Clinical Evaluation of an AI System for Streamlined Variant Interpretation in Genetic Testing

Jiri Ruzicka[1,#], Jean-Marie Ravel[1,#], Jérôme Audoux[1], Alexandre Boulat[1], Julien Thévenon[2], Kévin Yauy[3], Marine Dancer[4], Laure Raymond[4], Yannis Lombardi[5], Nicolas Philippe[1], Michael GB Blum[1,2], Nicolas Duforet-Frebourg[1,*], Laurent Mesnard[5,6,7]

[1] SeqOne Genomics, Montpellier, France
[2] Institute of Advanced Biosciences, Univ Grenoble Alpes, CNRS UMR 5309, Grenoble, France.
[3] University of Montpellier, LIRMM, CNRS, France
[4] Service de Génétique, Eurofins Biomnis, Lyon, France
[5] Service des Soins intensifs Néphrologiques et Rein Aigu, Hôpital Tenon, Assistance Publique- Hôpitaux de Paris, Paris, France
[6] Centre Maladie rare MAHREA and ERKNET, Hôpital Tenon, Paris, France
[7] INSERM CORAKID, Hopital Tenon, Paris, France
# These authors contributed equally to this work
* Corresponding author

Abstract

Background

The growing use of whole exome/genome sequencing for diagnosing hereditary diseases has increased the interpretive workload for clinical laboratories. Efficient methods are needed to identify pathogenic variants and maximize diagnostic yield without overwhelming resources.

Methods

We developed DiagAI, an AI-powered system trained on 2.5 million ClinVar variants to predict ACMG pathogenicity classes. DiagAI ranks variants, proposes diagnostic shortlists, and identifies probands likely to receive molecular diagnoses. It integrates molecular features, inheritance patterns, and phenotypic data when available. We retrospectively analyzed 966 exomes from a nephrology cohort, including 196 with causal variants and 770 undiagnosed cases.

Results

DiagAI identified 94.9% of causal variants in diagnostic exomes with HPO terms, compared to 90.8% without, with median shortlist sizes of 12 and 9 variants, respectively. It achieved a sensitivity of 57.1% and a specificity of 92.6% in tagging exomes likely to contain a diagnostic variant. With HPO terms, 74% of top-ranked (top 1) variants were diagnostic, versus 42% without, and DiagAI outperformed Exomiser and AIMARRVEL in this setting.

Conclusion

DiagAI generates accurate shortlists of variants that streamline the variant interpretation process. It provides a scalable solution for managing growing diagnostic test volumes without compromising quality.

Introduction

Hereditary diseases are a significant health concern worldwide, and exome sequencing (ES) and whole genome sequencing (WGS) have become essential for their diagnosis[1,2]. Efficient genomic variant interpretation is a critical step in clinical genomics[3].

To provide a molecular diagnosis, a large number of variants detected by high throughput sequencing needs to be interpreted. The American College of Medical Genetics and Genomics (ACMG) offers standardized guidelines for variant classification, grouping variants into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign, and benign. However, these guidelines can lead to discordant classifications among laboratories for a given variant[4] due to the absence of an universal algorithm or precisely defined numerical thresholds, although updated recommendations proposed more quantitative criteria for pathogenicity classification[5].

To address these challenges, artificial intelligence (AI) has emerged as a promising solution. Several studies have demonstrated AI's potential to improve the efficiency of variant interpretation workflows, reduce analysis times, and alleviate the human workload[6–9].

We have developed DiagAI, a commercially available AI-powered system designed to predict ACMG variant classifications and prioritize the most likely causal variants. DiagAI relies on a classifier trained on 2.5 million variants annotated in ClinVar[10]. It integrates expert knowledge, curated datasets, and machine learning models to comprehensively analyze genomics data. Furthermore, when phenotypic information coded with Human Phenotype Ontology (HPO) terms is available, DiagAI upweights genes associated with reported phenotypes to enhance diagnostic precision.

To evaluate the performance of DiagAI, we conducted a retrospective analysis of 966 exomes from patients admitted to an adult nephrology unit. DiagAI successfully shortlisted diagnostic variants in 94.9% of cases when phenotypic information, encoded with HPO terms, was included. In comparison, the success rate was 90.8% when HPO terms were not included. The average shortlist contained 9 variants when using HPO terms and 12 variants without them.

