 1	20p12.2 del	118450	AD	DN	Chr20:10471400-13459331del				80	M	U
6030	T	rWGS	Neurofibromatosis 1; Left ventricular noncompaction 10	*NF1 & MYBPC3*	162200, 615396	AD, AD	DN, I	c.5118delT, p.Val1707PhefsTer	c.3184delG p.Val1062LeufsTer13	LP	LP	227	M	No
6031	T	rWGS	Catecholaminergic polymorphicVentricular tachycardia 1	*RYR2*	604772	AD	DN	c.1646C>T; p.Ala549Val				6087	F	No
6037	T	rWGS	Neonatal cholestasis; Extrahepatic biliary atresia	none	none	n.a.	n.a.	n.a.				60	M	U
6041	T	rWGS	EIEE7	*KCNQ2*	613720	AD	DN	c.875T>C; p.Leu292Pro				2	F	No
6044	S	rWGS	Pleuropulmonary blastoma	*DICER*	601200	AD	n.d.	c.2771T>G; p.Leu924*				564	M	U
6045	S	rWGS	Medulloblastoma	none	none	n.a.	n.a.	n.a.				5475	M	U
6051	S	rWGS	Glioma	none	none	n.a.	n.a.	n.a.				2555	M	U
6052	T	rWGS	MECRCN	*TANGO2*	616878	AR	I	c.605+1G>A	33 kb del *TANGO2* exons 3-9			898	F	U
6066	D	rWGS	Neonatal cholestasis; Cleft lip and palate	none	none	n.a.	n.a.	n.a.				60	F	U
6117	D	rWGS	Neonatal cholestasis	none	none	n.a.	n.a.	n.a.				60	F	U

Abbreviations: EIEE: Early Infantile Epileptic Encephalopathy; AD: Autosomal Dominant; DN: de novo; P: Pathogenic; LP: Likely Pathogenic; M: Male; F: Female; S: Singleton; D: Duo; T: Trio; I: Inherited; XLD: X-linked dominant; MECRN: Metabolic encephalomyopathic crises, recurrent, with rhabdomyolysis, cardiac arrhythmias, and neurodegeneration; U: undetermined; OMIM: Online Mendelian Inheritance in Man.

table S4. Precision and recall of phenotypic features extracted by CNLP from EHRs in ten children with genetic diseases (see tables S5-S14)**.** Precision=tp/tp+fp. Recall=tp/tp+fn**.**

Family	S or T	rWES or rWGS	Disease	Affected Gene	OMIM ID	Inher-itance	*de novo* or inher-ited	Variant 1 (V1)	Variant 2 (V2)	V1 P/ LP	V2 P/ LP	Age at enroll-ment (days)	Sex	Consan-guinity	CNLP Features	CNLP Precision	CNLP Recall	OMIM CF detected by CNLP
201	T	rWES	Prader Willi Syndrome	15q11-q13 del	176270	AD	DN	Chr15:23684685-26108259del				3	♀	U	26	0.88	n.d.	3%
205	T	rWGS	Dursun Syndrome	*G6CP3*	612541	AR	I	c.207dupC, p.Ile70HisfsTer17	c.199_218+1delCTCAACCTCATCTTCAAGTGG	P	P	2	♂	No	96	0.80	n.d.	15%
213	S	rWGS	Visceral Heterotaxy 5	*NODAL*	270100	AD	I	c.778G>A, p.Gly260Arg				3	♂	U	95	0.67	0.91	56%
233	T	rWGS	Tuberous Sclerosis 1	*TSC1*	191100	AD	DN	c.1498C>T, p.Arg500Ter				3	♀	No	158	0.51	0.91	14%
243	T	rWGS	Pyridoxine dependent seizures	*ALDH7A1*	266100	AR	I	c.328C>T, p.Arg110Ter	c.1279G>C, p.Glu427Gln			7	♂	No	85	0.82	0.93	21%
6094	T	rWGS	Argininosuccinic Aciduria	*ASL*	207900	AR	I	c.706C>T, p.Arg236Trp	c.706C>T, p.Arg236Trp	P	P	7	♀	Yes	90	0.83		11%
6098	T	rWGS	Gaucher disease	*GBA*	230800	AR	I	c.1503C>G, p.Asn501Lys	c.1448T>C , p.Leu483Pro			214	♀	No	96	0.9		21%
6108	T	rWGS	Tuberous Sclerosis 2	*TSC2*	613254	AD	DN	c.935_936delTC, p.Leu312GlnfsTer25				3	♂	No	83	0.76		5%
7003	T	rWGS	EIEE6	*SCN1A*	607208	AD	DN	c.5555T>C, p.Met1852Thr				424	♂	U	44	0.84	0.93	25%
7004	T	rWGS	Hypertrophic cardiomyopathy type 1	*MYH7*	192600	AD	I	c.746G>A, p.Arg249Gln				5171	♂	U	71	0.94	0.96	44%
Mean															86.7	0.80	0.93	22%
		Standard Deviation													32.8	0.13	0.02	0.17

