The authors’ full names, academic degrees, and affiliations are listed in the Appendix. Dr. Caulfield can be contacted at m.j.caulfield@qmul.ac.uk or at Genomics England, William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, United Kingdom.

Drs. Smedley and Smith, Mr. Martin, and Drs. E.A. Thomas, McDonagh, Cipriani, Ellingford, Arno, Tucci, Vandrovcova, Chan, and H.J. Williams and Drs. Scott, Fowler, Rendon, and Caulfield contributed equally to this article.


BACKGROUND
The U.K. 100,000 Genomes Project is in the process of investigating the role of genome sequencing in patients with undiagnosed rare diseases after usual care and the alignment of this research with health care implementation in the U.K. National Health Service. Other parts of this project focus on patients with cancer and infection.

METHODS
We conducted a pilot study involving 4660 participants from 2183 families, among whom 161 disorders covering a broad spectrum of rare diseases were present. We collected data on clinical features with the use of Human Phenotype Ontology terms, undertook genome sequencing, applied automated variant prioritization on the basis of applied virtual gene panels and phenotypes, and identified novel pathogenic variants through research analysis.

RESULTS
Diagnostic yields varied among family structures and were highest in family trios (both parents and a proband) and families with larger pedigrees. Diagnostic yields were much higher for disorders likely to have a monogenic cause (35%) than for disorders likely to have a complex cause (11%). Diagnostic yields for intellectual disability, hearing disorders, and vision disorders ranged from 40 to 55%. We made genetic diagnoses in 25% of the probands. A total of 14% of the diagnoses were made by means of the combination of research and automated approaches, which was critical for cases in which we found etiologic noncoding, structural, and mitochondrial genome variants and coding variants poorly covered by exome sequencing. Cohortwide burden testing across 57,000 genomes enabled the discovery of three new disease genes and 19 new associations. Of the genetic diagnoses that we made, 25% had immediate ramifications for clinical decision making for the patients or their relatives.

CONCLUSIONS
Our pilot study of genome sequencing in a national health care system showed an increase in diagnostic yield across a range of rare diseases. ( Funded by the National Institute for Health Research and others.)
RARE DISEASES ARE A WORLDWIDE
health care challenge, with approximately
10,000 disorders affecting 6% of the
population in Western societies. More
than 80% of rare diseases have a genetic component,
and these conditions are disabling and expensive
to manage. One third of children with a rare dis-
ease die before their fifth birthday. The adoption
of next-generation sequencing has improved rates
of diagnosis of rare diseases over the past de-
cade. However, the majority of patients with
rare diseases remain without a molecular diag-
nosis after standard diagnostic testing. To
address this lack of diagnosis, the U.K. govern-
ment launched the 100,000 Genomes Project in 2013
to apply whole-genome sequencing to the study
of rare diseases, cancers, and infections in a
national health care setting.

To assess the effect of the whole-genome–
sequencing approach on the genetic diagnosis of
rare diseases in the National Health Service (NHS) in the United Kingdom, we carried out a
pilot study in which we enrolled families and
undertook detailed clinical phenotyping of the
proband. We collected electronic health records
from all participants and stored these together
with the genomic and clinical data in a com-
puter environment with multi-petabytes of stor-
age (the Genomics England research environ-
ment). When necessary, we validated diagnostic
variants in the laboratory and performed com-
putational analyses.

METHODS

PARTICIPANTS

After approval from the national research ethics
committee was obtained, we recruited partici-
ants who had been identified by health care
professionals and researchers as having rare
diseases (across a broad range of categories) that
had not been diagnosed after receipt of usual
care in the NHS, which included either no diag-
nostic tests (because none were available) or
approved diagnostic tests that did not include ge-
nome sequencing. The participants were recruited
at nine English hospitals, and written informed
consent was obtained from the participants by
the National Institute for Health Research (NIHR)
BioResource for Rare Diseases.

To test the broad applicability of genome se-
quencing, we determined that participants were
eligible if they had a rare disease (as defined in
the United Kingdom as a disorder affecting ≤1
in 2000 persons), were likely to have a single-
gene or oligogenic cause, and had not received a
genomic diagnosis. Data on previous testing in
proband panels when possible; testing included single-gene tests, karyotyping, single-
nucleotide polymorphism arrays, next-generation
sequencing panels, and exome sequencing. Pro-
bands and, when feasible, parents or other fam-
ily members were enrolled across multiple clinical specialties in the NHS. Standardized baseline
clinical data were recorded with the use of Hu-
man Phenotype Ontology (HPO) terms guided by
disease-specific data models, and whole blood
samples were obtained for DNA extraction. In
the 100,000 Genomes Project, participants are
followed over their life course with the use of
electronic health records (all hospital episodes,
registry entries, and cause of death).

This pilot study was undertaken in partner-
ship with the NIHR BioResource and is part of
the portfolio of translational research at the
NIHR Biomedical Research Centres at Barts,
Cambridge University Hospitals NHS Foundation
Trust, Great Ormond Street Hospital for Chil-
dren NHS Foundation Trust, Manchester Univer-
sity NHS Foundation Trust, Moorfields Eye Hos-
pital NHS Foundation Trust, Newcastle upon Tyne
Hospitals NHS Foundation Trust, Oxford Uni-
versity Hospitals NHS Foundation Trust, and
University College London Hospitals NHS Foun-
dation Trust. Clinical data from the NHS and
NHS Digital were used in this work.

