

Scaling genomic reanalysis to unlock diagnoses and transform rare disease care

Shira Rockowitz,^{1,2,3,4,5} Wanqing Shao,¹ Courtney French,¹ Tina K. Truong,^{1,2,5} Jacob Hagen,^{1,3} Rylee McGonigle,¹ Alexa Geltzeiler,³ Beth Sheidley,⁶ Lacey Smith,⁶ Alissa M. D’Gama,^{1,3,4,7} Mira Irons,^{1,2,3,4,8} Janet Chou,^{1,3,4,9} Joan Stoler,^{1,2,3,4} Amy Kritzer,^{2,3,4} Lance Rodan,^{4,5,6} Akiko Shimamura,^{1,3,4,8,10,11} Olaf Bodamer,^{1,2,3,4,5,8} Stephanie Sacharow,^{2,3,4} Janet S. Soul,^{1,4,6} Siddharth Srivastava,^{1,4,6,12} Amy Roberts Kennedy,^{1,2,3,4,13} Aya Abu-El-Haija,^{2,3,4} Abbe Lai,⁶ Heather Olson,^{4,5,6} Jane Juusola,¹⁴ Erin Ryan,¹⁴ Bethany Friedman,¹⁴ Anupama Singh,¹ Cliff Li,¹ Rittika Mallik,¹ Gwendolyn Strickland,^{1,2} Gillian Prinzing,⁶ Alisa Mo,^{4,6} Anne O’Donnell-Luria,^{1,2,3,4,5,8} Jeff Bolton,^{4,6} Philip M. Boone,^{1,2,3} William Brucker,^{3,4} Michael Duyzend,^{2,3,4} Sonal Mahida,⁶ David T. Miller,^{2,3,4} Jacklyn Omorodion,^{2,3,4} Jeanette Petit,² Jonathan Picker,^{3,4,15} Annapurna Poduri,^{1,4,6} Colleen Carlston,^{2,3,8} Monica H. Wojcik,^{1,2,3,4,5,7,8} Piotr Sliz,^{1,3,4,5,16,17} and Wendy K. Chung^{1,2,3,4,8,18,*}

Summary

Genomic reanalysis can identify causative variants for rare diseases as patient phenotypes evolve and gene-disease knowledge expands. Despite its diagnostic value, routine reanalysis is limited by clinician capacity, lack of patient follow-up, data silos, cost, and lack of availability of clinical data to testing laboratories that are not obligated to conduct reanalysis. The Children’s Rare Disease Collaborative at Boston Children’s Hospital (BCH) has integrated genomic and phenotypic data from over 15,500 patients into a clinician-facing platform. Leveraging this infrastructure, we developed a Proactive Genomic Reanalysis (PGR) workflow at BCH for clinical sequencing data that is centralized, semi-automated, and clinically integrated. Here, we report initial results and outline required resources and transferable insights applicable to other healthcare settings. Initial pilot implementation, applied to a subset of clinical sequencing patients’ data, revealed practical challenges, notably clinician turnover and patient recontact difficulties. Of 42 patients’ candidate variants discovered by the PGR bioinformatics pipeline and returned to treating clinicians, 33 were determined to have a high suspicion of disease causality and an additional 3 were determined to be candidate variant of uncertain significance. A process to generate reports and return results to patients was initiated when applicable. Although the initial pilot implementation was limited, the PGR bioinformatics pipeline is designed to be utilized iteratively, making reanalysis a continuing process. This work highlights the feasibility and impact of centralized PGR processes and the potential for healthcare institutions to scale genomic reanalysis.

Introduction

Genomic medicine is enhancing patient care, particularly since methodologies such as exome sequencing (ES) and genome sequencing (GS) have become a part of standard clinical practice and cost-effective diagnostic tools.^{1–5} Precision medicine interventions depend on accurate genetic diagnoses,⁶ but a substantial portion (55%–67%) of patients with a presumed genetic condition remain without a definitive diagnosis despite ES/GS.⁷ This diagnostic gap is often due to incomplete understanding of genomic variation responsible for the disease, challenges regarding the detection of certain variant types due to the nature of

short read ES/GS modalities as well as bioinformatics limitations, limitations in phenotyping younger patients whose disease manifestation is still evolving at the time of testing, and the static nature of traditional laboratory workflows.⁸

Reanalysis of genomic data has emerged as a powerful strategy to address this gap, diagnosing between 2% and 31% of previously non-diagnostic cases after 1–5 years.^{9–23} Clinician involvement in reanalysis has a large (11.9%) diagnostic impact, improving the interpretation of the clinical context for genes considered to be associated with the patient phenotype.²⁴ Although it is well known that reanalysis increases diagnoses, current approaches

¹Children’s Rare Disease Collaborative, Boston Children’s Hospital, Boston, MA, USA; ²Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA, USA; ³Department of Pediatrics, Boston Children’s Hospital, Boston, MA, USA; ⁴Harvard Medical School, Boston, MA, USA; ⁵The Manton Center for Orphan Disease Research, Boston Children’s Hospital, Boston, MA, USA; ⁶Boston Children’s Hospital, Department of Neurology, Boston, MA, USA; ⁷Boston Children’s Hospital, Division of Newborn Medicine, Boston, MA, USA; ⁸Broad Institute of MIT and Harvard, Cambridge, MA, USA; ⁹Boston Children’s Hospital, Division of Immunology, Boston, MA, USA; ¹⁰Department of Hematology and Oncology, Boston Children’s Hospital, Boston, MA, USA; ¹¹Dana Farber Cancer Institute, Boston, MA, USA; ¹²Rosamund Stone Zander Translational Neuroscience Center, Boston Children’s Hospital, Boston, MA, USA; ¹³Department of Cardiology, Boston Children’s Hospital, Boston, MA, USA; ¹⁴GeneDx, LLC, Gaithersburg, MD, USA; ¹⁵Division of Developmental Medicine, Boston Children’s Hospital, Boston, MA, USA; ¹⁶Division of Molecular Medicine, Boston Children’s Hospital, Boston, MA, USA; ¹⁷Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA

¹⁸Lead contact

*Correspondence: wendy.chung@childrens.harvard.edu

<https://doi.org/10.1016/j.xhgg.2026.100582>.

© 2026 The Authors. Published by Elsevier Inc. on behalf of American Society of Human Genetics.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



to clinical testing remain largely static based upon the payment model where reanalysis is not uniformly covered by health insurance but may be offered free of cost according to individual laboratory policies.^{25,26} Genomic data are analyzed initially at the diagnostic laboratory; subsequently the locus of control is with the clinician to initiate lab reanalysis since they can best assess the patient's clinical state and relevance of receiving a diagnosis.²⁷ Raw genomic data, which could be reanalyzed by clinicians as information is amassed, typically stays with reference laboratories, which generally do not initiate reanalysis unless prompted by a request from the treating clinician. Even reclassification of variants that were already included in the clinical report is not always performed due to implementation complexities.²⁸ Proactive reanalysis, defined here as institution-initiated and systematically applied reanalysis, as opposed to clinician-initiated reanalysis, has been shown to have a large additional impact on initially unsolved ES.²¹ Although clinicians perceive automated or semi-automated reanalysis, referring to computational approaches that reduce manual effort, as imparting many clinical and workflow benefits^{29–31} several challenges impede adoption including: data are separated, with changing clinical information residing at hospitals with clinicians while genomic data generally reside at reference laboratories; there is little documented about potential insurance reimbursement for recurrent reanalysis²⁵; and there are no turnkey commercial or research tools available to make the process of reanalysis simple, automatic and scalable.

To date, systematic reanalysis efforts have been primarily focused on cohorts of research participants.^{9,11,15} Many of these efforts focused on specific diseases and performed a single bulk reanalysis rather than an ongoing iterative process that reaches individual patients with unique phenotypes. Various approaches to streamline algorithms involved in genomic reanalysis have been reported, but manual processes and large review groups are still core to the workflow.^{9,10} Although there is a large body of literature regarding the utility of reanalysis in the research space, automatic clinical reanalysis—re-evaluation of existing genomic data within a clinical framework—is rarely practiced. Only one commercial or academic laboratory out of 28 surveyed by Frees et al., practiced any automated reanalysis.²⁵ Most (86%) of the laboratories solely relied on clinician-initiated reanalysis and, as a result, clinical reanalysis uptake has been low (<10%).^{25,29} No reports have described the systematic and standard application of reanalysis to clinical data as part of clinical processes after initial sequencing, nor closed the loop regarding the impact of the evolving clinical findings for patients. While clinician-initiated laboratory reanalysis and research-based reanalysis have contributed to diagnoses at Boston Children's Hospital (BCH), these approaches remain episodic, resource-intensive, and dependent on individual clinician initiative or research enrollment. In this manuscript, these existing reanalysis pathways are presented to establish the

institutional baseline and unmet need that motivated the development of a centralized Proactive Genomic Reanalysis (PGR) workflow.

Here, we describe the development and deployment of a PGR workflow at BCH for clinical sequencing data that is centralized, semi-automated, and clinically integrated. We were able to leverage our established process for the return of raw genomic data to the institution after clinical ES/GS testing and research technologies utilized by the Children's Rare Disease Collaborative (CRDC) to enable clinicians to use BCH analytic tools to review and interpret prioritized candidate variants for their patients. We describe the development of the PGR bioinformatics pipeline to semi-automatically identify variants associated with patients' evolving phenotypes at scale, the processes developed to confirm and return those findings to patients, and complexities discovered along the way. We outline necessary resources and transferable insights applicable in other healthcare settings. PGR is designed to empower clinicians, as those most familiar with the patient's evolving phenotype, to regularly and systematically review possibly relevant variants for all patients. By bridging the gap between laboratory infrastructure and frontline clinical care, this model facilitates a dynamic and responsive approach to genomic medicine. Our approach demonstrates how a centralized, dynamic institution-initiated reanalysis system can be deployed at hospitals that service many patients with genetic conditions to democratize access, enhance clinical utility, and serve as a scalable model for genomic medicine.

