Published in final edited form as:

Curr Opin Lipidol. 2020 April ; 31(2): 85–93. doi:10.1097/MOL.0000000000000662.

Using the electronic health record for genomics research

Maya S. Safarova,

Iftikhar J. Kullo

Atherosclerosis and Lipid Genomics Laboratory and Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota, USA

Abstract

Purpose of review—Although primarily designed for medical documentation and billing purposes, the electronic health record (EHR) has significant potential for translational research. In this article, we provide an overview of the use of the EHR for genomics research with a focus on heritable lipid disorders.

Recent findings—Linking the EHR to genomic data enables repurposing of vast phenotype data for genomic discovery. EHR data can be used to study the genetic basis of common and rare disorders, identify subphenotypes of diseases, assess pathogenicity of novel genomic variants, investigate pleiotropy, and rapidly assemble cohorts for genomic medicine clinical trials. EHR-based discovery can inform clinical practice; examples include use of polygenic risk scores for assessing disease risk and use of phenotype data to interpret rare variants. Despite limitations such as missing data, variable use of standards and poor interoperablility between disparate systems, the EHR is a powerful resource for genomic research.

Summary—When linked to genomic data, the EHR can be leveraged for genomic discovery, which in turn can inform clinical care, exemplifying the virtuous cycle of a learning healthcare system.

Keywords

electronic health record; electronic phenotyping; familial hypercholesterolemia; genetics and genomics; informatics; phenome-wide association studies; translational research

INTRODUCTION

Electronic health record (EHR) systems are digital platforms used by healthcare service providers to maintain medical information of patients and for billing purposes. In the United States, the Department of Veterans Affairs introduced the Veterans Health Information Systems and Technology Architecture in the 1970s and academic centers started using EHRs

Correspondence to Iftikhar J. Kullo, MD, Department of Cardiovascular Medicine and the Gonda Vascular Center, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. Kullo.Iftikhar@mayo.edu.

Conflicts of interest

M.S.S. has no conflicts to disclose. I.J.K. is on the advisory board for InformedDNA and received unrestricted research funding from 2bPrecise.

in the 1990s. However, it was only after the passage of the Health Information Technology for Economic and Clinical Health Act in 2009 that EHRs came into widespread use [1].

The EHR includes demographics and admission/discharge data, provider notes, patient entered information, procedures, medications, laboratory results, histopathology reports, and radiology reports. The accumulation of data over time allows the study of temporal evolution of diseases/traits of interest. The abundance of both cross-sectional and longitudinal data in EHRs on diverse populations offers significant opportunities for observational research as well as rapid assembly of cohorts for genomic medicine trials and investigation of rare diseases.

Much of the work demonstrating the potential of EHRs for genomics research was done as part of the Electronic Medical Records and Genomics (eMERGE) Network [2,3] initiated by the National Human Genome Research Institute in 2007. Since its inception, the network has played a major role in developing methods for EHR-based genomic research, including methods for extracting and validating phenotypic data using semiautomated algorithms, and conducting phenome-wide association studies (PheWAS) [3,4]. Subsequently, several biobanks that are linked to EHR data (e.g., UK Biobank, China Kadoorie Biobank, Danish National Biobank, Estonia Biobank, FinnGen Biobank, The Canadian Partnership for Tomorrow Project, EuroBioBank Network, Qatar Biobank) have been established across the world [5].

In this article, we provide an overview of the use of EHR systems for genomics research, particularly research related to heritable lipid disorders. We discuss examples related to familial hypercholesterolemia, the prototypical heritable lipid disorder. We describe how data from an EHR are exported to a data warehouse, the use of standards and common data models (CDM), various EHR data types and approaches to mining EHR data, EHR-based genomics research focusing on lipid traits, and challenges in using EHR data for research. We regret any inadvertent omission of relevant contributions.

ELECTRONIC HEALTH RECORD DATA WAREHOUSES, STANDARDS AND COMMON DATA MODELS

Typically data is ‘wrangled’ from one or more disparate sources in the EHR into a data warehouse where it can be used for diverse purposes including research (Fig. 1) [6]. The Extract-Transform-Load process aggregates and transforms data for ware-housing by reading desired EHR data, converting it into a usable form, and then writing it into a searchable relational database. Hadoop, a programing framework that supports the handling of large datasets across many computers, is often used to process, store and manage voluminous EHR data.

