Characterizing trends in clinical genetic testing: A single-center analysis of EHR data from 1.8 million patients over two decades

Authors

Lisa Bastarache, Rory J. Tinker, Bryce A. Schuler, ..., Gillian W. Hooker, Josh F. Peterson, Douglas M. Ruderfer

Correspondence

lisa.bastarache@vumc.org

A study of electronic health records (EHRs) from 1.8 million patients at Vanderbilt University Medical Center highlights the growing role of genetic testing in clinical medicine. The study demonstrates an increase in the number of patients receiving tests and the expanding capability of genetic diagnoses to explain the medical phenome.

Bastarache et al., 2025, The American Journal of Human Genetics *112*, 1029–1038

May 1, 2025

https://doi.org/10.1016/j.ajhg.2025.03.009

CellPress

Characterizing trends in clinical genetic testing: A single-center analysis of EHR data from 1.8 million patients over two decades

Lisa Bastarache,[1,2,*] Rory J. Tinker,[3] Bryce A. Schuler,[2,3,4] Lucas Richter,[2] John A. Phillips III,[4] William W. Stead,[1,5] Gillian W. Hooker,[3,6] Josh F. Peterson,[1,5] and Douglas M. Ruderfer[1,2,7]

Summary

A lack of structural data in electronic health records (EHRs) makes assessing the impact of genetic testing on clinical practice challenging. We extracted clinical genetic tests from the EHRs of more than 1.8 million patients seen at Vanderbilt University Medical Center from 2002 to 2022. With these data, we quantified the use of clinical genetic testing in healthcare and described how testing patterns and results changed over time. We assessed trends in types of genetic tests, tracked usage across medical specialties, and introduced a new measure, the genetically attributable fraction (GAF), to quantify the proportion of observed phenotypes attributable to a genetic diagnosis over time. We identified 104,392 tests and 19,032 molecularly confirmed diagnoses. The proportion of patients with genetic testing in their EHRs increased from 1.0% in 2002 to 6.1% in 2022, and testing became more comprehensive with the growing use of multi-gene panels. The number of unique diseases diagnosed with genetic testing increased from 51 in 2002 to 509 in 2022, and there was a rise in the number of variants of uncertain significance. The phenome-wide GAF for 6,505,620 diagnoses made in 2022 was 0.46%, and the GAF was greater than 5% for 74 phenotypes, including pancreatic insufficiency (67%), chorea (64%), atrial septal defect (24%), microcephaly (17%), paraganglioma (17%), and ovarian cancer (6.8%). Our study provides a comprehensive quantification of the increasing role of genetic testing at a major academic medical institution and demonstrates its growing utility in explaining the observed medical phenome.

Introduction

Over the last 20 years, our ability to interrogate the human genome has become more robust, efficient, and cost effective, resulting in more opportunities to use genetic testing for clinical purposes.[1–3] However, informatics solutions to integrate these tests into electronic health records (EHRs) have lagged behind.[4] Researchers have long envisioned a system where genetic test results are stored in a searchable format and easily flow between systems, similar to other laboratory test results.[5] In reality, genetic test results are often stored in the EHR as scanned documents that are not machine readable and are hidden from easy accounting.[6–8]

The lack of EHR integration makes it challenging to measure the impact of genetic testing in healthcare. Most studies of genetic testing utilization have focused on specific clinical contexts or testing indications and often rely on imperfect datasets, such as coded billing data, which lack specificity and are inconsistently applied, or data from testing laboratories, which lack medical context and longitudinality.[9–12] Despite these limitations, prior studies have highlighted key trends, including the increased use and enhanced diagnostic capability of genetic testing within certain clinical populations,[13–15] as well as current challenges related to inclusive testing results and inefficiencies of the testing process.[16–19] However, a full accounting of how genetic testing has been adopted across subspecialties and indications has not yet been performed.

To address this gap, we generated a clinical genetics database (CGdb) by extracting genetic testing results from the EHRs of 1.86 million patients at the Vanderbilt University Medical Center (VUMC) from 2002 to 2022. Using both automated parsing and manual chart review, we uncovered a substantial amount of genetic testing hidden in unstructured text of the EHRs. Our study tracks the cumulative growth of clinical genetic testing and illustrates the complexity of extracting such information, given the current lack of structure. We also analyze the evolution of genetic testing over the past two decades and note a shift toward more comprehensive testing. Additionally, we catalog diagnoses made through genetic testing, including the many rare conditions that are enriched in a tertiary healthcare population, and quantify the accumulation of uncertain test results. Finally, we assess patient exposure to genetic testing across specialties and measure the morbidity explained by genetic diagnoses across phenotypes. Our results demonstrate that genetic testing is now pervasive in medicine, and we provide examples of the

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [2]Center for Digital Genomic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [3]Division of Medical Genetics and Genomic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA; [5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [6]Concert Genetics, Nashville, TN, USA; [7]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
*Correspondence: lisa.bastarache@vumc.org
https://doi.org/10.1016/j.ajhg.2025.03.009.

value of structured clinical genetic testing data for both operational and research purposes.

Subjects and methods

Study population

Our study focused on patients seen at the VUMC from January 1, 2002, to December 31, 2022. All data were derived from the research derivative (RD), which is an image of the structured and textual data elements of the EHRs and was organized by the observational medical outcomes partnership (OMOP) common data model. It includes demographics, claims data (e.g., International Classification of Disease [ICD] codes), clinical notes, problem lists, visit information, vital signs, and laboratory measures. The cohort included individuals with at least three encounters spaced a minimum of 90 days apart or those who were born at the VUMC. To avoid an influx of thinly phenotyped patients during the COVID pandemic, we excluded ICDs for vaccination, testing, or prophylactic treatment of infectious disease. Outpatient visit dates and clinic type/subspecialty information were obtained using the care site identifier linked to each patient's visit. Outpatient visits included office visits, outpatient evaluations, and telemedicine visits. Race and gender were extracted from the demographics table in the EHRs. State of residence and insurance type were recorded on the basis of the most recent entry for each patient. All ICD codes were mapped to phecodeX.[20] See Table S1 for data element definitions. This study was approved by Vanderbilt's institutional review board (IRB; #171011).

Identifying and processing genetic test results

Our goal was to comprehensively extract and curate germline genetic tests from the EHRs, including single-variant or single-gene tests, multi-gene panels (tests that analyze two or more genes), exome and genome sequencing (ES/GS), nucleotide repeat expansions, chromosomal microarrays (CMAs), and karyotypes. We also included methylation tests for syndromic disorders (e.g., Beckwith-Wiedemann syndrome and Angelman syndrome). We did not attempt to extract pharmacogenetic testing, somatic tumor testing, or circulating fetal DNA tests.