Subjects and methods

Underlying infrastructure and agreements

The implementation of a collaborative, clinically integrated PGR workflow at BCH was built on previously established pipelines for genomic data return and data processing and analysis. In early 2018, raw sequencing data from Claritas Genomics was returned to the hospital en masse and included 576 probands and 493 relatives. Beginning in 2019, as per the recommendation from the BCH Genomic Consent Taskforce, all raw sequencing data and metadata from the current clinical sequencing provider, GeneDx, has been returned to BCH via establishment of a robust data transfer and integration framework (Figure S1A). This dataset comprises a total of 6,279 probands and 10,825 relatives through May 1, 2025 (47 probands and 75 relatives had sequencing performed at both Claritas Genomics and GeneDx). Leveraging the previously published CRDC genomic analysis infrastructure^{32,33} enabled rapid deployment of existing methodologies for collecting Human Phenotype Ontology (HPO) terms from the electronic health record (EHR) via natural language processing (NLP) on a monthly basis, harmonizing variant calling data via secondary analysis pipeline DRAGEN, and loading data into established BCH analytic tools for clinicians (Figure S1B). The BCH analytic tools integrate dynamic phenotypic and genomic data and include a suite of genomic analysis tools hosted by GeneDx

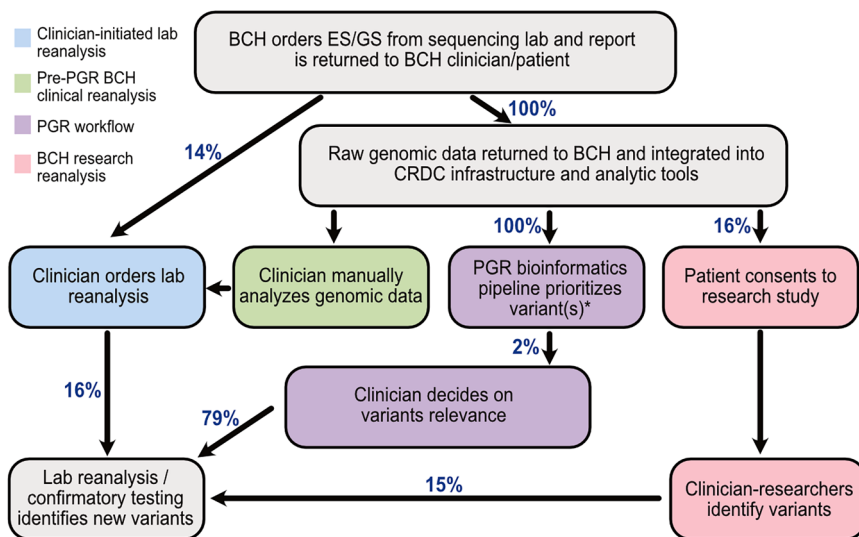


Figure 1. Overview of reanalysis workflows at BCH

Clinicians at BCH order ES/GS from the sequencing lab and the initial report is returned to the patient. In 13.6% of patients, a clinical reanalysis from the lab is ordered (blue). A total of 16% of patients with such a reanalysis ordered have findings. The raw data for all patients and relatives included in clinical sequencing are returned from the sequencing laboratories (Claritas Genomics and GeneDx) to BCH and integrated into the CRDC infrastructure and institutional analytic tools. We do not have an estimate of how often clinicians review genomic data (green) as this activity is not routinely tracked but 21% of clinician respondents (7/33) who took the seqr training at BCH (a group highly biased to interest in genomics) said they had previously looked at a patient's raw genomic

data. This activity reflects existing, decentralized clinician expertise and access to analytic infrastructure at BCH; while not a scalable or systematic reanalysis workflow, it provides essential clinical context and human interpretation that inform and complement the centralized PGR process. The PGR bioinformatics pipeline is applied to all data from patients with no reported variants and will soon be expanded to all patients with data returned to BCH (purple), identifying variants in 2% (42/2,144) of patients. A total of 79% (33/42) variants identified were considered by clinicians to be diagnostic. *Limited variants were discussed in this article due to the specific nature of the initial pilot implementation of the VS-NN algorithm. We expect that increasing the sensitivity of the PGR bioinformatics pipeline will vastly increase the percentage of patients with identified variants. A total of 16% of patients consent to research studies (e.g., CRDC studies such as the Manton Center³⁶) (pink); these research efforts are not evaluated as part of this analysis and are shown solely to illustrate the broader institutional genomic research environment in which PGR operates. French et al.³³ report that 15% of the patients participating in the CRDC have variants clinically reported.

(distinct from the GeneDx tools used for internal clinical analysis and hereafter called “the GeneDx research platform”), an instance of seqr,³⁴ which is hosted at BCH, and REDCap.^{34,35}

Genomic data analysis training and quality improvement survey

To facilitate clinician access to their patients' data for reanalysis (Figure 1, green box), a training session was held to introduce the process. Following this training session, clinicians who attended were granted access to patient genomic data and invited to attend monthly office hours to bring their cases and ask for help. An offer was also extended via email to help with any reanalysis cases. Following these interventions, a survey was sent to those who had signed up for the training session as well as other genetics and neurology department members to assess interest in accessing their patients' data, comfort level doing so, frequency of ordering genetic testing, previous experience with genomic analysis, and any barriers or concerns with the process (Figure S2).

Transfer of genomic data and metadata from Claritas

Genomic data (ES hg19 BAM files) were transferred as a one-time data deposit via Amazon Web Services (AWS) to BCH in January of 2018. Metadata were contained in an excel spreadsheet and included Claritas accession to identifier information, test ordered (bone marrow failure panel, neurology panel, ES) for all individuals ($n = 1,069$) and HPO terms for some patients (42/576 probands had HPO terms annotated in the metadata). Variants reported by Claritas were not returned to BCH in a structured format.

Transfer of genomic data and metadata from GeneDx

Data from 6,279 probands and 10,825 relatives were returned to BCH by GeneDx between April 11, 2019, and May 1, 2025. Genomic data (ES and GS hg19 CRAMs and VCFs) were delivered to BCH via AWS as an ongoing process as sequencing was completed. Metadata returned via REDCap included predicted genetic similarity to reference groups determined by principal-component analysis, mean coverage, sex, kinship scores, GeneDx case identifiers, HPO terms, percent 10× coverage, heterozygosity/homozygosity ratios (including X chromosome-specific metrics), and which family members were sequenced. For each individual, accession numbers, file locations, and file names were recorded. Variant-level data encompassed gene names, predicted effects, HGVS cDNA and protein annotations, clinical interpretations, zygosity, inheritance patterns, and genomic coordinates (hg19), along with variant-specific zygosity across all sequenced family members. A custom extract-transform-load process was developed, deployed, and iteratively refined to translate case-level metadata, where each additional family member or reanalysis review generated a case, into an individual-centric format compatible with the BCH database architecture.

Analysis of variants reported by GeneDx

Variants were classified by GeneDx in accordance with the American College of Medical Genetics and Genomics (ACMG) guidelines for variant interpretation from Richards et al.³⁷: causative pathogenic or likely pathogenic (P/LP), possibly associated P/LP, possibly associated variant of uncertain significance (VUS), and candidate VUS. Causative variants refer to P/LP variants in genes associated with human disease where the disease phenotype aligns with the patient's reported clinical findings. Possibly

Table 1. Baseline clinical variant findings, clinician-initiated reanalysis, and PGR pilot results

	Patients with annotated variants in original analysis	Patients reanalyzed by clinical sequencing provider	Patients with changes to their reports after reanalysis by clinical sequencing provider	Patients with variants upgraded through clinician-initiated clinical laboratory reanalysis	Patients with variants downgraded through clinician-initiated clinical laboratory reanalysis	Patients with variants downgraded and upgraded through clinician-initiated clinical laboratory reanalysis	No change to variant after clinician-initiated clinical laboratory reanalysis	Patients reanalyzed by PGR	Patients with variants from PGR	Patients with variants from PGR, high suspicion of disease causality or candidate VUS
Causative P/LP	1,303	34	5	3	2	0	29	N/A	N/A	N/A
Possibly associated P/LP	765	56	10	2	7	1	46	N/A	N/A	N/A
Possibly associated VUS	1,050	92	17	16	1	0	75	N/A	N/A	N/A
Candidate VUS	250	23	6	5	0	1	17	N/A	N/A	N/A
No reported variants (2023+)	1,572	19	5	5	0	0	14	606	4	4
No variant reported (2021–2022)	866	146	20	20	0	0	126	825	12	9
No variant reported (2019–2020)	239	70	13	13	0	0	57	208	4	3
Unknown (Claritas)	529	N/D	N/D	N/D	N/D	N/D	N/D	505	22	20

This table shows the different types of patients with data input into the pilot implementation of the PGR pipeline. N/D, not determined; N/A, not applied.

associated variants refer to a VUS or P/LP variant in genes that may be associated with the patient’s phenotype but are not clearly diagnostic. Some examples of possibly associated variants include a single P/LP variant in an autosomal recessive (AR) gene where a second variant was not identified, variants classified as VUS in genes that are a moderate fit for the phenotype, and single P/LP variants in a gene that has both AR and autosomal dominant (AD) inheritance but the phenotype is not exactly aligned with the known AD phenotype. Candidate variants refer to a variant in a gene that needs additional evidence to establish whether there is a relationship between the gene and human disease. A total of 234 probands was removed from the analysis because they were not easily mappable to the above categories including instances where the variant was annotated as previously reported ($n = 183$) and 51 probands who were removed due to unclear variant reporting in reanalysis, e.g., a variant appearing in a reanalysis as previously reported that had not been included in the initial analysis. Following these exclusions, 6,045 probands were included in the subsequent analyses. Counts of probands with reported variants are included in [Table 1](#), labeled based on the most impactful variant hierarchically. Probands with no reported variants or variants that do not belong to the above categories, including patients who only had ACMG secondary findings reported, were labeled as “no reported variants.”

EHR metadata

Date of birth, sex, self-reported race and ethnicity, zip code, and type of health insurance were pulled from the EHR for each pro-

band on May 27, 2025. Age at sequencing was calculated as the year of sequencing minus the date of birth. Family structure was calculated as proband only if only the proband had data returned by GeneDx, duo if the proband and one parent had data returned by GeneDx, and trio if the proband and both parents had data returned by GeneDx. Revised combined race and ethnicity category was calculated from self-reported race and ethnicity as defined in Chopra and co-workers.³⁸ Zip code was used to categorize by distance from the hospital and was integrated with the Agency for Healthcare Research and Quality’s database on Social Determinants of Health zip code data from 2020³⁹ to obtain the median income for the zip code of origin when the zip code was in in the United States. NLP of clinical notes was performed using Clinithink’s CLiX Focus software to extract HPO terms on a recurring monthly basis as described in Rockowitz et al.³²

Statistical analysis of EHR data

Two-sided Fisher’s exact tests were run on each categorical variable within the following demographics: age, sex, family structure, revised combined race and ethnicity category, distance to BCH, type of health insurance, median income of the zip code of residence, and variant type (P/LP, VUS, no reported variants) and reanalysis status to identify variables that had significant associations. A total of 102 tests was performed comparing categorical variables to variant type and 34 tests were performed comparing categorical variables to reanalysis status. Benjamini-Hochberg correction was performed on all resulting p values to adjust for multiple hypotheses testing.