To enable integration, sharing, and retrieval of such data, standards such as the Unified Medical Language System (UMLS) are necessary. UMLS consists of more than 100 clinical terminologies and coding systems that can map-related textual terms and ‘regular expressions’ to structured concepts [7,8], thereby enabling recording, formatting, and retrieval of phenotype information from the EHR [9]. Examples of coding systems include

Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine–Clinical Terms, RxNorm and International Classification of Diseases (ICD).

Another strategy to enable integration and sharing of EHR data across healthcare systems is to use a CDM, a standard collection of schemas (relationships, concepts) with well defined semantics which enables information from different systems to be organized in a common standardized format. Examples include the National Patient-Centered Clinical Research Network, informatics for integrating biology and the bedside (i2b2), and the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [10]. The current OMOP-CDM release consists of 15 clinical data tables that include demographics, observation period, drug exposure, and so on. To map medical terms to standardized vocabularies, the OMOP-CDM makes use of the UMLS. The eMERGE Network used OMOP-CDM and phenotyping tools from the Phenotype KnowledgeBase (PheKB) to facilitate definition of patient cohorts at the network sites, resulting in more effective EHR data integration. The All of us research program, a nationwide cohort that will eventually include EHR and genomic data of one million patients, is also using the OMOP data model [11■■].

ELECTRONIC HEALTH RECORD DATA TYPES

Individual data elements in the EHR can be derived from structured datasets or unstructured clinical text and in some instances both. A brief description of data types is provided below.

International statistical classification of diseases, also known as ICD, is a classification system developed by the WHO for describing patient diagnosis. A list of ICD codes are assigned to a particular hospital admission, and eventually used for reimbursement purposes. ICD-10 contains 68 000 diagnosis codes and provides increased granularity compared with ICD-9 which contained 13 000 diagnosis codes. To reduce data dimensionality, ICD/Current Procedural Terminology (CPT) codes can be clustered into a smaller subset of ‘phecodes’. The Monarch Merged Disease Ontology [12] employs automated clustering of ICD codes into phecodes with subsequent expert curation. Codes can be used alone or in combination with the other ICD/CPT codes and EHR data to ascertain traits of interest.

Medication information can be obtained using standard drug nomenclature such as the National Drug Codes coding system consisting of 10 or 11 digits that uniquely identifies drugs approved by the US Food and Drug Administration and RxNorm, developed by the US National Library of Medicine, which represents medications by ingredients, strength, and dose form. Each combination of ingredient, strength, and dose form has a common unique identifier called RxCUI (concept unique identifier). To identify medication usage prior to the implementation of such standard nomenclature, mining of clinical notes by natural language processing (NLP) may be necessary.

Laboratory tests are often standardized by Logical Observation Identifiers Names and Codes [13] (athree to seven-character long system with >80 000 entries) and transmitted using fast healthcare interoperability resource (FHIR) standards. LOINC-coded laboratory tests can be converted into human phenotype ontology (HPO)-coded terms using a SMART on FHIR

resource containing patient information, test identification, test result, normal reference ranges, and interpretations [14■]. Changes in the codes for a particular lab test over time as well as differing LOINCs for the same test across labs pose challenges in extracting laboratory data from the EHR.

Clinic notes are a rich source of information about medical history, presenting symptom/signs and management plans. The narrative text is processed into structured features by NLP techniques such as named-entity recognition which classifies named entities from unstructured text into predefined categories such as symptom, diagnosis, medication, treatment, procedure, and anatomy. Identification of the medical terms and mapping those terms to concepts can be aided with specific databases of medical terms dictionaries and ontologies. Text-mining strategies have to account for synonyms, acronyms, abbreviations, misspellings, negations (e.g., condition not present), modifiers (e.g., family history of the condition, condition present in the relative).

Lifestyle and environmental measures as well as social determinants of health are not routinely captured [1], prompting an Institute of Medicine recommendation that these and other domains be integrated into the EHR, including four existing (i.e., race/ethnicity, current address, alcohol use, and tobacco use) and eight new domains (e.g., stress, social isolation, physical activity). Researchers can link geocoded patient addresses to location-specific data and use Geographic Information Systems to study an individual's proximity to hazards related to disease susceptibility. This process can be used to study negative health impacts from both direct exposure, for example, air pollution, and contextual exposure, for example, residential zip code poverty rates.