Materials and Methods

Study design

To validate DiagAI, we conducted a retrospective analysis from exome sequencing (ES) data generated from adult participants (n=966) with nephropathy of unknown origin, sequenced from March 2018 to July 2022 (Supplemental Table 1). Of these, 196 (24%) were considered positive cases, defined as containing a causal variant previously identified by a geneticist. The remaining 770 (76%) were considered negative cases, where no diagnosis could be established by a geneticist based on the exome sequencing data.

Exome sequencing

DNA was extracted from peripheral blood using the QIAsymphony DSP DNA Mini Kit on a QIAsymphony instrument following the manufacturer's (QIAGEN) guidelines. Library preparation and capture was performed with Twist reagents (Human Comprehensive Exome or Human Exome 2.0 Plus Comprehensive Exome Spike-in). Sequencing was performed on the Illumina NovaSeq6000 in paired-end mode (2×150 bp reads). Raw data (bcl format) were converted to fastq format using BCL Convert. Reads were aligned to the human reference genome (UCSC Genome Browser build hg37) with Burrows-Wheeler Aligner for maximal exact matches aligner. Calling was performed with an internal procedure, the GermVar pipeline, of SeqOne Genomics.

The GermVar pipeline implements a comprehensive variant detection strategy utilizing multiple variant calling algorithms. For panel-based analyses, the pipeline integrates Freebayes, GATK, GRIDSS, AluMEI, and GATK-Mutect2-Mitochondrial (versions ≥2.0), whereas exome analysis employs Freebayes, GRIDSS, AluMEI, and GATK-Mutect2-Mitochondrial. This strategic combination of callers enables robust detection of various variant types, including single nucleotide variants (SNVs), multiple nucleotide variants (MNVs), and insertions/deletions (indels).

Detection sensitivity parameters are optimized according to the analysis type. In panel-based analyses, the Freebayes algorithm is configured to detect variants with allele frequencies ≥5%, while exome analysis maintains a more stringent threshold of ≥10% for SNVs. Notably, GRIDSS, AluMEI, and GATK operate independently of variant allele frequency thresholds for small variant detection, allowing for maximum sensitivity in structural variant identification..

DiagAI priorisation algorithm

SeqOne DiagAI is a suite of tools that prioritizes genetic variants (both SNVs and indels) using molecular and phenotypic information.

A molecular pathogenicity score UP[2] (Universal Pathogenicity Predictor) is computed using a machine-learning classifier that predicts the ACMG class for each variant observed in the sequencing data. The training dataset for the classifier comprised 2.5 millions variants from ClinVar, labeled according to the five ACMG pathogenicity categories. Each variant is represented by 72 molecular features derived from various evidence sources and related to the ACMG criteria (PVS1, PS1, *etc.*). The feature values were calculated using information from several databases including ClinVar for variant annotations and interpretations, Gnomad v4.1 for population frequencies and constraint metrics, dbNSFP dbscSNV for *in silico* predictors, CI-SpliceAI for in silico splice prediction[11], VEP transcript annotations and effect prediction[12], and RepeatMasker for the distance to repetitive genomic regions (related to the PM4 criterion).

A comprehensive prioritization score, referred to as the DiagAI score, was built upon the UP[2] molecular score by incorporating additional contextual information when available. This includes phenotypic data, variant inheritance mode consistency from databases such as PanelApp[13], OMIM, and MedGen, variant calling data (DP, AO, VAF) and quality (base quality phred score), and parental variant data in cases where trio sequencing was performed. Phenotypic information was incorporated using HPO terms, with gene prioritization performed by Phenogenius, a tool that leverages gene-HPO term associations from literature-based matrices[14]. To effectively integrate ACMG-related molecular features with HPO-based phenotypic features, a machine learning regression model was trained using a dataset of 307 diagnosed exomes from multiple cohorts, encompassing a total of 26 million variants.

To build a shortlist of variants most likely to be causal, two thresholds—one for the molecular UP[2] score and another for the comprehensive DiagAI score—were applied to define the shortlist. The DiagAI threshold values varied depending on whether HPO terms were incorporated. A third threshold was applied to discard variants with a frequency above 1.5% in the phenotypically matched cohort.

In addition to generating a shortlist of variants, DiagAI identifies exomes or genomes most likely to result in a molecular diagnosis. An exome is tagged only if it contains at least one

variant, referred to as a 'smartpick,' with both DiagAI and UP^2 scores exceeding predefined thresholds. While a shortlist is generated for every exome or genome, only some are tagged as likely to lead to a molecular diagnosis, while others remain untagged.