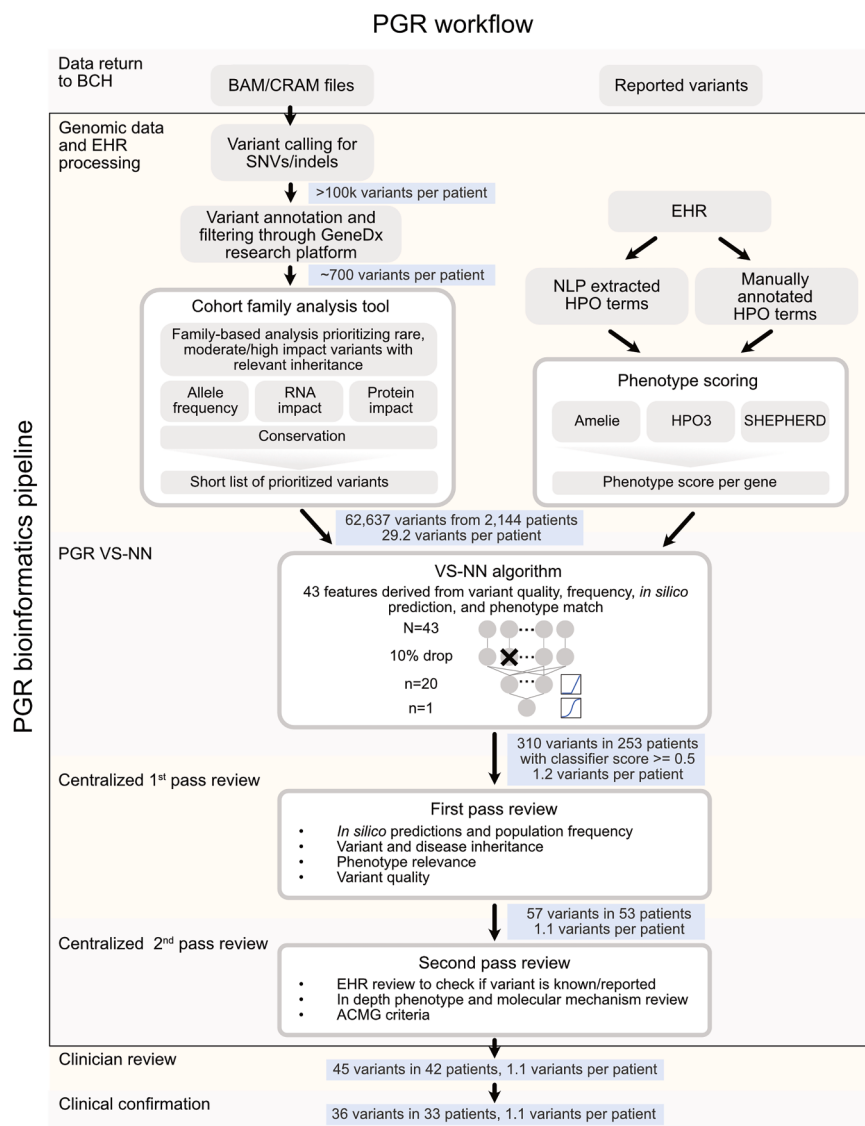


Figure 2. PGR workflow

This diagram visualizes the overall PGR workflow, which includes the PGR bioinformatic pipeline (VS-NN algorithm followed by first- and second pass review processes). This figure contains the number of variants remaining after each stage of the workflow.

from GeneDx's internal database of reanalysis led to the exclusion of another of 89 probands. For the remaining 440 probands who underwent clinical reanalysis, the labeling system described in [analysis of variants reported by GeneDx](#) was applied to the reanalyzed data. Patients were then categorized as having (1) variants that were upgraded or added through reanalysis, (2) variants that were downgraded or removed through reanalysis, (3) variants upgraded/added and variants downgraded/removed, and (4) no change after reanalysis.

Secondary and tertiary analysis

All ES and GS data were reprocessed using DRAGEN read mapping and variant calling pipelines on human reference genome build hg38 ([Figure 2](#)). SNVs and small indels from the standard DRAGEN VCF output were integrated into the GeneDx research platform, which provides mode of inheritance based on joint genotype calling when parental samples are available, variant annotations using Ensemble's Variant Effect Predictor (VEP),⁴¹ gnomAD v.2,⁴² dbnsfp,⁴³ ClinVar,³⁴ OMIM,⁴⁴ HPO,⁴⁵ and a custom variant classification prediction based on the ACMG scoring,³³ in line with the ap-

proaches described in InterVar⁴⁶ and the Bayesian classification framework by Tavtigian et al.⁴⁷ The GeneDx research platform does not provide allele frequency, variant classifications, or phenotype annotations from GeneDx's proprietary databases, but instead relies on the publicly available sources listed above. Initial filtering restricts variants in the VCF based on variant effect (exonic ± 20 base pairs or P/LP in ClinVar⁴⁸), allele frequency across reference databases and internal BCH cohorts, and variant quality (to filter out low-quality variants and regions with low-quality mapping).

Cohort family analysis

This study used the default cohort family analysis (CFA) filters, which includes the following: HIGH (loss of function) and MODERATE (other protein-changing) VEP impact, allele frequency $\leq 1.5\%$, heterozygous count ≤ 30 , homozygous count ≤ 2 , not annotated benign (i.e., not ClinVar benign/likely benign nor GeneDx research platform ACMG auto-classification benign/likely benign or VUS with only benign criteria), and inheritance pattern consistent with *de novo*, compound heterozygous, homozygous, or hemizygous inheritance or otherwise segregating with

Reanalysis annotations in clinical notes

HoundDog,⁴⁰ an internally developed tool at BCH for searching and indexing free text within clinical notes, was used to scan clinical notes to identify patients who had terms related to sequencing ("WES" or "WGS" or "exome" or "genome") and terms related to reanalysis ("reanalysis" or "re-analysis"). If a patient had a single clinical note with both sets of terms, they were classified as having reanalysis annotated in their clinical notes. A total of 4,975 patients had reanalysis annotated in their clinical notes; 88 patients were excluded because they were not easily mappable to defined categories of GeneDx analysis results (see [analysis of variants reported by GeneDx](#)). Of the remaining 4,887 patients, 407 were included in the 440 who underwent clinician-initiated clinical laboratory reanalysis.

Annotation of reanalyses reported by GeneDx

Probands with clinical reanalysis were identified as those with the original test ordered after April 11, 2019, and having multiple GeneDx case identifiers. The BCH schema only allowed for one proband per family. Cases with changes in family structure or other uncertainty were excluded. Secondary analysis of data

affected members of the family. For cases where parent data are not available (non-trios), additional filters were used including stronger frequency and impact filters (e.g., high CADD, constraint, or splicing scores) and requiring either patient-gene phenotype overlap or pathogenic annotation of the variant.

HPO match and phenotype scoring

To speed manual variant review, HPO term matches between the patient's terms and the terms associated with the gene containing the variant were extracted. Exact term matches along with matches between terms with one ontology level difference were pulled. Additionally, a term information content^{49–51} cutoff of two and four was employed to highlight more specific and rare term matches. Furthermore, phenotype scoring between the patient's terms and the terms associated with the gene was performed using HPO3,^{49,50} AMELIE,⁵² and SHEPHERD.⁵³ The resulting term match features and phenotype scores were used as input to a variant-scoring algorithm.

Variant-scoring algorithm

A total of 2,575 variants (2,154 from clinically ordered sequencing, 421 clinically confirmed findings from research) from 1,844 probands was used to train the variant-scoring neural network (VS-NN) algorithm. Variants that were categorized as pathogenic ($n = 434$), likely pathogenic ($n = 811$), and VUS ($n = 1,330$), were labeled as positive instances. A total of 53,223 variants that passed the CFA filters but were absent from clinical reports for these cases served as negative instances for model training. Each variant was annotated with 43 features, including variant allele fraction, variant depth, variant inheritance, gene inheritance, VEP,⁴¹ ClinVar clinical significance,⁴⁸ gnomAD v.4 frequency,⁴² CRDC internal frequency, HPO3,^{49,50} AMELIE,⁵² SHEPHERD,⁵³ HPO match,^{49–51} CADD PHRED,⁵⁴ REVEL,⁵⁵ AlphaMissense,⁵⁶ SpliceAI,⁵⁷ MisFit,⁵⁸ S_{coef} ,⁵⁹ S_{het} ,⁵⁹ pLI,⁶⁰ and GeneDx research platform ACMG classification.³⁷ These features were numerically encoded, standardized, and used as input for the VS-NN algorithm, which outputs a pathogenicity score ranging from 0 to 1. For initial model training, 30% of the dataset was reserved for validation and, after evaluating model performance, the final model was retrained using the full dataset. After training, the VS-NN algorithm was applied to 62,637 variants that passed the CFA filters from 2,144 patients with no reported variants as of July 26, 2024, and variants with a score ≥ 0.5 underwent manual bioinformatician review.

Pruning HPO analysis

For each patient analyzed, the ancestral hierarchies of the manually collected HPO terms were traced up to the third highest level in the ontological tree. Terms with the same third-level ancestor were grouped into branches representing related phenotypes. The pruning algorithm was run repeatedly, removing one branch each time, to simulate phenotype clusters (e.g., neurological) not yet present earlier in the patient's history. Following each pruning iteration, the VS-NN algorithm was rerun to assess the effect of branch-level term reduction on downstream variant prioritization and model performance. p values comparing different sets of HPO terms in true positives (TPs) and false negatives (FNs) from the validation dataset were calculated using two-sided Wilcoxon rank-sum tests and corrected using the Benjamini-Hochberger method.

Manual bioinformatician review of the VS-NN algorithm results

Variants from the VS-NN algorithm with a score ≥ 0.5 were screened by a PhD bioinformatician with variant curation training. The bioinformatician verified that *in silico* predictions did not suggest a benign or likely benign classification. Population frequency was evaluated using gnomAD v.4, including a review of all possible variants at the coordinate and summing allele counts (AC) and frequencies (AF) for variants with similar impact. For recessive variants, strict thresholds (gnomAD AF ≤ 0.001 and homozygous/hemizygous count ≤ 2) were applied, only considering variants with homozygous/hemizygous presence if the phenotype match was strong. For dominant variants not associated with neurodevelopmental disorders, a gnomAD AC ≤ 2 was required unless incomplete penetrance was supported by literature or phenotype match was compelling. For dominant variants linked to neurodevelopmental conditions, variant presence in gnomAD v.3 controls and non-neurodevelopmental populations were examined, removing those with AC > 2 in controls and retaining those with up to 1 in non-neurodevelopmental populations, unless incomplete penetrance was supported by literature. Homozygous/hemizygous presence in gnomAD was not allowed for dominant variants. The variant's mode of inheritance consistency with the disease inheritance pattern in OMIM was confirmed, if an OMIM record was available. If an OMIM record was not available, AMELIE literature matches and a PubMed search were used to gather information about the mode of inheritance for the disease. The phenotype relevance using OMIM, AMELIE literature matches and the patient's EHR were then assessed. Variants that only meet OMIM susceptibility to diseases were not returned. Finally, all candidate variants underwent manual quality confirmation in IGV.⁶¹

Manual second pass review of VS-NN algorithm results

Variants from the VS-NN algorithm that also passed bioinformatician review were then manually reviewed by a licensed genetic counselor or PhD variant curator with significant ACMG criteria application training for deeper review of genotype-phenotype correlation, gene-disease association, and ACMG variant criteria.^{47,62} This review was based on available literature, the Clinical Genome Resource (ClinGen⁶³) and Gene Curation Coalition (GenCC⁶⁴) for gene-disease validity as well as the Sequencing Variant Interpretation (SVI) frameworks. VUS leaning P/LP per SVI and ACMG criteria and/or had an identified functional assay that could potentially be pursued by the ordering physician and thereby upgrading the pathogenicity were determined to be returnable.