Family history is poorly recorded in EHRs, often not meeting the standards endorsed by the US Agency for Healthcare Research and Quality [15]. Family history of certain diseases (e.g., coronary heart disease) requires age of onset in male or female relatives, and this information is often absent. We and others have attempted to address this limitation of EHRs by mining family history from clinical notes using NLP [16,17]. To improve documentation of family history in the EHR several efforts have been initiated including the use of a patient-facing web-based tool such as MeTree [18■] which employs a SMART-FHIR interface.

Imaging data are usually stored in an ancillary system (e.g., picture archiving and communications systems for radiology images) and, therefore, not directly accessible for interrogation. However, interpretive text reports can be mined by NLP. For example in one study carotid ultrasound reports of 2562 individuals were mined using NLP to identify cases with carotid artery disease defined on the basis of peak systolic and end-diastolic velocities and respective ratios [19].

MINING ELECTRONIC HEALTH RECORD DATA

Electronic phenotyping algorithms

An electronic phenotypic algorithm typically mines structured data elements from an EHR data warehouse alone or in combination with NLP of clinical text, to identify cases

and controls for a disease of interest [20,21]. Such algorithms enable semiautomated phenotyping for large-scale case–control studies and can also identify phenotype subgroups and longitudinal changes in phenotypes of interest [8,22,23]. In the eMERGE Network, electronic phenotyping algorithms are developed typically as a pseudocode by one site, iteratively refined till satisfactory metrics of accuracy are obtained, validated at another site and finally deployed across the remaining sites to identify cases and controls for genetic association analyses. The network has assembled a publicly accessible archive of validated algorithms for EHR phenotyping, that are portable across healthcare systems, in a centralized database (PheKB) [23,24]. A data dictionary with information on the covariates encoded using UMLS accompanies each algorithm. Figure 2 illustrates an example of an electronic phenotyping algorithm to ascertain cases of primary severe hypercholesterolemia from the EHR and Fig. 3 illustrates the EHR-derived data elements that were used to ascertain cases of familial hypercholesterolemia [16].

Phenotype risk scores

Phenotype Risk Scores (PheRS) were developed to detect undiagnosed Mendelian diseases by mapping the relevant clinical features into phecodes annotated with HPO terms [26■■]. The scores, derived from EHR data, are able to distinguish cases and controls for several Mendelian disorders. Recent enhancements to PheRS include integration of ICD-10 codes, linkage of custom groupings to HPO terms, and addition of laboratory measurements [27,28]. PheRS may be helpful in identifying patients with rare lipid disorders who are yet to be diagnosed.

Machine learning

The vast phenotypic data in EHRs provides an ideal platform for machine learning approaches to identify phenotypes of interest and phenotype subgroups as well as develop predictive models. In contrast to the rule-based algorithms described above, machine learning models learn from examples that are provided in the form of inputs (features) and outputs (labels). In ‘supervised learning’ (annotated labels), computers learn how to go from features to labels and then are able to process previously unseen set of inputs [29]. In ‘unsupervised learning’, analyses seek to find hidden patterns within the data, including phenotypic structure within or across categories. Deep learning is a class of machine-learning algorithms that use artificial neural networks that can learn highly complex relationships between features and labels and can mine ‘big data’ in EHRs for a variety of applications including predictive modeling [30].

Banda *et al.* [31■■] used a machine learning-derived algorithm to identify familial hypercholesterolemia cases (defined on the basis of the modified Dutch Lipid Clinic Network criteria) from the EHR with positive predictive values of 88 and 85% in the training and test sets, respectively. The familial hypercholesterolemia Foundation’s FIND FH initiative subsequently utilized this approach to automate the detection of familial hypercholesterolemia from millions of EHRs from multiple centers. Using structured data derived from patients with a confirmed diagnosis by an FH expert, the algorithm identified 75 relevant medical features. The FIND FH tool is portable across *Epic* and was applied in 170 416 201 residents from a national database of healthcare encounters and 173 733

individuals from the Oregon Health and Science University healthcare system identifying 78 and 50% of 'yet-to-be-diagnosed' cases of familial hypercholesterolemia, respectively [32■■].

ELECTRONIC HEALTH RECORD-BASED GENOMICS RESEARCH RELATED TO HERITABLE LIPID DISORDERS

In this section, we discuss examples of using EHR-based research related to heritable lipid disorders, ranging from epidemiology, genomic association studies and PheWAS, to estimates of penetrance and variant annotation.