Interpretability features

To assess the relative importance of different ACMG criteria in our classification model, we calculated Shapley values for all 72 features used in the molecular pathogenicity score. These Shapley values were then aggregated in order to be mapped to the known ACMG criteria. This approach allowed us to quantify the contribution of each ACMG tag to the molecular UP^2 score, providing insight into which criteria were most influential in determining variant pathogenicity according to our model. To calculate a Shapley value for each ACMG criterion, we sum the Shapley values of all features associated with that specific criterion.

Evaluation of performance

To benchmark the performance of DiagAI, we compared its ranking performance to AI-MARRVEL and Exomiser (v13), which prioritizes genes or variants by leveraging information on variant frequency, predicted pathogenicity, inheritance modes, and gene-phenotype association [6,15].

Exomiser13 was run using default pathogenicity sources MVP and REVEL, and failedVariantFilter, inheritanceFilter, frequencyFilter and pathogenicityFilter with the keepNonPathogenic option set to true. After filtering the OmimPrioritizer and hiPhivePrioritizer steps were used. AI-MARRVEL ran with the lite version, and no access to the HGMD resource. Because DiagAI uses in its scoring a filter profile based on quality, variant allele frequency, depth and number of observed alternate alleles, we applied the same filtering prior to the run of Exomiser13 and AI-MARRVEL. We evaluated the rankings at the gene level. For Exomiser13 we used the gene ranks from the json output file. For both DiagAI and AI-MARRVEL genes were ranked based on the highest-priority variant among all their associated variants.

Results

Proportion of Causal Variants Identified in Shortlists

For exomes with a confirmed molecular diagnosis, 94.9% (186/196) of causal variants were contained in the shortlist when HPO terms were used, compared to 90.8% (176/196) when

HPO terms were not used. The median shortlist size was 9 variants when HPO terms were not used (min=2, max=44) and 12 variants when HPO terms were used (min=4, max=29).

Sensitivity and Specificity of DiagAI in Diagnosing Exomes

DiagAI demonstrated a sensitivity of 57.1% (112/196) in tagging probands likely to result in a molecular diagnosis. Among 770 exomes that did not result in a diagnosis, DiagAI correctly excluded 713 exomes while incorrectly tagging 57 as likely diagnosable. This corresponds to a specificity of 92.6%.

Variant ranking

We compared DiagAI's variant ranking accuracy with and without utilizing HPO-based clinical information (Figure 1) on the 196 exomes with a confirmed diagnostic variant. Incorporating clinical data significantly improved the ranking accuracy for the top-ranked variant. Specifically, 42% of top-ranked variants were diagnostic when HPO terms were not included, whereas this percentage increased to 74% when HPO terms were accounted for. The improvement was less pronounced when considering a larger list of 20 genes, with 93% of diagnostic variants included without HPO terms versus 97% with HPO terms. DiagAI achieved improved ranking performance compared to Exomiser v13 and AI-MARRVEL when HPO terms were provided, and for top-ranked lists of three or more variants when HPO terms were not included (Figure 1).