Establishment of a centralized genomic reanalysis workflow

A centralized genomic reanalysis process was established in 2024, when a strategic working group was convened to develop more scalable approaches to support genetic diagnoses, clinical trials, and clinical implementation of genetic therapies within the Department of Pediatrics and more broadly across BCH. The BCH Genomic Medicine/Genetic Therapies Working Group included clinicians across a wide range of specialties and identified areas of focus where pilot projects could have the biggest impact on patient care. At the time, several reanalysis options were already in place at BCH, including clinician-initiated lab

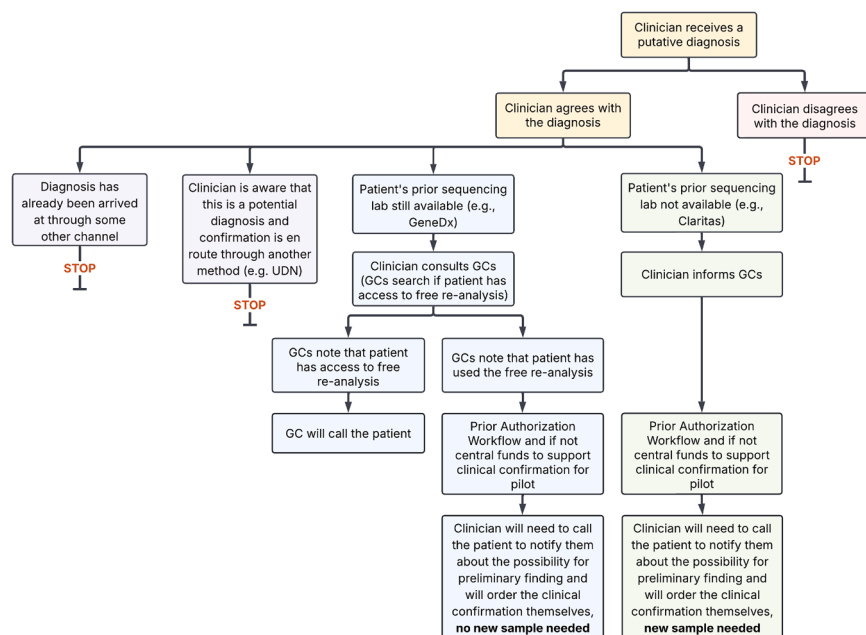


Figure 3. Clinical workflow diagram
This flowchart depicts the process followed by the clinician after receipt of a variant of interest from the centralized working group. Genetic counselor indicated by GC in the workflow diagram.

cians to coordinate confirmation of the identified variants. An internal report, which contained one to two variants, was communicated to the clinician who was most familiar with the patient, frequently this was the ordering provider for the initial ES/GS. If the ordering clinician and/or the clinician most familiar with the patient had left BCH, a clinician was identified by their Division/Department. To address provider turnover, the Division of Genetics and Genomics and Department of Neurology, who respectively ordered 64% and 33% of the ES/GS at BCH, established workflows to

reanalysis (Figure 1, blue box), where the laboratory reanalysis was requested at the clinician's discretion directly by the clinician who was seeing the patient; clinician manual analysis of genomic data (Figure 1, green box), which was empowered by structured training for clinicians who routinely ordered genetic testing on the BCH genomic analysis tools (Figure S2); and research reanalysis (Figure 1, pink box), which consisted of investigator-initiated Mendelian disease research efforts performed by research teams and included CRDC-affiliated studies such as the Manton Center,³⁶ rather than a single standardized research pipeline or institutional workflow. One of the recommendations that emerged from this process was for a pilot project focused on a centralized, automated reanalysis process within the hospital that was clinically meaningful and scalable across varied care settings, leading to the development of the PGR workflow. The initial pilot implementation phase included the development of the semi-automated BCH PGR bioinformatics pipeline to identify potentially diagnostic variants, application of the bioinformatics pipeline on currently available clinical genomic data, and definition of the clinical processes to return the identified variants to clinicians and patients. The initial pilot implementation was conducted between July 2024 and May 2025 and was a multidisciplinary effort among clinicians in multiple departments/divisions, institutional leadership, and the technical team from the CRDC. The centralized working group consisted of clinical and bioinformatics leadership, including PhD bioinformaticians, MD and MD/PhD clinicians with expertise in clinical and molecular genetics and licensed genetic counselors, who together led the development of the PGR bioinformatics pipeline, established variant review and result dissemination workflows and supported clinician engagement. The procedures followed were part of a clinical quality improvement program.

PGR workflow implementation and pilot context

During the pilot phase, the PGR workflow was clinically integrated and operationalized as part of routine clinical processes. Internal reports generated by the centralized working group were reviewed by clinicians and workflows were defined for clini-

reassign responsibility to a provider. The expectation was that the clinician would review the internal report and the patient's medical record holistically to make a determination about whether the variant was a good phenotypic match and cause for the patient's phenotype (Figure 3). The report (Figure S6) included information to inform the clinician about the variant, such as published allele frequencies in gnomAD, PubMed IDs for relevant publications, and internal ACMG scoring conducted by the second pass reviewer. If the clinician determined that the variant fit the patient's phenotype and a diagnosis had not already been delivered to the patient, then the clinician would reach out to the patient to coordinate and initiate the variant confirmation by the CLIA lab. The report also contained guidance on how to request confirmation of the variant(s) if one was needed. Patients would be informed of the specific variants only after clinical reanalysis or targeted gene testing was completed. When possible, variants would be confirmed by lab reanalysis rather than targeted gene testing, as lab reanalysis triggered no financial charge to the patient. In this case, the specific variant for confirmation would be communicated to the reference laboratory that completed the initial ES/GS. If the patient's reanalysis had already been utilized, the clinician would initiate a prior authorization process for the specific variant confirmation following a standardized workflow. BCH's International Office was engaged by clinicians in cases for whom the patient's primary address was outside of the United States. Pilot funds were allocated to support variant confirmation in cases when prior authorization was denied or unavailable.

Results

Diagnostic trends and correlates in clinically ordered ES/GS at BCH

We reviewed the results from the diagnostic testing conducted by GeneDx ("testing laboratory"). From 2021 through 2024, an average of 996 probands per year had clinical sequencing genomic data returned to BCH

(ranging from 866 to 1,157) (Figure S1C). In 2025, the return of genomic data began growing more rapidly, with 1,134 probands' data returned by May 1, accompanied by an increase in GS vs. ES. Review of reported variants included ES/GS cases for which variants had been returned in a structured and interpretable format ($n = 6,045$) and did not include other cases such as those sequenced at Claritas (Table 1). A total of 21.6% ($n = 1,303$) of these probands received a causative P/LP variant from the analysis; 43.2% ($n = 2,677$) of probands had no reported variants and 34.2% ($n = 2,065$) had variants that were VUS or P/LP but not causative for the patient's disease. The 78.4% ($n = 4,742$) of probands without a causative P/LP variant could potentially benefit from reanalysis over time, which may identify a diagnosis, upgrade the interpretation of a variant(s) from VUS to P/LP or resolve a candidate or possibly associated gene into the causative gene for the patient's disease. Additionally, while some studies demonstrated that ~5% of diagnosed patients would typically be expected to have dual diagnoses involving two or more disease loci,⁶⁵ in this cohort only 2.8% ($n = 37$) of probands with causative variants had causative variants in multiple genes, which may reflect underdiagnosis or ascertainment differences in this cohort.

We tested the association between the type of result and the patient's age, sex, family structure, revised combined race and ethnicity category calculated by integrating self-reported race and ethnicity from the EHR,³⁸ distance to BCH, type of health insurance, and median income of the zip code of residence to see if any variables were associated with having a finding on the clinical testing report (Figure S3). Across a total of 136 comparisons, 7 showed statistically significant differences (two-sided Fisher's exact test, Benjamini-Hochberg corrected $p < 0.001$) (Table S1). For example, VUS were less often identified in trios (19% had VUS) compared with duos and singletons (29% had VUS). VUS were also less often identified in patients with private insurance (18%) compared with government insurance (24%), a difference likely influenced by associations between types of health insurance and the other variables (e.g., revised combined race and ethnicity category, family structure).

Institutional baseline: Clinician-initiated reanalysis prior to PGR

To contextualize the clinical need addressed by PGR, we examined the frequency, uptake, and diagnostic impact of clinician-initiated laboratory reanalysis at BCH prior to implementation of the centralized PGR workflow. We reviewed the frequency of and impact from clinician-initiated reanalyses conducted at the behest of BCH clinicians by the testing laboratory and research reanalyses conducted by BCH researchers. Prior to the work described in this article, given the burden of reanalysis on individual clinicians, clinician-initiated lab reanalysis was performed infrequently (in 13.6% of

patients, $n = 852$) and usually upon patient request and/or after a subsequent clinical encounter (Figure 1, blue box). Despite broad consensus among clinicians in the strategic working group that genomic data should be reanalyzed over time, real-world uptake remains low, with only 15% (754/4,975) of patients at BCH with documented plans for reanalysis (including probands whose original analysis was completed prior to the agreement with BCH, but had data returned after) actually receiving it. Overall, for probands with original testing performed in the last 6 years (since April 2019), clinician-initiated lab reanalysis had been completed for 6.9% of probands ($n = 440$), initiated by 279 different clinicians. This occurred on average 2.37 years (0.75 SD) after the initial analysis. These findings highlight the limited reach and scalability of clinician-initiated reanalysis under current practice and underscore the need for a centralized, proactive approach.

White, non-Hispanic/non-Latino revised combined race and ethnicity category patients were more likely (9%) than patients with other revised combined race and ethnicity categories (5%) to have had a clinician-initiated lab reanalysis, while patients with race/ethnicity not reported (unknown or decline to report) were less likely (3%) to have had a clinician-initiated lab reanalysis (Table S1). A total of 17% ($n = 76$) of those with clinician-initiated lab reanalysis had their report changed over time, with the majority ($n = 64$) having variants added or variants upgraded. Ten probands had variants that were downgraded or removed and two patients had a variant upgraded/added and a variant downgraded/removed (Table 1). Additionally, 16% of patients ($n = 953$) were enrolled in a Mendelian disease research study by 103 clinician-scientists, where their data were reanalyzed by research teams (3.6% of patients, $n = 205$, had both clinical and research reanalysis performed). Although research reanalysis can help close the gap, a substantial fraction of clinically sequenced patients could benefit from a centralized proactive approach.