Our efforts to comprehensively curate genetic test results in the EHRs began with a survey of the data to locate where genetic test results were recorded and discussed. To locate tests reported in clinical notes, we indexed the text content of the OMOP notes table for 4,207 gene names from the Online Mendelian Inheritance in Man (OMIM) (Table S2) and the names of the genetic testing laboratories. A review of gene mentions in the notes revealed that genetic test results were recorded in three distinct formats: pathology reports, templated reports, and free text. Pathology reports were produced by the local pathology laboratory and recorded in the EHR in a standard format. We identified these structured test reports by searching the OMOP note table for keywords related to genetics (e.g., genetics, gene, and DNA) in the pathology report note types (note type concept ID 44814642). After locating these reports—which included CMAs, karyotypes, and single-gene or -variant tests for *HFE*, *F2*, *MTHFR*, etc.—pathology reports were excluded from further searches for test results. Templated reports was discovered during a review of gene mentions in the clinical notes, which revealed instances of recurring text surrounding gene mentions. These templates were created by clinicians to organize genetic testing results and were most commonly used in clinics with a high volume of genetic testing such as medical genetics and the hereditary cancer clinics,. Templated reports were identified using regular expressions to find the key field. Custom parsers were developed to extract information from pathology reports and templated results (see Note S1A for examples of pathology reports and templates).

A review of gene mentions also revealed a plethora of test results that were recorded in clinical notes as natural language. Many of these unstructured mentions pertained to tests from external laboratories, which were not stored in a structured format but were rather linked to the EHR as non-searchable PDF reports. The first step in processing these results was to distinguish between gene mentions that pertained to the target patient's test results and those that were false positives, which included synonymy, mentions of genetic variants found in family members, discussions of possible genetic testing, and somatic variants (see Note S1B for examples of false positives). A 200-character window around each gene name was screened for false positive results using recurrent phrases like "*CCND1* amplification." We used a locally developed tool to manually review patient records with gene name mentions that were not already included in the database; this interface allowed the reviewer to process one patient record at a time, highlighting gene mention context to enable rapid review. Positive mentions (i.e., those referring to germline testing for the patient) were added to the database through the terminal interface, including information about the test type (e.g., single-gene or multi-gene panel), indication (e.g., diagnostic, cascade testing, or carrier screening), and variants returned from testing. PDFs were reviewed to cross-check information recorded in the notes and retrieve information missing from clinical notes. Each test instance was annotated by the test type, name, date, and any variants returned. Variants were annotated based on the interpretation provided by the clinical testing laboratory.

Test results were annotated as diagnostic if they detected pathogenic variants that were consistent with a diagnosis documented in the EHR. Diagnostic test results were linked to unique disease identifiers from OMIM or Orphanet if OMIM lacked a specific disease identifier.[21,22] Twenty-seven diagnoses were defined as risk factors as they contributed to a multifactorial disorder with both genetic and environmental components (e.g., factor V Leiden thrombophilia, MIM: 188055). (Table S3) Tests were labeled as "carrier" if they returned a pathogenic or likely pathogenic variant linked to an autosomal recessive gene, "inconclusive" if they returned one or more variants of uncertain significance (VUSs), and "negative" if they returned no variants. A clinical geneticist grouped genetic panel tests into broad disease categories (e.g., cancer, endocrinology, or cardiology).

Linking diseases and genes to phenotypes

We linked genes and genetic diagnoses to phecodeX; phecodes are ICD-based phenotypes that capture diagnoses, signs, and symptoms across the medical phenome.[23] To do so, we used the Human Phenotype Ontology (HPO) annotations for rare diseases (v.2024-04-24) that link genetic diseases and genes to phenotypes encoded as HPO terms.[24] We developed an HPO-to-phecodeX map such that the phenotypic manifestations of each disease and gene could be described as a set of phecodes (per methods described in prior publications).[25,26] To link diagnoses and test results to medical subspecialties, we mapped 12 of the HPO high-level phenotype categories to corresponding subspecialties (e.g., neoplasm → oncology) (Tables S4–S6).

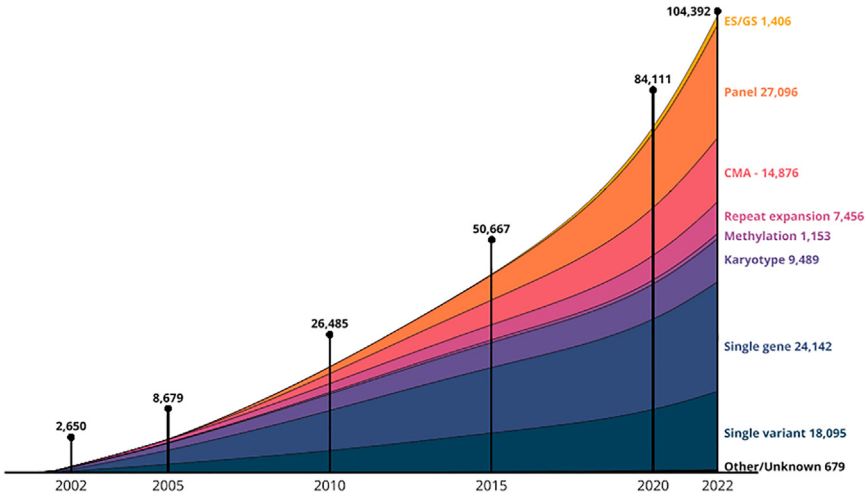

Figure 1. The cumulative growth of genetic testing recorded in the EHRs
The cumulative number of tests by year, stratified by test type. Cumulative counts of tests are shown for years 2002, 2005, 2010, 2015, 2020, and 2022. Cumulative totals for each test type are labeled on the right side of the figure.

Descriptive statistics for tests and diagnoses

To assess changes in genetic testing utilization over time, we summarized the cumulative and yearly counts for the number of tests, diagnoses, and variants returned, stratified by test type and/or indication. We also summarized the test types and panel disease categories by diagnostic yield (i.e., the percentage of tests that returned a diagnostic result), as well as the fraction of tests that returned carrier status (i.e., a single pathogenic variant for an autosomal recessive disease), uncertain variants, or no results (negative).

Assessing testing rates across subspecialty clinics

We assessed the percentage of patients exposed to genetic testing over time, with the denominator as the number of patients who were seen at the VUMC for each target year and the numerator as the number of patients who had genetic testing in their EHRs. We stratified the percent-tested measure across all outpatient clinics that saw at least 200 patients in our cohort in 2022. For 12 subspecialties, we further assessed the relevance of the genetic testing present in patients' records by linking their diagnoses and results to subspecialties.

