

ORIGINAL ARTICLE

Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland

Caroline F. Wright, Ph.D., Patrick Campbell, M.B., B.S., Ruth Y. Eberhardt, Ph.D., Stuart Aitken, Ph.D., Daniel Perrett, M.Phil., Simon Brent, B.Sc., Petr Danecek, Ph.D., Eugene J. Gardner, Ph.D., V. Kartik Chundru, Ph.D., Sarah J. Lindsay, Ph.D., Katrina Andrews, M.B., B.Ch., Juliet Hampstead, B.Sc., Joanna Kaplanis, Ph.D., Kaitlin E. Samocha, Ph.D., Anna Middleton, Ph.D., Julia Foreman, Ph.D., Rachel J. Hobson, Ph.D., Michael J. Parker, Ph.D., Hilary C. Martin, Ph.D., David R. FitzPatrick, M.D., Matthew E. Hurles, Ph.D., and Helen V. Firth, D.M., for the DDD Study*

ABSTRACT

BACKGROUND

Pediatric disorders include a range of highly penetrant, genetically heterogeneous conditions amenable to genomewide diagnostic approaches. Finding a molecular diagnosis is challenging but can have profound lifelong benefits.

METHODS

We conducted a large-scale sequencing study involving more than 13,500 families with probands with severe, probably monogenic, difficult-to-diagnose developmental disorders from 24 regional genetics services in the United Kingdom and Ireland. Standardized phenotypic data were collected, and exome sequencing and microarray analyses were performed to investigate novel genetic causes. We developed an iterative variant analysis pipeline and reported candidate variants to clinical teams for validation and diagnostic interpretation to inform communication with families. Multiple regression analyses were performed to evaluate factors affecting the probability of diagnosis.

RESULTS

A total of 13,449 probands were included in the analyses. On average, we reported 1.0 candidate variant per parent-offspring trio and 2.5 variants per singleton proband. Using clinical and computational approaches to variant classification, we made a diagnosis in approximately 41% of probands (5502 of 13,449). Of 3599 probands in trios who received a diagnosis by clinical assertion, approximately 76% had a pathogenic de novo variant. Another 22% of probands (2997 of 13,449) had variants of uncertain significance in genes that were strongly linked to monogenic developmental disorders. Recruitment in a parent-offspring trio had the largest effect on the probability of diagnosis (odds ratio, 4.70; 95% confidence interval [CI], 4.16 to 5.31). Probands were less likely to receive a diagnosis if they were born extremely prematurely (i.e., 22 to 27 weeks' gestation; odds ratio, 0.39; 95% CI, 0.22 to 0.68), had in utero exposure to antiepileptic medications (odds ratio, 0.44; 95% CI, 0.29 to 0.67), had mothers with diabetes (odds ratio, 0.52; 95% CI, 0.41 to 0.67), or were of African ancestry (odds ratio, 0.51; 95% CI, 0.31 to 0.78).

CONCLUSIONS

Among probands with severe, probably monogenic, difficult-to-diagnose developmental disorders, multimodal analysis of genomewide data had good diagnostic power, even after previous attempts at diagnosis. (Funded by the Health Innovation Challenge Fund and Wellcome Sanger Institute.)

From the Department of Clinical and Biomedical Sciences, University of Exeter Medical School, Royal Devon and Exeter Hospital, Exeter (C.F.W.), the Wellcome Sanger Institute (P.C., R.Y.E., D.P., S.B., P.D., E.J.G., V.K.C., S.J.L., K.A., J.H., J.K., K.E.S., J.F., R.J.H., H.C.M., M.E.H., H.V.F.), European Molecular Biology Laboratory, European Bioinformatics Institute (D.P., S.B., J.F.), and Wellcome Connecting Science (A.M.), Wellcome Genome Campus, and Cambridge University Hospitals Foundation Trust, Addenbrooke's Hospital (P.C., H.V.F.), Cambridge, the Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh (S.A., D.R.F.), and the Wellcome Centre for Ethics and Humanities, Ethox Centre, Oxford Population Health, Big Data Institute, University of Oxford, Oxford (M.J.P.) — all in the United Kingdom. Dr. Wright can be contacted at caroline.wright@exeter.ac.uk or at the Department of Clinical and Biomedical Sciences (Medical School), Faculty of Health and Life Sciences, University of Exeter, RILD Building, Barrack Rd., Exeter EX2 5DW, United Kingdom.

*The scientists and collaborators in the DDD study are listed in the Supplementary Appendix, available at NEJM.org.

This article was published on April 12, 2023, and updated on April 27, 2023, at NEJM.org.

N Engl J Med 2023;388:1559-71.

DOI: 10.1056/NEJMoa2209046

Copyright © 2023 Massachusetts Medical Society.

GENOMIC SEQUENCING HAS MADE extraordinary strides toward identifying novel molecular causes for rare monogenic disorders and is becoming increasingly available in diagnostic clinics throughout the world.^{1,2} Pediatrics has particularly benefited from the use of high-throughput next-generation sequencing technologies, partly because of the high clinical need and potential for lifelong benefit with diagnosis and treatment.³ In addition, an early presentation of a patient with severe disease makes genetic diagnosis more feasible because causal variants are largely absent from control data sets.⁴

Progress in the genomic study of rare pediatric diseases has been spearheaded by numerous diagnostic research groups across the world.^{5,6} One of the first studies to combine large-scale genomic research with individual patient feedback was the Deciphering Developmental Disorders (DDD) study,⁷⁻⁹ which recruited more than 13,500 families and generated exome sequencing and microarray data, which were complemented by rich clinical phenotypes recorded by more than 200 clinicians across the United Kingdom and Ireland. Here, we describe analytic strategies developed over a decade in the DDD study to identify and classify thousands of new molecular diagnoses and report the factors affecting the probability of receiving a diagnosis.

METHODS

STUDY OVERVIEW

The DDD study was approved by the Cambridge South Research Ethics Committee in the United Kingdom and the National Research Ethics Committee in the Republic of Ireland. A multicenter research collaboration was set up with all 24 regional genetics services, and a management committee (comprising clinicians, scientists, and a bioethicist) was created to provide ongoing ethical oversight (Table 1). In addition to genomic and data scientists, a social scientist was employed to perform ethics research.¹⁰

COHORT

A total of 13,610 probands (88% of whom were recruited in family [parent–offspring] trios) were ascertained and recruited from April 2011 through April 2015 by consultant clinical geneticists, assisted by research nurses and genetic counselors.

Written informed consent for participation was obtained from the families. Eligibility criteria included the presence of any of the following: neurodevelopmental disorders, congenital anomalies, abnormal growth measurements (>4 SDs above the mean for a single measurement or >3 SDs above the mean for ≥ 2 measurements), dysmorphic features, unusual behavioral phenotypes, and genetic disorders that have large effects but for which the molecular basis was unknown. The study was initially limited to probands younger than 16 years of age at the date of recruitment, but this age limit was later removed (except in Scotland).

