Published in final edited form as:

Annu Rev Genomics Hum Genet. 2021 August 31; 22: 219–238. doi:10.1146/annurev-genom-121120-125204.

The Role of Electronic Health Records in Advancing Genomic Medicine

Jodell E. Linder[1], **Lisa Bastarache**[2], **Jacob J. Hughey**[2], **Josh F. Peterson**[2,3]

[1]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, Tennessee 37203, USA

[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37203, USA

[3]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37203, USA

Abstract

Recent advances in genomic technology and widespread adoption of electronic health records (EHRs) have accelerated the development of genomic medicine, bringing promising research findings from genome science into clinical practice. Genomic and phenomic data, accrued across large populations through biobanks linked to EHRs, have enabled the study of genetic variation at a phenome-wide scale. Through new quantitative techniques, pleiotropy can be explored with phenome-wide association studies, the occurrence of common complex diseases can be predicted using the cumulative influence of many genetic variants (polygenic risk scores), and undiagnosed Mendelian syndromes can be identified using EHR-based phenotypic signatures (phenotype risk scores). In this review, we trace the role of EHRs from the development of genome-wide analytic techniques to translational efforts to test these new interventions to the clinic. Throughout, we describe the challenges that remain when combining EHRs with genetics to improve clinical care.

Keywords

translational genomics; electronic health records; GWAS; PheWAS; PheRS; phenome

INTRODUCTION

Genomic medicine is an emerging multidisciplinary specialty that aims to improve human health through the application of genomic research findings to clinical care (63, 119). Arguably, it is the component of precision medicine that is most salient to clinical practice,

jodell.jackson@vumc.org .

DISCLOSURE STATEMENT

Vanderbilt University Medical Center has licensed BioVU PheWAS data to Nashville Biosciences, a Vanderbilt University Medical Center–owned entity, and L.B. receives a portion of those royalty payments. J.F.P. is a consultant for Color.

as it builds upon the decades-long field of medical genetics and leverages well-established and increasingly affordable laboratory technologies to provide clinical-grade sequencing at the point of care. Genomic medicine is distinguished from traditional genetics in that it considers the functions and interactions of all genes in the genome (42). Thus, the field expands on the clinical model of using pedigrees to inform the diagnosis and treatment of monogenic or Mendelian disease, creating a model where polygenic effects address the hereditary components of common complex diseases, enable targeted therapy, and improve understanding of the molecular basis for all disease.

One prominent example of this evolution is in the care of patients with breast cancer. For several decades, physicians have modeled genetic susceptibility to breast cancer with *BRCA1* and *BRCA2* variants to characterize women's predisposition to breast cancer incidence and recurrence (7, 8). However, over the last 10 years, the use of genetic data has greatly expanded to include panels of somatic and germline variants or indicators of gene expression to personalize treatment. Patients with estrogen-dependent (ER+) or human epidermal growth factor receptor 2 (HER2) oncoprotein–expressing breast cancers in particular have benefited from treatment de-escalation from chemotherapy to well-tolerated targeted therapy and hormonal prophylaxis (62, 75). Precision medicine hopes to achieve similar gains across a wide spectrum of diseases. The dramatic reduction in the cost of sequencing has enabled the study of genetic variation at the population level; the rate-limiting resource is the availability of the large populations of diverse, well-phenotyped individuals that are needed to unravel the associations between complex disease and genomic variation.

It is not surprising, then, that central to the emergence of genomic medicine is the marriage of genetic data to rich sources of phenotypic data, particularly comprehensive electronic health records (EHRs) (1). In contrast to disease-specific cohorts, EHRs provide data on a complete spectrum of human disease, treatment effects, and outcomes. EHRs are foundational for phenome science, defined as the study of phenotypic characteristics across large populations (Figure 1). Both genome science and phenome science have required the development of large-scale analytic methods and resources to extract and organize vast amounts of data and draw meaningful conclusions. The development of these techniques, including phenome-wide association studies (PheWASs) (23, 24), genome-wide association studies (GWASs) (38, 43, 56, 74), and electronic phenotyping (e-phenotyping), is the subject of this review, along with the derivative translational methods of phenotype risk scores (PheRSs) and polygenic risk scores (PRSs), which are promising new interventions that may further influence clinical practice.

UTILIZING ELECTRONIC HEALTH RECORDS TO ENABLE GENOMIC SCIENCE

EHRs, which were introduced more than 50 years ago (34, 69), did not gain substantial popularity in the United States until 2007, when federal funding from the Health Information Technology for Economic and Clinical Health (HITECH) Act (47) prompted rapid adoption. Within eight years, more than 80% of federally incentivized hospitals had

adopted EHRs, and clinical data for a large majority of the US population (>95%) began accumulating (3); similar trends have occurred in other countries that have implemented national record systems.

Since EHRs automate the collection of clinical data as they are generated, they provide a unique opportunity to define disease incidence, trajectory, and outcomes across an entire health system or, in international settings, a national population. The array of data available within EHRs (Table 1) also provides a broader and potentially more nuanced representation of the phenome than is found in most clinical research cohorts. For example, findings from radiographic, laboratory, and procedural reports provide objective confirmatory evidence of disease that complements administrative codes and problem list entries, and also provide clinical details to allow disease staging and other metrics of disease severity. In addition, longitudinal EHR data enable investigators to examine how risk factors and disease are interwoven over the course of an individual's life span. Both of these EHR features allowed investigators from a US-based integrated health system to study the impact of familial hyperlipidemia–related variants over patients' lifetimes (2). Specifically, they were able to demonstrate the association between familial hyperlipidemia–related variants and low-density lipoprotein (LDL) cholesterol, and then the cumulative effect of elevated LDL values on the lifetime risk of ischemic heart disease. In this study, the risk for premature coronary artery disease (defined as having the disease at age 55 or younger in males and age 65 or younger in females) among familial hyperlipidemia variant carriers was particularly notable, with an odds ratio of 3.7 compared with noncarriers, reinforcing the prognostic importance of knowing one's familial hyperlipidemia status early in life. Without the availability of decades of EHRs across a large, sequenced population, such studies are not feasible.

However, EHR advantages are balanced by common limitations of using EHR data for clinical research. The primary challenge is the completeness of patient records; some records may be fragmented across different health systems or interrupted when new EHRs are implemented or migrated to a new vendor (110). Records may be "left-censored" prior to the date the patient begins receiving care at an institution and "right-censored" at the point the patient exits the care of that institution (13). EHR-based cohorts are also not population-based samples and represent only those populations that have access to and can afford care at that institution. Both of these limitations indicate that the lack of a phenotypic signal within patient records does not always constitute strong evidence for the absence of that phenotype; sufficient detail may simply be missing from the available EHRs. Investigators have developed several strategies for mitigating these sources of bias, including the use of a "medical home" population that is likely to receive longitudinal primary care at the institution hosting the EHR (13). This strategy narrows the retrospective study cohort to those with repeated historical visits at specific clinics. A second strategy is to cross-link registry, state, or other external data sources to fill in gaps in local EHR data and/or provide corroborating signals. Overall, the limitations of EHRs are outweighed by the wealth of clinical information that is available. The ability to use these data in a high-throughput mechanism and link to genomic data is critical to the advancement and practice of genomic medicine.

Access to EHR data for research purposes requires the development of a parallel resource, the clinical data warehouse, which provides data to investigators in formats conducive to large-scale research (70). Though clinical data warehouses derived from EHRs can be costly to build and maintain, the investment can facilitate rapid translational and discovery-based research. At this time, there is no unified approach to constructing a clinical data warehouse; a recent comprehensive review found approximately 29 separate data architectures for these data repositories (38). This heterogeneity complicates the pooling of data across institutions and is part of the reason that the development of e-phenotyping algorithms requires validation at multiple institutions to demonstrate portability.

Development of Electronic Phenotyping

As EHR data accumulated over decades, researchers began to utilize highly structured data types to represent phenotypes, or the observable characteristics of an individual resulting from the interaction of one's genotype with the environment. The earliest e-phenotyping methods are founded on the common denominator of the administrative coding that underlies the process of billing for healthcare. In the United States, the Medicare program was instrumental in requiring diagnostic and procedural codes in machine-readable formats, which initially allowed researchers to determine causes of hospitalization in elderly populations over time (67). This schema comprises two key code sets: the International Classification of Diseases (ICD) codes and the Current Procedural Terminology (CPT) codes. Now in its 10th revision, the ICD diagnostic codes are used in the majority of disease-based phenotype algorithms developed in the last decade. As researchers began to utilize electronic code data, issues with accuracy began to arise (50, 60), and grouping and collapsing codes to increase diagnostic reliability was recommended (89). These efforts grew into early e-phenotyping (15, 57), where researchers utilized combinations of billing codes and discharge data to define cases and controls for diseases and clinical outcomes. As institutions began to organize their data and create integrated data warehouses (38), the breadth of data available for research grew beyond standard codes and administrative data to include laboratory results, medications, vital signs, medical notes, and reports (115). Table 1 lists commonly utilized data elements gathered from EHRs and the associated utility for phenome science.