Epidemiology

Electronic phenotyping algorithms can be useful in assessing prevalence, awareness, detection and control of heritable lipid disorders in the population. We deployed an familial hypercholesterolemia phenotyping algorithm in a cohort of individuals who receive care at Mayo Clinic Rochester [16] and noted a prevalence of 1 : 310. A relevant billing code was present in only 55% (indicating low awareness), statin use was noted in 70% and LDL cholesterol (LDL-C) levels were at goal in 80% and in only 20% of those with premature coronary heart disease (CHD) (indicating inadequate control). Subsequently, similar findings were noted in the Geisinger DicovEHR cohort, where a 'genotype first' approach based on exome sequencing of 50 726 individuals established the prevalence of familial hypercholesterolemia variants to be 1 : 256 in a primarily white population [33]. Only 15% of familial hypercholesterolemia-variant carriers had an ICD-10 diagnosis code and only 58% were on a statin.

Genome-wide association studies (GWAS)

The eMERGE Network pioneered the use of high-density genotype data linked to EHR-derived phenotypes for discovery of genomic variants associated with inter-individual variation in medically relevant traits or disease susceptibility [3,34]. Subsequently, several EHR-based studies have identified genomic variants associated with lipid traits [35■■,36,37■■]. The discovery of common variants influencing LDL-C levels has enabled development of polygenic scores for LDL-C and it has become apparent that a polygenic cause is more common in families with hypercholesterolemia than a monogenic cause.

Genome sequence data

Linkage of genomic sequence data to EHR can facilitate discovery of rare variants that influence lipid traits, interpretation of rare putatively pathogenic variants as well as estimation of penetrance of such variants. In eMERGE phase III, sequence data for 106 medically relevant genes was linked to EHR data [38■■], with the goal of assessing variant pathogenicity, penetrance and outcomes after returning results. From among ~25 000 participants, 203 had pathogenic or likely pathogenic (P/LP) variant in one of the three familial hypercholesterolemia genes. While penetrance of P/LP variants related to familial hypercholesterolemia was high, the same was not observed for P/LP variants in arrhythmia or cardiomyopathy genes. In another study of patients with variants predicted to

be pathogenic in *SCN5A* or *KCNH2,* only 35% had an arrhythmia or ECG phenotype on EHR review [39]. EHR phenotype data can be useful in clarifying pathogenicity of genetic variants. For example in the Return of Actionable Variants Empirical study [40■], several variants of uncertain significance in familial hypercholesterolemia genes were reclassified after an in depth review of EHR data (unpublished data).

Pleiotropy

The wide spectrum of traits captured in EHRs provides an opportunity to assess pleiotropic effects of genetic variants [41–44]. In contrast to GWAS, PheWAS typically examines a limited set of target genotypes and their associations with the entire array of phenotypes in the EHR (phenome) (Fig. 4) [43,45]. To reduce the dimensionality of the phenotypic space, related diagnostic codes are often clustered into a smaller set of 'phecodes'. To investigate pleiotropic effects of LDL-C-modifying variants in the familial hypercholesterolemia genes (*PCSK9, APOB,* and *LDLR*) Safarova *et al.* [46■] conducted a PheWAS with 1232 phecodes in 51 700 European-Americans and 585 phecodes in 10 276 African-Americans. The investigators found expected associations with lipid and CHD phecodes but none with diabetes, neurocognitive disorders, or cataract, allaying concerns related to potential side effects of drugs that target these genes/gene products.

Pharmacogenomics

EHR-based studies can identify the genetic basis of interindividual variation in drug response as well as the basis of adverse drug reactions [47]. The PheWAS approach may help identify off-target effects of drugs targeting genes in the lipoprotein metabolism pathway. For example associations with diabetes, neurocognitive impairment or cataract were not found in a PheWAS study of *PCSK9* variants that lower LDL-C. In the eMERGE-PGx project [48], actionable pharmacogenetic variants identified by sequencing of 85, very important pharmacogenes, were placed in the EHR with linkage to clinical decision support. For example, an alert went to the provider to avoid simvastatin in patients carrying one or two copies of the *SLCO1B1* allele that is associated with statin-induced myopathy.