GENOME SEQUENCING

Genome sequencing was performed with the use
of the TruSeq DNA polymerase-chain-reac-
tion (PCR)–free sample preparation kit (Illumina)
on a HiSeq 2500 sequencer, which generates a
mean depth of 32× (range, 27 to 54) and a depth
greater than 15× for at least 95% of the reference
human genome. Whole-genome sequencing reads
were aligned to the Genome Reference Consor-
tium human genome build 37 (GRCh37) with
the use of Isaac Genome Alignment Software.
Family-based variant calling of single-nucleotide
variants (SNVs) and insertion or deletions (in-
dels) for chromosomes 1 to 22, the X chromo-
some, and the mitochondrial genome (mean cov-
verage, 2814×; range, 142 to 16,581) was performed
with the use of the Platypus variant caller.
DIAGNOSTIC PIPELINE

We constructed an automated analytic pipeline to filter the genome down to rare, segregating, and predicted damaging candidate variants in coding regions. To limit the possibility of overlooking or inefficiently prioritizing diagnoses, we focused initially on applied virtual gene panels (applied panels) that were based on both the recruited clinical indication or disease and the submitted HPO terms. To address the issue of which genes have sufficient evidence to show causation and be included in these applied panels, we used our PanelApp software to enable expert, crowdsourced review and curation of genes with diagnostic-grade evidence for each of our disease categories (e.g., evidence in at least three unrelated families). Loss-of-function or de novo protein-altering variants affecting genes in the applied panels were classified as tier 1, other variant types such as missense variants affecting these genes were classified as tier 2, and all other filtered variants were classified as tier 3 (Fig. S1 in the Supplementary Appendix, available with the full text of this article at NEJM.org).

To further reduce the possibility of missing or inefficiently prioritized diagnoses, we used a phenotype-based approach with the Exomiser application to search across all genes in the genome for a diagnosis. Exomiser prioritizes rare, segregating, and predicted pathogenic variants in genes in which the patient phenotypes match previously referenced knowledge from human disease or model organism databases. The ontology-driven phenotype matching can identify patients who have an atypical profile for a disease. Additional details regarding the Exomiser are provided in the Diagnostic Pipeline section in the Supplementary Appendix.

Prioritization of variants and return of candidate variants for presentation to the 13 NHS Genomic Medicine Centres (GMCs) were performed with the use of decision-support systems and with assistance from clinical genetics teams from Congenica and Fabric Genomics. These variants were reviewed by NHS clinical scientists and clinicians using the guidelines of the American College of Medical Genetics and Genomics, and a diagnostic report was issued for each proband. Final clinical outcomes included whether a genetic diagnosis was obtained, identification of the variant or variants involved, whether the variant or variants explained all or some of the phenotypes, and whether an intervention was used.

Recruitment of the participants in the pilot study and sequencing were performed during the period from January 2014 through December 2016, while the infrastructure to collect, quality check, process, and return data was being established. Results were returned to the GMCs from May 2016 through April 2019. Now that the information pipeline has been established (post-pilot phase), results are returned to the GMCs within 6 weeks after the sample is obtained.

NOVEL PATHOGENIC VARIANTS

Researchers investigated coding and noncoding regions to detect novel diagnostic variants in genes matching the patients’ phenotypes, including the presence of de novo variants in highly constrained coding regions in the 95th percentile. We use the term novel to describe diagnostic variants we have detected that have not previously been described in the literature as causative. This is distinct from de novo variants, which are present for the first time in a family member due to either a new variant in an egg or sperm or a new mutation at conception. The variant may have been previously described. We used a new method described by Wei et al. to analyze mitochondrial DNA that accounts for heteroplasmasy, the Genomiser to detect noncoding pathogenic variants, and the ExpansionHunter software tool to detect simple tandem repeat expansions. Finally we used a new random forest method to analyze Canvas and Manta calls and to identify potentially pathogenic copy-number and structural variants.

Gene-based burden testing to detect enrichment of rare, predicted pathogenic, and segregating variants in novel genes in specific disease cohorts relative to controls was performed on the genomes in the pilot study as well as on additional genomes from the rest of the 100,000 Genomes Project to increase power (57,002 genomes; see the Supplementary Methods in the Supplementary Appendix). The genomic and clinical data from the pilot study are freely accessible to members of a Genomics England Clinical Interpretation Partnership domain (https://www.genomicsengland.co.uk/about-gecip/).

STATISTICAL ANALYSIS

Testing was performed with the use of the R software, version 3.6.0 (R Foundation for Statistical Computing), and Stata software, version 16 (StataCorp). Further details on the individual
methods used in the study are provided in the Supplementary Appendix.

## Results

### Participants

We enrolled 4660 participants (2183 probands and 2477 family members), among whom 161 disorders across a broad spectrum of rare diseases were present (Table 1). Neurologic, ophthalmologic, and tumor syndromes were commonly represented (Table 2). Participants were recruited with varying numbers of affected and unaffected family members. We aimed to recruit family trios (both parents and a proband) or larger family structures to facilitate more effective variant prioritization, and our efforts were met with varying degrees of success. Among the recruited probands with multiple bowel polyps, 93% were singletons (i.e., probands for whom no other family member was recruited). In contrast, 12% of the probands with intellectual disability were singletons. Adult probands were more commonly enrolled than pediatric probands (age ≤18 years at recruitment) (74% vs. 26%), which is in line with the percentage of children and adults in the general population in England and Wales (79% vs. 21% [2011 census of England and Wales]).