In 2010, Ritchie et al. (84) developed e-phenotypes for five conditions of interest and examined genetic associations across a large biobank population; by replicating known associations and discovering new ones, the team demonstrated the utility of e-phenotyping for establishing genome–phenome associations. The need to improve phenotype fidelity prompted investigators to develop tools to extract more complex data (96). What started as more structured, rule-based algorithms moved to methods such as natural language processing (95, 120), deep data mining, machine learning (121), and artificial intelligence (51). These techniques have allowed researchers over the last decade to scan both the unstructured and structured components of EHRs. For example, a combination of natural language processing–derived disease concepts, administrative codes, and laboratory results can define a broad spectrum of ischemic heart disease risk factors; analysis of these longitudinal data using machine learning greatly increases the discrimination of cardiovascular disease predictivity (121). The portability of these techniques (93) across

institutions and data systems is critical to move from research on custom cohorts and populations to large-scale, cross-institutional, translational research. The organization of EHR data into integrated research warehouses allowed for high-quality phenotypes in large cohorts, and the standardization of these data warehouses into common data models precipitated the era of large-scale data sharing.

Transitioning from Local Data to Large-Scale Collaborations

To facilitate cross-institutional analyses, electronic health data must be standardized to minimize the bias of local institution data storage, terminology, and formats. Over the last two decades, the use of common data models has increased in many research programs, enabling researchers to develop analyses locally and rapidly implement them across external institutions. Examples include the Observational Medical Outcomes Partnership, established after the Food and Drug Administration Amendments Act of 2007 (35), which required the Food and Drug Administration to collaborate with public and private partners and access disparate data sources to increase safety data analyses (100). The Observational Medical Outcomes Partnership has grown into the Observational Health Data Sciences and Informatics program, which has more than 2,500 users from 19 countries and half a billion patient records from more than 100 different databases. The National Patient-Centered Clinical Research Network (PCORnet) (17) has demonstrated the ability to collect large quantities of strictly curated EHR data across more than 70 million people and 11 research networks and to create a coordinating center using a common data model. This rigorous structure allows for more rapid data collection at lower costs, effectively giving researchers access to a large, nationwide EHR data set. Other examples include the Shared Health Research Information Network (SHRINE), which aims to enable population-based research through large-scale data sharing and is key to bridging the gap between small discovery-based cohorts and larger translational studies, and the Informatics for Integrating Biology and the Bedside (i2b2) tool, which aims to enable precision medicine through open source data sharing, standardizations, and integration.

These large networks allow researchers to quickly respond to emerging diseases. For instance, the National COVID Cohort Collaborative was quickly set up through a partnership between the National Center for Advancing Translational Sciences and the Clinical and Translational Science Awards program in the spring of 2020 and rapidly established an infrastructure for accepting, aggregating, and providing expedited access to EHR data on coronavirus disease 2019 (COVID-19) patients to support cutting-edge research during the pandemic. The data structures, based on the i2b2, PCORnet, Observational Medical Outcomes Partnership, and TriNetX common data models acquire data twice monthly from more than 50 different institutions. Without standardization, analyses would be limited to local instances or come with high costs in effort and funds to transform analyses across data warehouses.

Common data models and large institutional data warehouses have facilitated the increase in high-throughput research over the last decade and enabled large-scale clinical research. However, it is the linkage of these data to genomics that enables precision medicine and the success of translational genomics research.

Large-Scale Biobanking-Enabled Genomic Research

Preserving biospecimens for later research is a fundamental component of both discovery and translational studies. Biobanks can range from small, study-specific repositories to large, institution-wide efforts. The establishment of institutional biobanks in hospital settings has allowed researchers to preserve specimens collected through routine clinical care. This allows for banking of specimens already being sampled for clinical purposes, reducing the burden on patients. For example, groups such as the National Cancer Institute have developed strategies and operational procedures to maximize the creation of standardized, sustainable resources (58, 106). Biobanks with prospective enrollments offer the ability to use germline DNA for large-scale genomics. Examples like the Vanderbilt University Medical Center bank BioVU (87) enroll participants and then obtain discarded blood collected through routine care, allowing the participant to contribute to the DNA bank without requiring an additional blood draw. Large-scale genomic biobanks include the *All of Us* Research Program (4, 18), which focuses on enrolling a million participants with an emphasis on underrepresented populations; the UK Biobank (19, 76), which followed and collected data on 500,000 participants across the United Kingdom, tied these data to genomic data, and made the data available to researchers across the world; and the Electronic Medical Records and Genomics (eMERGE) network, which collected EHR and genomic data on more than 130,000 participants across the United States.

Hundreds of biobanks exist across the world (77, 101), setting the stage for advances in a variety of diseases and overall health. These specimens can be used to answer genomic, epigenomic, proteomic, and metabolomic research questions. Biobanks storing blood or extracted DNA have the potential to examine genomics across large cohorts and diverse populations. The ability to tie these biobanks to longitudinal EHRs is critical when it comes to examining large-scale genomic research. A few data points collected at time of enrollment, or a snapshot of an EHR, does not allow for large-scale data mining, longitudinal data, or the ability to assess disease outcomes. Interfacing with participants requires time, study staff effort, and participant education and can be costly, making the ability to tie genomics to on-the-shelf large-scale EHR data paramount. As reviewed by Stark et al. (101), governments around the world are making investments in genomic medicine initiatives to help bridge the gap between discovery research and translational medicine, and these initiatives aim to collect genomic data tied to clinical health records.

Several factors contribute to the ability to move from local disease-specific analysis to large-scale translational genomics work. First, the ability to effectively mine large-scale EHR data utilizing electronic methods and tools allows researchers to look across their data warehouse for associations in local patient populations over time. Second, institution-wide efforts to share data in structured formats across institutions facilitate researchers' ability to nationally and globally share data. This paves the way for research focused on rare and common disease, increases the ability to examine conditions over diverse populations, and contributes to national and global efforts. Finally, the ability to tie these large-scale EHR data to genomics empowers investigators to take the next step in translational and precision medicine, allowing the field of genomic medicine to rapidly increase and diversify.

This linkage launches myriad tools and techniques that today are leading to a new era of translational genomics research.

FROM DISCOVERY TO CLINICAL TRANSLATION

Genome-Wide Association Studies to Polygenic Risk Scores

The theory behind using large-scale genomic associations to understand common complex disease was proposed by Risch & Merikangas (83) in 1996. GWAS technology was developed in 2002 (74) to agnostically search for genetic associations with a single trait, and was implemented shortly thereafter, in 2005, to examine genome-wide associations for single-nucleotide polymorphisms (SNPs) involved in age-related macular degeneration (43, 56) (see Figure 2). The success of early GWASs incentivized researchers to find more efficient ways to study large cohorts. By 2007, both the Wellcome Trust Case Control Consortium (111) and the Framingham Heart Study (11) were publishing studies that addressed multiple phenotypes in the same cohort. Using shared controls for multiple phenotypes streamlined GWASs by reducing the number of subjects who needed to be genotyped. GWASs set the stage for heritability estimates of SNPs to be associated with common complex diseases. GWAS analyses with imputation have shown that the heritability of many common diseases can be explained by common variants with small effect sizes across the genome (104, 108). Utilization of GWASs expanded rapidly, with 4,771 publications and 214,295 associations in the GWAS Catalog (12) as of November 2020.

GWASs also allowed researchers to examine polygenic associations of diseases across the genome. The initial association studies that used GWAS data in humans to examine polygenic risk focused on risk of psychiatric disorders, cancers, and cardiovascular disease (49, 97, 116). This work paved the way for the field of PRSs. Over the last several years, PRSs have facilitated the transition from genomic discovery research using GWASs to clinical translational work, associating risk values, odds ratios, and statistical confidence with genomic associations. SNP selection and weighting (how likely it is that the SNPs are associated with the condition of interest) are used to model and validate polygenic risk (14).