Computing the genetically attributable fraction

Genetic diagnoses can explain a wide range of signs and symptoms. For instance, a diagnosis of cystic fibrosis clarifies why a patient may experience bronchiectasis and pancreatitis, while hereditary transthyretin (TTR) amyloidosis accounts for a patient's heart failure and neuropathy. We developed a metric called the genetically attributable fraction (GAF) to measure the proportion of a phenotype that can be linked to a genetic diagnosis. The GAF is calculated as follows: the denominator is the total number of patients with a specific phenotype (affected individuals), while the numerator is the number of these affected individuals that have a genetic diagnosis associated with that phenotype (diagnosed affected individuals). We defined affected individuals as those with a specific phecode recorded on at least two distinct dates within the target year. We linked phecodes from patient EHRs to genetic diagnoses by using the HPO-to-phecodeX map described earlier. We computed phenome-wide GAFs for all phenotypes observed in 2022 (excluding phecodes in the genetic diagnosis chapter) as well as for individual phenotypes. Additionally, we assessed the change in GAFs from 2002 to 2022 for 25 phecodes with at least 100 cases, a GAF $> 5\%$, and linkage to more than one genetic disease. Only one phecode per parent code was included. We similarly linked non-diagnostic test results to the target phenotype using the HPO-to-gene mapping.

Results

Cohort characteristics

Our cohort included 1,836,752 patients with a median age of 36 at their last encounter at the VUMC (interquartile range [IQR]: 16–60), with a median of 12 encounters (IQR: 6–28) and a median of 5.0 years (IQR: 1.7–10.9) between the first and last encounters. 5.1% of patients were born at the VUMC, and 33.9% had at least one encounter as a pediatric patient (age < 18). The cohort was linked to 314,313,905 clinical notes from the EHRs. A full description of cohort demographics can be found in Table S7.

The EHR contains a significant and expanding repository of clinical genetic tests and diagnoses

Collectively, information relating to genetic tests, results, and diagnoses was compiled into a resource we call the CGdb, which includes a cumulative total of 104,392 genetic tests for 77,033 patients (Figure 1). Overall, multigene panels were the most common test type (26.0%), followed by single-gene tests (23.1%), single-variant tests (17.3%), CMA (14.3%), karyotype tests (9.7%), repeat-expansion tests (7.1%), exome or genome sequencing (ES/GS, 1.3%), and methylation tests (1.1%). By indication, most tests were for diagnostic purposes (85.8%); carrier screening and familial variant testing accounted for 9.6% and 4.4%, respectively. In terms of results, the majority of tests were negative (63.1%); 15.7% were diagnostic, 2.7% returned a risk factor, 5.9% returned carrier status, and 10.1% returned one or more VUSs and no diagnosis (Table S8).

Overall, the CGdb comprises 19,032 diagnoses and risk factors for 18,476 patients; 533 patients (2.8%) had more than one diagnosis, similar to rates reported in other studies.[27,28] The CGdb includes diagnoses for 1,564

Table 1. The prevalence of genetic disorders as estimated by Orphanet and measured in the CGdb

OrphaCode	Disease	No. affected in CGdb	Prevalence estimates (per 100,000)		Estimated population biobank size
			Orphanet	CGdb	
870	Down syndrome	1046	57	56.9	1,835,088
586	Cystic fibrosis	962	11.1	52.4	8,641,831
399	Huntington disease	335	2.7	18.2	12,407,407
881	Turner syndrome	325	5.5	17.7	5,909,091
567	DiGeorge syndrome	286	37.5	15.6	762,667
363700	Neurofibromatosis, type 1	234	21.3	12.7	1,098,592
98896	Duchenne muscular dystrophy	195	2.8	10.6	6,964,286
558	Marfan syndrome	167	15	9.1	1,113,333
98878	Hemophilia A	103	4.9	5.6	2,123,711
273	Myotonic dystrophy 1	97	12.5	5.3	776,000

The table includes the ten most common genetic diagnoses in CGdb with available prevalence data available from Orphanet.

different diseases, including 20% of the 7,528 disorders currently cataloged in OMIM (Tables S9A and S9B). Most diseases were uncommon in the CGdb: 85% ($n = 1,330$) were diagnosed in fewer than 10 patients, and 43% ($n = 673$) were diagnosed in a single patient. Importantly, the prevalence of diagnoses in the CGdb depends on the likelihood that a genetic disorder is diagnosed with molecular confirmation. Patients who were not tested, received a test that did not interrogate the causal gene, or were clinically diagnosed without genetic confirmation were not included in the prevalence figures. Several diseases were enriched in the CGdb. On the basis of Orphanet's "Prevalence and incidence of rare diseases" (November 2023), a population-based biobank would need over 5 million participants to have the equivalent of patients diagnosed with cystic fibrosis, Huntington disease, and Duchenne muscular dystrophy[29] (Table 1). The relatively high prevalence of these three diseases may, in part, be related to referrals to VUMC clinics that treat these specific disorders.

No consistent convention for documenting genetic testing or diagnosis in the EHR

Genetic testing and diagnostic information was scattered throughout the EHRs in multiple formats, and building the CGdb necessitated both custom automated extraction methods and substantial manual review. We identified 65,836 structured reports generated by the pathology lab at the VUMC, 80.8% of which were for CMAs, karyotypes, and single-gene or -variant tests. The remaining 38,556 tests were embedded in clinical notes, including 5,881 tests recorded in templated language pasted within the body of a note and 32,745 tests that lacked any consistent formatting. To locate genetic testing mentioned in free text only, we indexed all notes for gene mentions, yielding 3,526,486 mentions for 432,581 patients and 3,994 unique genes. Only 10% of gene mentions indicated the results of a genetic test. False positive mentions included synonymy for short gene names (e.g., *ESPN* and *OTC*), mentions of genetic variants found in family members, discussions of possible genetic testing, and somatic variants. Overall, 36.9% of tests were recorded in the clinic notes with no structure, including most multi-gene panels and ES/GS (81.6% and 64.2%, respectively). The fraction of tests reported in free text only increased over time, from 14.8% in 2002 to 48.9% in 2022 (Figure S1).

Genetic tests originated from 112 different laboratories, accounting for 32,218 of the tests in the CGdb. Eight external labs contributed more than 1,000 tests each, accounting for 25.1% of the tests overall. 103 additional external laboratories contributed an additional 6,029 tests.

Growth of genetic testing over 21 years

An increasing percentage of the cohort received genetic testing over the course of the study period. 6-fold more patients had genetic testing recorded in their EHRs in 2022 (6.1%) compared to 2002 (1.0%), as illustrated in Figure S2. The number of new tests increased over time except for in 2020, where there was a notable decrease in genetic testing, likely due to healthcare disruptions during the COVID-19 pandemic (Figure S3). In 2002, the percentage of patients with a molecularly confirmed diagnosis or genetic risk factor was 0.2%, which increased to 1.4% in 2022.

A shift toward more comprehensive testing increased the variety of diagnoses made with genetic testing

In 2002, the most common test types were single-gene (33.0%) and single-variant (28.0%) tests. Multi-gene panels grew in popularity over the study period. In 2015, multi-gene panels overtook single-gene tests as the most popular test type and accounted for 46.3% of testing in 2022 (Figure 2A). The first instance of ES/GS was reported in 2011, and 349 new instances of ES/GS were reported in 2022, surpassing the number of karyotypes that year ($n = 314$).