The preponderance of adults was unusual as compared with previous sequencing projects and reflects the eligibility criterion that probands had to have undergone usual care; in many cases, usual care involved standard genetic testing (mostly single-gene or panel-based). A lower percentage of recruited probands were female than male owing to the difference among pediatric probands (232 girls and female adolescents [11%] vs. 339 boys and male adolescents [16%], P<0.001); the expected percentage of female probands was 51% (on the basis of the 2011 census of England and Wales) across most disease categories. The greater susceptibility of males than of females to recessive X-linked conditions may account for this sex bias: more than 6% of all diagnoses involved variants on the X chromosome (which represents approximately 5% of the genome). The inferred ancestry of the probands (see the Supplementary Appendix) was in line with what was expected on the basis of the general population, in which 86% of children and adults were White, 8% Asian, 3% Black, 2% mixed, and 1% other (2011 census of England and Wales). However, South Asian ancestry was significantly more common among pediatric probands than among adult probands (16% vs. 4%, P<0.001); our results indicated potential consanguinity in

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All Probands (N = 2183)</th>
<th>Pediatric Probands (N = 571)</th>
<th>Adult Probands (N = 1612)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex — no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1138 (52)</td>
<td>339 (59)</td>
<td>799 (50)</td>
</tr>
<tr>
<td>Female</td>
<td>1045 (48)</td>
<td>232 (41)</td>
<td>813 (50)</td>
</tr>
<tr>
<td>Total</td>
<td>2183 (100)</td>
<td>571 (100)</td>
<td>1612 (100)</td>
</tr>
<tr>
<td>Median age at recruitment (IQR) — yr</td>
<td>35 (18–54)</td>
<td>9 (5–14)</td>
<td>45 (31–60)</td>
</tr>
<tr>
<td>Inferred ancestry — no. (%) [% consanguinity suggested in record]†</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>50 (2) [0]</td>
<td>25 (4) [0]</td>
<td>25 (2) [0]</td>
</tr>
<tr>
<td>American with mixed ancestry</td>
<td>26 (1) [23]</td>
<td>12 (2) [25]</td>
<td>14 (1) [21]</td>
</tr>
<tr>
<td>East Asian</td>
<td>8 (&lt;1) [23]</td>
<td>2 (&lt;1) [0]</td>
<td>6 (&lt;1) [0]</td>
</tr>
<tr>
<td>European</td>
<td>1931 (88) [&lt;1]</td>
<td>438 (77) [&lt;1]</td>
<td>1493 (93) [&lt;1]</td>
</tr>
<tr>
<td>South Asian</td>
<td>163 (7) [36]</td>
<td>93 (16) [43]</td>
<td>70 (4) [25]</td>
</tr>
<tr>
<td>Not determined</td>
<td>5 (&lt;1) [0]</td>
<td>1 (&lt;1) [0]</td>
<td>4 (&lt;1) [0]</td>
</tr>
<tr>
<td>Total</td>
<td>2183 (100)</td>
<td>571 (26) [8]</td>
<td>1612 (74) [2]</td>
</tr>
</tbody>
</table>

*Pediatric probands were younger than 18 years of age at recruitment, and adult probands were 18 years of age or older at recruitment. IQR denotes interquartile range.
† Ancestry was inferred on the basis of a random forest model trained on principal component analysis of the five super populations described for the 1000 Genomes Project.
43% of the 93 pediatric South Asian probands and in 1% of the other 478 pediatric probands (Table 1).

**Clinical Data and Sequencing**

We collected clinical data with the use of HPO terms for each affected participant (a median of 4 [range, 1 to 61] present terms, and a median of 4 [range, 0 to 144] absent terms [phenotypes that were assessed and confirmed as definitely not observed in the proband]). We then performed genome sequencing, followed by quality assurance to check coverage, sequence quality, presence of repeat sample submissions or sample swaps, and consistency with reported family structures (see the Supplementary Appendix).

**Diagnostic Yield**

We made genetic diagnoses in 25% of the probands and deposited the genotypes into the ClinVar repository (accession numbers, SCV001759972 to SCV001760540). Of these diagnoses, 60% were made on the basis of coding SNVs or indels in the applied panels; 26% were made on the basis of coding SNVs or indels affecting well-established disease genes not included in the applied panels (diagnoses were made through phenotype-based prioritization or expert review by the study clinicians or the clinical genetics teams from Congenica or Fabric Genomics); and 14% were made on the basis of genomewide, phenotype-agnostic research analysis that investigated beyond SNVs and indels, coding regions, and disease genes in the applied panels (Fig. 1). On the basis of international guidelines, an additional 10% of the probands were classified as having variants of unknown significance in genes that were considered to be consistent with the phenotype on clinical review at the study site but that required further functional validation. Fewer can-