PRSs allow investigators to utilize the cumulative effect of relatively common genetic variants that may contribute to common complex diseases. They have gained traction in neurological disorders like schizophrenia (65, 105) and Alzheimer’s disease (33), as well as many common complex diseases, such as colorectal cancer (103), prostate cancer (118), coronary artery disease (54, 102), atrial fibrillation, inflammatory bowel disease, type 2 diabetes (54), type 1 diabetes (94), and breast cancer (54, 66, 90). Several of these studies have demonstrated that PRS risk can be equivalent to monogenic risk (54, 66), suggesting that PRSs will also have clinical utility for predicting incident disease and tailoring preventative care. The initial work on PRSs has led to randomized controlled trials to examine their utility in clinical settings. A 2017 trial and meta-analysis of two other randomized controlled trials on statin usage for individuals with atherosclerosis risk found that those in the highest genetic risk categories derived greater relative and absolute benefit from the statins and reduction in coronary heart disease events (72) than those in other risk categories.

PRSs are an example of how the integration of large-scale genomics to examine multiple components of disease development can drive translational research. Though the evidence base for the clinical utility of PRSs is growing, and some (such as breast cancer and cardiovascular disease) have been incorporated into clinical risk scores (14, 104), several pitfalls have emerged. The performance of PRSs across males and females (44) and across ancestry groups is not always maintained, and applying them without adjustment may exacerbate health disparities in underrepresented populations (65). This is partially because the first generation of PRSs are derived from GWAS data sets that do not have sufficient numbers of non–European ancestry individuals. Recent work has moved to the generation and validation of PRSs in more diverse cohorts (5, 28, 73) and trans-ancestry modeling (26, 61) to help mitigate issues of translatability to clinical populations. The capabilities of PRS research and its ultimate clinical utility rely heavily not only on the genomic data available but also on links to the phenome. Because polygenic risk does not necessarily demonstrate the whole picture of disease development (30, 71), clinical factors, family history, and monogenic risks must also be considered. The ability to determine which individuals have conditions and traits of interest and the connection of these conditions and traits to the genomic information has taken place through advancements in biobanking, EHR mining, and data capitalization as well as the capability for large-scale data sharing over the last 10 years.

Development of Phenome-Wide Association Studies to Conduct Large-Scale Analysis of Human Phenomes

GWASs were enabled by genotype array technology that allowed researchers to sample genetic variation across the human genome. Similarly, the adoption of EHRs enabled querying of a broad spectrum of signs, symptoms, diagnoses, and laboratory and radiographic findings across the human phenome. The breadth of phenomic data in EHRs motivated the introduction of a large-scale method representing the analytic inverse of a GWAS: a PheWAS. PheWASs scan a large set of diagnoses or other clinical findings to identify phenomic features associated with single genetic loci (24). One initial application included the exploration of genetic pleiotropy—the phenomenon where a single gene influences multiple traits (23). For example, the PheWAS technique has been used to identify potential functions for the highly polymorphic human leukocyte antigen (HLA) genes encoding major histocompatibility complexes involved in immune processes (45, 52).

Underlying the PheWAS technique is a knowledge base of diagnostic codes that can characterize a cohort on a phenome-wide scale; manually grouped administrative codes are binned to create phecodes that each represent a single disease entity. Phecode mappings (109, 117) can be found at https://phewascatalog.org (23). Currently, phecodes are defined for more than 1,800 diseases, symptoms, and clinical findings (109, 117).

PheWASs have been validated in part by replicating known genotype–phenotype associations; for example, a PheWAS exploring genetic associations with seven diverse diseases replicated four of seven previously established findings (24). Subsequently, a larger study replicated 51 of 77 associations reported in the GWAS Catalog for which there was a matching phecode (23). These studies showed that phenome-wide characterizations of EHR

cohorts could be used for both validation and discovery. However, the results also showed that phecodes and large-scale analytic methods such as PheWAS trade some precision for breadth. Replicated PheWAS associations often exhibit an attenuated effect size compared with the original GWAS. While some attenuation is expected due to regression to the mean, some loss of signal may also be related to the drawbacks of scalable billing code–based phenotypes, which are subject to loss of sensitivity and specificity.

Application of Large-Scale Phenomic Analyses

The development of PheWASs has inspired more recent work to leverage EHRs to identify genetic syndromes that have a complex phenotypic expression. The PheRS method was initially created to study the impact of rare genetic variants and Mendelian disease (see the sidebar titled Creating a Phenotype Risk Score along with Figure 3). PheRSs use clinical descriptions taken from the Online Mendelian Inheritance in Man (OMIM) database and annotated using the Human Phenotype Ontology (HPO) to create phenotype profiles for thousands of Mendelian diseases. Each individual in a cohort is assigned to a score based on the presence or absence of matching features for the target Mendelian disease. The HPO provides a standardized vocabulary of characteristic abnormalities encountered in human disease. HPO terms can then be mapped to consolidated billing codes (phecodes), individual codes in ICD9 or ICD10, or other information extractable from the EHR, establishing well-coded disease definitions. By then assessing the presence of these features within the record of a patient of interest, one can apply a predictive lens. Specifically, the PheRS for a given Mendelian disease is defined as the sum of clinical features observed in a given subject weighted by the log inverse prevalence of the feature—essentially a disease likelihood based on tractable canonical disease symptom overlap. After initially being used to assess the pathogenicity of rare genetic variants, PheRSs were refined to serve as a scalable approach to identifying undiagnosed disease and assessing gene expression (9, 10, 91, 122).

EHRs have also been used to interpret clinical genetic sequences more efficiently. Interpreting clinical genetic data often requires manual chart abstraction to help prioritize and interpret genetic variants. Tools like ClinPhen have been designed to automate this process, using natural language processing techniques to extract clinical concepts relevant to Mendelian disease diagnosis and map them to the HPO. Clark et al. (16) described an automated pipeline that extracts features from an EHR and pairs them with whole-exome sequencing results.

Applying Discovery Research Methods to Translational Medicine

Utilizing genomic data in a clinical setting can be associated with many barriers, including operational issues, physician comprehension of the results and attitudes toward genetic data, determining how to effectively utilize results, clinical decision support, integration of the result data into the EHR itself, and concerns about associated costs (29, 53).
As an example, Vanderbilt University Medical Center developed a research program in 2010 aimed at determining the effectiveness of preemptive pharmacogenomic testing of high-risk patient populations to decrease medication-related adverse events. This program, called the Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) (81), combined genomic testing, integration of the results into the EHR,

and associated clinical decision support for physicians. Physician attitudes were studied, and while the majority agreed that immediate notification of significant drug–genome interaction was beneficial, there were divisions regarding the responsibility of the physician, which physicians should be notified, and whether patients should be notified directly (79). Nationwide surveys supported the findings, suggesting that physicians did not feel prepared regarding pharmacogenomic testing (99).

As pharmacogenomic testing has become more common, physicians have become more familiar with resources such as the Pharmacogenomics Knowledge Base (PharmGKB) (112), the Pharmacogenomics Research Network, and the Clinical Pharmacogenetics Implementation Consortium guidelines (82). While acceptance and utilization of pharmacogenomics are still challenges (37, 40), recent studies have shown that education and even having physicians undergo personal genomic testing can greatly alter attitudes and understanding (59). Examples from the pharmacogenomics field can inform barriers to genomic medicine in general. Clinical education, understanding, and support are key for the successful integration of genomics into a healthcare setting.

Moving from discovery research to translational medicine and ultimately informing changes in patient care has been the focus of billions of dollars of research in countries across the whole world for the last decade (101). Many of these countries nationally fund networks and research programs whose main goals are to overcome barriers to implementing genomic medicine in clinical practices and determine best practices and lessons for translational medicine as a whole. Lessons from these networks inform the integration of genomics into healthcare research. Two networks funded by the National Human Genome Research Institute have focused on large-scale EHR research and integrating genomic results into translational research and clinical practice: the eMERGE network and the Implementing Genomics in Practice (IGNITE) network.

The eMERGE network moved from discovery research focused on GWASs (27, 85), PheWASs (23, 24, 107), and e-phenotyping (48, 92) in its earlier phases to returning and integrating actionable genomic variants (31, 32). The network is currently investigating how genomic and polygenic risk factors integrate and associate with development of common complex diseases. It led the field in the reuse of EHR data for secondary research (20, 41, 55, 68, 78), in addition to developing methods for the integration of genomic results and assessment of clinical uptake and utilization of genomics over the last several years (6, 21, 36, 46, 88, 113). Lessons from the network (32) include up-front data requirements; rapid sharing of data across EHR systems by using standardized common data models and collection with local expertise; strong centralized communication, policies, and project management; consistency in methods; harmonization of data flow and integration utilizing Health Level 7 International (HL7) and Fast Healthcare Interoperability Resources (FHIR) standards when possible; and a specific study design with the identification of attainable short- and long-term goals for downstream analysis of clinical utilization and uptake across sites for future downstream analyses (31).