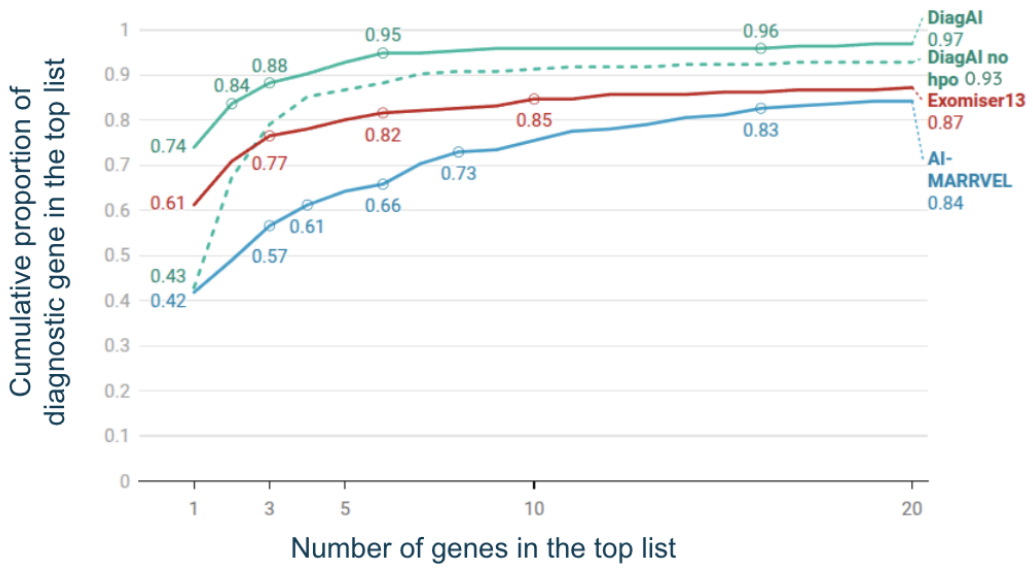

Figure 1: **Performance of variant ranking evaluated on the 196 exomes with a confirmed molecular diagnosis.** The ranking accuracy of top-ranked genes was assessed using two approaches when evaluating DiagAI: the molecular pathogenicity score alone (no

HPO) and a comprehensive score integrating molecular pathogenicity and HPO-based clinical information (with HPO). Variants with a cohort frequency above 1.5% were excluded from the DiagAI ranking. Exomiser v13 and AI-MARRVEL, which also use HPO for gene ranking, were used for comparison.

Interpretability of the classifier

We applied our interpretability framework to identify the ACMG criteria most influential in determining the ACMG classifications of 176 diagnostic variants, which were identified across 196 exomes with a molecular diagnosis, noting that some variants were shared between exomes. Among the 85 diagnostic variants with ClinVar submissions, features related to the number of pathogenic and benign submissions were the most influential (Figure 2). However, for 13 variants, other ACMG criteria were more impactful, including 6 with PP3/BP4 (in-silico predictors of pathogenicity and benignity), 5 with PVS1 (predicted impact by VEP), and 2 with PM2/BA1 (absent or at extremely low frequency in general population).

For the 91 diagnostic variants without ClinVar submissions, PP3/BP4 features were instrumental for 66% (60/91) of the variants, while PVS1 features were key for 29% (26/91). The scores for the remaining five variants were primarily determined by a combination of other ACMG features.

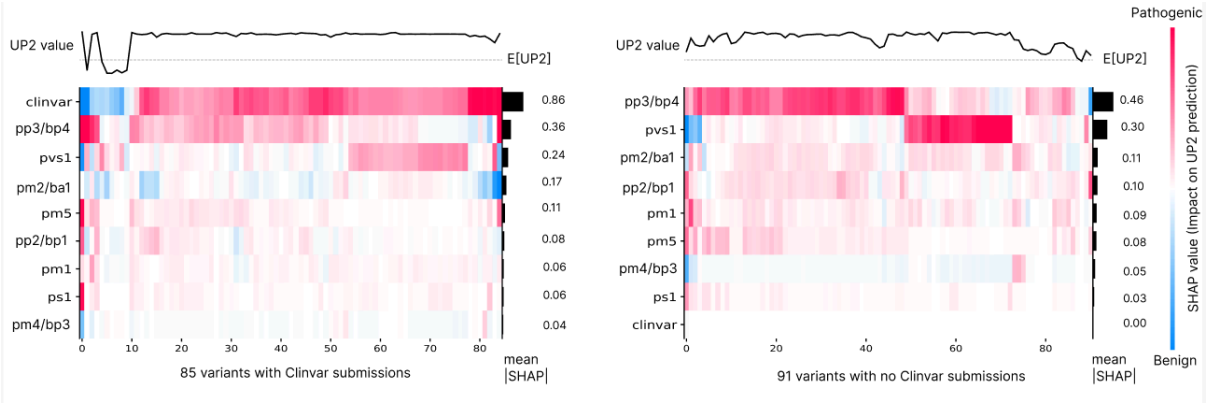

Figure 2: Explicability profile of the 176 diagnostic variants for the Universal Pathogenicity Predictor. The heatmap shows for each variant if the impact of the features related to an ACMG criteria is rather pathogenic (red) or benign (blue). Variants that have ClinVar Submissions (Panel A, left) have UPP predictions that are mostly influenced by the ClinVar submission predictors with an average absolute SHaP value of .86 for ClinVar predictors.

Variants that have no ClinVar Submission (Panel B, right) have UPP predictions that are mostly influenced by features from PP3/BP4 or PVS1 criteria. The right side histogram

shows that the average absolute SHaP value is .46 for PP3/BP4 and .30 for PVS1 on the UP^2 value for variants without ClinVar submission. The top plots show UP^2 values that are between -1 and 1.