Most probands had previously undergone clinical chromosomal microarray analysis (85%) or single-gene testing with or without chromosomal microarray analysis (53%) but had not received a diagnosis. Probands were assigned pseudonymized identifications, and basic clinical information, quantitative growth data, developmental milestones, and Human Phenotype Ontology (HPO)¹¹ terms were recorded for all participants by way of a bespoke standardized interface in DECIPHER (Database of Genomic Variants and Phenotype in Humans Using Ensembl Resources; <https://www.deciphergenomics.org>).¹²

GENOMIC ANALYSES

Detailed assay protocols^{13,14} and variant-filtering pipelines^{7,15} have been described elsewhere (see the Supplementary Appendix, available with the full text of this article at NEJM.org). Briefly, the following three independent genomic assays were performed: exome sequencing of DNA samples from complete family trios and singleton probands (i.e., those who had not been recruited in a family trio), exon-focused array comparative genomic hybridization (aCGH) of DNA from probands, and genomewide single-nucleotide polymorphism (SNP) genotyping of DNA from probands. Assay designs remained largely unchanged throughout the study.

Data sets were processed in batches, and multiple different algorithms were used to detect and annotate sequence and structural variants (Fig. 1). The inheritance status of variants in the proband was determined by means of comparison with parental data.¹⁶ For clinical reporting, we selected high-quality, rare, nonsynonymous variants overlapping genes in the Developmental Disorders Gene2Phenotype (DDG2P) database¹⁷

with the appropriate zygosity and inheritance pattern when available. We augmented this pipeline with additional analyses to find missing, probably causal variants, including “pathogenic” or “likely pathogenic” variants in the ClinVar database¹⁸ and de novo variants that were mosaic,¹⁹ created upstream open reading frames,²⁰ affected splicing,²¹ or were coding insertions or deletions of intermediate size²² or were caused by mobile element insertions²³ or mosaic chromosomal alterations.²⁴

DEFINING A DIAGNOSIS

Candidate diagnostic variants that had been identified bioinformatically were evaluated for analytic and clinical validity by a central clinical review panel before being reported in batches to regional genetics teams by way of DECIPHER (April 2014 through April 2022) (Fig. 2). The referring clinician then evaluated the reported variant or variants, requested diagnostic laboratory confirmation when warranted, and communicated diagnoses to the families.

At the time of this analysis, clinical classifications of variant pathogenicity (i.e., “benign,” “likely benign,” “uncertain,” “likely pathogenic,” or “pathogenic”) and contribution to the phenotype (full, partial, unknown, or none) were recorded in DECIPHER for 84% of variants. These were supplemented by automated predictions for selected variant classification criteria (see the Supplementary Appendix) based on published guidelines from the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG–AMP)²⁵ and from the Association for Clinical Genetic Science (ACGS).²⁶ A provisional variant classification was calculated with the use of a log-additive Bayesian framework described elsewhere (see the Supplementary Appendix).²⁷ Variants with a posterior probability of greater than 0.90 to 0.99 were classified as likely pathogenic, and those with a posterior probability of greater than 0.99 to 1.00 were classified as pathogenic; variants with a posterior probability of 0.001 to less than 0.100 were classified as likely benign, and those with a posterior probability of 0 to less than 0.001 were classified as benign. For genes with 10 or more pathogenic or likely pathogenic variants, computational phenotype matching was performed with the use of a Bayesian classifier called

IMPROVE-DD (Integrating Multiple Phenotype Resources Optimizes Variant Evaluation in Genetically Determined Developmental Disorders),²⁸ with the application of the same Bayesian framework to combine variant classifications and gene–disease models; phenotype-based likelihoods were scaled appropriately and used at the evidence equivalent of “strong.”²⁷ The number of diagnoses per gene is provided in Table S6 in the Supplementary Appendix.

Probands were classified as having received a diagnosis if one or more variants or two or more compound heterozygous variants were annotated as pathogenic or likely pathogenic by either the proband’s referring clinician or according to the predicted classification and if the contribution to the phenotype was not clinically annotated as “none.” Factors influencing the chance of receiving a diagnosis (based on clinical annotation only) were investigated with the use of multivariable logistic regression among 13,368 probands for whom complete clinical and demographic data were available (see the Supplementary Appendix).

RESULTS

COHORT CHARACTERISTICS

The DDD study included 13,449 probands (9859 in family trios) who had severe, previously undiagnosed developmental disorders and had available exome sequencing, exon-focused aCGH, and SNP genotyping data. The probands were recruited from across the United Kingdom and Ireland (median recruitment per center, 216 probands per million population; range, 69 to 588). The median age of the probands was 7 years (range, 0 to 63) at the time of recruitment, and for parents, the median age was 31 years (range, 15 to 70) at the time of the proband’s birth. Among the probands, 42% were female, and 16% were of non-European ancestry (Table S1). The median number of HPO terms recorded per proband, including 65% of the probands with global developmental delay or intellectual disability, was 6 (range, 1 to 36), and 72% of the probands were the only affected member of their family.

GENETIC FINDINGS

At the time of this analysis, 19,285 potentially pathogenic sequence and structural variants have been identified among the probands in the study

Table 1. Ethical Considerations in the DDD Study.*

Ethical Domain	Key Issues	Resolution of Issues in the Study
Building and maintaining partnerships at clinical–research interface	Trust among researchers, clinicians, and patients and their families should be ensured. A trade-off should be found between creating a large research data set and maintaining small clinical cohorts. Practical ethical considerations should be managed throughout the life cycle of the study.	Scientific scope was limited to understanding the causes of developmental disorders. Local training sessions and regular discussions with stakeholders were held with respect to project planning and decisions. Regular study management committee meetings and annual national collaborator meetings were held.
Recruitment, consent, capacity, and eligibility	Most probands in the study lacked the capacity to provide informed consent, because of either young age or intellectual disability. Recruitment of underrepresented ethnic groups in research studies is a challenge. Initial study eligibility was limited to persons <16 years of age, which created inequity; however, recruitment of adults who lack the capacity to provide informed consent is extremely challenging, especially in Scotland. Confidentiality of study participants should be protected when possible.	Detailed consent materials and a website were developed for families and guardians. Study consent materials were translated into several different languages. New consent materials were written, and recruitment was opened to adults with or without the capacity to provide informed consent in England, Wales, Northern Ireland, and Ireland. Pseudonymized identifications were used throughout the study; minimum data required for research were stored in DECIPHER, and personal identifiable data (e.g., date of birth) were not stored in DECIPHER.
Sample inclusion, collection, and verification	Balance should be found between the scientific benefit of obtaining DNA samples from parents and clinical concerns about the scope of genetic information and data management. Many children with developmental disorders are very distressed by hospital visits to have blood samples drawn. Sample mix-ups could occur in the study (e.g., within families, at recruitment centers, or at the Wellcome Sanger Institute). The potential exists for the detection of incest or misattributed parentage.	Parents were recruited in the study with the agreement that their data would only be used when relevant to the understanding of their child's disorder. Saliva sample kits were used to collect child and parental samples, thereby allowing sample collection at home. Genetic “barcodes” were created for all samples by genotyping 60 SNPs; individual and family data were cross-checked. Potential safeguarding issues were flagged to the referring clinical team; discordant samples or biologically unrelated parents were excluded from further analysis.
Sharing clinically relevant variants	Public opinion about feedback of incidental findings from genomics research was largely unknown and unexplored. Balance should be found between the benefits and harms of returning different types of clinically actionable findings. Pertinent findings (i.e., potentially relevant to the child's developmental disorder) were deemed to be within the scope of research study and clinical testing, in which case the benefits are likely to outweigh the harms to the family. Incidental findings and findings with unclear relevance, particularly in children, were deemed to be outside the scope and expertise of clinicians and researchers, in which case the harms to the family are likely to outweigh the benefits.	An ethics and social science researcher was embedded in the study to investigate attitudes among the public, patients, scientists, and health care professionals toward the feedback of incidental findings in genomics. Study documentation stated that pertinent findings would be reported to clinical teams for communication with families but incidental findings would not. The DDG2P database and variant-filtering rules were developed to select plausibly pathogenic variants for reporting into linked DECIPHER records; genes in the DDG2P database that were also associated with adult-onset diseases were flagged for review. Pathogenic variants and phenotypes were shared openly by way of DECIPHER once a family had been informed.
Sharing genomewide variants	Requests were received from parents in the study for genomic data to be returned directly to them. Access to research data should be prioritized for the hundreds of scientists and clinicians involved in the recruitment of and provision of care for families in the study. Research data should be shared widely with external researchers to advance research, but data sets are sensitive, since they relate to severely unwell children, and consent to access data is limited to the purpose of understanding the causes of developmental disorders.	Requests for genomic data regarding the individual or family were declined on the basis of concerns about sample identity, a lack of resources to provide informatics-related support, and potential inability to mitigate against unintended consequences. The Collaborative Analysis Project system was created, in which research plans were reviewed by the management committee and data were shared by means of a secure file-transfer protocol. Genomic data were shared with bona fide researchers under managed access by way of the European Genome–Phenome Archive; anonymized variants of potential relevance were shared through DECIPHER as research variants.