The IGNITE network focuses on accelerating genomic medicine utilization by developing methods for incorporating genomics into clinical care across diverse settings. With an

emphasis on implementation science, the lessons from the IGNITE network highlight the importance of having transdisciplinary teams to ensure appropriate expertise during implementation, understanding the educational needs of clinic providers and staff and having appropriate tools to address these needs, carrying out patient education and engagement, and (as mentioned above for the eMERGE network) having specific study designs for the outcomes of interest and strong IT support and data flow standards (39, 98). The IGNITE network also identified that increasing the priority at the institution of integrating genomics within the health system EHRs by utilizing data warehouses can assist with overcoming integration challenges (98). The IGNITE network's Clinical Informatics Working Group recently published a framework data flow for germline genomic result generation and integration into an institution's EHRs from both external and internal vendors (25). The framework, validated through a survey at both IGNITE and eMERGE institutions, highlights the importance of the automation and standardization of genomic information and reporting across the pipeline to enhance utility and streamlined integration, since the knowledge bases associated with genomic medicine are constantly evolving (25). These lessons are applicable not only to other consortia but also to the transition from discovery to translational research.

The Challenges of Clinical Utility and Implementation

The growth of large-scale genomic and phenomic analyses and the resulting genomic risk scores (GRSs) have greatly increased the opportunities for translating genomics to clinical practice. Publications referencing GWASs, PheWASs, and GRSs were introduced between 2000 and 2010 but exponentially increased early in the last decade and have continued to grow (Figure 4). The majority of these publications describe applications of these three methods to new clinical domains. However, there have been few published studies of translation, implementation, or clinical utility. Such studies have been initiated and include the latest phase of eMERGE, which began in mid-2020 and is investigating genomic risk assessment and management. During this phase, the network is focusing on returning an integrated risk to participants that incorporates PRSs, family history, clinical risk factors, and monogenic risks.

In addition, clinical trials to study early-phase commercial products featuring PRSs are recruiting participants. Both the commercial and academic environment have recognized the challenge of using PRSs in practice; many are based on GWAS data that do not include diverse ethnic representation and consequently perform poorly for people with non-European ancestry (28). The actionability of PRSs for diseases with a long latency (such as many cancers and cardiovascular disease) is also not established for all age ranges, and outcomes are difficult to study given the need for long follow-up. Given this state, most experts are cautious regarding the utility of applying PRSs to clinical care. The National Comprehensive Cancer Network guidelines specifically referenced PRSs in a 2020 update to discourage clinical use outside of clinical trials or until their interpretation and therapeutic implications could be clarified (22). The ability of PheRSs to identify patients with genomic syndromes is also not yet established, in part because the technique requires very large populations with detailed EHRs to accrue a sufficient number of patients with a latent or unrecognized genomic syndrome.

Barriers to the adoption of genomic medicine, including complex interventions such as PRSs, have been identified at multiple institutions. Chief among these are provider uptake, education, and willingness to integrate into clinical care (79, 99). As longitudinal EHR databases become more accessible across diverse populations and tied to banked genetic data, the field of clinical genomics will rapidly expand. Networks that aim to increase diversity in recruited population cohorts, such as the *All of Us* Research Program and eMERGE, will provide researchers with the diverse genetic samples that have previously been lacking from studies of large cohorts. Studies demonstrating the feasibility and clinical utility of these new techniques in diverse populations are critical for widespread adoption. Despite these reservations, it is likely that both PRS and PheRS interventions will follow the path of more established genomic medicine inventions, such as pharmacogenomics and the diagnosis of unknown diseases, which gradually gained acceptance in the clinic as clinical trials and implementation cohort studies were completed (64, 86, 114). Clinical trials have begun on PRSs over the last few years (72), and more trials are expected as the PRS model is vetted across multiple populations. Successful completion of current and future consortia will be needed to formally test clinical use. Since genomic medicine is a relatively young field, common approaches to outcome assessment for polygenic risk will need to be reconciled across studies and established, similarly to the approach taken for monogenic disease (80).

CONCLUSIONS

High-throughput and large-scale methods for associating genomic and phenomic data have accelerated the discovery of large sets of genetic markers with potential prognostic clinical value. The methods are increasingly dependent on the availability of comprehensive and longitudinal EHR data using structured data linked to sequence data on very large populations. These innovations have fueled the development of new risk stratification and predictive tools that have proven value for discovery, particularly the ability to characterize rare variants and pleiotropy, and have promising but unproven clinical value. For clinical use, there is a need to define the actionability of the predictive information, perform additional validation across ethnicities, and perform outcome-based studies. These advances in translational genomic medicine are founded on the collaborative nature of cross-institutional and global data sharing made possible by the advances in EHR utilization over the past few decades.

ACKNOWLEDGMENTS

This work was supported by grants U01 HG011166 and U01 HG010232 from the National Human Genome Research Institute and award UL1 TR002243 from the Clinical and Translational Science Awards program of the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. This work was also supported by grant R01 LM010685-09.

Glossary

EHR	electronic health record
PheWAS	phenome-wide association study

GWAS	genome-wide association study
PheRS	phenotype risk score
PRS	polygenic risk score

LITERATURE CITED

1. Abul-Husn NS, Kenny EE. 2019. Personalized medicine and the power of electronic health records. Cell 177:58–69 [PubMed: 30901549]
2. Abul-Husn NS, Manickam K, Jones LK, Wright EA, Hartzel DN, et al. 2016. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science 354:aaf7000 [PubMed: 28008010]
3. Adler-Milstein J, Jha AK. 2017. HITECH Act drove large gains in hospital electronic health record adoption. Health Aff. 36:1416–22
4. All of Us Res. Program Investig. 2019. The “All of Us” Research Program. N. Engl. J. Med 381:668–76 [PubMed: 31412182]
5. Allman R, Dite GS, Hopper JL, Gordon O, Starlard-Davenport A, et al. 2015. SNPs and breast cancer risk prediction for African American and Hispanic women. Breast Cancer Res. Treat 154:583–89 [PubMed: 26589314]
6. Antommaria AHM, Brothers KB, Myers JA, Feygin YB, Aufox SA, et al. 2018. Parents’ attitudes toward consent and data sharing in biobanks: a multisite experimental survey. AJOB Empir. Bioeth 9:128–42 [PubMed: 30240342]
7. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, et al. 2008. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br. J. Cancer 98:1457–66 [PubMed: 18349832]
8. Baretta Z, Mocellin S, Goldin E, Olopade OI, Huo D. 2016. Effect of BRCA germline mutations on breast cancer prognosis: a systematic review and meta-analysis. Medicine 95:e4975 [PubMed: 27749552]
9. Bastarache L, Hughey JJ, Goldstein JA, Bastarache JA, Das S, et al. 2019. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J. Med. Inform. Assoc 26:1437–47
10. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, et al. 2018. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 359:1233–39 [PubMed: 29590070]
11. Benjamin EJ, Dupuis J, Larson MG, Lunetta KL, Booth SL, et al. 2007. Genome-wide association with select biomarker traits in the Framingham Heart Study. BMC Med. Genet 8(Suppl. 1):S11 [PubMed: 17903293]
12. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47:D1005–12 [PubMed: 30445434]
13. Casey JA, Schwartz BS, Stewart WF, Adler NE. 2016. Using electronic health records for population health research: a review of methods and applications. Annu. Rev. Public Health 37:61–81 [PubMed: 26667605]
14. Chatterjee N, Shi J, García-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet 17:392–406 [PubMed: 27140283]
15. Cherkin DC, Deyo RA, Volinn E, Loeser JD. 1992. Use of the International Classification of Diseases (ICD-9-CM) to identify hospitalizations for mechanical low back problems in administrative databases. Spine 17:817–25 [PubMed: 1386943]
16. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, et al. 2019. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci. Transl. Med 11:eaat6177 [PubMed: 31019026]
17. Collins FS, Hudson KL, Briggs JP, Lauer MS. 2014. PCORnet: turning a dream into reality. J. Am. Med. Inform. Assoc 21:576–77 [PubMed: 24821744]