### Table 2. Disease Categories among the Probands in the 100,000 Genomes Project Pilot Study.†

<table>
<thead>
<tr>
<th>Disease Category</th>
<th>All Family Structures (2183 Probands)</th>
<th>Singleton (881 Probands)</th>
<th>Family Duo (343 Probands)</th>
<th>Family Trio (797 Probands)</th>
<th>Larger Family Structure (162 Probands)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number of probands (percent)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular disorder</td>
<td>147 (7)</td>
<td>56 (6)</td>
<td>24 (7)</td>
<td>49 (6)</td>
<td>18 (11)</td>
</tr>
<tr>
<td>Ciliopathy</td>
<td>69 (3)</td>
<td>34 (4)</td>
<td>14 (4)</td>
<td>16 (2)</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Dermatologic disorder</td>
<td>38 (2)</td>
<td>9 (1)</td>
<td>5 (1)</td>
<td>22 (3)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Dysmorphic or congenital abnormality</td>
<td>20 (1)</td>
<td>10 (1)</td>
<td>2 (1)</td>
<td>7 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Endocrine disorder</td>
<td>87 (4)</td>
<td>57 (6)</td>
<td>14 (4)</td>
<td>12 (2)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Gastroenterologic disorder</td>
<td>32 (1)</td>
<td>0</td>
<td>0</td>
<td>18 (2)</td>
<td>14 (9)</td>
</tr>
<tr>
<td>Growth disorder</td>
<td>3 (&lt;1)</td>
<td>3 (&lt;1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hematologic or immunologic disorder</td>
<td>5 (&lt;1)</td>
<td>2 (&lt;1)</td>
<td>3 (1)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hematologic disorder</td>
<td>7 (&lt;1)</td>
<td>0</td>
<td>3 (1)</td>
<td>2 (&lt;1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Hearing or ear disorder</td>
<td>35 (2)</td>
<td>6 (1)</td>
<td>5 (1)</td>
<td>17 (2)</td>
<td>7 (4)</td>
</tr>
<tr>
<td>Metabolic disorder</td>
<td>93 (4)</td>
<td>24 (3)</td>
<td>12 (3)</td>
<td>48 (6)</td>
<td>9 (6)</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>130 (6)</td>
<td>10 (1)</td>
<td>24 (7)</td>
<td>78 (10)</td>
<td>18 (11)</td>
</tr>
<tr>
<td>Neurologic or neurodevelopmental disorder</td>
<td>521 (24)</td>
<td>193 (22)</td>
<td>93 (27)</td>
<td>194 (24)</td>
<td>41 (25)</td>
</tr>
<tr>
<td>Ophthalmologic disorder</td>
<td>348 (16)</td>
<td>74 (8)</td>
<td>62 (18)</td>
<td>199 (25)</td>
<td>13 (8)</td>
</tr>
<tr>
<td>Renal and urinary tract disorder</td>
<td>176 (8)</td>
<td>125 (14)</td>
<td>21 (6)</td>
<td>26 (3)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Respiratory disorder</td>
<td>2 (&lt;1)</td>
<td>1 (&lt;1)</td>
<td>0</td>
<td>1 (&lt;1)</td>
<td>0</td>
</tr>
<tr>
<td>Rheumatologic disorder</td>
<td>48 (2)</td>
<td>14 (2)</td>
<td>6 (2)</td>
<td>25 (3)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Skeletal disorder</td>
<td>62 (3)</td>
<td>15 (2)</td>
<td>11 (3)</td>
<td>23 (3)</td>
<td>13 (8)</td>
</tr>
<tr>
<td>Tumor syndrome</td>
<td>293 (13)</td>
<td>231 (26)</td>
<td>31 (9)</td>
<td>27 (3)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Other</td>
<td>67 (3)</td>
<td>17 (2)</td>
<td>12 (3)</td>
<td>34 (4)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Total</td>
<td>2183 (100)</td>
<td>881 (100)</td>
<td>343 (100)</td>
<td>797 (100)</td>
<td>162 (100)</td>
</tr>
</tbody>
</table>

* Singleton refers to a proband for whom no other family member was recruited, family duo to a parent–proband pair, family trio to both parents and a proband, and family quad to a proband, sibling, and parents.
† Neurologic and neurodevelopmental symptoms excluded intellectual disability.
candidate variants were returned to the GMCs after filtering (i.e., the removal of extremely unlikely candidates) in larger family structures (Table 3), which made it easier to identify causative variants and in turn led to higher diagnostic yields for family trios and quads (proband, sibling, and parents) and more complex family structures (Fig. 2A), even within a disorder (e.g., the diagnostic yield for hereditary ataxia was 21% among singletons and 32% among persons in family trios) (Table S4).

We obtained a higher diagnostic yield for diseases that we considered likely to have a monogenic cause than those we considered likely to have a complex cause (35% vs. 11%) (Fig. 2A).
Diseases were considered likely to have a monogenic cause if they were present in the Online Mendelian Inheritance in Man database, involved genetic testing as part of the standard diagnostic workup, and had a consensus of opinion among three clinical geneticists (who were unaware of each other’s assessments) that they had monogenic cause. Diagnostic yield was highly varied across diseases (Fig. 2B and Table S3); the diagnostic yield ranged from 40 to 55% for intellectual disability and various vision and hearing disorders and was 6% for tumor syndromes.

We obtained data on the presence or absence of previous genetic testing in 1177 participants. The number of tests per proband ranged from 0 to 16, with a median of 1 (interquartile range, 0 to 2), and approximately half the probands in this subgroup had been tested at least once. The overall diagnostic yield with the use of genome sequencing in this subgroup increased by 32%, and there was only a slight difference depending on whether previous testing had been performed (33%) or not (31%).

With the use of phenotype-based prioritization with the Exomiser to score and rank the most likely causative variants, diagnoses were detected in 77% of the top-ranked candidate variants, in 86% of the top three candidates, and in 88% of the top five candidates (Fig. 2D). Use of the Exomiser and applied panels was complementary — 92% of the 490 diagnoses were made with the applied panels or the Exomizer top five candidates.