Table 1. (Continued.)

Ethical Domain	Key Issues	Resolution of Issues in the Study
Managing withdrawal	<p>Study participants were allowed to withdraw from the study at any time; therefore, a range of actions was required to manage samples, data, and associated records.</p> <p>Previously shared data cannot be withdrawn and may be required to support published findings.</p>	<p>If a withdrawal request was received, related samples were destroyed, unshared data were removed, and individual DECIPHER records were deleted to break any link to the family.</p> <p>Data in previously published data sets were not withdrawn, as stated in consent materials.</p>

* The Deciphering Developmental Disorders (DDD) study depended on the integration of ethics into decision making and collaboration building, both upfront and throughout the project, a factor that allowed important ethical questions to be identified and ethical policies to be developed through a consensual process. DECIPHER denotes Database of Genomic Variants and Phenotype in Humans Using Ensembl Resources, DDG2P Developmental Disorders Gene2Phenotype, and SNP single-nucleotide polymorphism.

and returned to referring clinicians through up to six rounds of iterative reanalysis and batch reporting involving 18 different variant-detection algorithms (Fig. 1 and Table S2).^{7,15} The majority of variants were identified with the use of the DDG2P database, a clinically curated database of 1840 genes associated with developmental disorders,¹⁷ which was updated at a rate of approximately 100 genes per year through literature curation and cohortwide enrichment analyses. Included in DDG2P were 60 genes associated with developmental disorders that had been newly identified in the study (Fig. 2).^{5,13,14,29,30}

A total of 44% of the reported variants were in genes that had been added to the DDG2P database after the first round of reporting in 2014. The majority of reported variants were single-nucleotide variants and small insertions or deletions that were detected with the use of exome sequencing data (71% were protein-altering, 19% protein-truncating, and 3% noncoding variants); structural variants were identified through a combination of microarray and exome sequencing analyses (6% were copy-number variants and 1% other structural variants) (Fig. 2). On average, 1.0 variant was reported per family trio and 2.5 variants were reported per singleton proband (Fig. S1), and probands with non-European ancestry had more variants reported (Fig. S2). Each new round of analysis resulted in approximately 1 additional variant being reported for every six family trios (Fig. S3).

In findings consistent with those of similar studies,^{1,3,31} de novo variants in the proband and variants inherited from a mosaic parent (i.e., postzygotic parental de novo variants) in dominant genes provided the highest diagnostic yield, with 79% of reported variants being clinically classified as pathogenic or likely pathogenic. In

contrast, 32% of variants in autosomal recessive conditions, 23% of those that were maternally inherited on the X chromosome, and 11% of those in autosomal dominant conditions inherited from an affected parent or with unknown inheritance were clinically classified as pathogenic or likely pathogenic (Fig. S4).

Concordance between clinical and predicted classifications of variant pathogenicity and benignity was high, with 4425 concordant variants — a result that corresponds with a sensitivity of our hybrid approach of 99.5%, a specificity of 85.0%, a positive predictive value of 96.5%, and a negative predictive value of 97.9% (Fig. S6).²⁵⁻²⁷ Discrepancies (149 nonconcordant variants [3%]) were due to false positive variant calls, incorrect clinical classifications (e.g., atypical disease presentations), or inappropriate ACMG–AMP or ACGS criteria assignment (e.g., incorrect disease mechanisms). On the basis of concordance between clinical and predicted classifications of variant pathogenicity, we estimated that a minimum of 3347 of the 13,449 probands (25%) received a diagnosis; this proportion increased to 4363 (32%) for probands who received a predicted diagnosis, to 4484 (33%) for probands with a diagnosis through clinical assertion, and to 5502 (41%) for probands with a clinical or predicted diagnosis (Fig. 3).

Among the 4484 probands who received a diagnosis by means of clinical assertion, 3599 were in family trios, 2750 of whom (76%) had a pathogenic de novo variant (Table S5). Of these 4484 probands, 561 (13%) received a partial diagnosis, and 121 (3%) received two or more different genetic diagnoses, potentially resulting in a composite phenotype (with the inclusion of computational predictions, this proportion increased to 359 of 5502 probands [7%]).³³ Al-