18. Collins FS, Varmus H. 2015. A new initiative on precision medicine. N. Engl. J. Med 372:793–95 [PubMed: 25635347]
19. Collins R 2012. What makes UK Biobank special? Lancet 379:1173–74 [PubMed: 22463865]
20. Crawford DC, Crosslin DR, Tromp G, Kullo IJ, Kuivaniemi H, et al. 2014. eMERGEing progress in genomics—the first seven years. Front. Genet 5:184 [PubMed: 24987407]
21. Crosslin DR, Robertson PD, Carrell DS, Gordon AS, Hanna DS, et al. 2015. Prospective participant selection and ranking to maximize actionable pharmacogenetic variants and discovery in the eMERGE Network. Genome Med. 7:67 [PubMed: 26221186]
22. Daly MB, Pilarski R, Yurgelun MB, Berry MP, Buys SS, et al. 2020. NCCN Guidelines Insights: Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, version 1.2020. J. Natl. Compr. Cancer Netw 18:380–91
23. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol 31:1102–10 [PubMed: 24270849]
24. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26:1205–10 [PubMed: 20335276]
25. Dexter P, Ong H, Elsey A, Bell G, Walton N, et al. 2020. Development of a genomic data flow framework: results of a survey administered to NIH-NHGRI IGNITE and eMERGE consortia participants. AMIA Annu. Symp. Proc 2019:363–70 [PubMed: 32308829]
26. Genet Diabetes. Replication Meta-Anal. (DIAGRAM) Consort., Asian Genet. Epidemiol. Netw. Type 2 Diabetes (AGEN-T2D) Consort., South Asian Type 2 Diabetes (SAT2D) Consort., Mex. Am. Type 2 Diabetes (MAT2D) Consort., Type 2 Diabetes Genet. Explor. Next-Gener. Seq. Multi-Ethnic Samples (T2D-GENES) Consort., et al. 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet 46:234–44 [PubMed: 24509480]
27. Dumitrescu L, Ritchie MD, Denny JC, El Rouby NM, McDonough CW, et al. 2017. Genome-wide study of resistant hypertension identified from electronic health records. PLOS ONE 12:e0171745 [PubMed: 28222112]
28. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, et al. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun 10:3328 [PubMed: 31346163]
29. Dunnenberger HM, Crews KR, Hoffman JM, Caudle KE, Broeckel U, et al. 2015. Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers. Annu. Rev. Pharmacol. Toxicol 55:89–106 [PubMed: 25292429]
30. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, et al. 2020. Predictive accuracy of a polygenic risk score-enhanced prediction model versus a clinical risk score for coronary artery disease. JAMA 323:636–45 [PubMed: 32068818]
31. eMERGE Consort. 2019. Harmonizing clinical sequencing and interpretation for the eMERGE III Network. Am. J. Hum. Genet 105:588–605 [PubMed: 31447099]
32. eMERGE Consort. 2021. Lessons learned from the eMERGE Network: balancing genomics in discovery and in practice. Hum. Genet. Genom. Adv 2:100018
33. Escott-Price V, Myers AJ, Huentelman M, Hardy J. 2017. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Ann. Neurol 82:311–14 [PubMed: 28727176]
34. Evans RS. 2016. Electronic health records: then, now, and in the future. Yearb. Med. Inform. Suppl 1:S48–61
35. Food Drug Adm. 2018. Food and Drug Administration Amendments Act (FDAAA) of 2007. Food and Drug Administration. https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/food-and-drug-administration-amendments-act-fdaaa-2007
36. Fossey R, Kochan D, Winkler E, Pacyna JE, Olson J, et al. 2018. Ethical considerations related to return of results from genomic medicine projects: the eMERGE Network (phase III) experience. J. Pers. Med 8:2

37. Frigon M-P, Blackburn M-È, Dubois-Bouchard C, Gagnon A-L, Tardif S, Tremblay K. 2019. Pharmacogenetic testing in primary care practice: opinions of physicians, pharmacists and patients. Pharmacogenomics 20:589–98 [PubMed: 31190623]
38. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. 2020. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. JMIR Form. Res 4:e17687 [PubMed: 32852280]
39. Ginsburg GS, Horowitz CR, Orlando LA. 2019. What will it take to implement genomics in practice? Lessons from the IGNITE Network. Pers. Med 16:259–61
40. Giri J, Curry TB, Formea CM, Nicholson WT, Vitek CRR. 2018. Education and knowledge in pharmacogenomics: still a challenge? Clin. Pharmacol. Ther 103:752–55 [PubMed: 29417560]
41. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, et al. 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med 15:761–71 [PubMed: 23743551]
42. Guttmacher AE, Collins FS. 2002. Genomic medicine—a primer. N. Engl. J. Med 347:1512–20 [PubMed: 12421895]
43. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–21 [PubMed: 15761120]
44. Hajek C, Guo X, Yao J, Hai Y, Johnson WC, et al. 2018. Coronary heart disease genetic risk score predicts cardiovascular disease risk in men, not women. Circ. Genom. Precis. Med 11:e002324 [PubMed: 30354305]
45. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. 2013. A PheWAS approach in studying HLA-DRB1*1501. Genes Immun 14:187–91 [PubMed: 23392276]
46. Herr TM, Peterson JF, Rasmussen LV, Caraballo PJ, Peissig PL, Starren JB. 2019. Pharmacogenomic clinical decision support design and multi-site process outcomes analysis in the eMERGE Network. J. Am. Med. Inform. Assoc 26:143–48 [PubMed: 30590574]
47. HIPAA J 2020. What is the HITECH Act? HIPAA Journal. https://www.hipaajournal.com/what-is-the-hitech-act
48. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, et al. 2019. Facilitating phenotype transfer using a common data model. J. Biomed. Inform 96:103253 [PubMed: 31325501]
49. Int. Schizophr. Consort. 2009. Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. Nature 460:748–52 [PubMed: 19571811]
50. Jencks SF. 1992. Accuracy in recorded diagnoses. JAMA 267:2238–39 [PubMed: 1556801]
51. Juhn Y, Liu H. 2020. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J. Allergy Clin. Immunol 145:463–69 [PubMed: 31883846]
52. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, et al. 2017. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. Sci. Transl. Med 9:eaai8708 [PubMed: 28490672]
53. Karnes JH, Van Driest S, Bowton EA, Weeke PE, Mosley JD, et al. 2014. Using systems approaches to address challenges for clinical implementation of pharmacogenomics. Wiley Interdiscip. Rev. Syst. Biol. Med 6:125–35 [PubMed: 24319008]
54. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet 50:1219–24 [PubMed: 30104762]
55. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, et al. 2012. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J. Am. Med. Inform. Assoc 19:212–18 [PubMed: 22101970]
56. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308:385–89 [PubMed: 15761122]
57. Kluger MD, Sofair AN, Heye CJ, Meek JI, Sodhi RK, Hadler JL. 2001. Retrospective validation of a surveillance system for unexplained illness and death: New Haven County, Connecticut. Am. J. Public Health 91:1214–19 [PubMed: 11499106]
58. Krishnamurthy S 2015. Biospecimen repositories and cytopathology. Cancer Cytopathol. 123:152–61 [PubMed: 25524469]

59. Lee KH, Min BJ, Kim JH. 2019. Personal genome testing on physicians improves attitudes on pharmacogenomic approaches. PLOS ONE 14:e0213860 [PubMed: 30921347]
60. Lezzoni LI. 1990. Using administrative diagnostic data to assess the quality of hospital care. Pitfalls and potential of ICD-9-CM. Int. J. Technol. Assess. Health Care 6:272–81 [PubMed: 2203703]
61. Mahajan A, Spracklen CN, Zhang W, Ng MC, Petty LE, et al. 2020. Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. medRxiv 2020.09.22.20198937. 10.1101/2020.09.22.20198937
62. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. 2020. Molecular profiling for precision cancer therapies. Genome Med. 12:8 [PubMed: 31937368]
63. Manolio TA, Chisholm RL, Ozenberger B, Roden DM, Williams MS, et al. 2013. Implementing genomic medicine in the clinic: The future is here. Genet. Med 15:258–67 [PubMed: 23306799]
64. Manolio TA, Rowley R, Williams MS, Roden D, Ginsburg GS, et al. 2019. Opportunities, resources, and techniques for implementing genomics in clinical care. Lancet 394:511–20 [PubMed: 31395439]
65. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet 51:584–91 [PubMed: 30926966]
66. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, et al. 2019. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am. J. Hum. Genet 104:21–34 [PubMed: 30554720]
67. May DS, Kelly JJ, Mendlein JM, Garbe PL. 1988. Surveillance of major causes of hospitalization among the elderly, 1988. Morb. Mortal. Wkly. Rep. Surveill. Summ 40(SS-1):7–21
68. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, et al. 2011. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genom 4:13
69. McDonald CJ, Murray R, Jeris D, Bhargava B, Seeger J, Blevins L. 1977. A computer-based record and clinical monitoring system for ambulatory care. Am. J. Public Health 67:240–45 [PubMed: 842761]
70. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. 2017. Clinical data reuse or secondary use: current status and potential future progress. Yearb. Med. Inform 26:38–52 [PubMed: 28480475]
71. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, et al. 2020. Predictive accuracy of a polygenic risk score compared with a clinical risk score for incident coronary heart disease. JAMA 323:627–35 [PubMed: 32068817]
72. Natarajan P, Young R, Stitziel NO, Padmanabhan S, Baber U, et al. 2017. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. Circulation 135:2091–101 [PubMed: 28223407]
73. Onengut-Gumuscu S, Chen W-M, Robertson CC, Bonnie JK, Farber E, et al. 2019. Type 1 diabetes risk in African-ancestry participants and utility of an ancestry-specific genetic risk score. Diabetes Care 42:406–15 [PubMed: 30659077]
74. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, et al. 2002. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet 32:650–54 [PubMed: 12426569]
75. Paik S, Shak S, Tang G, Kim C, Baker J, et al. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med 351:2817–26 [PubMed: 15591335]
76. Palmer LJ. 2007. UK Biobank: bank on it. Lancet 369:1980–82 [PubMed: 17574079]
77. Paskal W, Paskal AM, Dębski T, Gryziak M, Jaworowski J. 2018. Aspects of modern biobank activity – comprehensive review. Pathol. Oncol. Res 24:771–85 [PubMed: 29728978]
78. Pathak J, Wang J, Kashyap S, Basford M, Li R, et al. 2011. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J. Am. Med. Inform. Assoc 18:376–86 [PubMed: 21597104]