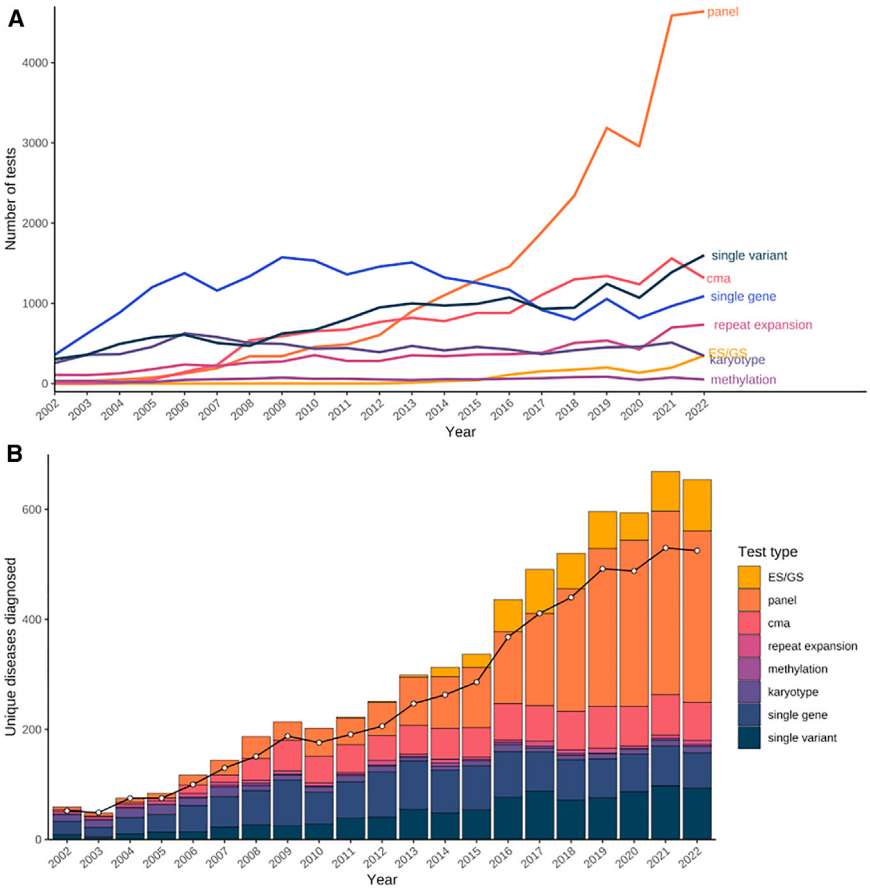

Figure 2. Trend toward more comprehensive tests and unique diagnoses

(A) The number of new genetic tests per year, stratified by test type.

(B) The unique number of diagnoses per year made with different test types. Some diseases were diagnosed with different test types in the same year; the black line represents the overall number of unique diagnoses made in that year, regardless of test type.

As genetic testing became more comprehensive, the number of different diseases diagnosed with genetic testing increased from 51 in 2002 to 509 in 2022 (Figure 2B). Multi-gene panels diagnosed 984 unique genetic diseases, or 63% of all diseases observed. Despite making up only 1.5% of tests overall, ES/GS yielded diagnostic results for 410 different diseases, 168 of which were not diagnosed with any other test type.

More comprehensive testing associated with increased diagnostic yield as well as uncertainty

Overall, 28,646 pathogenic variants and 25,623 VUSs were returned from genetic testing. While the cumulative number of pathogenic variants was greater than VUSs, by 2018, VUSs began accumulating at a faster rate than pathogenic variants, with 1.7 VUSs per pathogenic variant by 2022 (Figure S4). The increase in the rate of VUSs corresponded with the increase in multi-gene panels and ES/GS testing, which returned an average of 2.6 and 1.9 VUSs per pathogenic variant, respectively. While ES/GS was the most likely test to return a diagnosis (36.6%), it was also the most likely to return inconclusive results (29.8%; Figure S5). We also observed differences in the diagnostic yield across panel test categories, with tests for vision disorders having the highest diagnostic yield (49.5%) and obesity the lowest (1.5%; Figure S6).

Genetic testing is pervasive across medical specialties

To determine the pervasiveness of genetic testing, we assessed the genetic testing rates among patients who visited various outpatient clinics. 6.5% of patients with an outpatient visit in 2022 had one or more genetic test in their EHR. The percentage of patients tested varied by clinic type, ranging from 1.1% for student health to 80.5% for medical genetics (Figure S7; Table S10). Because not all genetic tests are relevant to every subspecialty, we further analyzed the relevance of genetic testing across 12 medical specialties. For patients with outpatient visits with these subspecialties, an average of 11.9% (SD: 4.4%) had testing in their EHR; 2.6% (SD: 1.7%) had a relevant diagnosis, 2.8% (SD: 1.0%) had a relevant inconclusive test, and 0.7% (SD: 1.5%) had a relevant negative test (Figure 3).

A substantial and increasing proportion of the observed medical phenome can be attributed to genetic diagnoses

Finally, we assessed the proportion of diagnoses in our population attributable to a diagnosed genetic disease using the GAF (see subjects and methods). In 2022, patients received 6,462,640 diagnoses as encoded by 3,201 unique phecodes, of which 29,545 were linked to a genetic diagnosis, resulting in a phenome-wide GAF of 0.46%. The GAF increased over time across a variety of unrelated phenotypes (Figure 4). Of the 2,049 phecodes with at least 100 cases, 788 had a GAF $>$ 0%. The phecodes with the highest GAF were exocrine pancreatic insufficiency (67.1%), chorea (63.5%), and long QT syndrome (56.4%). Developmental delay had a GAF of 6.2% (659 of 10,704 patients) and was linked to 301 different genetic diagnoses, the largest number of any phenotype. Nine neoplasm phecodes had a GAF $>$ 4%, including paragangliomas (16.6%), ovarian cancer (8.6%), and breast cancer (4.8%; Figure S8).