Genomic medicine clinical trials

The EHR can be used to overcome hurdles in recruiting patients to clinical trials and collection of data at baseline and follow-up. Eligibility can be assessed using EHR-based algorithms, those eligible contacted for e-consent and provided necessary information related to genomics research thereby reducing need for face-to-face genetic counseling [49]. The MIGENES trial is an example of an EHR-based genomic medicine clinical trial [50]. Individuals at intermediate risk for CHD were identified using an EHR-based algorithm and randomized to the disclosure of a 10-year risk of CHD using conventional risk equations or conventional risk equations and a polygenic risk score for CHD. Risk was disclosed by a genetic counselor using pictograms embedded in the EHR followed by the use of an EHR-based shared decision making tool to consider use of a statin with a physician.

CHALLENGES

EHR and administrative datasets are primarily meant for medical documentation and billing, not research. Thus, there are inherent limitations in their adaptation for genomics research [51] including missing data, imprecision in trait ascertainment, bias and lack of interoperability (Table 1). Ethical, legal, and social issues related to the use of EHRs for genomics research have been reviewed previously [52] and include the need to balance data security and privacy with data sharing [53■]. Patient re-identification is possible using genomic data and even through unique combinations of diagnostic codes [53■]. Another concern is the underrepresentation of minorities in genomic studies [54,55■■,56] and the potential to exacerbate healthcare disparities by uneven implementation of genomic medicine. EHR-based studies can potentially redress this imbalance. The possibility of bias in algorithms (algorithmic bias) particularly in machine learning algorithms has also recently been highlighted [57].

CONCLUSION

As a repository of phenotype data over the life span at a population scale, the EHR is a powerful resource for genomic research. Capture of social and behavioral determinants of health, improved interoperability between systems, and integration of data related to lifestyle and behavior as well as other environmental factors will further enhance the value of EHRs for research. Adoption of machine/deep learning and artificial intelligence to analyze ‘big data’ in EHRs will increase accuracy of risk prediction and prognostication models. Patients will be able to contribute survey, outcomes and wearable sensor data to EHRs through the ‘Sync for Science’ mechanism. EHR-based genomic discovery can inform clinical care, exemplifying the virtuous cycle of a learning healthcare system. Consequently, healthcare systems and EHR vendors are attempting to build a ‘genome enabled EHR’ for genomic medicine implementation [58]. The availability of big data from EHRs, ‘omics’, and wearables coupled with machine learning/deep algorithms will accelerate discovery and implementation, thereby revolutionizing the practice of medicine in the coming years.

Acknowledgements

The authors would like to acknowledge Ms. Luanne Wussow for assistance with article preparation.

Financial support and sponsorship

I.J.K. is supported by the National Human Genome Research Institute’s electronic Medical Records and Genomics (eMERGE) Network through grant HG006379, and grants R01 HL135879 and K24 HL137010 from the National Heart Lung and Blood Institute. M.S.S. was supported bythe American Heart Association Postdoctoral Fellowship Award 16POST27280004; and the American Heart Association grant 17IG33660937. The National Human Genome Research Institute and American Heart Association had no role in the content of the work; preparation, review, or approval of the article; and decision to submit the article for publication.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

■ of special interest

■■ of outstanding interest

1. Adler-Milstein J, Holmgren AJ, Kralovec P, et al. Electronic health record adoption in US hospitals: the emergence of a digital 'advanced use' divide. J Am Med Inform Assoc 2017; 24:1142–1148. [PubMed: 29016973]
2. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011; 4:13. [PubMed: 21269473]
3. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013; 15:761–771. [PubMed: 23743551]
4. Crawford DC, Crosslin DR, Tromp G, et al. eMERGEing progress in genomics – the first seven years. Front Genet 2014; 5:184. [PubMed: 24987407]
5. Stark Z, Dolman L, Manolio TA, et al. Integrating genomics into healthcare: a global responsibility. Am J Hum Genet 2019; 104:13–20. [PubMed: 30609404]
6. Horton I, Lin Y, Reed G, et al. Empowering Mayo Clinic Individualized Medicine with genomic data warehousing. J Pers Med 2017; 7:pii: E7. doi: 10.3390/jpm7030007.
7. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012; 13:395–405. [PubMed: 22549152]
8. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015; 350:h1885. [PubMed: 25911572]
9. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32:D267–D270. [PubMed: 14681409]
10. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015; 216:574–578. [PubMed: 26262116]
11■■. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: transforming i2b2 data into the OMOP common data model. PLoS One 2019; 14:e0212463. [PubMed: 30779778] Authors discuss imporatance and challenges in making data interoperable in the All of Us Program; demonstrate and validate use of the common data model (CDM) and data warehousing across the network. Implications: successful application of the i2b2-to-OMOP concept. Harmonization of the native EHR data structures to a CDM and a research data warehouse to enable effective genomics research.
12. Mungall CJ, McMurry JA, Kohler S, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 2017; 45:D712–D722. [PubMed: 27899636]
13. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003; 49:624–633. [PubMed: 12651816]
14■. Zhang XA, Yates A, Vasilevsky N, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med 2019; 2:pii: 32. doi: 10.1038/s41746-019-0110-4. The authors validated an approach to annotate commonly used Logical Observation Identifiers Names and Codes-encoded laboratory tests with human phenotype ontology terms. Implications: development of a SMART on FHIR application for the use within EHR systems to repurpose available laboratory tests for phenotyping.
15. Polubriaginof F, Tatonetti NP, Vawdrey DK. An assessment of family history information captured in an electronic health record. AMIA Annu Symp Proc 2015; 2015:2035–2042. [PubMed: 26958303]
16. Safarova MS, Liu H, Kullo IJ. Rapid identification of familial hypercholesterolemia from electronic health records: the SEARCH study. J Clin Lipidol 2016; 10:1230–1239. [PubMed: 27678441]
17. Mehrabi S, Wang Y, Ihrke D, Liu H. Exploring gaps of family history documentation in EHR for precision medicine – a case study of familial hypercholesterolemia ascertainment. AMIA Jt Summits Transl Sci Proc 2016; 2016:160–166. [PubMed: 27570664]