79. Peterson JF, Field JR, Shi Y, Schildcrout JS, Denny JC, et al. 2016. Attitudes of clinicians following large-scale pharmacogenomics implementation. Pharmacogenom. J 16:393–98
80. Peterson JF, Roden DM, Orlando LA, Ramirez AH, Mensah GA, Williams MS. 2019. Building evidence and measuring clinical outcomes for genomic medicine. Lancet 394:604–10 [PubMed: 31395443]
81. Pulley JM, Denny JC, Peterson JF, Bernard GR, Vnencak-Jones CL, et al. 2012. Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. Clin. Pharmacol. Ther 92:87–95 [PubMed: 22588608]
82. Relling MV, Klein TE. 2011. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin. Pharmacol. Ther 89:464–67 [PubMed: 21270786]
83. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–17 [PubMed: 8801636]
84. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am. J. Hum. Genet 86:560–72 [PubMed: 20362271]
85. Ritchie MD, Verma SS, Hall MA, Goodloe RJ, Berg RL, et al. 2014. Electronic medical records and genomics (eMERGE) network exploration in cataract: several new potential susceptibility loci. Mol. Vis 20:1281–95 [PubMed: 25352737]
86. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, et al. 2019. Pharmacogenomics. Lancet 394:521–32 [PubMed: 31395440]
87. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, et al. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther 84:362–69 [PubMed: 18500243]
88. Rohrer Vitek CR, Abul-Husn NS, Connolly JJ, Hartzler AL, Kitchner T, et al. 2017. Healthcare provider education to support integration of pharmacogenomics in practice: the eMERGE Network experience. Pharmacogenomics 18:1013–25 [PubMed: 28639489]
89. Roos LL, Roos NP, Cageorge SM, Nicol JP. 1982. How good are the data? Reliability of one health care data bank. Med. Care 20:266–76 [PubMed: 7078285]
90. Rudolph A, Song M, Brook MN, Milne RL, Mavaddat N, et al. 2018. Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. Int. J. Epidemiol 47:526–36 [PubMed: 29315403]
91. Salvatore M, Beesley LJ, Fritsche LG, Hanauer D, Shi X, et al. 2021. Phenotype risk scores (PheRS) for pancreatic cancer using time-stamped electronic health record data: discovery and validation in two large biobanks. J. Biomed. Inform 113:103652 [PubMed: 33279681]
92. Shang N, Liu C, Rasmussen LV, Ta CN, Caroll RJ, et al. 2019. Making work visible for electronic phenotype implementation: lessons learned from the eMERGE network. J. Biomed. Inform 99:103293 [PubMed: 31542521]
93. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, et al. 2019. Developing a portable natural language processing based phenotyping system. BMC Med. Inform. Decis. Mak 19(Suppl. 3):78 [PubMed: 30943974]
94. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, et al. 2019. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. Diabetes Care 42:200–207 [PubMed: 30655379]
95. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. 2019. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med. Inform 7:e12239 [PubMed: 31066697]
96. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, et al. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. J. Am. Med. Inform. Assoc 21:221–30 [PubMed: 24201027]
97. Simonson MA, Wills AG, Keller MC, McQueen MB. 2011. Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. BMC Med. Genet 12:146 [PubMed: 22029572]

98. Sperber NR, Carpenter JS, Cavallari LH, Damschroder LJ, Cooper-DeHoff RM, et al. 2017. Challenges and strategies for implementing genomic services in diverse settings: experiences from the Implementing GeNomics In pracTicE (IGNITE) network. BMC Med. Genom 10:35
99. Stanek EJ, Sanders CL, Taber KAJ, Khalid M, Patel A, et al. 2012. Adoption of pharmacogenomic testing by US physicians: results of a nationwide survey. Clin. Pharmacol. Ther 91:450–58 [PubMed: 22278335]
100. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, et al. 2010. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann. Intern. Med 153:600–6 [PubMed: 21041580]
101. Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, et al. 2019. Integrating genomics into healthcare: a global responsibility. Am. J. Hum. Genet 104:13–20 [PubMed: 30609404]
102. Tada H, Melander O, Louie JZ, Catanese JJ, Rowland CM, et al. 2016. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. Eur. Heart J 37:561–67 [PubMed: 26392438]
103. Thomas M, Sakoda LC, Hoffmeister M, Rosenthal EA, Lee JK, et al. 2020. Genome-wide modeling of polygenic risk score in colorectal cancer risk. Am. J. Hum. Genet 107:432–44 [PubMed: 32758450]
104. Torkamani A, Wineinger NE, Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet 19:581–90 [PubMed: 29789686]
105. Toulopoulou T, Zhang X, Cherny S, Dickinson D, Berman KF, et al. 2019. Polygenic risk score increases schizophrenia liability through cognition-relevant pathways. Brain J. Neurol 142:471–85
106. Vaught J, Rogers J, Myers K, Lim MD, Lockhart N, et al. 2011. An NCI perspective on creating sustainable biospecimen resources. J. Natl. Cancer Inst. Monogr 2011(42):1–7 [PubMed: 21672889]
107. Verma A, Verma SS, Pendergrass SA, Crawford DC, Crosslin DR, et al. 2016. eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. BMC Med. Genom 9(Suppl. 1):32
108. Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. Am. J. Hum. Genet 90:7–24 [PubMed: 22243964]
109. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLOS ONE 12:e0175508 [PubMed: 28686612]
110. Wei W-Q, Denny JC. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 7:41 [PubMed: 25937834]
111. Wellcome Trust Case Control Consort. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–78 [PubMed: 17554300]
112. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, et al. 2012. Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther 92:414–17 [PubMed: 22992668]
113. Williams JL, Chung WK, Fedotov A, Kiryluk K, Weng C, et al. 2018. Harmonizing outcomes for genomic medicine: comparison of eMERGE outcomes to ClinGen outcome/intervention pairs. Healthc. Basel Switz 6:83
114. Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, et al. 2019. Genomic medicine for undiagnosed diseases. Lancet 394:533–40 [PubMed: 31395441]
115. Wood GC, Still CD, Chu X, Susek M, Erdman R, et al. 2008. Association of chromosome 9p21 SNPs with cardiovascular phenotypes in morbid obesity using electronic health record data. Genom. Med 2:33–43
116. Wray NR, Goddard ME, Visscher PM. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17:1520–28 [PubMed: 17785532]
117. Wu P, Gifford A, Meng X, Li X, Campbell H, et al. 2019. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med. Inform 7:e14325 [PubMed: 31553307]

118. Yu H, Shi Z, Lin X, Bao Q, Jia H, et al. 2020. Broad- and narrow-sense validity performance of three polygenic risk score methods for prostate cancer risk assessment. Prostate 80:83–87 [PubMed: 31634418]

119. Zeggini E, Gloyn AL, Barton AC, Wain LV. 2019. Translational genomics and precision medicine: moving from the lab to the clinic. Science 365:1409–13 [PubMed: 31604268]

120. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. 2019. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Trans. Comput. Biol. Bioinform 16:139–53 [PubMed: 29994486]

121. Zhao J, Zhang Y, Schlueter DJ, Wu P, Kerchberger VE, et al. 2019. Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: cardiovascular disease case study. J. Biomed. Inform 98:103270 [PubMed: 31445983]

122. Zhong X, Yin Z, Jia G, Zhou D, Wei Q, et al. 2020. Electronic health record phenotypes associated with genetically regulated expression of CFTR and application to cystic fibrosis. Genet. Med 22:1191–200 [PubMed: 32296164]

RELATED RESOURCES

Clinical Trials.gov: https://www.clinicaltrials.gov

Electronic Medical Records and Genomics (eMERGE) network: https://emerge-network.org

GWAS Catalog: https://www.ebi.ac.uk/gwas

Informatics for Integrating Biology and the Bedside (i2b2): https://www.i2b2.org

National COVID Cohort Collaborative: https://covid.cd2h.org

National Patient-Centered Clinical Research Network (PCORnet): https://pcornet.org

Observational Health Data Sciences and Informatics: https://www.ohdsi.org

Pharmacogenomics Knowledge Base (PharmGKB): https://www.pharmgkb.org

Pharmacogenomics Research Network: http://www.pgrn.org

PheWAS Resources: https://phewascatalog.org

CREATING A PHENOTYPE RISK SCORE

A PheRS is a numeric value assigned to an individual based on the number of features they share with the clinical description of a disease.