Causal variants absent from the shortlists

A total of 10 diagnostic variants were absent from the DiagAI shortlists, despite being considered diagnostic by the geneticists. Of these, five were classified as benign in ClinVar, and three were labeled as variants of uncertain significance (VUS) or had conflicting classifications in ClinVar, with no additional evidence supporting pathogenicity. The remaining two variants were missed by DiagAI due to specific limitations: one was a low-quality variant, and the other was located in a recessive gene without a second heterozygous variant with a sufficiently high DiagAI score to support a composite diagnosis.

Discussion:

Our study demonstrates DiagAI's effectiveness in prioritizing causal variants in exomes from nephrology patients, analyzed as single cases rather than trios. Depending on the availability of phenotypic data, 90-96% of shortlists contained the causal variant. DiagAI also outperformed Exomiser v13 and AI-MARRVEL in gene-level ranking comparisons.

We examined some of the causal variants missed by the shortlist of DiagAI. We identified one variant in the PODXL gene, whose pathogenic effect has been linked to kidney diseases[16]. We also identified a variant in *CFI*, which dysregulates the complement alternative pathway. However, *CFI* variants are challenging to interpret; they are difficult to classify using ACMG criteria[17], and due to their incomplete penetrance, they are often considered risk factors rather than causative in a Mendelian disease[18]. Additionally, a synonymous variant in *NPHP3* with a splicing effect, not captured by computational approaches, was also overlooked[19]. Overall, these variants were missed by DiagAI but were identified as diagnostic by geneticists thanks to recent research documenting their pathogenicity, combined with expert knowledge of genes involved in kidney diseases.

These cases illustrate the inherent challenges in variant interpretation, particularly for variants with emerging or complex pathogenic mechanisms that are not yet well captured by computational models. While DiagAI improves variant prioritization by leveraging multivariate ACMG evidence tags and machine learning trained on ClinVar ACMG classifications, it remains limited by the available knowledge and data used for training. By integrating diverse

evidence sources and phenotypic data encoded with HPO terms, DiagAI enhances variant ranking, but expert review remains essential for capturing novel or particularly challenging cases.

DiagAI contributes to the growing landscape of computational solutions for variant prioritization, joining both open-source and commercial offerings such as AI-MARRVEL, InVitae MOON, Fabric GEM, and the Emedgne software from Illumina[6,8,9,20]. Direct comparisons between these tools are challenging due to differences in their evaluation cohorts. Nonetheless, we found that DiagAI outperformed AI-MARRVEL in gene ranking and demonstrated comparable performance in identifying diagnosable cases, with both tools automatically detecting 50–60% of such cases[6]. For a fair comparison, the Critical Assessment of Genome Interpretation (CAGI) challenge offers a standardized benchmarking framework; however, DiagAI has yet to be evaluated within this context.

DiagAI's accuracy in variant ranking, particularly when integrating clinical data, highlights its potential to streamline genomic diagnostics by reducing the number of variants requiring manual review. However, the path to full automation remains long, with less than 60% of diagnosed cases detected automatically. These findings suggest that AI-powered tools like DiagAI can significantly reduce the interpretive workload in clinical genomics while maintaining high diagnostic accuracy. This assessment should be evaluated beyond the specific case of nephrology and further tested in the context of whole genome sequencing analysis.

Bilbiography:

1. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *N Engl J Med*. 2021;385(20):1868-1880. doi:10.1056/NEJMoa2035790
2. Doreille A, Lombardi Y, Dancer M, et al. Exome-First Strategy in Adult Patients With CKD: A Cohort Study. *Kidney Int Rep*. 2023;8(3):596-605. doi:10.1016/j.ekir.2022.12.007
3. Austin-Tse CA, Jobanputra V, Perry DL, et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. *NPJ Genomic Med*. 2022;7(1):27.
4. Amendola LM, Muenzen K, Biesecker LG, et al. Variant Classification Concordance using the ACMG-AMP Variant Interpretation Guidelines across Nine Genomic Implementation Research Studies. *Am J Hum Genet*. 2020;107(5):932-941. doi:10.1016/j.ajhg.2020.09.011
5. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*. 2022;109(12):2163-2177.
6. Mao D, Liu C, Wang L, et al. AI-MARRVEL — A Knowledge-Driven AI System for Diagnosing Mendelian Disorders. *NEJM AI*. 2024;1(5). doi:10.1056/AIoa2300009
7. Owen MJ, Lefebvre S, Hansen C, et al. An automated 13.5 hour system for scalable diagnosis and acute management guidance for genetic diseases. *Nat Commun*.