**Table 3. Candidate Variants Returned for Presentation to the NHS Genomic Medicine Centres per Proband with the Automated Virtual Panel–Based Analysis Pipeline.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>All Family Structures (2183 Probands)</th>
<th>Singleton (881 Probands)</th>
<th>Family Duo (343 Probands)</th>
<th>Family Trio (797 Probands)</th>
<th>Larger Family Structure (162 Probands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median no. of variants returned after filtering (IQR)†</td>
<td>221 (49–288)</td>
<td>292 (258–327)</td>
<td>149 (117–213)</td>
<td>29 (17–136)</td>
<td>22 (9–71)</td>
</tr>
<tr>
<td>Median no. of variants in virtual panels returned after filtering (IQR)</td>
<td>1 (0–2)</td>
<td>1 (0–2)</td>
<td>1 (0–3)</td>
<td>1 (0–2)</td>
<td>0 (0–1)</td>
</tr>
</tbody>
</table>

* Singleton refers to a proband for whom no other family member was recruited, family duo to a parent–proband pair, family trio to both parents and a proband, and family quad to a proband, sibling, and parents.
† Filtering is the removal of extremely unlikely candidate variants.

**DIAGNOSTIC PIPELINE**

The aim of the automated diagnostic pipeline is to identify a few potentially causative candidate variants, among the millions in a whole genome, through the removal of extremely unlikely candidates (filtering) and the identification of the most likely candidates in the remainder (prioritization). This approach facilitates manual clinical interpretation and diagnostic reporting by clinicians at the GMCs.

A total of 322 (66%) of the 490 diagnoses that were based on SNVs or indels from the genomes were made with the virtual panel–based pipeline, and the positive predictive value was high given the millions of variants in the whole genomes — 291 of 1041 candidate variants (28%) returned to the GMCs proved to be diagnostic. We re-ran this analysis in December 2019 to assess the effects of updated versions of the applied panels with the latest disease gene discoveries, improved selection of the applied panel or panels on the basis of the patient’s phenotype, and advances in variant-filtering strategies (e.g., allowance for incomplete penetrance when suspected). With the use of these updated versions, the number of genetic diagnoses increased from 322 to 377 of the 490 diagnoses (77% sensitivity), and the positive predictive value was 15% (Fig. 2D). This result shows effective filtering and prioritization of the variants, with a median number of only 1 candidate variant (interquartile range, 0 to 2) included in the panels returned to the GMCs per proband (Table 3). Ongoing evolution of the applied panels with new disease genes is expected to continue to increase the diagnostic yield with this approach.

With the use of phenotype-based prioritization with the Exomiser to score and rank the most likely causative variants, diagnoses were detected in 77% of the top-ranked candidate variants, in 86% of the top three candidates, and in 88% of the top five candidates (Fig. 2D). Use of the Exomiser and applied panels was complementary — 92% of the 490 diagnoses were made with the applied panels or the Exomizer top five candi-
dates (last blue bar in Fig. 2D). Precision phenotyping in our participants was essential for both the Exomiser and the selection of additional applied panels; without such phenotyping, only 54% of these diagnoses would have been prioritized in the virtual panel for the recruited disease and presented to the GMCs as a likely candidate (first blue bar in Fig. 2D).

**RESEARCH-BASED DIAGNOSES**
A total of 14% of the genetic diagnoses required further research outside the diagnostic pipeline (Fig. 1). This research involved combined analysis of the genome sequences and clinical data in our research environment and validation with the use of wet-bench orthogonal tests and computational approaches (Table S7). Additional diagnoses were made by screening for the presence of de novo variants in highly constrained coding regions. These diagnoses included a de novo EBF3 missense variant in a patient with hereditary ataxia. A mitochondrial genome analysis that accounted for heteroplasmia led to four new diagnoses, as well as the nine that had already been made by means of the main pipeline. Twelve probands had intronic splicing variants that were prioritized by Exomiser owing to the known pathogenic status of these variants in the ClinVar database. Nine diagnoses involving novel, previously undescribed noncoding variants required exploration of the whole genome and in vitro functional validation by means of reverse transcriptase–PCR, minigene, or luciferase assays. For these diagnoses, unsolved cases in probands had been queried for noncoding variants that affect genes, either alone or in compound heterozygosity with loss-of-function variants, included in the applied panels. These cases were identified with the use of Genomiser or, for probands with retinal disorders, systematic analysis of the untranslated regions, promoter, or introns. The cases in 43 additional probands were fully or partially explained by structural variants or simple tandem repeat expansions in the genes HTT or FXN in the probands with hereditary spastic paraplegia.

**NEW DISEASE–GENE ASSOCIATIONS**
We performed burden testing to identify new mendelian disease–gene associations and make potential genetic diagnoses in probands with unsolved cases; 828 significant disease–gene associations (Q value of <0.1) were identified, including 249 known and 579 novel genes (novel with respect to their association with disease), with a mean (±SD) number of associations of only 0.03±0.2 (range, 0 to 3) from 10,000 permutations in which the cases and controls were assigned randomly. A total of 22 candidates represent the most probable new, fully penetrant, mendelian disease genes (Table S8; ClinVar accession numbers, SCV001759972 to SCV001760540) with three recently independently confirmed diagnoses: UBAPI in hereditary spastic paraplegia, FOXJ1 in non–cystic fibrosis bronchiectasis, and SORD in Charcot–Marie–Tooth disease. Diagnostic reports were issued for three probands with these genes (Fig. 1), and we are currently investigating others with the use of the online tool GeneMatcher and with functional validation studies in model organisms.

**HEALTH CARE OUTCOMES AFTER DIAGNOSIS**
The findings from our approach ended long diagnostic odysseys for some participants and their families (the median duration of such an odyssey was 75 months, and the median number of hospital visits was 68) (Table S1), and we speculate that they will mitigate NHS resource costs (the combined cost for 183,273 episodes of hospital care among the affected participants was £87 million [$122 million]) (Table S3). In addition, 134 of the 533 genetic diagnoses (25%) were reported by clinicians to be of immediate clinical actionability — only 11 (0.2%) were described as having no benefit. As of now, the remainder of the diagnoses are of unknown usefulness. The benefits in terms of health care included 4 diagnoses that led to a suggested change in medication, 26 that led to suggested additional surveillance of the proband or relatives, 13 that allowed for clinical trial eligibility, 59 that informed future reproductive choices, and 32 that had other benefits (Table S9).