Abbreviations: EIEE: Early Infantile Epileptic Encephalopathy; AD: Autosomal Dominant; AR: Autosomal Recessive; DN: de novo; P: Pathogenic; LP: Likely Pathogenic; S: Singleton; T: Trio; I: Inherited; U: undetermined; OMIM: Online Mendelian Inheritance in Man; CF: Clinical Feature.

table S5. Precision of and recall 26 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 201. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S6. Precision and recall of 96 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 205. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S7. Precision and recall of 95 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 213. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S8. Precision and recall of 158 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 233. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. TS: Tuberous Sclerosis. (Available as Excel)

table S9. Precision and recall of 85 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 243. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S10. Precision and recall of 90 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 6094. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S11. Precision and recall of 96 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 6098. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S12. Precision and recall of 83 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 6108. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S13. Precision & recall of 44 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 7003. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S14. Precision & recall of 71 phenotypic features extracted and proportion of OMIM clinical features detected by CNLP from the EHR of patient 7004. Abbreviations: CF: clinical feature. Citation: Citation from the EHR. TP: true positive. OMIM: Online Mendelian Inheritance in Man. (Available as Excel)

table S15. The test cohort diagnosed manually by rapid genome or exome sequencing and interpreted retrospectively with an autonomous system. The cohort comprised 101 affected children with 105 genetic diseases. Phenotypic features identified by expert manual review of the EHR are shown in data file S2. Phenotypic features extracted by CNLP are shown in data file S3. Where a patient had two molecular diagnoses, they are indicated as (1) and (2). Abbreviations**:** Concordant: manual and autonomous diagnoses agree. Abbreviations: S: singleton; D: duo; T: trio; WGS: rapid whole genome sequencing; WES: rapid whole exome sequencing; AD: autosomal dominant; AR: autosomal recessive; XLD: X-linked dominant; XLR: X-linked recessive; DN: de novo; I: Inherited; n.d.: not done; SOM: somatic; P: Pathogenic; LP: Likely Pathogenic; n.k.: not known; F: female; M: male; Inc.: Incidental finding; ACDMPV: Alveolar capillary dysplasia with misalignment of pulmonary veins; EIEE: Early infantile epileptic encephalopathy; CLIFAHDD: Congenital contractures of the limbs and face, hypotonia, and developmental delay; CHARGE: coloboma, heart anomaly, choanal atresia, retardation, genital and ear anomalies; IPEX: Immunodysregulation, polyendocrinopathy, and enteropathy, X-linked; SMARD1: spinal muscular atrophy with respiratory distress 1. *Alternatively, patient 6059 could be considered diagnosed with OMIM#114480 (dominant breast cancer susceptibility). (Available as Excel)

table S16. Variant characteristics in rapid genome or exome sequencing of the 101 children with 105 genetic diseases. Abbreviations: SNV: single nucleotide variant; V: variants; CD: coding domain; Het: heterozygous; Hom: homozygous; P: pathogenic; LP: likely pathogenic; x̄: Mean. (Available as Excel)

table S17. Number of nucleotide variants shortlisted by MOON and rank of the causal variant in MOON in 84 children with 86 genetic diseases. Abbreviations: S: Singleton; T: Trio; D: Duo. (Available as Excel)

table S18. Number of structural variants shortlisted by MOON and rank of the causal variant in MOON in 11 children with genetic diseases. All samples were run as singletons.

Family	rWES / rWGS	# SV calls in gVCF	# SV shortlisted by MOON	Causal SV rank in MOON
201	rWES	6	2	1
259	rWES	16	9	1
286	rWES	7	3	1
319	rWES	12	4	1
217	rWGS	21	8	1
223	rWGS	16	9	5
302	rWGS	22	17	13
6140	rWGS	11	8	1
6146	rWGS	23	15	9
6164	rWGS	25	15	12
7023	rWGS	17	12	12
Mean, rWES		10.3	4.5	Median rWGS, rWES 1.0
Mean, rWGS		19.3	12.0	

Abbreviations: gVCF: Genomic variant call file; rWES: rapid whole exome sequencing; rWGS: rapid whole genome sequencing; SV: structural variant.

table S19. Summary statistics of provisional diagnoses reported for rapid clinical genome sequencing. Total probands refers to children tested through January 2019.

Total Probands	Provisional Reports Returned	Mean Time to Provisional Report (Sample Accession to Preliminary Results Communicated), Days
684	114 (16.7%)	3.6

data file S1. Mapping of HPO terms to SNOMED expressions. (Available as Excel)

data file S2. Phenotypic features of 101 children with genetic diseases that were manually extracted by experts from the EHR at time of rWGS or rWES interpretation. (Available as Excel)

data file S3. Phenotypic features of 101 children with genetic diseases that were automatically extracted from the EHR by CNLP at time of enrollment. (Available as Excel)