18■. Wu RR, Myers RA, Sperber N, et al. Implementation, adoption, and utility of family health history risk assessment in diverse care settings: evaluating implementation processes and impact with an implementation framework. Genet Med 2019; 21:331–338. [PubMed: 29875427] The authors investigated integrating a family health history-based risk assessment and clinical decision support platform (MeTree). Implications: when implemented in the primary care setting, the yield was highest in minorities and those with less education.

19. Khaleghi M, Isseh IN, Jouni H, et al. Family history as a risk factor for carotid artery stenosis. Stroke 2014; 45:2252–2256. [PubMed: 25005442]

20. Pendergrass SA, Crawford DC. Using electronic health records to generate phenotypes for research. Curr Protoc Hum Genet 2019; 100:e80. [PubMed: 30516347]

21. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc 2016; 23:731–740. [PubMed: 27107443]

22. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE Network. J Am Med Inform Assoc 2013; 20:e147–e154. [PubMed: 23531748]

23. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016; 23:1046–1052. [PubMed: 27026615]

24. Henderson J, Ke J, Ho JC, et al. Phenotype Instance Verification and Evaluation Tool (PIVET): a scaled phenotype evidence generation framework using web-based medical literature. J Med Internet Res 2018; 20:e164. [PubMed: 29728351]

25. Safarova MS, Kullo IJ. My approach to the patient with familial hypercholesterolemia. Mayo Clin Proc 2016; 91:770–786. [PubMed: 27261867]

26■■. Bastarache L, Hughey JJ, Hebbring S, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 2018; 359:1233–1239. [PubMed: 29590070] The study describes an approach that aggregates phenotypes into phenotype risk scores for Mendelian disorders. Implications: The method can improve rare-variant interpretation and allow identification of patients with Mendelian disorders who have yet to be diagnosed.

27. Bastarache L, Hughey JJ, Goldstein JA, et al. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J Am Med Inform Assoc 2019; 26:1437–1447. [PubMed: 31609419]

28. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med Inform 2019; 7:e14325. [PubMed: 31553307]

29. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med 2016; 46:2455–2465. [PubMed: 27406289]

30. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019; 380:1347–1358. [PubMed: 30943338]

31■■. Banda JM, Sarraju A, Abbasi F, et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. NPJ Digit Med 2019; 2:23. [PubMed: 31304370] The study describes the familial hypercholesterolemia Foundation's FIND FH initiative to developed a machine-learning algorithn to identify potential familial hypercholesterolemia patients using EHR data. Implications: a supervised classifier effectively finding candidate patients for further familial hypercholesterolemia screening.

32■■. Myers KD, Knowles JW, Staszak D, Shapiro MD. Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data. Lancet Digital Health 2019; 1:e393–e402. [PubMed: 33323221] Investigators describe development of a machine-learning algorithm trained to identify familial hypercholesterolemia. This article demonstrates the portability of the algorithm and a significant gap in case identification in clinical practice.