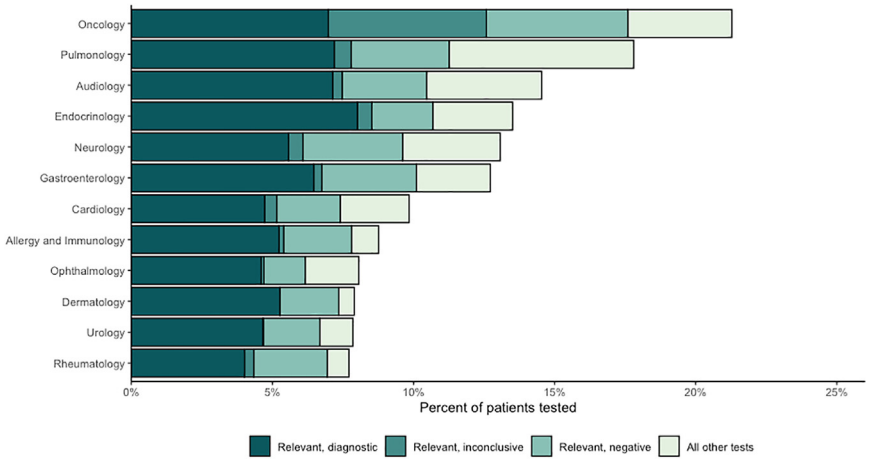

Figure 3. Percentage of patients with genetic testing in their EHRs, by clinic specialty

The bars show the fraction of patients with genetic testing in the EHRs, stratified by clinical specialty. Only outpatient or telemedicine visits in 2022 were included. Genetic tests were deemed relevant to the subspecialty if they were linked with a phenotype related to the subspecialty (e.g., breast cancer is relevant to oncology and hypothyroidism is relevant to endocrinology). The definition of the test categories are as follows: "relevant, diagnostic" includes tests diagnostic for a condition linked to the subspecialty, "relevant, inconclusive" includes non-diagnostic tests that returned a VUS or carrier status linked to the subspecialty, and "relevant negative" includes all other unlinked tests.

Discussion

By comprehensively curating genetic testing from the EHRs, we were able to characterize its use and value across medical specialties and over time. Our findings were broadly consistent with previous studies, showing an increase in clinical genetic testing over the last 20 years, a trend toward more comprehensive testing, a corresponding increase in the total number and diversity of diagnoses, and an increased number of inconclusive test results. Unlike in prior studies, here we show how these trends manifest across test types and indications for an entire medical population.

Our study demonstrates the increasing role of genetic testing in clinical medicine. The rate of clinical genetic testing increased year over year from 2002 to 2022, consistent with prior studies that looked at genetic testing for particular populations, including hereditary cancer screening and preimplantation screening.[11,13] In contrast to prior studies, we provide information on testing across the entire medical enterprise that allows for a more comprehensive understanding of how widespread clinical genetic testing has become since the completion of the Human Genome Project. By 2022, nearly 1 in 16 patients had received some form of genetic testing during their clinical care. Importantly, this number continues to grow, and expansion has occurred in nearly every area of medicine. This includes more than 20% of oncology patients, similar to prior estimates for patients with breast or ovarian cancer,[30] but less than 7% of urology and rheumatology patients. These findings suggest that a growing number of clinicians are faced with the challenge of managing patients with genetic testing; at the same time, these tests are becoming more complex to interpret and utilize.

Our results also show the real-world impact of the increasing diagnostic power of genetic testing. Over the course of the study, the number of unique diseases diagnosed with genetic testing each year increased 10-fold—from 51 in 2002 to 509 in 2022—representing great progress in the ability to detect a wide variety of diseases with genetic testing. These genetic diagnoses explained a significant and growing fraction of the observed phenome. Among the nearly 6.5 million phenotypes observed in 2022, 0.46% could be linked to a genetic diagnosis. This measure is one example of how EHR-linked genetic test data can be used to track the utility of genetic testing at the population level and measure the effect of interventions.

Our findings also reflect the known tension between the improved diagnostic resolution of comprehensive testing and the rise in VUSs.[31] We found that the highest rate of uncertain tests came from ES/GS, followed by panel testing. Here, our results contrast with a recent study of over 1.5 million genetic test results from 19 clinical laboratories, which found that the rate on inconclusive tests was higher for multi-gene panels compared with ES/GS.[12] These differences likely reflect variations in variant reporting guidelines and local testing practices, highlighting the importance of assessing the utility of genetic testing across a variety of real-world settings. The shift toward more comprehensive testing is driving a growing demand for improved tools to support variant annotation and interpretation[32,33] and calls to reexamine reporting practices.[34]

Systematically organizing genetic test data throughout a healthcare system can be valuable for various purposes. Operationally, it can be utilized to monitor emerging trends in the fast-evolving field of genetic testing, providing valuable insights that aid in the development of best practices in areas where they have yet to be established, such as follow-up on inconclusive ES results.[35] Measuring these trends can also help assess the impact of new guidelines on testing practices. Additionally, a comprehensive genetic database may help identify gaps in testing. For example, comparing the use of genetic testing across clinical contexts might reveal where current practices could be altered to increase diagnostic yield and reduce diagnostic delay.[36] Furthermore, measuring diagnostic yield across different test types and clinical contexts may inform cost-benefit analyses, which are essential for expanding the reach of genetic testing.[37,38] In