Semi-automated identification of prioritized variants

We developed a scalable PGR bioinformatics pipeline and centralized review process (Figure 2). The pipeline utilizes the established CRDC infrastructure,^{32,33} which includes variant calling, variant annotation and filtering, and collecting HPO terms. The in-house tool, CFA, was developed on the GeneDx research platform^{66,67} and applies further annotation and stringent filtering criteria. Specifically, CFA incorporates filters based on variant allele frequency and count, automated ACMG scoring,³⁷ ClinVar⁴⁸ classification, predicted variant effects, and inheritance patterns. Initial exploratory analyses of the clinical sequencing data reviewed variants and genes that had been added to ClinVar or OMIM⁴⁴ after the original analysis date and had some phenotype overlap, as well as variants in genes that scored highly in association with the patient's phenotype according to AMELIE⁵² and HPO3.^{49,50}

To formalize and automate variant prioritization, we developed and trained a VS-NN algorithm, with 43 characteristics (Figure 2). We evaluated the performance of the VS-NN algorithm using variants reported by the testing laboratory as a validation dataset. Variants with prediction scores of 0.5 or higher were flagged as a threshold of ≥ 0.5 yielded the optimal balance of precision and recall (Figure S4A). TP variants were defined as variants reported by the testing laboratory that received a score above 0.5 from the VS-NN algorithm. FN variants were defined as variants reported by the testing laboratory that received a score below 0.5 from the VS-NN algorithm. False positive variants were defined as variants not reported by the testing laboratory that received a score above 0.5 from the VS-NN algorithm. We observed an area under the receiver operating characteristic curve (AUC) of 0.99 for P/LP variants and 0.94 for VUS (Figure S4B). While AUC provides a general measure of discriminative ability, it can be disproportionately influenced by the majority class, particularly in imbalanced datasets like ours where negative cases significantly outnumber positive ones. To address this, we further assessed model specificity and sensitivity. The model demonstrated strong precision and recall for P/LP variants at 74% and 83%, respectively. However, precision and recall for VUS were notably lower at 57% and 32%, suggesting limited performance in detecting VUS and highlighting an area for potential model improvement.

To better understand the performance of the VS-NN algorithm, we further examined the distinguishing characteristics among the TP and FN variants (Figure S5; Table S2). Compared with FN variants, TP variants had a lower allele frequency in gnomAD and across CRDC participants; a higher S_{coef} , S_{het} , HPO3 score, and AMELIE score; more HPO matches with information content cutoff of two; more likely to be classified as P/LP by the automated ACMG scoring, exhibiting *de novo* inheritance in dominant genes, they are less likely to have conflicting classification in ClinVar. TP P/LP variants also had a higher pLI, were more likely to be in OMIM dominant genes and pathogenic in ClinVar, and less likely to be part of a compound heterozygous variation or biallelic in proband only cases than FN P/LP variants. Whereas TP VUS had higher CADD PHRED scores, REVEL scores, AlphaMissense scores, MisFit D, MisFit S, HPO match, and HPO match with information content cutoff of four, they were less likely to be absent in ClinVar and more likely to be ClinVar VUS. Overall, these results show that rare variants with stronger *in silico* predictions and better phenotype matches are more likely to receive higher VS-NN scores, consistent with expectations. At the same time, the findings highlight the complexity of automating variant classification, pointing to opportunities for further model refinement.

We also reviewed the impact of providing the VS-NN algorithm with different subsets of HPO terms using the TP/FN validation cases, as a proxy for simulating pheno-

type changes over time. We compared using only HPO terms manually annotated by clinicians and researchers with including HPO terms automatically extracted from the clinical notes with NLP (Table S3). The advantage of including NLP-extracted HPO terms is that patient phenotypes can change over time and that NLP extraction enables a scalable method to capture this change, which is why we trained the classifier on the combined HPO terms. There was no statistically significant change for including the NLP-extracted HPO terms on the VS-NN algorithm score, supporting the feasibility of integrating dynamic phenotype data at scale. Furthermore, we simulated the impact of changing phenotypes over time by analyzing how pruning HPO terms from the manually annotated HPO terms changed the VS-NN algorithm score. We observed a significant decrease (10%) in the score for TP and no change for FN variants with pruning the HPO terms. When evaluating the general performance of the scoring algorithm on all variants input for these patients, there was no difference between scores for any group of HPO terms (combined, manual only, pruned), indicating that the change in performance was specific for relevant, high-scoring variants rather than a global score shift.

An internal two-pass review was utilized by the centralized working group. Flagged variants were reviewed by a bioinformatician who assessed phenotypic match, mode of inheritance, variant quality, *in silico* predictions, and frequency in gnomAD v.4.⁴² Given the high specificity of the VS-NN algorithm implemented during the initial pilot, only 253 patients had results (310 variants). Manual review by the bioinformatician winnowed this down to 53 patients with 57 variants. Variants were rejected because there was a mismatch between the phenotypes reported for the gene and the patient (54%, $n = 169$), the mode of inheritance was not consistent with the disease inheritance pattern reported in OMIM (12%, $n = 37$), the reads in the BAM file did not pass manual quality assessment (8%, $n = 25$), or *in silico* predictions suggested a benign or likely benign classification (7%, $n = 22$). This review took on average 5 min per variant. A second pass review determined ACMG criteria and classification and performed a more in-depth review of phenotype match and molecular mechanism as needed, either adhering to published gene-disease validation and variant curation guidelines (e.g., ClinGen) when available or critical review of available literature on genotype-phenotype correlation to ensure variants communicated to clinicians would be most likely to be relevant to the proband's phenotype. This second review averaged 3 min per variant. A total of 45 variants in 42 patients passed the second pass review; 16% of the variants that passed the first pass review ($n = 9$) were rejected as they were already documented in the EHR; these were all in the unknown classification/Claritas dataset and there was no structured data for variants returned for these cases. Five percent ($n = 3$) of the variants were rejected due to phenotype, with the second pass reviewer not considering the disease phenotype and

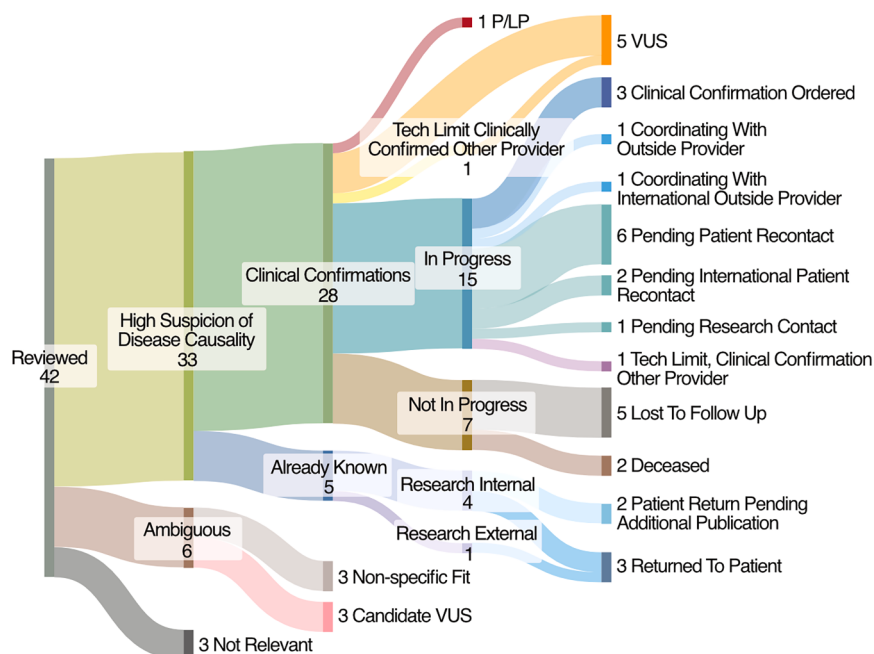


Figure 4. Early results

This Sankey diagram visualizes the results from the initial batch of analyses that were shared with clinicians. Variants were reviewed in the context of the patient's phenotype and clinicians determined whether each variant was a potential clinical diagnosis, not relevant to the phenotype or ambiguous. Variants were further subtyped to already known or requiring clinical confirmation. Results were annotated P/LP, VUS, in progress, or not in progress. In progress cases were subtyped to clinical confirmation ordered, coordinating with outside providers, pending patient recontact, and technical limitation from GeneDx/reflexed to alternative reference laboratory. Not in progress cases were subtyped to lost-to-follow-up and deceased. This figure was created using SankeyMATIC.³³

patient phenotype closely enough matched to warrant reporting. An internal report was written by the centralized working group (Figure S6), which averaged 20 min per proband.

Centralized PGR pilot results

For the initial pilot implementation of the PGR workflow, downstream variant prioritization and manual review focused on 1,639 patients who did not have reported variants identified by the testing laboratory and 505 patients where the variant reports were unknown using the PGR workflow (PGR bioinformatics pipeline followed by the clinical process plan, Figure 2), we then tracked the results of this round of analyses. Twenty-three unique providers were involved in communications about the variants that passed the internal two-pass review process, reflecting downstream clinical follow-up within the defined workflow (Figure 3). Clinicians were only asked to review the small number of prioritized variants which required modest effort and was similar to the skills required to assess a clinical report. Clinicians identified 33/42 probands who had variants identified by the centralized working group as having variants with a high suspicion of disease causality. Of the other nine probands, three had variants that were considered to be candidate VUS predicted not to positively impact patient care while the rest did not have a strong enough fit to the phenotype (Figure 4). Different providers had to be nominated due to provider turnover in 24 of the cases. Twenty-two patients had been sequenced at Claritas, which closed in 2017 and did not offer subsequent reanalysis. Even among GeneDx cases ($n = 20$), five had already utilized their clinical reanalysis. Of the GeneDx cases in which the variant was considered by the clinician to have a high suspicion

of disease causality ($n = 14$), nine needed clinical reanalysis for confirmatory testing. Two had already used the clinical reanalysis, triggering the need for single-gene confirmation and prior authorization. Reasons that confirmatory testing was not pursued for some probands included variants already being known due to research ($n = 5$) and not having reports uploaded to the EHR (three were returned to patients via research studies, two were awaiting publication of natural history studies, which would enable stronger variant classification); patients lost to follow-up ($n = 5$); and patients who died ($n = 2$). Additional complexities included technical limitations with specific loci preventing confirmation and reporting of certain variants by one laboratory, but were able to be tested at another reference laboratory ($n = 2$). To date, one proband received a P/LP variant report, 5 received reports with VUS in candidate genes that were likely to positively impact patient care, and the 15 remaining probands where the clinician agreed with the diagnosis remain in progress. Under the conservative criteria used in this initial pilot, the PGR workflow identified candidate variants in 42 of 2,144 reviewed cases, with clinicians determining that 33 represented a high suspicion of disease causality.

To examine whether any variables were associated with having a finding through the PGR workflow, we tested if there was any difference in the patient's age, sex, family structure, revised combined race and ethnicity category, distance to BCH, type of health insurance, or median income of the zip code of residence between patients who had variants that were input into the VS-NN algorithm, patients with variants that passed the VS-NN algorithm score threshold, patients with variants that passed either step of review, and patients with variants considered to

be high suspicion of disease causality or candidate VUS. Given that such a large fraction of the patients with pilot PGR process results were sequenced at Claritas between 2015 and 2017 and the changing nature of clinical guidelines around sequencing, we stratified these analyses by laboratory. With this stratification, there were no statistically significant differences between the ratios of any of these variables.

Identified variants were analyzed according to Robertson et al., who categorized mechanisms driving the increases in diagnostic yield in reanalysis studies.⁸ Two of the VUS were not included in the original report due to limited clinical overlap at the time of the original test (“gathering of detailed phenotype information and/or evolution of the patient phenotype”). Three variant identifications were classified as due to technical limitations (“generation of additional genomic data and/or examination of additional variant types”): the two that were reflexed to another reference laboratory and one VUS diagnosis that was in a region that was not called by the clinical laboratory’s pipelines as it had been identified as a low coverage region. Thirteen had variant identifications due to the “discovery of gene-disease/variant-disease associations” that were reported in OMIM,⁴⁴ GenCC,⁶⁴ or ClinGen⁶⁸ after the date of original analysis, including nine that had been sequenced at Claritas and four that had been sequenced at GeneDx.