Calculating the Phenotype Risk Score

PheRSs are calculated by summing up the weights of each feature present in the EHR (Figure 3a). The weights are defined as the $-\log_{10}$ of the prevalence of the phenotype in a large cohort. Different scores can be calculated for different sets of phenotypes. The phenotype sets can be defined based on the clinical manifestations of a particular disease or created de novo to describe a particular patient. The score is intended to reflect the degree of similarity between a patient and the feature set.

Phenotype Risk Scores for a Mendelian Disease

OMIM provides clinical descriptions for thousands of Mendelian diseases. These descriptions have been annotated with HPO terms, and a map has been created between HPO terms and phecodes; thus, every disease in OMIM can be described as a set of phecodes. Some HPO terms are mapped to phecodes that match exactly, while others are mapped to broader phecodes. Figure 3b shows an abbreviated version of OMIM's clinical description for cystic fibrosis. When PheRSs are applied to a cohort, they can distinguish between cases and controls for a particular disease without relying on the disease label itself. PheRSs have been used to identify pathogenic variants in HER-linked biobanks.

Finding Undiagnosed Patients

Beyond using PheRSs to study rare genetic variants, preliminary work suggests that they may help identify undiagnosed patients. For example, in Figure 3c, a patient diagnosed with cystic fibrosis late in life had a PheRS in the 99th percentile prior to diagnosis.

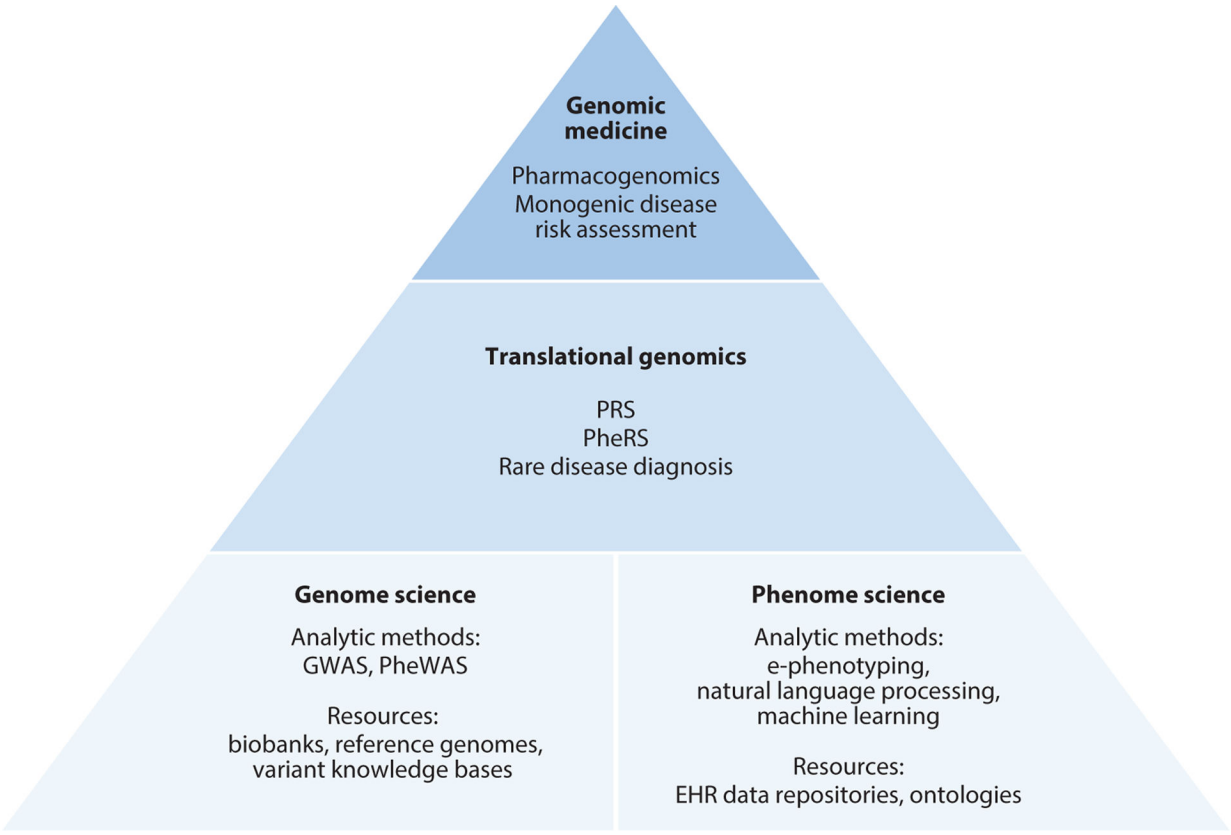

Figure 1.

Advancing translational genomics relies on research across the genome and phenome. Progress relies both on enabling resources and on analytic methods and tools to capitalize on those resources. Discovery research utilizing new technologies built off large-scale EHR and genomic data has led to clinical translation and implementation and to eventual changes in clinical practice. Abbreviations: EHR, electronic health record; e-phenotyping, electronic phenotyping; GWAS, genome-wide association study; PheRS, phenotype risk score; PheWAS, phenome-wide association study; PRS, polygenic risk score.

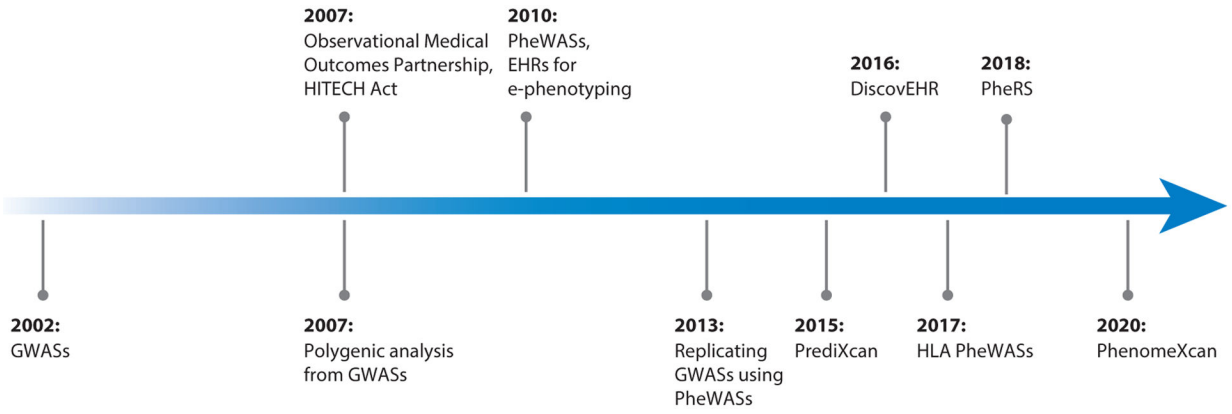

Figure 2.

Milestones enabling translational research. EHR (*top*) and genomic data (*bottom*) technologies facilitated advancements in medical genomics, increasing the understanding of common complex diseases. Abbreviations: EHR, electronic health record; e-phenotyping, electronic phenotyping; GWAS, genome-wide association study; HITECH, Health Information Technology for Economic and Clinical Health; HLA, human leukocyte antigen; PheRS, phenotype risk score; PheWAS, phenome-wide association study.