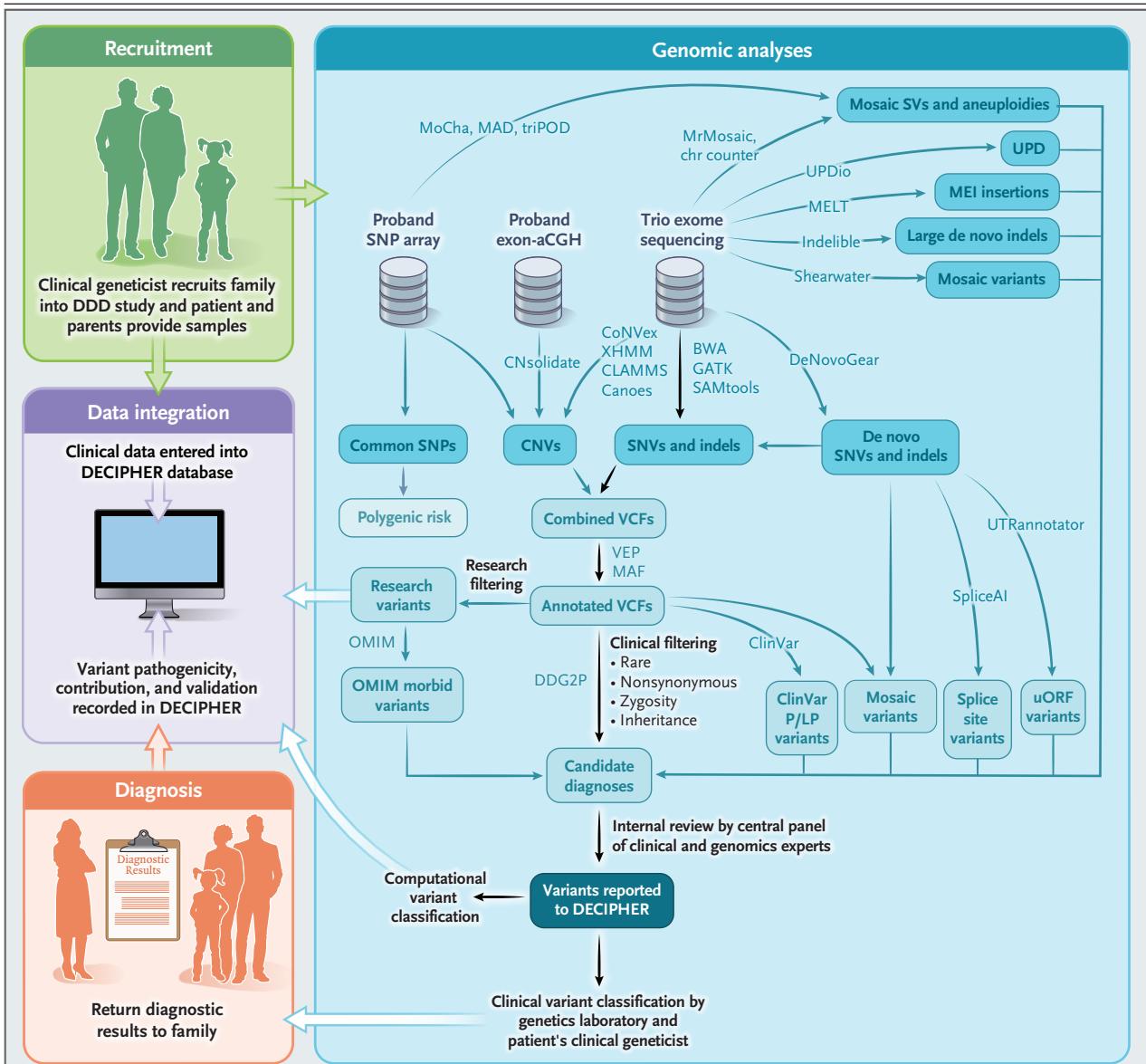


Figure 1. Overview of Variant-Detection and Variant-Filtering Pipelines Used in the Study.

Physician–patient interactions in the Deciphering Developmental Disorders (DDD) study were supported through the use of DECIPHER (Database of Genomic Variants and Phenotype in Humans Using Ensembl Resources), including recruitment, barcoded sample collection and phenotyping at the start, and variant reporting, diagnostic interpretation, and discussion of results at the end. Genomic assays are shown above the three gray icons. Variant callers and analytic processes are annotated on arrows (further details and references are provided in the Supplementary Appendix). Once candidate variants were deposited in DECIPHER, clinical judgment was used to assess whether a patient’s phenotype was consistent with the genotype before returning confirmed diagnoses to families. The abbreviation aCGH denotes array comparative genomic hybridization, CNVs copy-number variants, DDG2P Developmental Disorders Gene2Phenotype, indels insertions or deletions, MAF minor allele frequency, MEI mobile element insertion, OMIM Online Mendelian Inheritance in Man database, P/LP “pathogenic” or “likely pathogenic” (variants in the ClinVar database), SNP single-nucleotide polymorphism, SNVs single-nucleotide variants, SVs structural variants, UPD uniparental disomy, uORF upstream open reading frame, VCFs variant call files, and VEP variant effect predictor.

though 4020 of the 13,449 probands (30%) had no reported variants, 4945 (37%) either had benign or likely benign variants only (866 probands) or had variants of uncertain significance (4079 probands). With the inclusion of computational predictions, the number of probands with vari-

ants of uncertain significance decreased to 2997 (22%), among whom 99 (0.7%) had variants with a predicted Bayesian posterior probability of pathogenicity of 0.8 to 0.9. High concordance was also seen in the subset of variants for which we were able to derive a phenotype-based gene-disease model with the use of IMPROVE-DD,²⁸ and the classification of a further 18 variants of uncertain significance was raised to likely pathogenic on the basis of the individual patient's phenotype.

FACTORS INFLUENCING DIAGNOSTIC YIELD

We performed multivariable logistic regression to investigate how demographic, clinical, phenotypic, prenatal, and ancestral factors affected the chance of receiving a clinical diagnosis in the DDD study (Fig. 4). The model explained approximately 14% of the variance. Recruitment in a family trio had the largest effect on the chance of receiving a diagnosis (odds ratio, 4.70; 95% confidence interval [CI], 4.16 to 5.31). Other factors that considerably increased the chance of receiving a diagnosis included the presence of severe intellectual disability or developmental delay (odds ratio, 2.41; 95% CI, 2.10 to 2.76), longer time interval since recruitment (increased odds of receiving a diagnosis, 1.25 per additional year; 95% CI, 1.20 to 1.30), being the only affected family member (odds ratio, 1.74; 95% CI, 1.57 to 1.92) or having fewer affected first-degree relatives (Fig. S7), the presence of features suggestive of a syndrome (odds ratio, 1.23; 95% CI, 1.12 to 1.34), and greater number of organ systems affected (increased odds of receiving a diagnosis, 1.08 per additional organ system; 95% CI, 1.06 to 1.11).