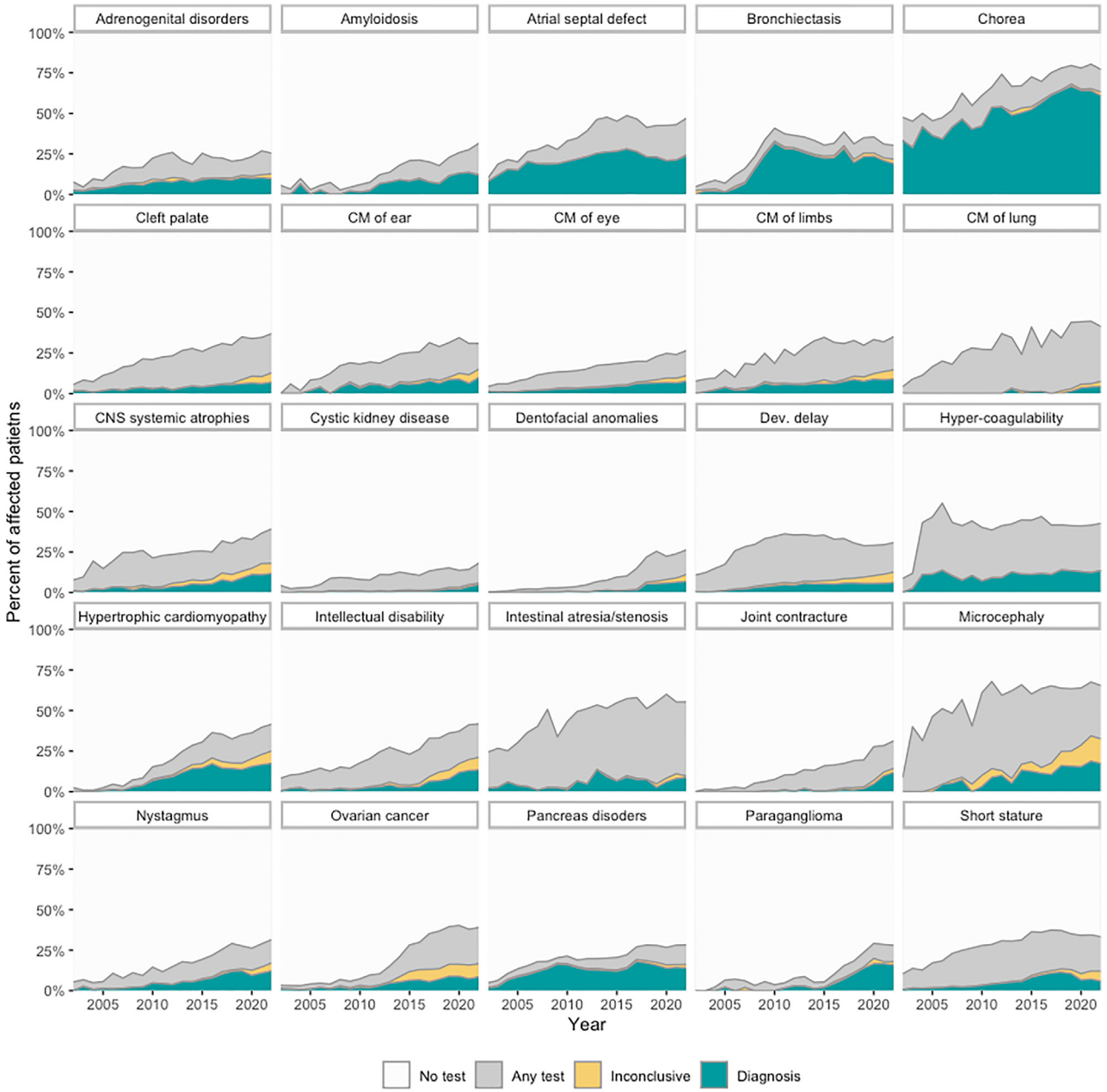

Figure 4. The percentage of phenotypes attributable to genetic diagnoses
Each square represents data for a single phenotype with the GAF (green), the fraction of patients with an inconclusive test (yellow), and the fraction of patients with genetic testing not linked to the phenotype (gray). Phecode labels were abbreviated for space. CM, congenital malformation; CNS, central nervous system.

summary, these data can be used to guide more effective use of genetic testing, enabling studies of medical practice and promoting a virtuous cycle of health improvement as envisioned by proponents of the learning healthcare system.[39]

Our study also reveals that the EHRs of a single academic medical center contain a wealth of genetic testing and diagnoses that could be used to improve knowledge and treatment of genetic diseases. Prior studies show the value of EHR-based cohorts to increase our understanding of the phenotypic manifestations and natural history of rare disorders,[40] improve our understanding of the epidemiology of these conditions,[41] identify patients for clinical trials,[42] and train machine learning algorithms to facilitate early detection.[43] Indeed, an EHR-linked CGdb constitutes a kind of biobank, one that is highly enriched for rare genetic diseases and continues to grow over time. With nearly 20,000 patients with a genetic diagnosis across over 1,500 diseases at the VUMC alone, a collaborative resource spanning multiple medical centers—a "diagnosed disease network"—could revolutionize the study of rare diseases.

A major challenge in curating clinical genetic testing is the poor organization of EHRs, necessitating substantial efforts to extract these data. More than one-third of tests were not recorded in a structured format but instead were documented as natural language in narrative text. The proportion of these unstructured tests increased over the study period, driven by the growing use of send-out tests originating from more than 100 different external laboratories. While indexing the medical records for potential gene mentions enabled us to locate unstructured tests, manual review was used to identify true positives and extract the relevant information. We found that only 10% of potential gene mentions in the narrative text referred to germline test results pertinent to the target patient. The high incidence of reporting somatic variants, test results for family members, and hypothetical results complicated our efforts to automatically extract test results using traditional natural language processing methods. Manually curating these data with a simple, locally developed tool was time consuming. More sophisticated language models that leverage larger context windows may prove more effective.

The lack of searchable, machine-readable test results is a widespread problem that not only limits the potential to use these data for analytics but is also fraught with clinical risk.[3] Standards have been developed to facilitate EHR integration, but progress has been slow and restricted to a small number of institutions.[44,45] Improvements in integration would make it easier for clinicians to locate genetic test results and enable more sophisticated clinical decision support.[46]

Our study has several key limitations. First, this work was conducted at a single academic medical center with a substantial clinical genetic footprint. Genetic testing trends and patterns may differ across institutions and clinical specialties, such that our results may not completely generalize to other medical settings. Relatedly, the way genetic tests are recorded in the EHRs is likely to vary between institutions. Consequently, we cannot provide a universally applicable methodology. Instead, the approach detailed in this paper serves as a guideline. Implementing a similar system in other settings will necessitate tailored programming to accommodate specific institutional requirements. Second, we identified testing by indexing medical records for keywords, including gene names and genetic testing companies. This process might have missed genetic testing that was not recorded in the clinical notes, and this might have biased our dataset against negative tests. Third, our dataset is restricted to germline testing, except for pharmacogenomics, fetal DNA, and tumor testing. We anticipate that these forms of testing could be incorporated via a slightly modified approach.

In their current unstructured state, clinical genetic test results represent a vast and largely untapped resource. Structured test results can be used operationally to measure and monitor genetic test utilization and clinical utility. This dataset can be leveraged for research to improve understanding of how genetic diseases present in the EHRs for improved detection and recognition. As clinical genetic testing becomes an ordinary part of healthcare, it has become necessary to capture testing data in a computable format to improve genetic testing and diagnostics in the future.

Data and code availability

We are unable to make the CGdb a publicly available resource due to privacy concerns. However, all resources used to conduct the study, including the extraction of genetic data from the EHRs and all analyses, are available in the supplemental tables and files of this manuscript. The phecodeX mapping is available on our public website, phewascatalog.org.

Acknowledgments

This study was supported by a grant from the National Human Genome Research Institute (HG012657). Figure 1 was based on an example from the R Graph Gallery (https://r-graph-gallery.com) from Gilbert Fontana.

Author contributions

L.B. conceived of the study and participated in all aspects of its execution. R.J.T. advised on the annotation of genetic disorders and panel types. R.J.T. and B.A.S. assisted in the drafting and editing of the manuscript. B.A.S. assisted with the interpretation of results. J.A.P., W.W.S., and G.H. advised on the content of the manuscript and provided comments. J.F.P. assisted with the conception and implementation of the study and provided feedback on the manuscript. D.M.R. participated in the design of the study and all aspects of assembling the manuscript.

Declaration of interests

L.B. receives royalties from Nashville Biosciences.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2025.03.009.

Web resources

OMIM, http://www.omim.org

Received: October 21, 2024
Accepted: March 12, 2025
Published: April 16, 2025

References

1. Durmaz, A.A., Karaca, E., Demkow, U., Toruner, G., Schoumans, J., and Cogulu, O. (2015). Evolution of genetic techniques: past, present, and beyond. BioMed Res. Int. *2015*, 461524. https://doi.org/10.1155/2015/461524.
2. Halbisen, A.L., and Lu, C.Y. (2023). Trends in Availability of Genetic Tests in the United States, 2012–2022. J. Pers. Med. *13*, 638. https://doi.org/10.3390/jpm13040638.