In several specific probands, diagnoses have had important clinical actionability. In a 36-year-old man with suspected choroideremia, we detected a novel CHM promoter variant causing loss of gene expression, a diagnosis that enabled eligibility for a gene-replacement trial. A male neonate proband presented with severe infection and transient neurologic symptoms immediately after birth and died at 4 months of age with no diagnosis but with health care costs of approximately £80,000 ($112,000) (Table S10). A diagnosis of transcobalamin II deficiency due to a homo-

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITY COLLEGE LONDON on November 15, 2021. For personal use only. No other uses without permission.
Copyright © 2021 Massachusetts Medical Society. All rights reserved.
A  Yield According to Family Structure and Disease Cause

B  Yield According to Disease Category

C  Yield According to Previous Genetic Testing and Test Type

D  Sensitivity and Positive Predictive Value of Strategies

The New England Journal of Medicine

Downloaded from nejm.org at UNIVERSITY COLLEGE LONDON on November 15, 2021. For personal use only. No other uses without permission.  
Copyright © 2021 Massachusetts Medical Society. All rights reserved.
zygous frameshift in TCN2 was made from this study, which enabled predictive testing to be offered to the younger brother within 1 week after birth. The younger child, who received a positive result, received weekly hydroxocobalamin injections to prevent metabolic decompensation.

A 10-year-old girl was admitted to the intensive care unit with life-threatening chicken pox. She had undergone a diagnostic odyssey over a period of 7 years at a total cost of £356,571 (€70,000, $98,000), and predictive testing in her proband pair, family trio to both parents and a proband, and family quad to a proband, sibling, and parents. Panel A shows diagnostic yield among probands according to previous genetic testing and most extensive testing type: chromosomal (karyotyping, array-based comparative genomic hybridization, single-nucleotide polymorphism arrays), targeted (targeted single-gene tests), next-generation sequencing (NGS) panels, or whole-exome sequencing (WES). The values above the bars are the numbers of probands. Panel D shows the performance of virtual panel-based and Exomiser-based prioritization for identifying the diagnoses. “Disease panel only” indicates the use of a single virtual panel for the recruited disease category. “Applied panels” indicates the use of all applied virtual gene panels used in the pipeline, including the recruited disease-associated panel as well as 0 or more additional panels selected on the basis of the patient’s phenotypes (Human Phenotype Ontology terms). “Exomiser top” indicates Exomiser use in the top-ranked candidate variants, “Exomiser top 3” use in the top three candidates, and “Exomiser top 5” use in the top five candidates. Sensitivity is the percentage of true positive diagnoses based on SNVs or indels that were identified, and the positive predictive value is the percentage of prioritized variants that led to a positive diagnosis. The values above the bars are percentages. In this analysis, the diagnosed variant or variants are true positives, and the other candidate variants that were returned are false positives.

zgyous frameshift in TCN2 was made from this study, which enabled predictive testing to be offered to the younger brother within 1 week after birth. The younger child, who received a positive result, received weekly hydroxocobalamin injections to prevent metabolic decompensation.

A 10-year-old girl was admitted to the intensive care unit with life-threatening chicken pox. She had undergone a diagnostic odyssey over a period of 7 years at a total cost of £356,571 (€70,000, $98,000), and predictive testing in her proband pair, family trio to both parents and a proband, and family quad to a proband, sibling, and parents. Panel A shows diagnostic yield among probands according to previous genetic testing and most extensive testing type: chromosomal (karyotyping, array-based comparative genomic hybridization, single-nucleotide polymorphism arrays), targeted (targeted single-gene tests), next-generation sequencing (NGS) panels, or whole-exome sequencing (WES). The values above the bars are the numbers of probands. Panel D shows the performance of virtual panel-based and Exomiser-based prioritization for identifying the diagnoses. “Disease panel only” indicates the use of a single virtual panel for the recruited disease category. “Applied panels” indicates the use of all applied virtual gene panels used in the pipeline, including the recruited disease-associated panel as well as 0 or more additional panels selected on the basis of the patient’s phenotypes (Human Phenotype Ontology terms). “Exomiser top” indicates Exomiser use in the top-ranked candidate variants, “Exomiser top 3” use in the top three candidates, and “Exomiser top 5” use in the top five candidates. Sensitivity is the percentage of true positive diagnoses based on SNVs or indels that were identified, and the positive predictive value is the percentage of prioritized variants that led to a positive diagnosis. The values above the bars are percentages. In this analysis, the diagnosed variant or variants are true positives, and the other candidate variants that were returned are false positives.