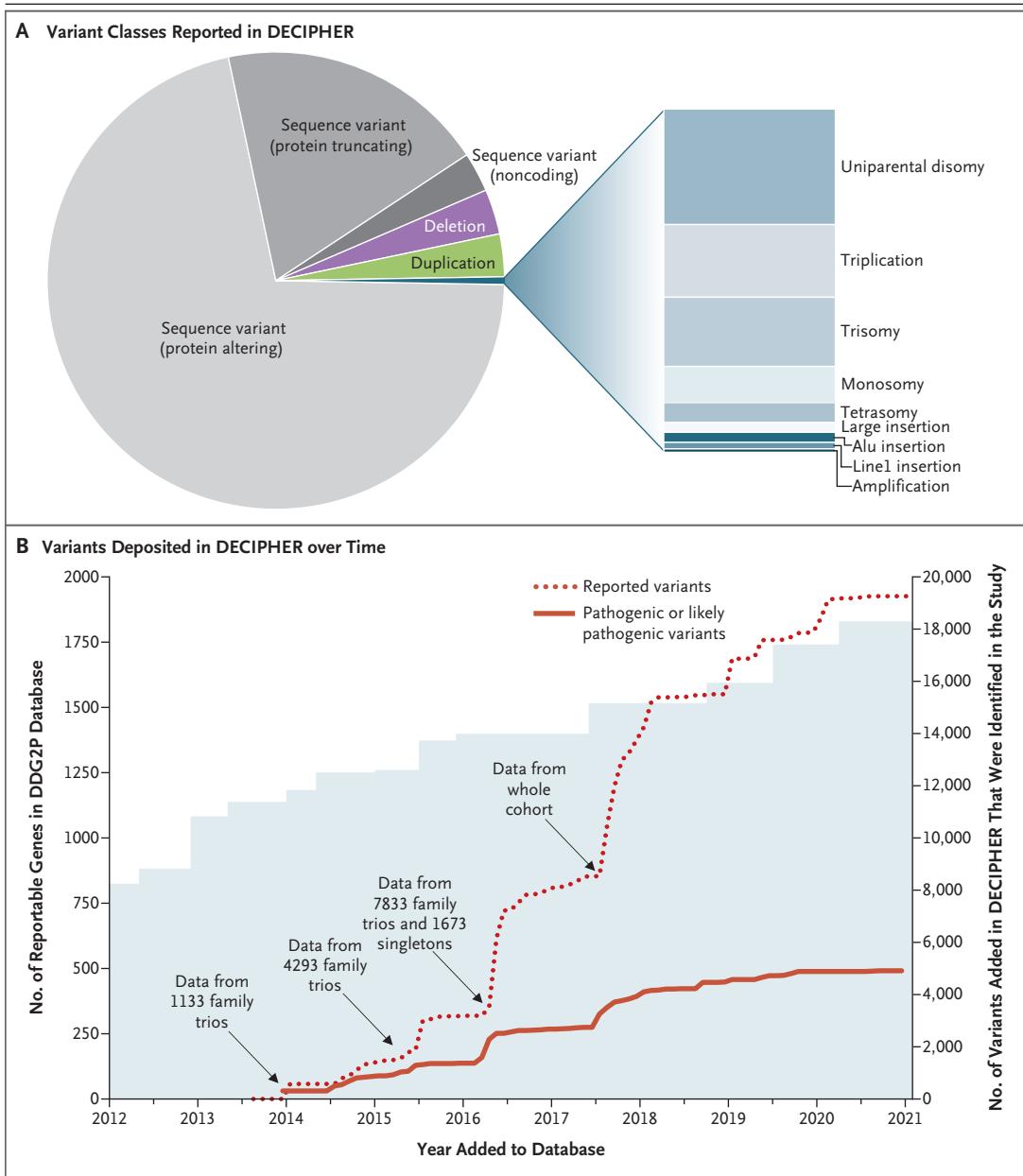
Probands were less likely to receive a diagnosis if they were born extremely prematurely (odds ratio, 0.39; 95% CI, 0.22 to 0.68), had in utero exposure to antiepileptic medications (odds ratio, 0.44; 95% CI, 0.29 to 0.67), or had mothers with diabetes (odds ratio, 0.52; 95% CI, 0.41 to 0.67). Other factors that reduced the odds of receiving a diagnosis were male sex (odds ratio, 0.72; 95% CI, 0.67 to 0.79) and greater degree of homozygosity due to consanguinity (decreased odds of diagnosis, 0.72 for each increase equivalent to the offspring of first cousins; 95% CI, 0.62 to 0.83). The diagnostic yield was lower among probands of African ancestry than among those of other ancestries (odds ratio, 0.51; 95% CI, 0.31 to 0.78), a finding that was driven by fewer diagnoses among singleton probands (Fig. S8).

DISCUSSION

In the DDD study, molecular diagnoses have been identified and communicated to thousands of families across the United Kingdom and Ireland that are affected by severe, previously undiagnosed developmental disorders. Despite the provision of clinical genetic and genomic testing services across the United Kingdom and Ireland, these probands show how a genome-driven approach in combination with detailed phenotyping can improve diagnostic yield over the previous standard of care. Our analysis highlights the value of using diverse and agnostic variant-detection methods in combination with stringent variant-filtering rules and repeated, iterative variant analysis and classification to enable new diagnoses to be made from existing data.¹⁵

The high burden of pathogenic de novo variants and the current diagnostic yield of approximately 41% are consistent with the findings in similar studies.³⁴ Our analysis supports clinical intuition about the likelihood of establishing a molecular diagnosis in patients with developmental disorders (with the likelihood being affected by attributes such as the availability of parental genotype data, as well as sex, ancestry, and degree of phenotypic severity of the proband) and moves toward quantifying the expectation of making such a diagnosis. The work also highlights groups with lower diagnostic yield in our cohort (e.g., those who were not recruited in a family trio, families with multiple affected members, probands of non-European ancestry, and probands with high consanguinity) and reinforces the imperative to increase participation in research involving underrepresented groups. Probands of African ancestry had a particularly low diagnostic yield, owing in part to the lack of ancestry-matched controls to estimate allele frequency and the lower likelihood of being recruited in a family trio.

With the exclusion of cohort-specific factors, our multivariable logistic-regression model predicted a diagnostic yield of 52% among probands in the top decile of probability of receiving a diagnosis, as compared with a yield of 16% among probands in the bottom decile. We hypothesize that the lower diagnostic yield observed among probands with certain prenatal risk factors reflects a larger role of environmental influences that affect them. Premature birth,³⁵ maternal diabetes,³⁶ and in utero exposure to antiepileptic



medications³⁷ are known risk factors for developmental disorders. Further exploration is needed to better understand the relative contributions and interplay of genetic and environmental influences in this cohort.

The genetic architecture of developmental disorders is heterogeneous; although the large burden of highly penetrant *de novo* variants facilitates both diagnosis and large-scale gene–disease discovery,⁵ the number of composite and partial diagnoses suggests that many probands are likely to have multiple contributing factors, including

both rare and common incompletely penetrant genetic variants and nongenetic causes. On the basis of a liability-threshold model of disease,³⁸ probands who have a substantial environmental contribution may require less severe or even no large-effect genetic variants for a neurodevelopmental disorder to develop. Nonetheless, statistical analyses of mutational burden suggest that many more diagnoses remain to be found in protein-coding genes than in noncoding elements.³⁹ These diagnoses will most likely be identified through the discovery of novel associations between genes

Figure 2 (facing page). Variants Identified in the Study and Deposited in DECIPHER.

Panel A shows variant classes reported in DECIPHER. Sequence variants were detected with the use of exome sequencing and included variants smaller than 100 bp among genes in the DDG2P database; structural variants range from those larger than 100 bp to whole chromosomes and were detected with the use of microarray and exome sequencing analyses. Panel B shows changes in the DDG2P database and the number of variants reported and annotated as “pathogenic” or “likely pathogenic” with time. Gene–disease entities were added to the DDG2P database after curation of the literature by consultant clinical geneticists or after burden analyses performed in the study. DNA samples from participants were sequenced and analyzed in batches on the basis of recruitment date, sample receipt, and family trio status. Variant filtering was repeated over the course of the study to enable evaluation of novel variants and variants in newly included genes. As a result of this iterative variant-filtering strategy, some probands were evaluated up to six times, and all were evaluated at least twice (Fig. S3 in the Supplementary Appendix). After evaluation, variants were deposited in DECIPHER, usually in batches, for evaluation by clinical teams. Clinical annotation of pathogenicity was not immediate on deposition, but once annotated, most (97%) of the variants did not change their annotation. Blue shading indicates the cumulative number of reportable genes in the DDG2P database, the red dotted line the cumulative number of total variants identified in the study that were deposited in DECIPHER, and the red continuous line the cumulative number of clinically annotated pathogenic or likely pathogenic variants identified in the study that were deposited in DECIPHER.

and developmental disorders (especially for dominant disorders), evaluation of incompletely penetrant variants, and functional assays to improve interpretation of existing candidate variants. Ultimately, clinical interpretation remains indispensable for determining the relevance of genomic findings for an individual patient.