3. Fullerton, S.M., and Brothers, K.B. (2024). Expanding Applications of Clinical Genetic Testing - Ethical Challenges. N. Engl. J. Med. *390*, 1349–1351. https://doi.org/10.1056/NEJMp2311466.
4. Walton, N.A., Johnson, D.K., Person, T.N., and Chamala, S. (2019). Genomic Data in the Electronic Health Record. Adv. Mol. Pathol. *2*, 21–33. https://doi.org/10.1016/j.yamp.2019.07.001.
5. Fiol, G.D., Williams, M.S., Maram, N., Rocha, R.A., Wood, G.M., and Mitchell, J.A. (2006). Integrating Genetic Information Resources with an EHR. AMIA Annu. Symp. Proc. *2006*, 904.
6. Kho, A.N., Rasmussen, L.V., Connolly, J.J., Peissig, P.L., Starren, J., Hakonarson, H., and Hayes, M.G. (2013). Practical challenges in integrating genomic data into the electronic health record. Genet. Med. *15*, 772–778. https://doi.org/10.1038/gim.2013.131.
7. Kannry, J.L., and Williams, M.S. (2013). The undiscovered country: the future of integrating genomic information into the EHR. Genet. Med. *15*, 842–845. https://doi.org/10.1038/gim.2013.130.
8. Carter, A.B., Abruzzo, L.V., Hirschhorn, J.W., Jones, D., Jordan, D.C., Nassiri, M., Ogino, S., Patel, N.R., Suciu, C.G., Temple-Smolkin, R.L., et al. (2022). Electronic Health Records and Genomics: Perspectives from the Association for Molecular Pathology Electronic Health Record (EHR) Interoperability for Clinical Genomics Data Working Group. J. Mol. Diagn. *24*, 1–17. https://doi.org/10.1016/j.jmoldx.2021.09.009.
9. Lynch, J.A., Berse, B., Dotson, W.D., Khoury, M.J., Coomer, N., and Kautter, J. (2017). Utilization of genetic tests: analysis of gene-specific billing in Medicare claims data. Genet. Med. *19*, 890–899. https://doi.org/10.1038/gim.2016.209.
10. Mackenzie, S.J., Lin, C.C., Todd, P.K., Burke, J.F., and Callaghan, B.C. (2020). Genetic testing utilization for patients with neurologic disease and the limitations of claims data. Neurol. Genet. *6*, e405. https://doi.org/10.1212/NXG.0000000000000405.
11. Kurian, A.W., Ward, K.C., Abrahamse, P., Bondarenko, I., Hamilton, A.S., Deapen, D., Morrow, M., Berek, J.S., Hofer, T.P., and Katz, S.J. (2021). Time Trends in Receipt of Germline Genetic Testing and Results for Women Diagnosed With Breast Cancer or Ovarian Cancer, 2012-2019. J. Clin. Oncol. *39*, 1631–1640. https://doi.org/10.1200/JCO.20.02785.
12. Rehm, H.L., Alaimo, J.T., Aradhya, S., Bayrak-Toydemir, P., Best, H., Brandon, R., Buchan, J.G., Chao, E.C., Chen, E., Clifford, J., et al. (2023). The landscape of reported VUS in multi-gene panel and genomic testing: Time for a change. Genet. Med. *25*, 100947. https://doi.org/10.1016/j.gim.2023.100947.
13. Hipp, H.S., Crawford, S., Boulet, S., Toner, J., Sparks, A.A.E., and Kawwass, J.F. (2022). Trends and Outcomes for Preimplantation Genetic Testing in the United States, 2014-2018. JAMA *327*, 1288–1290. https://doi.org/10.1001/jama.2022.1892.
14. Phillips, K.A., Deverka, P.A., Hooker, G.W., and Douglas, M.P. (2018). Genetic Test Availability And Spending: Where Are We Now? Where Are We Going? Health Aff. *37*, 710–716. https://doi.org/10.1377/hlthaff.2017.1427.
15. Ceyhan-Birsoy, O., Jayakumaran, G., Kemel, Y., Misyura, M., Aypar, U., Jairam, S., Yang, C., Li, Y., Mehta, N., Maio, A., et al. (2022). Diagnostic yield and clinical relevance of expanded genetic testing for cancer patients. Genome Med. *14*, 92. https://doi.org/10.1186/s13073-022-01101-2.
16. Michaels-Igbokwe, C., McInnes, B., MacDonald, K.V., Currie, G.R., Omar, F., Shewchuk, B., Bernier, F.P., and Marshall, D.A. (2021). (Un)standardized testing: the diagnostic odyssey of children with rare genetic disorders in Alberta, Canada. Genet. Med. *23*, 272–279. https://doi.org/10.1038/s41436-020-00975-0.
17. Streff, H., Uhles, C.L., Fisher, H., Franciskovich, R., Littlejohn, R.O., Gerard, A., Hudnall, J., and Smith, H.S. (2023). Access to clinically indicated genetic tests for pediatric patients with Medicaid: Evidence from outpatient genetics clinics in Texas. Genet. Med. *25*, 100350. https://doi.org/10.1016/j.gim.2022.11.018.
18. Childers, K.K., Maggard-Gibbons, M., Macinko, J., and Childers, C.P. (2018). National Distribution of Cancer Genetic Testing in the United States. JAMA Oncol. *4*, 876–879. https://doi.org/10.1001/jamaoncol.2018.0340.
19. Kurian, A.W., Abrahamse, P., Furgal, A., Ward, K.C., Hamilton, A.S., Hodan, R., Tocco, R., Liu, L., Berek, J.S., Hoang, L., et al. (2023). Germline Genetic Testing After Cancer Diagnosis. JAMA *330*, 43–51. https://doi.org/10.1001/jama.2023.9526.
20. Shuey, M.M., Stead, W.W., Aka, I., Barnado, A.L., Bastarache, J.A., Brokamp, E., Campbell, M., Carroll, R.J., Goldstein, J.A., Lewis, A., et al. (2023). Next-generation phenotyping: introducing phecodeX for enhanced discovery research in medical phenomics. Bioinformatics *39*, btad655. https://doi.org/10.1093/bioinformatics/btad655.
21. Amberger, J.S., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res. *47*, D1038–D1043. https://doi.org/10.1093/nar/gky1151.
22. Weinreich, S.S., Mangon, R., Sikkens, J.J., Teeuw, M.E. en, and Cornel, M.C. (2008). Orphanet: a European database for rare diseases. Ned. Tijdschr. Geneeskd. *152*, 518–519.
23. Bastarache, L. (2021). Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu. Rev. Biomed. Data Sci. *4*, 1–19. https://doi.org/10.1146/annurev-biodatasci-122320-112352.
24. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res. *45*, D865–D876. https://doi.org/10.1093/nar/gkw1039.
25. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science *359*, 1233–1239. https://doi.org/10.1126/science.aal4043.
26. Bastarache, L., Hughey, J.J., Goldstein, J.A., Bastraache, J.A., Das, S., Zaki, N.C., Zeng, C., Tang, L.A., Roden, D.M., and Denny, J.C. (2019). Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J. Am. Med. Inform. Assoc. *26*, 1437–1447. https://doi.org/10.1093/jamia/ocz179.
27. Smith, E.D., Blanco, K., Sajan, S.A., Hunter, J.M., Shinde, D.N., Wayburn, B., Rossi, M., Huang, J., Stevens, C.A., Muss, C., et al. (2019). A retrospective review of multiple findings in diagnostic exome sequencing: half are distinct and half are overlapping diagnoses. Genet. Med. *21*, 2199–2207. https://doi.org/10.1038/s41436-019-0477-2.
28. Balci, T.B., Hartley, T., Xi, Y., Dyment, D.A., Beaulieu, C.L., Bernier, F.P., Dupuis, L., Horvath, G.A., Mendoza-Londono, R., Prasad, C., et al. (2017). Debunking Occam's razor: Diagnosing