**Figure 2 (facing page). Diagnoses in the Rare Disease Pilot Study.**

Panel A shows diagnostic yield for any disease and according to family structure and cause of disease. The diagnostic yield was 35% for diseases likely to have a monogenic cause and 11% for diseases likely to have a complex cause. The values above the bars are the numbers of probands. Singleton refers to a proband for whom no other family member was recruited, family duo to a parent–proband pair, family trio to both parents and a proband, and family quad to a proband, sibling, and parents. Panel B shows diagnostic yield according to disease category. The values above the bars are the numbers of probands. Panel C shows the diagnostic yield among probands according to previous genetic testing and most extensive testing type: chromosomal (karyotyping, array-based comparative genomic hybridization, single-nucleotide polymorphism arrays), targeted (targeted single-gene tests), next-generation sequencing (NGS) panels, or whole-exome sequencing (WES). The values above the bars are the numbers of probands. Panel D shows the performance of virtual panel-based and Exomiser-based prioritization for identifying the diagnoses. “Disease panel only” indicates the use of a single virtual panel for the recruited disease category. “Applied panels” indicates the use of all applied virtual gene panels used in the pipeline, including the recruited disease-associated panel as well as 0 or more additional panels selected on the basis of the patient’s phenotypes (Human Phenotype Ontology terms). “Exomiser top” indicates Exomiser use in the top-ranked candidate variants, “Exomiser top 3” use in the top three candidates, and “Exomiser top 5” use in the top five candidates. Sensitivity is the percentage of true positive diagnoses based on SNVs or indels that were identified, and the positive predictive value is the percentage of prioritized variants that led to a positive diagnosis. The values above the bars are percentages. In this analysis, the diagnosed variant or variants are true positives, and the other candidate variants that were returned are false positives.

**Discussion**

Our findings show a substantial increase in yield of genomic diagnoses made in patients with the use of genome sequencing across a broad spectrum of rare disease. The enhanced diagnostic benefit was observed regardless of whether participants had undergone previous genetic testing (diagnostic yields were 31% among those who had undergone testing and 33% among those who had not). In 25% of those who received a genetic diagnosis, there was immediate clinical actionability. The standardization of procedures — from the enrollment of patients to the return of NHS-validated results to clinicians — was critical to our success. For example, the collection of clinical data with the use of diseasespecific data models and HPO terms enabled diagnoses, which confirmed the value of standardization with the use of ontology terms and clinical annotation in precision medicine. These additional diagnoses, beyond the 264 (49% of total diagnoses) observed with the use of the single-disease virtual panel, came from the use of Exomiser and additional applied panels. The diagnostic discoveries derived by combining research, decision support, and clinical validation and assessment leveraged an additional 72 diagnoses.

Diagnostic yield was influenced by family structure, and for disorders likely to have mendelian inheritance and a single-gene etiologic factor, our yield increased to 35%: ophthalmologic, metabolic, and neurologic disorders yielded the greatest percentage of diagnoses. The scale of our data set enabled cohortwide burden testing, which identified numerous novel disease–gene associations, including three that have now
been confirmed and 19 with compelling evidence that are likely to be confirmed in independent data sets.

Of the diseases we diagnosed with the use of genome sequencing, 13% were caused by mutations in noncoding sequence or mitochondrial genomes, tandem repeat expansions in persons with Huntington’s disease, and a wide range of structural variants with nucleotide resolution of breakpoints (which were identified with the use of a new random forest method). An additional 2% of the diagnoses involved coding variants in regions of low coverage on exome sequencing. Our results provide new evidence of the value of genome sequencing and mirror the findings in a previous study in which 53% of the participants who received new diagnoses from genome sequencing had previously undergone exome sequencing.5

Previous studies have shown how next-generation sequencing can lead to diagnoses, with yields of 25 to 29% with the use of exome sequencing in persons who had received no previous genetic testing.32-34 The Undiagnosed Disease Network reported a diagnostic yield of 26% with the use of a mixture of whole-exome and whole-genome sequence analysis in 382 patients,5 and another study of genome sequencing showed a yield of 42% among 50 probands with intellectual disability who had previously undergone testing.35 Among probands with a broad range of disorders (161 in total) with an unmet diagnostic need, we obtained results that were similar to those in the previous studies. Our approach is limited to diagnoses that are readily made by means of short-read genome sequencing. Fully phased, long-read sequencing better detects structural variation and delivers sequence information from parts of the genome that are poorly captured by short-read sequencing.36

The findings from our pilot study support the case for genome sequencing in the diagnosis of certain specific rare diseases in the new NHS National Genomic Test Directory.37 In patients with specific disorders, such as intellectual disability, genome sequencing is now the first-line test in the NHS (Table S12). With a new National Genomic Medicine Service, the NHS in England is in the process of sequencing 500,000 whole genomes in rare disease and cancer in health care. We hope that our findings will assist other health systems in considering the role of genome sequencing in the care of patients with rare diseases.

The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the National Institute for Health Research (NIHR), or the Department of Health and Social Care.

Supported by the NIHR, the Welcome Trust, the Medical Research Council (MRC), Cancer Research U.K., the Department of Health and Social Care, and NHS England. The NIHR BioResource is funded by the NIHR, Drs. Caulfield and Ouwenuard are NIHR senior investigators. Dr. Chinnery is a Wellcome Trust Principal Research Fellow (212219/2/18/Z) and an NIHR Senior Investigator who receives support from the MRC Mitochondrial Biology Unit (MC_UU_00015/9), the MRC International Centre for Genomic Medicine in Neuromuscular Disease (MR/S005021/1), and the NIHR Biomedical Research Centre (BRC). Dr. Arno’s work is supported by grants from Versus Arthritis (21593), the NIHR BRC at Great Ormond Street Hospital, and the MRC (MR/R013926/1). Drs. Smedley, Cachero, and Cipriani receive support from the National Institutes of Health (NIH, grant 5-U1M-HG006370). Dr. Smedley’s team that performed much of the analysis was supported by grants from the NIH (1R24OD011883, US4 HG026370, and 1R01HD09805-01). Dr. Arno’s work is supported by a Fight for Sight (United Kingdom) Early Career Investigator Award (5045/46). NIHR BRC at Great Ormond Street Hospital Institute for Child Health, and Moorfields Eye Charity (Stephen and Elizabeth Archer in memory of Marion Woods). The Moorfields–University College London (UCL) Institute of Ophthalmology team is additionally funded by NIHR BRC at Moorfields Eye Hospital and UCL Institute of Ophthalmology.