Discussion

In this article we describe the development and implementation of a PGR workflow for clinical sequencing data at a single children’s hospital that is centralized, semi-automated and clinically integrated and designed to bridge clinical and laboratory data silos and empower clinicians to engage in ongoing variant interpretation (Figure 1). Clinician-initiated reanalysis is included as an institutional comparator to contextualize the need for a centralized PGR workflow. Prior work shows that faster integration of comprehensive clinical data and up-to-date gene-disease associations can increase diagnostic yield by 15% while avoiding unnecessary testing.⁶⁹ Building on this evidence, we introduce an institutional model for clinical genomic medicine that embeds reanalysis directly into routine care. Leveraging infrastructure established by the CRDC, we integrated raw genomic data and evolving phenotype information from over 6,279 patients and their families, enabling scalable and clinician integrated reanalysis. The almost 80% of clinically sequenced patients at BCH without known causative pathogenic variants stand to benefit from systematic reanalysis that could identify variants, and all may benefit from automatically updating the classification of previously reported variants and identifying second loci that contribute to disease. The PGR workflow democratized access to reanalysis regardless of ordering

provider, as sequencing data from all clinicians at BCH who have sent ES/GS sequencing were included in the PGR bioinformatic pipelines. A centralized working group oversaw variant prioritization using a combination of automated machine learning tools and expert review, resulting in the identification of potential diagnoses, some with treatment implications. Working through the process of returning the initial 42 patients’ identified candidate variants to the clinicians and, where appropriate, the patients, revealed the challenges and impact of adopting centralized institutional genomic reanalysis. In this initial pilot, the observed reanalysis yield was 2% using conservative filtering and review criteria, reflecting an intentionally high specificity approach.

Implementation of the CRDC infrastructure at BCH enabled integration of clinically returned genomic data and metadata, exemplifying how the research framework can be scalably repurposed to support clinical care. The transfer of data from the testing laboratory to BCH required sustained coordination between the two organizations, including regular communication to ensure accurate transformation of genomic data into a BCH-aligned schema. As clinical process streamlining led to substantial increases in the frequency of ES and GS (84 probands sequenced per month from 2021 to 2024, 248 probands sequenced per month beginning in 2025), the data transfer processes were scalable. This harmonization enabled efficient querying of previously siloed reported variants⁷⁰ for IRB-approved research and preparatory to research activities. Critically, patients with existing clinical sequencing no longer required research-generated ES/GS, yielding substantial cost savings to the CRDC³³ and other research studies (the total amount spent on processing and storing clinical sequencing data in the CRDC infrastructure is only about a third of the cost savings from reflexing clinically ordered sequencing data to research use). The CFA platform, originally built for research, was repurposed to deploy a centralized variant prioritization system with minimal overhead, illustrating that institutions with embedded research capacity can provide clinical insights by repurposing tools already at hand, enabling real-time use of clinical genomic data to inform care.

Despite promising early results from the PGR bioinformatics pipeline, several limitations have emerged that inform ongoing refinement. The VS-NN algorithm demonstrated high specificity but limited sensitivity, identifying candidate variants in only 1.8% of patients (Figure 1; Table 1). This low yield can be attributed to conservative filtering thresholds, a deliberate prioritization of precision over recall, and dependence on strong algorithmic phenotype matching using tools such as AMELIE and HPO3. When stratifying diagnostic yield by the time since sequencing or data return to BCH, we observed an increase in the percentage of patients receiving a diagnosis from reanalysis as the interval lengthened (Table 1), consistent with prior reports. Notably, the current workflow excluded patients with previously reported variants

and did not include any algorithmic mechanism to identify those eligible for variant reclassification, both of which represent additional opportunities for diagnostic yield and potential impact on patient care.

Comparison with other large-scale efforts provides additional context. The diagnostic yield of the PGR pipeline is comparable with that of the 100,000 Genomes Project¹⁰ (100kGP), but lower than that of Solve-RD,²⁴ where more variants per individual (mean = 8) were returned to task forces for review. In contrast, we returned a mean of 1.2 variants per patient, whereas 100kGP returned approximately 1.6. When including VUS, the VS-NN algorithm achieved a recall of 56% and precision of 80%, compared with 84% and 57% in 100kGP's Exomiser pipeline, respectively, highlighting our more conservative approach. However, when restricted to variants classified as P/LP by both return status and automated ACMG scoring, similar to the optimal settings used by 100kGP, our recall and precision were comparable (recall: 90% vs. 82%; precision: 81% vs. 88%; VS-NN vs. 100kGP, respectively). These benchmarks suggest that integration of lessons from pipelines such as Exomiser could improve sensitivity, particularly for VUS detection.

Manual review during the two-pass prioritization process illuminated clear opportunities for algorithmic tuning. Future improvements will include pruning features that fail to distinguish between TPs and FNs, enhancing features that discriminate between true and false positives, and refining allele frequency and inheritance filters. Adjustments to the phenotype-matching process, especially the weighting of evolving clinical phenotypes, may also increase sensitivity. Notably, a pruning analysis revealed that dynamic phenotypic updates significantly influenced classifier scores, but we did not assess how often those patient phenotypes changed over time. Future work will explore longitudinal HPO term evolution and its impact on reanalysis. Additionally, the current pipeline also restricts analysis to SNVs/InDels from the DRAGEN pipeline. Development is underway to expand support for other variant types, including structural variants and repeat expansions, which could further improve diagnostic rates.

The gap between the number of patients warranting genomic reanalysis and those receiving genomic reanalysis is large, reflecting a critical gap in current workflows, which rely on individual clinicians to initiate and navigate lab reanalysis requests, often limited by time, genomic expertise, and requiring patient prompting and partnership. As this reanalysis is currently clinically available only a single time per patient, there is uncertainty about optimal timing of this request. A centralized reanalysis model ameliorates some of these obstacles, improving access for all patients and all providers, not just those with advanced genetics training. It also enables the potential for repeated, longitudinal reanalysis, which is something current models rarely support. For example, commercial labs may offer only one reanalysis as part of the original

ordered testing and require prior authorization for additional reanalysis. By embedding reanalysis into institutional infrastructure, genomic data become a living asset, capable of delivering diagnoses over time and reshaping care for previously undiagnosed patients.

The initial pilot implementation phase revealed critical real-world challenges in deploying centralized reanalysis. Many early cases originated from sequencing at Claritas (2015–2017) with patients who had not been seen clinically for several years at our institution and often no longer had an identifiable provider affiliated with the institution. This made recontact difficult. Unlike testing sent to other labs, these Claritas cases lacked access to free reanalysis and required sample collection for confirmation, adding logistical and financial hurdles. To address costs, limited funds were allocated to support single-gene confirmatory testing when reanalysis or insurance coverage was unavailable or denied. Additional complexities included classification discrepancies between the testing laboratory and the centralized working group, challenges confirming variants (necessitating alternative testing in two cases), and international patients or patients who had died, complicating recontact and result disclosure. Although the planned process was intended to involve the primary clinician who had followed the patient/ordered the initial ES/GS, in practice it was often the case that clinicians with genetics expertise had to become involved to facilitate the process. In the Department of Neurology, this led to substantial revision of the initially proposed workflow (see [Figure S7](#)). Overall, the PGR bioinformatics pipeline alleviates the need for clinicians to manually identify variants and provides baseline analysis for all cases; however, there still exists substantial work for clinicians and departments in vetting the clinical correlation of the variants, identifying a laboratory that can confirm the findings, obtaining insurance authorization for confirmation, communicating effectively with families throughout the confirmation process and managing the return of results. Ongoing optimization efforts include refining the communication protocols, and workflow logic through iterative case-by-case resolution to handle the variety of clinical scenarios encountered, laying the groundwork for a more resilient and scalable reanalysis process. Future efforts will evaluate diagnostic yield, clinical utility, and patient/provider impact, alongside an analysis of cost per diagnosis. Ensuring sustainability will require ongoing optimization, but the potential to close the diagnostic gap and streamline care makes this a powerful step forward in the evolution of genomic medicine. Our centralized process is designed to provide more equitable care which starts with an accurate diagnosis. As therapeutic options increase, the value of a diagnosis will increase, and it will be critical to make diagnoses as early as possible while therapies are most beneficial. We have recently entered the second phase of the project, initiating the iterative application of the PGR workflow across a broader cohort. Early analyses have already uncovered

additional candidate variants, both in patients from the initial cohort as well as in individuals whose data were only included in the second phase of the project, as well as patients with variants that are eligible for reclassification. Importantly, we do not propose that the PGR workflow replace existing clinician-initiated or research-based reanalysis workflows. Rather, it is designed to complement these efforts by proactively identifying and returning findings that may otherwise be overlooked. However, we acknowledge that this model may be uniquely suited to institutions with specific capabilities. BCH benefits from a confluence of enabling conditions: a high volume of clinical sequencing sent to a single commercial laboratory, an established raw data repatriation pipeline, robust in-house bioinformatics and variant interpretation expertise, and a mature research informatics infrastructure. These factors enabled integration of the PGR workflow with minimal overhead and significant cost savings by reducing the need for redundant sequencing in research settings. For institutions without comparable resources, particularly those lacking a centralized research infrastructure or internal genomics expertise, the costs of developing and maintaining such a system may be prohibitive.

Moreover, our experience reflects a healthcare model prevalent in the U.S. and some other countries, where clinical genomic testing is often outsourced to commercial laboratories. This stands in contrast to healthcare systems with centralized or in-house genomic testing infrastructure, which operate under different constraints and incentives. We also benefited from sending nearly all testing to a single sequencing provider, simplifying integration. Scaling this model across multiple testing laboratories would require substantial additional development, although ongoing standardization efforts within EHR platforms such as Epic may facilitate broader implementation in the future. While the full PGR model may be best suited to high-volume, research-oriented institutions with existing genomic infrastructure, key components are broadly transferable. Structured variant triage processes, integration of longitudinal phenotype data, and centralized, clinician-facing tools can improve scalability and clinician engagement across varied care settings. A core strength of this model is that it reduces the burden on frontline providers by embedding reanalysis into institutional workflows without requiring additional clinician effort beyond standard care. Looking ahead, the PGR pipeline will be deployed regularly, biannually for the general patient population and more frequently (e.g., monthly) for high-priority groups such as NICU patients, offering a scalable approach to integrating dynamic genomic interpretation into routine clinical practice.

In summary, our experience demonstrates that a PGR workflow that is centralized, semi-automated, and clinically integrated is both feasible and impactful in a real-world healthcare setting. By integrating genomic and clinical data within a scalable institutional framework, we were able to systematically identify variants of interest, support

reclassification of existing findings, and return diagnoses that may otherwise have been missed. This model addresses persistent obstacles to reanalysis, limited clinician capacity and time, and siloed data, by embedding reanalysis directly into the clinical workflow and supporting it with semi-automated tools and structured team-based processes. Looking ahead, other hospitals that care for large numbers of patients with rare and evolving phenotypes, and that embrace research-informed care, may also be positioned to implement similar centralized models. Reanalysis can be performed within single institutions or potentially can be supported centrally across institutions with federated data. Automated reanalysis has the potential to democratize access to genomic insight, shift genetic testing from a static to a dynamic and iterative process, and enable genomic medicine to scale in service of continuous diagnostic improvement and precision care.