The DDD study used a hybrid clinical–research approach that involved the development of new methods to facilitate both large-scale analysis and individual variant feedback, which has since become standard practice in genomic medicine. The study primarily recruited infants and children and hence pioneered a conservative approach to individual variant feedback that focused on diagnosis^{7–9} while exploring clinicians’ attitudes toward communicating incidental findings⁴⁰ that influenced subsequent approaches.¹ A large network of expert clinicians and researchers and the integration of ethics at a high level through-

out the project life cycle served to both facilitate collaboration and enable real-time ethical issues to be openly and responsibly addressed (Table 1). To date, in addition to making thousands of new diagnoses for patients, the study has resulted in more than 290 publications (<https://www.ddduk.org/publications.html>), identified approximately 60 new disorders, and enabled more than 350 genotype- or phenotype-specific projects led by clinician and researchers across all 24 recruitment sites. DECIPHER¹² was another key component in the study, enabling nationwide recruitment, systematic phenotyping, individual feedback, variant interpretation, and data sharing. DECIPHER is a live online platform that allows reported variants to be reevaluated with current data (e.g., gene–disease associations, population frequencies, and co-located variants reported in the ClinVar database, DECIPHER, or publications) each time a patient is assessed in the clinic, thereby facilitating new opportunities for diagnosis as knowledge grows.

Although many of our conclusions are widely applicable across a range of rare diseases, the generalizability is limited by a number of factors. Recruitment of families after clinician-led differential diagnosis and routine diagnostic testing (karyotyping, aCGH, and targeted single-gene testing) resulted in a cohort that was probably depleted of clinically recognizable syndromes, large pedigrees with segregating pathogenic variants, recessive conditions in consanguineous families, and large structural variants. These biases will reduce the estimated diagnostic yield relative to first-tier testing and skew the factors affecting the chance of receiving a diagnosis. The diagnostic yield in this study therefore represents a conservative estimate; higher yields would be anticipated if genomic sequencing had been offered as a first-line investigation. Our genotyping approach (exome sequencing and microarray analyses) did not assay most noncoding variants and could not detect all complex structural variants or tissue-specific mosaicism, and our analytic approach was insensitive to incomplete penetrance. Furthermore, the study was not funded to capture longitudinal phenotype data, evaluate parental phenotypes in detail, record the effect of diagnosis on subsequent clinical management in families, or comprehensively assess social or environmental contributions to developmental disorders — all of which, in retrospect, would

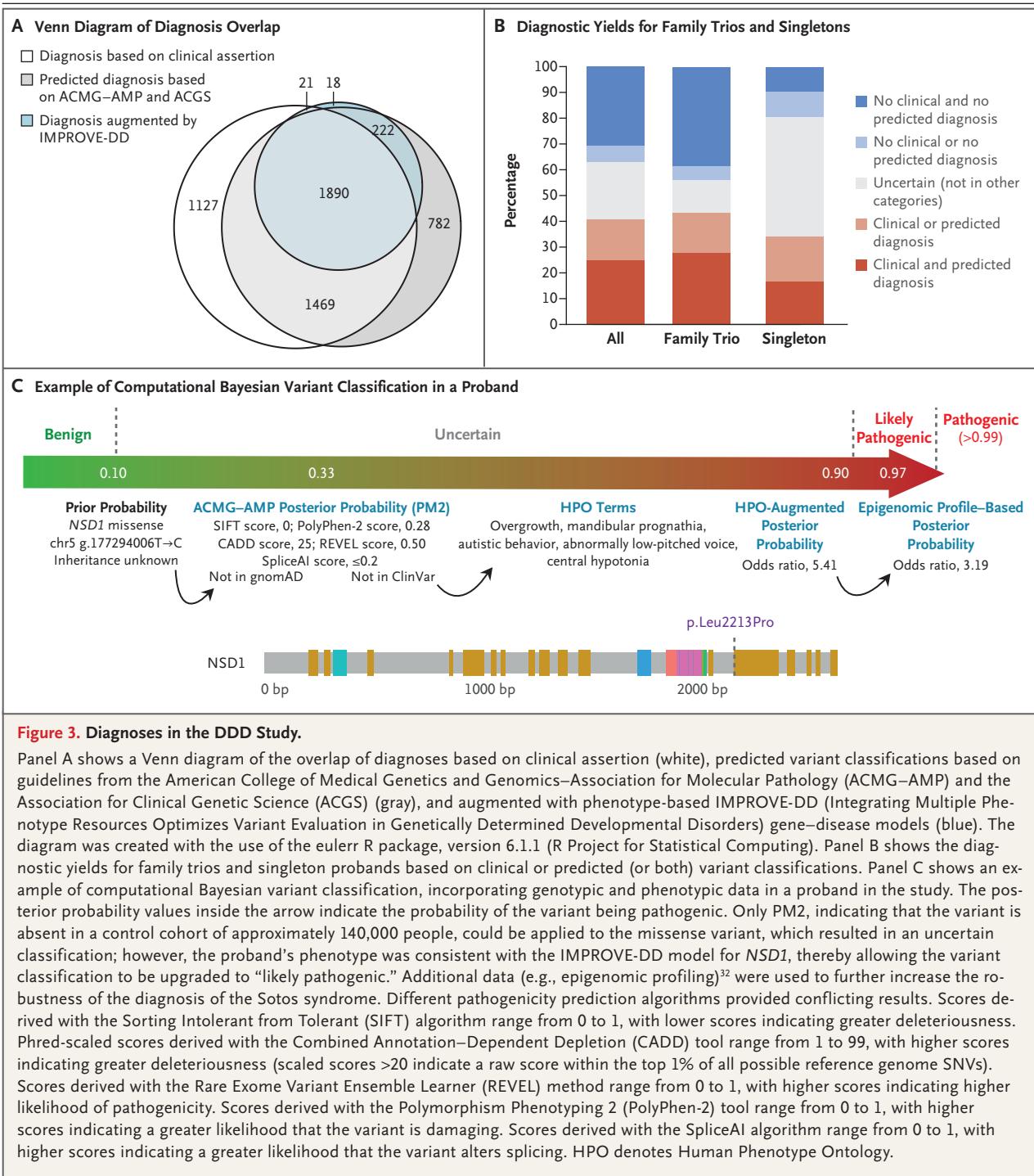


Figure 3. Diagnoses in the DDD Study.