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

We thank the personnel at NIHR BioResource for their partnership in this study; all the health care teams at Addenbrooke’s Hospital in Cambridge, Great Ormond Street Hospital NHS Foundation Trust, University College London NHS Foundation Trust, Guy’s and St. Thomas’ Hospital, Barts Health, Oxford University Hospitals NHS Foundation Trust, Manchester University NHS Foundation Trust, and the Newcastle Hospitals NHS Foundation Trust; the NHS patients and their families who made this work possible; all those across the world who have contributed to the PanelApp knowledge base and to the validation and reporting working group (Dr. Dom McMullan, Dr. Helen Firth, Dr. Steve Abbs, and Dr. Sian Ellard) for their role in supporting the development of the bioinformatics pipeline and reporting process; Dr. David Bick and Dr. Gil McVean for providing feedback on our work; Dr. Dame Sue Hill and the team at NHS England for the work to fund and establish the 13 GMCs, which enabled the NHS contribution that included the clinical return of results within the NHS in a standardized and validated format that led to the confirmation of the diagnoses, provided additional information, and led to the patient benefit reported; the Illumina Laboratory Services team at Hinxton for genome sequencing and secondary analysis; and the developers of the Human Phenotype Ontology (Monarch Initiative) and Exomiser (funded by the NIH Office of the Director (1R24OD011883)) for the support that was provided through these resources.

APPENDIX

The authors’ full names and academic degrees are as follows: Damian Smedley, Ph.D., Katherine R. Smith, Ph.D., Antonio Martin, M.Sc., Ellen A. Thomas, M.D., Ellen M. McDonagh, Ph.D., Valentina Cipriani, Ph.D., Jamie M. Ellingford, Ph.D., Gavin Arno, Ph.D., Arianna Tucci, M.D., Jana Vandrovcova, Ph.D., Georgia Chan, Ph.D., Thiloka Ratnaike, M.B., B.S., Ph.D., Wei M.Sc., Ellen A. Thomas, M.D., Ellen M. McDonagh, Ph.D., Valentina Cipriani, Ph.D., Jamie M. Ellingford, Ph.D., Gavin Arno, Ph.D., Michael R. Barnes, Ph.D., Kristina Ivanecz, Ph.D., Louise Moutianou, Ph.D., Matthias Wiltschker, Ph.D., Anna Need, Ph.D., Michael R. Barnes, Ph.D., Letizia Vestito, M.Sc., James Buchanan, D.Phil., Sarah Wordsworth, Ph.D., Sofie Ashford, B.Sc., Karola Rehnström, Ph.D., Emily Li, Ph.D., Gavin Fuller, M.Med.Sc., Philip Twiss, M.Sc., Olivera Spasic-Boskovic, M.Sc., Sally Halsall, Ph.D., R. Andres Floto, M.D., Ph.D., Kenneth Poole, M.D., Ph.D., Annette Wagner, M.D., Ph.D., Sarju G. Mehta, M.D., Mark Gunnell, M.D.,

The New England Journal of Medicine

Copyright © 2021 Massachusetts Medical Society. All rights reserved.
ute of Molecular Medicine (N.K., N.B.A.R., A.O.M.W.) and the Oxford Epilepsy Research Group (A.S.), Nuffield Department of Clinical Neurosciences (A.H.N.), University of Oxford, and the Department of Clinical Immunology (S.P.), John Radcliffe Hospital, Oxford, Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust (E.B.), and the University of Exeter Medical School (E.B., C.F.W.), Royal Devon and Exeter Hospital (S.E.), Exeter, Newcastle Eye Centre, Royal Victoria Inflammatory (A.C.B.), the Institute of Genetic Medicine, Newcastle University, International Centre for Life (V.S., P. Brennan), Wellcome Centre for Mitochondrial Research, Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University (G.S.G., R.H., A.M.S., D.M.T., R. Quinlan, R.W.T., J.A.S.), Highroyd Specialist (G.S.G., A.M.S.), Mitochondrial Service (G.S.G., A.M.S.), Royal Infirmary (J.A.S.), Northern Genetics Service (J.Burn), Newcastle upon Tyne Hospitals NHS Foundation Trust (J.A.S.), and NIHR Newcastle BRC (G.S.G., D.M.T., J.A.S.), Newcastle upon Tyne, the Institute of Cancer and Genomic Sciences, Institute of Biomedical Research, University of Birmingham (C. Palles), and Birmingham Women’s Hospitals’ Medical Research (D.M.), Birmingham, the Genomic Informatics Group (E.G.S.), University Hospital Southampton (I.K.T.), and the University of Southampton (I.K.T.), Southampton, Liverpool Women’s NHS Foundation Trust, Liverpool (A. Douglas), the School of Cellular and Molecular Medicine, University of Bristol, Bristol (A.M.D.), and Yorkshire and Humber, Sheffield Children’s Hospital, Sheffield (G.W.) — all in the United Kingdom; Fabric Genomics, Oakland (M. Babcock, M.G.R.), and the Ophthalmo- thology Department, University of California, San Francisco (A.T.M.) — both in California; the Jackson Laboratory for Genomic Medicine, Farmington, CT (P.N.R.); and the Center for Genome Research and Biocomputing, Environ- mental and Molecular Toxicology, Oregon State University, Corvallis (M.H.).

REFERENCES


Copyright © 2021 Massachusetts Medical Society.