multiple genetic diseases in families by whole-exome sequencing. Clin. Genet. *92*, 281–289. https://doi.org/10.1111/cge.12987.

29. Nguengang Wakap, S., Lambert, D.M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur. J. Hum. Genet. *28*, 165–173. https://doi.org/10.1038/s41431-019-0508-0.
30. Kurian, A.W., Ward, K.C., Howlader, N., Deapen, D., Hamilton, A.S., Mariotto, A., Miller, D., Penberthy, L.S., and Katz, S.J. (2019). Genetic Testing and Results in a Population-Based Cohort of Breast Cancer Patients and Ovarian Cancer Patients. J. Clin. Oncol. *37*, 1305–1315. https://doi.org/10.1200/JCO.18.01854.
31. Shirts, B.H., Pritchard, C.C., and Walsh, T. (2016). Family-Specific Variants and the Limits of Human Genetics. Trends Mol. Med. *22*, 925–934. https://doi.org/10.1016/j.molmed.2016.09.007.
32. Wang, J., Al-Ouran, R., Hu, Y., Kim, S.-Y., Wan, Y.-W., Wangler, M.F., Yamamoto, S., Chao, H.-T., Comjean, A., Mohr, S.E., et al. (2017). MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome. Am. J. Hum. Genet. *100*, 843–853. https://doi.org/10.1016/j.ajhg.2017.04.010.
33. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. https://doi.org/10.1126/science.adg7492.
34. Burke, W., Parens, E., Chung, W.K., Berger, S.M., and Appelbaum, P.S. (2022). The Challenge of Genetic Variants of Uncertain Clinical Significance : A Narrative Review. Ann. Intern. Med. *175*, 994–1000. https://doi.org/10.7326/M21-4109.
35. Wojcik, M.H., Reuter, C.M., Marwaha, S., Mahmoud, M., Duyzend, M.H., Barseghyan, H., Yuan, B., Boone, P.M., Groopman, E.E., Délot, E.C., et al. (2023). Beyond the exome: What's next in diagnostic testing for Mendelian conditions. Am. J. Hum. Genet. *110*, 1229–1248. https://doi.org/10.1016/j.ajhg.2023.06.009.
36. Zhao, Y., Weroha, S.J., Goode, E.L., Liu, H., and Wang, C. (2021). Generating real-world evidence from unstructured clinical notes to examine clinical utility of genetic tests: use case in BRCAness. BMC Med. Inform. Decis. Mak. *21*, 3. https://doi.org/10.1186/s12911-020-01364-y.
37. Guzauskas, G.F., Garbett, S., Zhou, Z., Schildcrout, J.S., Graves, J.A., Williams, M.S., Hao, J., Jones, L.K., Spencer, S.J., Jiang, S., et al. (2023). Population Genomic Screening for Three Common Hereditary Conditions : A Cost-Effectiveness Analysis. Ann. Intern. Med. *176*, 585–595. https://doi.org/10.7326/M22-0846.
38. Incerti, D., Xu, X.-M., Chou, J.W., Gonzaludo, N., Belmont, J.W., and Schroeder, B.E. (2022). Cost-effectiveness of genome sequencing for diagnosing patients with undiagnosed rare genetic diseases. Genet. Med. *24*, 109–118. https://doi.org/10.1016/j.gim.2021.08.015.
39. Friedman, C., Rubin, J., Brown, J., Buntin, M., Corn, M., Etheredge, L., Gunter, C., Musen, M., Platt, R., Stead, W., et al. (2015). Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J. Am. Med. Inform. Assoc. *22*, 43–50. https://doi.org/10.1136/amiajnl-2014-002977.
40. Havrilla, J.M., Zhao, M., Liu, C., Weng, C., Helbig, I., Bhoj, E., and Wang, K. (2021). Clinical Phenotypic Spectrum of 4095 Individuals with Down Syndrome from Text Mining of Electronic Health Records. Genes *12*, 1159. https://doi.org/10.3390/genes12081159.
41. Herr, K., Lu, P., Diamreyan, K., Xu, H., Mendonca, E., Weaver, K.N., and Chen, J. (2024). Estimating prevalence of rare genetic disease diagnoses using electronic health records in a children's hospital. HGGADVANCE *5*. HGG Adv. *5*, 100341. https://doi.org/10.1016/j.xhgg.2024.100341.
42. Horowitz, C.R., Sabin, T., Ramos, M., Richardson, L.D., Hauser, D., Robinson, M., and Fei, K. (2019). Successful recruitment and retention of diverse participants in a genomics clinical trial: a good invitation to a great party. Genet. Med. *21*, 2364–2370. https://doi.org/10.1038/s41436-019-0498-x.
43. Morley, T.J., Han, L., Castro, V.M., Morra, J., Perlis, R.H., Cox, N.J., Bastarache, L., and Ruderfer, D.M. (2021). Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. Nat. Med. *27*, 1097–1104. https://doi.org/10.1038/s41591-021-01356-z.
44. Dolin, R.H., Heale, B.S.E., Alterovitz, G., Gupta, R., Aronson, J., Boxwala, A., Gothi, S.R., Haines, D., Hermann, A., Hongsermeier, T., et al. (2023). Introducing HL7 FHIR Genomics Operations: a developer-friendly approach to genomics-EHR integration. J. Am. Med. Inform. Assoc. *30*, 485–493. https://doi.org/10.1093/jamia/ocac246.
45. Goehringer, J.M., Bonhag, M.A., Jones, L.K., Schmidlen, T., Schwartz, M., Rahm, A.K., Williams, J.L., and Williams, M.S. (2018). Generation and Implementation of a Patient-Centered and Patient-Facing Genomic Test Report in the EHR. EGEMS (Wash DC) *6*, 14. https://doi.org/10.5334/egems.256.
46. Roundtable on Translating Genomic-Based Research for Health, Board on Health Sciences Policy, and Institute of Medicine (2015). Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research: Workshop Summary (National Academies Press (US)).