Panel A shows a Venn diagram of the overlap of diagnoses based on clinical assertion (white), predicted variant classifications based on guidelines from the American College of Medical Genetics and Genomics–Association for Molecular Pathology (ACMG-AMP) and the Association for Clinical Genetic Science (ACGS) (gray), and augmented with phenotype-based IMPROVE-DD (Integrating Multiple Phenotype Resources Optimizes Variant Evaluation in Genetically Determined Developmental Disorders) gene–disease models (blue). The diagram was created with the use of the eulerr R package, version 6.1.1 (R Project for Statistical Computing). Panel B shows the diagnostic yields for family trios and singleton probands based on clinical or predicted (or both) variant classifications. Panel C shows an example of computational Bayesian variant classification, incorporating genotypic and phenotypic data in a proband in the study. The posterior probability values inside the arrow indicate the probability of the variant being pathogenic. Only PM2, indicating that the variant is absent in a control cohort of approximately 140,000 people, could be applied to the missense variant, which resulted in an uncertain classification; however, the proband's phenotype was consistent with the IMPROVE-DD model for *NSD1*, thereby allowing the variant classification to be upgraded to “likely pathogenic.” Additional data (e.g., epigenomic profiling)³² were used to further increase the robustness of the diagnosis of the Sotos syndrome. Different pathogenicity prediction algorithms provided conflicting results. Scores derived with the Sorting Intolerant from Tolerant (SIFT) algorithm range from 0 to 1, with lower scores indicating greater deleteriousness. Phred-scaled scores derived with the Combined Annotation–Dependent Depletion (CADD) tool range from 1 to 99, with higher scores indicating greater deleteriousness (scaled scores >20 indicate a raw score within the top 1% of all possible reference genome SNVs). Scores derived with the Rare Exome Variant Ensemble Learner (REVEL) method range from 0 to 1, with higher scores indicating higher likelihood of pathogenicity. Scores derived with the Polymorphism Phenotyping 2 (PolyPhen-2) tool range from 0 to 1, with higher scores indicating a greater likelihood that the variant is damaging. Scores derived with the SpliceAI algorithm range from 0 to 1, with higher scores indicating a greater likelihood that the variant alters splicing. HPO denotes Human Phenotype Ontology.

have enhanced the project. Finally, despite the large cohort size, because of the enormous genetic and phenotypic heterogeneity, we often had insufficient numbers of probands (particularly across different ancestries) with the same ultra-

rare condition to enable confident variant interpretation, a factor that highlights the need to aggregate phenotype information and structured electronic health data across cohorts internationally to improve variant interpretation.

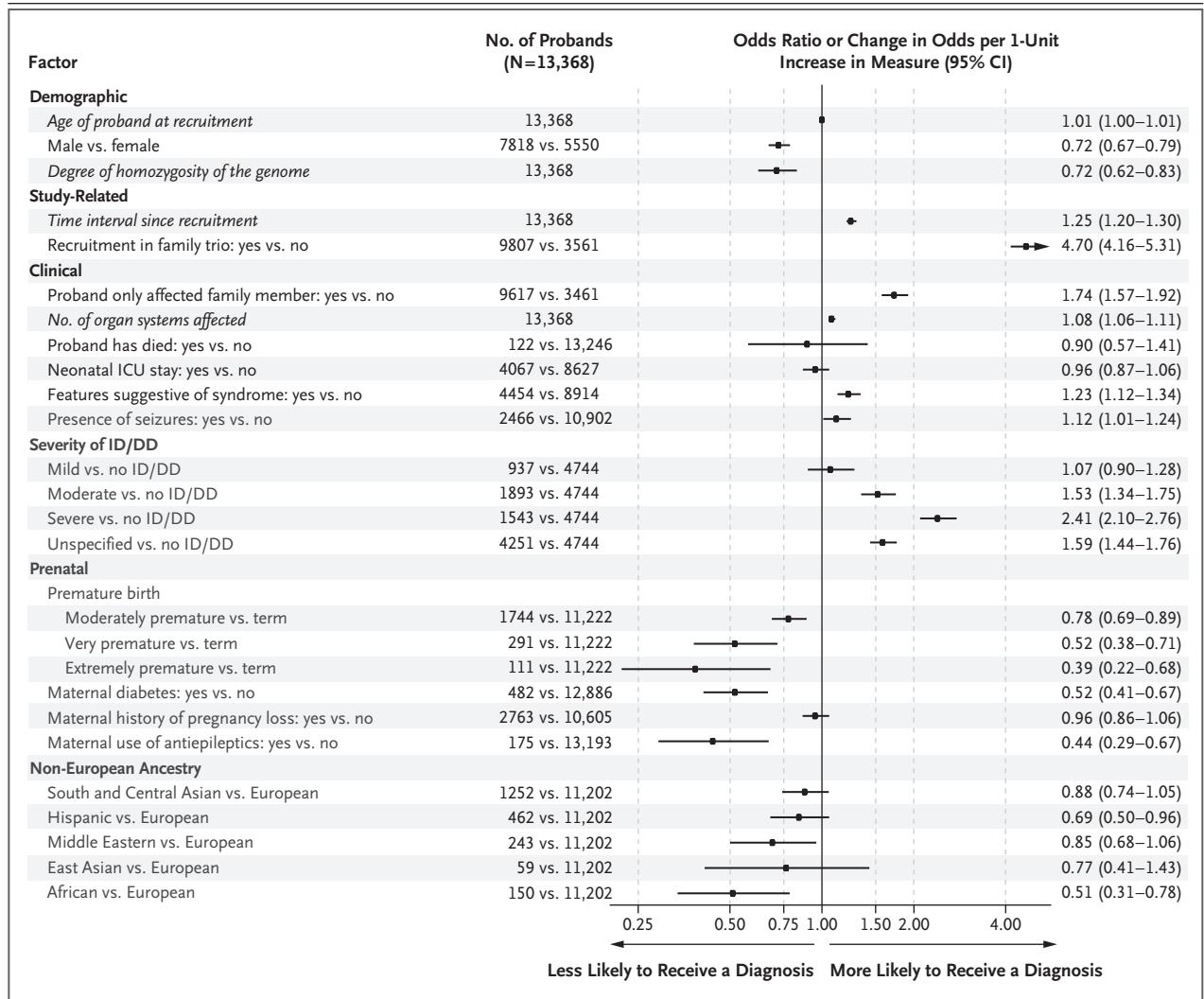


Figure 4. Factors Influencing the Probability of Receiving a Diagnosis.

The odds associated with receiving a full or partial diagnosis in the study (on the basis of clinician assertions of variant pathogenicity) are shown for covariates included in a multivariable logistic-regression model with adjustment for recruitment center and number of variants reported in DECIPHER. A total of 81 probands were excluded owing to missing information or suspected errors in phenotyping. A small group of probands whose status was classified as “unknown” for the factors of “proband only affected family member” and “neonatal intensive care unit (ICU) stay” are not shown in the plot but were included in the model as a categorical group. Odds ratios and 95% confidence intervals (CIs) are given for binary and categorical variables (nonitalicized), and changes in odds per 1-unit increase in measure and 95% CIs are given for quantitative variables (italicized); the widths of the 95% CIs have not been adjusted for multiplicity, and the intervals may not be used in place of hypothesis testing. In the premature birth vs. term category, term was defined as 37 or more weeks’ gestation, moderately premature as 32 to 36 weeks’ gestation, very premature as 28 to 31 weeks’ gestation, and extremely premature as 22 to 27 weeks’ gestation. Results of further analysis of the number of affected first-degree relatives are provided in Figure S7, and results of further analysis of ancestry are provided in Figure S8. ID/DD denotes intellectual disability or developmental delay.