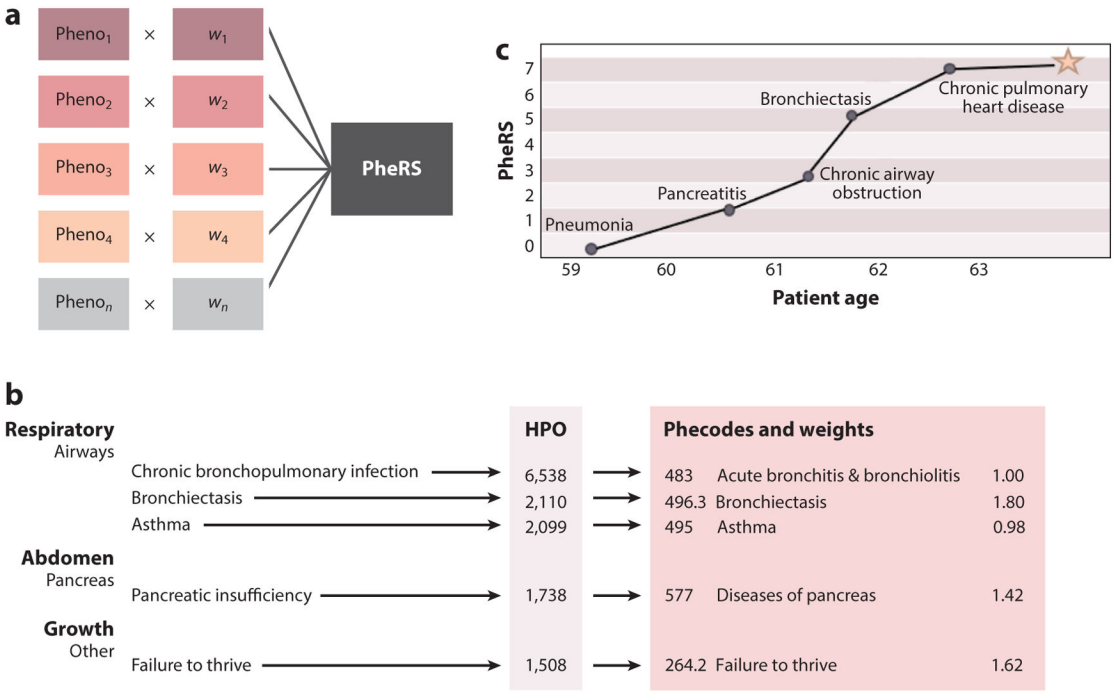

Figure 3.

Creating a PheRS (see also the sidebar titled Creating a Phenotype Risk Score). (*a*) Summing the weights of each feature present in an EHR to calculate the PheRS. (*b*) An abbreviated version of OMIM's clinical description for cystic fibrosis. (*c*) An example PheRS plot for a patient diagnosed with cystic fibrosis late in life. Before the diagnosis, this patient had a cystic fibrosis PheRS in the 99th percentile. Abbreviations: EHR, electronic health record; HPO, Human Phenotype Ontology; OMIM, Online Mendelian Inheritance in Man; PheRS, phenotype risk score.

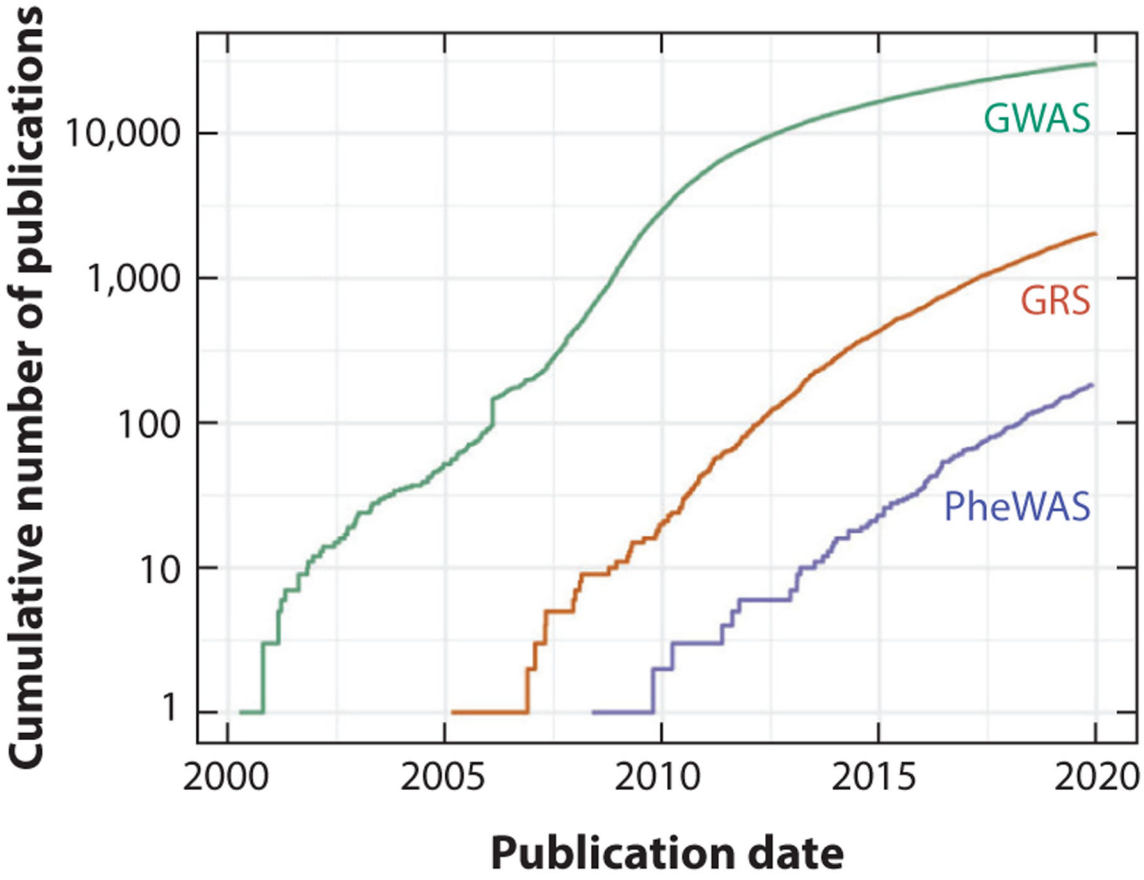

Figure 4.

Cumulative number of publications that included terms for common large-scale analytic methods (GWAS, genome-wide association study; PheWAS, phenome-wide association study, phenome wide; GRS, genomic risk score, genetic risk score, polygenic risk score) in the title, abstract, or MeSH term. Enabling methods such as GWAS and PheWAS in combination with the availability of large-scale EHR data laid the foundation for translational research such as PRSs and GRSs. Abbreviations: EHR, electronic health record; GRS, genomic risk score; GWAS, genome-wide association study; MeSH, Medical Subject Headings; PheWAS, phenome-wide association study; PRS, polygenic risk score.

Table 1

Commonly utilized data elements and tools for extracting phenotypes from EHRs for phenome-wide research

Data element	Description	Utility for phenome science
Claims data	Billing claims data used for diagnosis and procedures; examples include ICD, phecodes (derived from ICD), and CPT	Structured data to extrapolate patient diagnoses, symptoms, findings, and procedures
Demographics	Age, sex/gender, race, ethnicity, date of birth, date of death	Covariate adjustments, cohort definition, structured data
Indexed concepts from clinical narratives	Terms may be mapped to SNOMED-CT and the HPO	Standardizing phenotype concepts to index and merge narrative text
Semistructured documents	Problem lists, family history, flow sheets, radiology, pathology, procedures, cytology reports	Natural language processing for complex e-phenotyping
Encounters	Admission-discharge-transfer, provider and clinic assignments	Severity stratification, healthcare utilization
Laboratory	Laboratory name, value, unit, date; standardized by LOINC in some EHRs	Criteria for detecting diagnoses, cohort definitions, covariate adjustments
Medications	Medication name, dosing, frequency, route, duration, form, strength standardized to RxNorm standard	Criteria for detecting exposures, cohort definitions, covariate adjustments
Tumor (cancer) registry	Organization (e.g., North American Association of Central Cancer Registries) for cancer registry data across public and private organizations for standardization	Cancer-related e-phenotyping
Vital signs	Blood pressure, BMI, height, weight, temperature	Covariate adjustments, structured data

Abbreviations: BMI, body mass index; CPT, Current Procedural Terminology; EHR, electronic health record; e-phenotyping, electronic phenotyping; HPO, Human Phenotype Ontology; ICD, International Classification of Diseases; LOINC, Logical Observation Identifiers Names and Codes; SNOMED-CT, Systematized Nomenclature of Medicine–Clinical Terms.