Data and code availability

- The VS-NN algorithm code generated during this study is available on GitHub: https://github.com/BCH-RC/PGR_VS_NN.
- Genomic data from research-consented patients analyzed during this study are available to BCH researchers and clinicians via the CRDC infrastructure and are also shared externally through the Genomic Information Commons project (<https://www.genomicinformationcommons.org/>).
- This study did not generate new publicly shareable datasets requiring deposition in a public repository. Given that the procedures described were conducted as part of a clinical quality improvement program, data from patients not consented to research are not publicly available.

Acknowledgments

Funding provided by NIH/NHGRI R01HG010365, NIH/NHGRI U01HG011755, NIH/NINDS K08NS117891, and a Boston Children's Hospital Office of Faculty Development/Basic & Clinical Translational Research Executive Committees Faculty Career Development Fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions

Conception or design of the work, S.R., M.H.W., P.S., and W.K.C.; acquisition, S.R., A.G., B.S., L.S., A.M.D'G., M.I., J.C., J.S.S., A.K., L.R., A. Shimamura, O.B., S. Sacharow, J.S., S. Srivastava, A.R.K., A.A.E.H., A.L., H.O., J.J., E.R., B.F., A. Singh, C.L., R. Mallik, G.S., G.P., A.M., A.O'D.L., J.B., P.M.B., W.B., M.D., S.M., D.T.M., J.O., J. Petit, J. Picker, A.P., C.C., M.H.W., and W.K.C.; performance of analysis, S.R., W.S., C.F., T.K.T., J.H., R. McGonigle, and R. Mallik; figure generation, S.R., W.S., C.F., R. McGonigle, and C.C.; interpretation of data, S.R., W.S., C.F., and W.K.C.; creation of software, S.R., W.S., C.F., J.H., A. Singh, C.L., and R. Mallik; writing of the manuscript, S.R., C.F., P.S., and W.K.C.; analysis of reported variants, B.S., L.S., A.M.D'G., M.I., J.C., J.S.S., A.K., L.R., A. Shimamura, O.B., S. Sacharow, J.S., S. Srivastava, A.R.K., A.A.E.H., A.L., H.O., G.P.,

A.M., A.O'D.L., J.B., P.M.B., W.B., M.D., S.M., D.T.M., J.O., J. Petit, J. Picker, M.H.W., and W.K.C.; review and editing of the manuscript, S.R., W.S., C.F., T.K.T., J.H., R. McGonigle, A.G., B.S., L.S., A.M.D'G., M.I., J.C., J.S.S., A.K., L.R., A. Shimamura, O.B., S. Sacharow, J.S., S. Srivastava, A.R.K., A.A.E.H., A.L., H.O., J.J., E.R., B.F., A. Singh, C.L., R. Mallik, G.S., G.P., AM, A.O'D.L., J.B., P.M.B., W.B., M.D., S.M., D.T.M., J.O., J. Petit, J. Picker, A.P., C.C., M.H.W., P.S., and W.K.C. All authors read and approved the final manuscript.

Declaration of interests

A.O'D.L. has consulted for Addition Therapeutics and received research support in the form of reagents from Pacific Biosciences. M.H.W. has consulted for Illumina and Sanofi and receives speaking honoraria from Illumina and GeneDx. W.K.C. received in kind support for research from Illumina and GeneDx. E.R., J.J., and B.F. are employees of and may own stock in GeneDx.

Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.xhgg.2026.100582>.

Received: November 2, 2025

Accepted: February 13, 2026

References

- Platt, C.D., Zaman, F., Bainter, W., Stafstrom, K., Almutairi, A., Reigle, M., Weeks, S., Geha, R.S., Chou, J.; and International Consortium for Immunodeficiencies (2021). Efficacy and economics of targeted panel versus whole-exome sequencing in 878 patients with suspected primary immunodeficiency. *J. Allergy Clin. Immunol.* **147**, 723–726. <https://doi.org/10.1016/j.jaci.2020.08.022>.
- Iglesias, A., Anyane-Yeboah, K., Wynn, J., Wilson, A., Truitt Cho, M., Guzman, E., Sisson, R., Egan, C., and Chung, W.K. (2014). The usefulness of whole-exome sequencing in routine clinical practice. *Genet. Med.* **16**, 922–931. <https://doi.org/10.1038/gim.2014.58>.
- Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K.G., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* **18**, 696–704. <https://doi.org/10.1038/gim.2015.148>.
- Schwarze, K., Buchanan, J., Taylor, J.C., and Wordworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **20**, 1122–1130. <https://doi.org/10.1038/gim.2017.247>.
- Nurchis, M.C., Radio, F.C., Salmasi, L., Heidar Alizadeh, A., Raspolini, G.M., Altamura, G., Tartaglia, M., Dallapiccola, B., Pizzo, E., Gianino, M.M., and Damiani, G. (2024). Cost-Effectiveness of Whole-Genome vs Whole-Exome Sequencing Among Children With Suspected Genetic Disorders. *JAMA Netw. Open* **7**, e2353514. <https://doi.org/10.1001/jamanetworkopen.2023.53514>.
- Sisodiya, S.M. (2021). Precision medicine and therapies of the future. *Epilepsia* **62**, S90–S105. <https://doi.org/10.1111/epi.16539>.
- Nurchis, M.C., Altamura, G., Riccardi, M.T., Radio, F.C., Chillemi, G., Bertini, E.S., Garlasco, J., Tartaglia, M., Dallapiccola, B., and Damiani, G. (2023). Whole genome sequencing diagnostic yield for paediatric patients with suspected genetic disorders: systematic review, meta-analysis, and GRADE assessment. *Arch. Public Heal.* **81**, 93. <https://doi.org/10.1186/s13690-023-01112-4>.
- Robertson, A.J., Tan, N.B., Spurdle, A.B., Metke-Jimenez, A., Sullivan, C., and Waddell, N. (2022). Re-analysis of genomic data: An overview of the mechanisms and complexities of clinical adoption. *Genet. Med.* **24**, 798–810. <https://doi.org/10.1016/j.gim.2021.12.011>.
- Laurie, S., Steyaert, W., de Boer, E., Polavarapu, K., Schuermans, N., Sommer, A.K., Demidov, G., Ellwanger, K., Paramonov, I., Thomas, C., et al. (2025). Genomic reanalysis of a pan-European rare-disease resource yields new diagnoses. *Nat. Med.* **31**, 478–489. <https://doi.org/10.1038/s41591-024-03420-w>.
- Vestito, L., Jacobsen, J.O.B., Walker, S., Cipriani, V., Harris, N.L., Haendel, M.A., Mungall, C.J., Robinson, P., and Smedley, D. (2024). Efficient reinterpretation of rare disease cases using Exomiser. *NPJ Genom. Med.* **9**, 65. <https://doi.org/10.1038/s41525-024-00456-2>.
- Bullich, G., Matalonga, L., Pujadas, M., Papakonstantinou, A., Piscia, D., Tonda, R., Artuch, R., Gallano, P., Garrabou, G., González, J.R., et al. (2022). Systematic Collaborative Reanalysis of Genomic Data Improves Diagnostic Yield in Neurologic Rare Diseases. *J. Mol. Diagn.* **24**, 529–542. <https://doi.org/10.1016/j.jmoldx.2022.02.003>.
- Schobers, G., Schieving, J.H., Yntema, H.G., Pennings, M., Pfundt, R., Derks, R., Hofste, T., de Wijs, I., Wieskamp, N., van den Heuvel, S., et al. (2022). Reanalysis of exome negative patients with rare disease: a pragmatic workflow for diagnostic applications. *Genome Med.* **14**, 66. <https://doi.org/10.1186/s13073-022-01069-z>.
- Matalonga, L., Hernández-Ferrer, C., Piscia, D., Solve-RD SNV-indel working group, Schüle, R., Synofzik, M., Töpf, A., Vissers, L.E.L.M., de Voer, R., Solve-RD DITF-GENTURIS, et al. (2021). Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur. J. Hum. Genet.* **29**, 1337–1347. <https://doi.org/10.1038/s41431-021-00852-7>.
- James, K.N., Clark, M.M., Camp, B., Kint, C., Schols, P., Batalov, S., Briggs, B., Veeraghavan, N., Chowdhury, S., and Kingsmore, S.F. (2020). Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *npj Genom. Med.* **5**, 33. <https://doi.org/10.1038/s41525-020-00140-1>.
- Lassmann, T., Francis, R.W., Weeks, A., Tang, D., Jamieson, S.E., Broley, S., Dawkins, H.J.S., Dreyer, L., Goldblatt, J., Groza, T., et al. (2020). A flexible computational pipeline for research analyses of unsolved clinical exome cases. *NPJ Genom. Med.* **5**, 54. <https://doi.org/10.1038/s41525-020-00161-w>.
- Hiatt, S.M., Amaral, M.D., Bowling, K.M., Finnila, C.R., Thompson, M.L., Gray, D.E., Lawlor, J.M.J., Cochran, J.N., Bebin, E.M., Brothers, K.B., et al. (2018). Systematic reanalysis of genomic data improves quality of variant interpretation. *Clin. Genet.* **94**, 174–178. <https://doi.org/10.1111/cge.13259>.
- Wenger, A.M., Guturu, H., Bernstein, J.A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet. Med.* **19**, 209–214. <https://doi.org/10.1038/gim.2016.88>.