2022;13(1):4057.
8. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. *Genome Med*. 2021;13(1):153. doi:10.1186/s13073-021-00965-0
9. Meng L, Attali R, Talmy T, et al. Evaluation of an automated genome interpretation model for rare disease routinely used in a clinical genetic laboratory. *Genet Med*. 2023;25(6):100830.
10. Landrum MJ, Chitipiralla S, Brown GR, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res*. 2020;48(D1):D835-D844. doi:10.1093/nar/gkz972
11. Strauch Y, Lord J, Niranjan M, Baralle D. CI-SpliceAI—Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. *PLOS ONE*. 2022;17(6):e0269159. doi:10.1371/journal.pone.0269159
12. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4
13. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet*. 2019;51(11):1560-1565. doi:10.1038/s41588-019-0528-2
14. Yauy K, Duforet-Frebourg N, Testard Q, et al. Learning phenotypic patterns in genetic disease by symptom interaction modeling. Published online July 31, 2022:2022.07.29.22278181. doi:10.1101/2022.07.29.22278181
15. Cipriani V, Pontikos N, Arno G, et al. An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data. *Genes*. 2020;11(4):460. doi:10.3390/genes11040460
16. Blasco M, Quiroga B, García-Aznar JM, et al. Genetic Characterization of Kidney Failure of Unknown Etiology in Spain: Findings From the GENSEN Study. *Am J Kidney Dis Off J Natl Kidney Found*. 2024;84(6):719-730.e1. doi:10.1053/j.ajkd.2024.04.021
17. Schwotzer N, Fakhouri F, Martins PV, et al. Hot Spot of Complement Factor I Rare Variant p.Ile357Met in Patients With Hemolytic Uremic Syndrome. *Am J Kidney Dis*. 2024;84(2):244-249. doi:10.1053/j.ajkd.2023.12.021
18. Timmermans SAMEG, van Doorn DPC, van Paassen P. Rare Variants in Complement Genes May Not Be That Rare After All. *Kidney Int Rep*. 2023;8(10):1911-1913. doi:10.1016/j.ekir.2023.08.020
19. Molinari E, Decker E, Mabillard H, et al. Human urine-derived renal epithelial cells provide insights into kidney-specific alternate splicing variants. *Eur J Hum Genet EJHG*. 2018;26(12):1791-1796. doi:10.1038/s41431-018-0212-5
20. O'Brien TD, Campbell NE, Potter AB, Letaw JH, Kulkarni A, Richards CS. Artificial intelligence (AI)-assisted exome reanalysis greatly aids in the identification of new positive cases and reduces analysis time in a clinical diagnostic laboratory. *Genet Med*. 2022;24(1):192-200.

Data Availability

Data should be requested to Prof L Mesnard.

Conflict of Interest

J. Ruzicka, J.-M. Ravel, J. Audoux, A. Boulat, K. Yauy, N. Philippe, M. Blum, and N. Duforet-Frebourg are current or former employees of SeqOne Genomics and hold, or have received, stock or stock options from the company. M. Dancer and L. Raymond are employees of Eurofins Biomnis.

Acknowledgments

We would like to thank the technicians, scientists, bioinformaticians, operations staff, and biologists at Eurofins Biomnis for their valuable contributions to the production and interpretation of these data.

Funding

This study was conducted without additional funding.

Author Information

Conceptualization: M.G.B., N.P., N.D.-F., L.M.; Methodology: J.R., K.Y., N.D.-F., J.T.; Data Curation: J.R., A.B.; Data Analysis: J.R., N.D.-F.; Data Production: L.M., M.D., Y.L., L.R.; Software: J.R., N.D.-F., J.A., N.P.; Supervision: M.G.B., N. D.-F.; Writing – Original Draft: M.G.B.; Writing – Review and Editing: M.G.B., J.-M.R., L.M.

Ethics Statement

The cohort has been previously described[2]. The study was approved by an Institutional Review Board (Direction de la Recherche Clinique et de l'Innovation (APHP220461) and the Ethic board of Sorbonne Université (CER-2022-009).