Through its genomic analysis of a large clinical cohort using a hybrid clinical–research model, this study shows how the fusion of clinical expertise, genomic science, and bioinformatics can drive diagnosis and discovery in families in which standard, phenotypically driven diagnostic approaches have failed.

The views expressed in this publication are those of the authors and not necessarily those of the Wellcome Trust or the Department of Health.

Supported by a grant (HICF-1009-003) from the Health Innovation Challenge Fund, a parallel funding partnership between the Wellcome Trust and the Department of Health, and a grant (WT098051) from the Wellcome Sanger Institute. Dr. Campbell was supported by an academic clinical fellowship from the National Institute for Health and Care Research (NIHR).

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

A data sharing statement provided by the authors is available with the full text of this article at NEJM.org.

We thank the patients and families involved in the study, the per-

sonnel involved in genetics services at the U.K. NHS and the Irish Health Service Executive, the many scientists who have worked on the study data at the Wellcome Sanger Institute, and the personnel at the NIHR for their support provided through the Comprehensive Clinical Research Network.

REFERENCES

- 100,000 Genomes Project Pilot Investigators. 100,000 genomes pilot on rare-disease diagnosis in health care — preliminary report. *N Engl J Med* 2021;385:1868-80.
- Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders. *N Engl J Med* 2018;379:1353-62.
- Lalonde E, Rentas S, Lin F, Dulik MC, Skraban CM, Spinner NB. Genomic diagnosis for pediatric disorders: revolution and evolution. *Front Pediatr* 2020;8:373.
- Bennett CA, Petrovski S, Oliver KL, Berkovic SF. ExACtly zero or once: a clinically helpful guide to assessing genetic variants in mild epilepsies. *Neurol Genet* 2017;3(4):e163.
- Kaplanis J, Samocha KE, Wiel L, et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 2020;586:757-62.
- Beaulieu CL, Majewski J, Schwartzentruber J, et al. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am J Hum Genet* 2014;94:809-17.
- Wright CF, Fitzgerald TW, Jones WD, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 2015;385:1305-14.
- Wright CF, Hurler ME, Firth HV. Principle of proportionality in genomic data sharing. *Nat Rev Genet* 2016;17:1-2.
- Wright CF, Middleton A, Barrett JC, et al. Returning genome sequences to research participants: policy and practice. *Wellcome Open Res* 2017;2:15.
- Middleton A, Parker M, Wright CF, Bragin E, Hurler ME. Empirical research on the ethics of genomic research. *Am J Med Genet A* 2013;161A:2099-101.
- Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 2014;42:D966-D974.
- Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 2009;84:524-33.
- Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 2015;519:223-8.
- Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 2017;542:433-8.
- Wright CF, McRae JF, Clayton S, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med* 2018;20:1216-23.
- Ramu A, Noordam MJ, Schwartz RS, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* 2013;10:985-7.
- Thormann A, Halachev M, McLaren W, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun* 2019;10:2373.
- Wright CF, Eberhardt RY, Constantinou P, et al. Evaluating variants classified as pathogenic in ClinVar in the DDD study. *Genet Med* 2021;23:571-5.
- Wright CF, Prigmore E, Rajan D, et al. Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nat Commun* 2019;10:2985.
- Wright CF, Quaipe NM, Ramos-Hernández L, et al. Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. *Am J Hum Genet* 2021;108:1083-94.
- Lord J, Gallone G, Short PJ, et al. Pathogenicity and selective constraint on variation near splice sites. *Genome Res* 2019;29:159-70.
- Gardner EJ, Sifrim A, Lindsay SJ, et al. Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *Am J Hum Genet* 2021;108:2186-94.
- Gardner EJ, Prigmore E, Gallone G, et al. Contribution of retrotransposition to developmental disorders. *Nat Commun* 2019;10:4630.
- Eberhardt RY, Wright CF, FitzPatrick DR, Hurler ME, Firth HV. Detection of mosaic chromosomal alterations in children with severe developmental disorders recruited to the DDD study. April 7, 2022 (<https://www.medrxiv.org/content/10.1101/2022.03.28.22273024v2>). preprint.
- Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405-24.
- Ellard S, Baple EL, Callaway A, et al. ACGS best practice guidelines for variant classification in rare disease 2020. Association for Clinical Genomic Science, February 4, 2020 (<https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf>).
- Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med* 2018;20:1054-60.
- Aitken S, Firth HV, Wright CF, Hurler ME, FitzPatrick DR, Simple CA. IMPROVE-DD: integrating multiple phenotype resources optimizes variant evaluation in genetically determined developmental disorders. *HGG Adv* 2022;4:100162.
- Martin HC, Jones WD, McIntyre R, et al. Quantifying the contribution of recessive coding variation to developmental disorders. *Science* 2018;362:1161-4.
- Akawi N, McRae J, Ansari M, et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat Genet* 2015;47:1363-9.
- Vissers LELM, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. *Nat Genet* 2010;42:1109-12.
- Choufani S, Cytrynbaum C, Chung BHY, et al. NSD1 mutations generate a genome-wide DNA methylation signature. *Nat Commun* 2015;6:10207.
- Posey JE, Harel T, Liu P, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. *N Engl J Med* 2017;376:21-31.
- Srivastava S, Love-Nichols JA, Dies KA, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med* 2019;21:2413-21.
- Blencowe H, Lee ACC, Cousens S, et al. Preterm birth-associated neurodevelopmental impairment estimates at regional and global levels for 2010. *Pediatr Res* 2013;74:Suppl 1:17-34.
- Chen S, Zhao S, Dalman C, Karlsson H, Gardner R. Association of maternal diabetes with neurodevelopmental disorders: autism spectrum disorders, attention-deficit/hyperactivity disorder and intellectual disability. *Int J Epidemiol* 2021;50:459-74.
- Coste J, Blotiere P-O, Miranda S, et al. Risk of early neurodevelopmental disorders associated with in utero exposure to

- valproate and other antiepileptic drugs: a nationwide cohort study in France. *Sci Rep* 2020;10:17362.
38. Walsh R, Tadros R, Bezzina CR. When genetic burden reaches threshold. *Eur Heart J* 2020;41:3849-55.
39. Short PJ, McRae JF, Gallone G, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* 2018;555:611-6.
40. Middleton A, Morley KI, Bragin E, et al. Attitudes of nearly 7000 health professionals, genomic researchers and publics toward the return of incidental results from sequencing research. *Eur J Hum Genet* 2016;24:21-9.

Copyright © 2023 Massachusetts Medical Society.

JOURNAL ARCHIVE AT NEJM.ORG

Every article published by the *Journal* is now available at NEJM.org, beginning with the first article published in January 1812. The entire archive is fully searchable, and browsing of titles and tables of contents is easy and available to all. Individual subscribers are entitled to free 24-hour access to 50 archive articles per year. Access to content in the archive is also being provided through many institutional subscriptions.