18. Wright, C.F., McRae, J.F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T.W., Jones, P., Prigmore, E., Rajan, D., Lord, J., et al. (2018). Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.* 20, 1216–1223. <https://doi.org/10.1038/gim.2017.246>.
19. Baker, S.W., Murrell, J.R., Nesbitt, A.I., Pechter, K.B., Balciuniene, J., Zhao, X., Yu, Z., Denenberg, E.H., DeChene, E.T., Wilkens, A.B., et al. (2019). Automated Clinical Exome Reanalysis Reveals Novel Diagnoses. *J. Mol. Diagn.* 21, 38–48. <https://doi.org/10.1016/j.jmoldx.2018.07.008>.
20. Dai, P., Honda, A., Ewans, L., McGaughran, J., Burnett, L., Law, M., and Phan, T.G. (2022). 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. <https://doi.org/10.1016/j.gim.2022.04.021>.
21. Towne, M.C., Huang, J., Saliganan, S., Gasser, B., Holman, M., Webb, K., Wayburn, B., Radtke, K., and Farwell Hagman, K.D. (2025). Impact of Laboratory-Driven Proactive Reanalysis: Reclassification to Positive in 5% of Initially Negative or Uncertain Exome Sequencing Cases. *Genet. Med.* 27, 101464. <https://doi.org/10.1016/j.gim.2025.101464>.
22. Cuillerier, A., Goodman, A., Lawrence, C., Villeneuve-Cloutier, N., Armour, C.M., Bhola, P.T., Bourque, D.K., Carter, M.T., Lazier, J., Sawyer, S.L., et al. (2026). Diagnostic Utility of Exome Data Reanalysis After In Silico Multi-Gene Panels or Clinical Exome Testing for Patients With Epilepsy and Developmental Delay/Intellectual Disability: A Retrospective Cohort Study. *Clin. Genet.* 109, 40–49. <https://doi.org/10.1111/cge.70008>.
23. Hartley, T., Soubry, É., Acker, M., Osmond, M., Couse, M., Gillespie, M.K., Ito, Y., Marshall, A.E., Lemire, G., Huang, L., et al. (2023). Bridging clinical care and research in Ontario, Canada: Maximizing diagnoses from reanalysis of clinical exome sequencing data. *Clin. Genet.* 103, 288–300. <https://doi.org/10.1111/cge.14262>.
24. Basel-Salmon, L., Orenstein, N., Markus-Bustani, K., Ruhman-Shahar, N., Kilim, Y., Magal, N., Hubshman, M.W., and Bazak, L. (2019). Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. *Genet. Med.* 21, 1443–1451. <https://doi.org/10.1038/s41436-018-0343-7>.
25. Frees, M., Carter, J.N., Wheeler, M.T., and Reuter, C. (2025). The current landscape of clinical exome and genome reanalysis in the U.S. *J. Genet. Couns.* 34, e1968. <https://doi.org/10.1002/jgc4.1968>.
26. Clayton, E.W., Appelbaum, P.S., Chung, W.K., Marchant, G.E., Roberts, J.L., and Evans, B.J. (2021). Does the law require reinterpretation and return of revised genomic results? *Genet. Med.* 23, 833–836. <https://doi.org/10.1038/s41436-020-01065-x>.
27. Appelbaum, P.S., Berger, S.M., Brokamp, E., Brown, H.S., Burke, W., Clayton, E.W., Evans, B.J., Hamid, R., Marchant, G.E., Martin, D.M., et al. (2023). Practical considerations for reinterpretation of individual genetic variants. *Genet. Med.* 25, 100801. <https://doi.org/10.1016/j.gim.2023.100801>.
28. Mecky, J.E., Johansson, L., Plantinga, M., Fenwick, A., Lucassen, A., Dijkhuizen, T., van der Hout, A., Lyle, K., and van Langen, I. (2019). Reinterpretation, reclassification, and its downstream effects: challenges for clinical laboratory geneticists. *BMC Med. Genom.* 12, 170. <https://doi.org/10.1186/s12920-019-0612-6>.
29. Best, S., Fehlberg, Z., Richards, C., Quinn, M.C.J., Lunke, S., Spurdle, A.B., Kassahn, K.S., Patel, C., Vears, D.F., Goranitis, I., et al. (2024). Reanalysis of genomic data in rare disease: current practice and attitudes among Australian clinical and laboratory genetics services. *Eur. J. Hum. Genet.* 32, 1428–1435. <https://doi.org/10.1038/s41431-024-01633-8>.
30. Fehlberg, Z., Stark, Z., and Best, S. (2024). Reanalysis of genomic data, how do we do it now and what if we automate it? A qualitative study. *Eur. J. Hum. Genet.* 32, 521–528. <https://doi.org/10.1038/s41431-023-01532-4>.
31. van der Geest, M.A., Maeckelbergh, E.L.M., van Gijn, M.E., Lucassen, A.M., Swertz, M.A., van Langen, I.M., and Plantinga, M. (2024). Systematic reanalysis of genomic data by diagnostic laboratories: a scoping review of ethical, economic, legal and (psycho)social implications. *Eur. J. Hum. Genet.* 32, 489–497. <https://doi.org/10.1038/s41431-023-01529-z>.
32. Rockowitz, S., LeCompte, N., Carmack, M., Quitadamo, A., Wang, L., Park, M., Knight, D., Sexton, E., Smith, L., Sheidley, B., et al. (2020). Children's rare disease cohorts: an integrative research and clinical genomics initiative. *NPJ Genom. Med.* 5, 29. <https://doi.org/10.1038/s41525-020-0137-0>.
33. French, C.E., Andrews, N.C., Beggs, A.H., Boone, P.M., Brownstein, C.A., Chopra, M., Chou, J., Chung, W.K., D'Gama, A.M., Doan, R.N., et al. (2024). Hospital-wide access to genomic data advanced pediatric rare disease research and clinical outcomes. *NPJ Genom. Med.* 9, 60. <https://doi.org/10.1038/s41525-024-00441-9>.
34. Pais, L.S., Snow, H., Weisburd, B., Zhang, S., Baxter, S.M., DiTroia, S., O'Heir, E., England, E., Chao, K.R., Lemire, G., et al. (2022). seqr: A web-based analysis and collaboration tool for rare disease genomics. *Hum. Mutat.* 43, 698–707. <https://doi.org/10.1002/humu.24366>.
35. Patridge, E.F., and Bardin, T.P. (2018). Research Electronic Data Capture (REDCap). *J. Med. Libr. Assoc.* 106, 142–144. <https://doi.org/10.5195/jmla.2018.319>.
36. Schmitz-Abe, K., Li, Q., Rosen, S.M., Nori, N., Madden, J.A., Genetti, C.A., Wojcik, M.H., Ponnaluri, S., Gubbels, C.S., Picker, J.D., et al. (2019). Unique bioinformatic approach and comprehensive reanalysis improve diagnostic yield of clinical exomes. *Eur. J. Hum. Genet.* 27, 1398–1405. <https://doi.org/10.1038/s41431-019-0401-x>.
37. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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. <https://doi.org/10.1038/gim.2015.30>.
38. Frazier, Z.J., Brown, E., Rockowitz, S., Lee, T., Zhang, B., Sveden, A., Chamberlin, N.L., Dies, K.A., Poduri, A., Sliz, P., and Chopra, M. (2023). Toward representative genomic research: the children's rare disease cohorts experience. *Ther. Adv. Rare Dis.* 4, 26330040231181406. <https://doi.org/10.1177/26330040231181406>.
39. MD., A. for H.R. and Q.. Rockville, Social Determinants of Health Database. <https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html>.
40. Research Informatics Boston Children's Hospital (2018). HoundDog.

41. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122. <https://doi.org/10.1186/s13059-016-0974-4>.
42. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. <https://doi.org/10.1038/s41586-020-2308-7>.
43. Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 12, 103. <https://doi.org/10.1186/s13073-020-00803-9>.
44. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. <https://doi.org/10.1093/nar/gku1205>.
45. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglou, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 47, D1018–D1027. <https://doi.org/10.1093/nar/gky1105>.
46. Li, Q., and Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am. J. Hum. Genet.* 100, 267–280. <https://doi.org/10.1016/j.ajhg.2017.01.004>.
47. Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M., Nussbaum, R.L., Prabhu, S.A., Boucher, K.M., Biesecker, L.G.; and ClinGen Sequence Variant Interpretation Working Group ClinGen SVI (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* 20, 1054–1060. <https://doi.org/10.1038/gim.2017.210>.
48. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. <https://doi.org/10.1093/nar/gkt1113>.
49. Marcello, J. (2023). HPO3.
50. Schlicker, A., Domingues, F.S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinform.* 7, 302. <https://doi.org/10.1186/1471-2105-7-302>.
51. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am. J. Hum. Genet.* 83, 610–615. <https://doi.org/10.1016/j.ajhg.2008.09.017>.
52. Birgmeier, J., Haeussler, M., Deisseroth, C.A., Steinberg, E.H., Jagadeesh, K.A., Ratner, A.J., Guturu, H., Wenger, A.M., Diekhans, M.E., Stenson, P.D., et al. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* 12, eaau9113. <https://doi.org/10.1126/scitranslmed.aau9113>.
53. Alsentzer, E., Li, M.M., Kobren, S.N., Noori, A., Kohane, I.S., Zitnik, M., and Zitnik, M. (2024). Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. Preprint at medRxiv. <https://doi.org/10.1101/2022.12.07.22283238>.
54. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. <https://doi.org/10.1093/nar/gky1016>.
55. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* 99, 877–885. <https://doi.org/10.1016/j.ajhg.2016.08.016>.
56. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492. <https://doi.org/10.1126/science.adg7492>.
57. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535–548.e24. <https://doi.org/10.1016/j.cell.2018.12.015>.
58. Zhao, Y., Lan, T., Zhong, G., Hagen, J., Pan, H., Chung, W.K., and Shen, Y. (2025). A probabilistic graphical model for estimating selection coefficients of nonsynonymous variants from human population sequence data. *Nat. Commun.* 16, 4670. <https://doi.org/10.1038/s41467-025-59937-2>.
59. Zeng, T., Spence, J.P., Mostafavi, H., and Pritchard, J.K. (2024). Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* 56, 1632–1643. <https://doi.org/10.1038/s41588-024-01820-9>.
60. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., O'Donnell-Luria, A.H., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. <https://doi.org/10.1038/nature19057>.
61. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. <https://doi.org/10.1093/bib/bbs017>.
62. Tavtigian, S.V., Harrison, S.M., Boucher, K.M., and Biesecker, L.G. (2020). Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum. Mutat.* 41, 1734–1737. <https://doi.org/10.1002/humu.24088>.
63. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S.S., Goldstein, J., Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am. J. Hum. Genet.* 100, 895–906. <https://doi.org/10.1016/j.ajhg.2017.04.015>.
64. DiStefano, M.T., Goehringer, S., Babb, L., Alkuraya, F.S., Amberger, J., Amin, M., Austin-Tse, C., Balzotti, M., Berg, J.S., Birney, E., et al. (2022). The Gene Curation Coalition: A global effort to harmonize gene–disease evidence resources. *Genet. Med.* 24, 1732–1742. <https://doi.org/10.1016/j.gim.2022.04.017>.
65. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W.,

- Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* 376, 21–31. <https://doi.org/10.1056/nejmoa1516767>.
66. Guðbjartsson, H., Georgsson, G.F., Guðjónsson, S.A., Valdimarsson, R.þ., Sigurðsson, J.H., Stefánsson, S.K., Másson, G., Magnússon, G., Pálmason, V., and Stefánsson, K. (2016). GORpipe: a query tool for working with sequence data based on a Genomic Ordered Relational (GOR) architecture. *Bioinformatics* 32, 3081–3088. <https://doi.org/10.1093/bioinformatics/btw199>.
 67. Guðbjartsson, H., Ísleifsson, H.þ., Ragnarsson, B., Guimaraes, R., Wu, H., Ólafsdóttir, H., and Stefánsson, S.K. (2022). Ultra-fast joint-genotyping with SparkGOR. Preprint at bioRxiv. <https://doi.org/10.1101/2022.10.25.513331>.
 68. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* 372, 2235–2242. <https://doi.org/10.1056/nejmsr1406261>.
 69. Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., and Van den Brand, T. (2016). *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag).
 70. Veatch, O.J., Mathew, J., Rockowitz, S., Sliz, P., Baldrige, D., Wetzel, A., Niarchou, M., Clarke, M., Berger, S., German, K., et al. (2025). Finding Buried Genetic Test Results in the Electronic Health Record is Inefficient and Variable Across Institutions. *Ther. Adv. Rare. Dis.* 6. <https://doi.org/10.1177/26330040251356521>.