33. Abul-Husn NS, Manickam K, Jones LK, et al. Genetic identification of familial hypercholesterolemia within a single U.S. healthcare system. Science 2016; 354:pii: aaf7000. doi: 10.1126/science.aaf7000.

34. Zuvich RL, Armstrong LL, Bielinski SJ, et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE Network and quality control procedures to maintain high data quality. Genet Epidemiol 2011; 35:887–898. [PubMed: 22125226]

35■■. Hoffmann TJ, Theusch E, Haldar T, et al. A large electronic-health-record-based genome-wide study of serum lipids. Nat Genet 2018; 50:401–413. [PubMed: 29507422] The GWAS highlights the value of large-scale longitudinal EHR-derived data. Authors were able to explain variations in serum lipid levels based on genetic makeup among different races and sexes. Implications: discovery pertinent to lipid treatment and risk of coronary heart disease.

36. Rasmussen-Torvik LJ, Pacheco JA, Wilke RA, et al. High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. Clin Transl Sci 2012; 5:394–399. [PubMed: 23067351]

37■■. Klarin D, Damrauer SM, Cho K, et al. Genetics of blood lipids among ~300,000 multiethnic participants of the Million Veteran Program. Nat Genet 2018; 50:1514–1523. [PubMed: 30275531] Authors identified novel genome-wide significant loci in a meta-analysis with the Global Lipids Genetics Consortium and conducted PheWAS focusing on loss-of-function variants. Implications: Use of EHR-based data for genomic discovery and to reveal pleiotropic effects of lipid-related genes.

38■■. eMERGE Consortium. Harmonizing clinical sequencing and interpretation for the eMERGE III Network. Am J Hum Genet 2019; 105:588–605. [PubMed: 31447099] Investigators describe a platform established by the eMERGE III Network for integration of structured genomic results into multiple EHRs, informing genomic medicine. The report describes protocols and tools for harmonization of sequencing tests and standardization of interpretive aspects of genomic testing.

39. Van Driest SL, Wells QS, Stallings S, et al. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. JAMA 2016; 315:47–57. [PubMed: 26746457]

40■. Kullo IJ, Olson J, Fan X, et al. The Return of Actionable Variants Empirical (RAVE) study, a Mayo Clinic Genomic Medicine implementation study: design and initial results. Mayo Clin Proc 2018; 93:1600–1610. [PubMed: 30392543] The article describes design and initial results of a study to investigate penetrance and outcomes following the return of pathogenic variants in disease-related genes to inform best practices for genomic medicine. Investigators found relatively low penetrance for variants in arrhythmia and cardiomyopathy genes.

41. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013; 31: 1102–1110. [PubMed: 24270849]

42. Conway M, Berg RL, Carrell D, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc 2011; 2011:274–283. [PubMed: 22195079]

43. Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. Annu Rev Genomics Hum Genet 2016; 17:353–373. [PubMed: 27147087]

44. Verma A, Verma SS, Pendergrass SA, et al. eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. BMC Med Genomics 2016; 9(Suppl 1):32. [PubMed: 27535653]

45. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome–phenome relationship using phenome-wide association studies. Nature reviews. Genetics 2016; 17:129–145.

46■. Safarova MS, Satterfield BA, Fan X, et al. A phenome-wide association study to discover pleiotropic effects of PCSK9, APOB, and LDLR. NPJ Genom Med 2019; 4:3. [PubMed: 30774981] The EHR-based study using genotype data in European and African populations assessed phenotype–genotype associations for variants in the familial hypercholesterolemia-related genes. Implications: EHR-based PheWAS can be used to assess pleiotropic effects of variants in genes that regulate lipoprotein metabolism.

47. Wei WQ, Feng Q, Jiang L, et al. Characterization of statin dose response in electronic medical records. Clin Pharmacol Ther 2014; 95:331–338. [PubMed: 24096969]

48. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. Clin Pharmacol Ther 2014; 96:482–489. [PubMed: 24960519]

49. Sutton EJ, Kullo IJ, Sharp RR. Making pretest genomic counseling optional: lessons from the RAVE study. Genet Med 2018; 20:1157–1158. [PubMed: 29388941]

50. Kullo IJ, Jouni H, Austin EE, et al. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the MI-GENES Clinical Trial). Circulation 2016; 133:1181–1188. [PubMed: 26915630]

51. Hall JL, Ryan JJ, Bray BE, et al. Merging electronic health record data and genomics for cardiovascular research: a science advisory from the American Heart Association. Circ Cardiovasc Genet 2016; 9:193–202. [PubMed: 26976545]

52. Hazin R, Brothers KB, Malin BA, et al. Ethical, legal, and social implications of incorporating genomic information into electronic health records. Genet Med 2013; 15:810–816. [PubMed: 24030434]

53■. Malin B, Goodman K. Section editors for the IMIA yearbook special section. Between access and privacy: challenges in sharing health data. Yearb Med Inform 2018; 27:55–59. [PubMed: 30157505] The authors discuss the risks of sharing new data and strategies for publishing summary information about genome–phenome studies.

54. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med 2016; 375:655–665. [PubMed: 27532831]

55■■. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature 2019; 570: 514–518. [PubMed: 31217584] The Population Architecture using Genomics and Epidemiology study investigators conducted GWAS of clinical and behavioral phenotypes in a large cohort of non-European individuals. Implications: there was evidence of effect-size heterogeneity across ancestries highlighting the need for ancestry-specific GWAS.

56. Martin AR, Kanai M, Kamatani Y, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 2019; 51:584–591. [PubMed: 30926966]

57. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366:447–453. [PubMed: 31649194]

58. Kullo IJ, Jarvik GP, Manolio TA, et al. Leveraging the electronic health record to implement genomic medicine. Genet Med 2013; 15:270–271. [PubMed: 23018749]

KEY POINTS

- Although primarily designed for medical documentation and billing purposes, the EHR is a valuable tool for research.
- Linking the EHR to genomic data enables repurposing of vast phenotype data for genomic discovery.
- EHR-based genomic discovery can inform clinical care, exemplifying the virtuous cycle of a learning healthcare system.

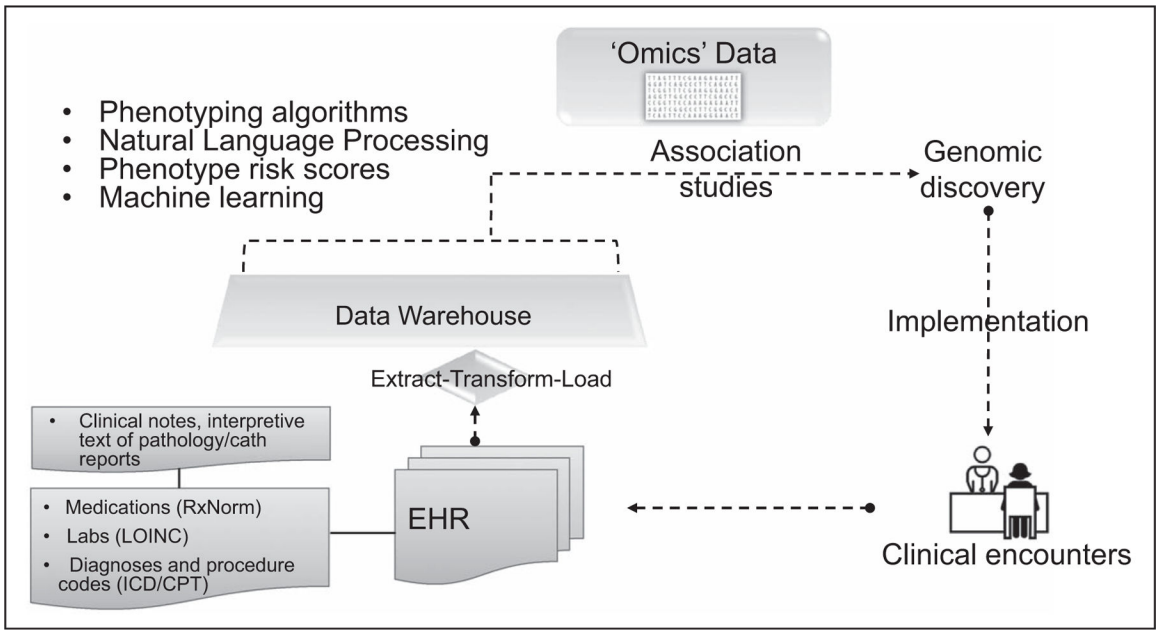

FIGURE 1.

Using electronic health record for genomic discovery. The concept of combining DNA biorepositories with electronic health record systems was pioneered by the electronic Medical Record and Genomic Network investigators. Genomic discovery can then feedback into the electronic health record for genomic medicine implementation exemplifying the 'learning healthcare system'.

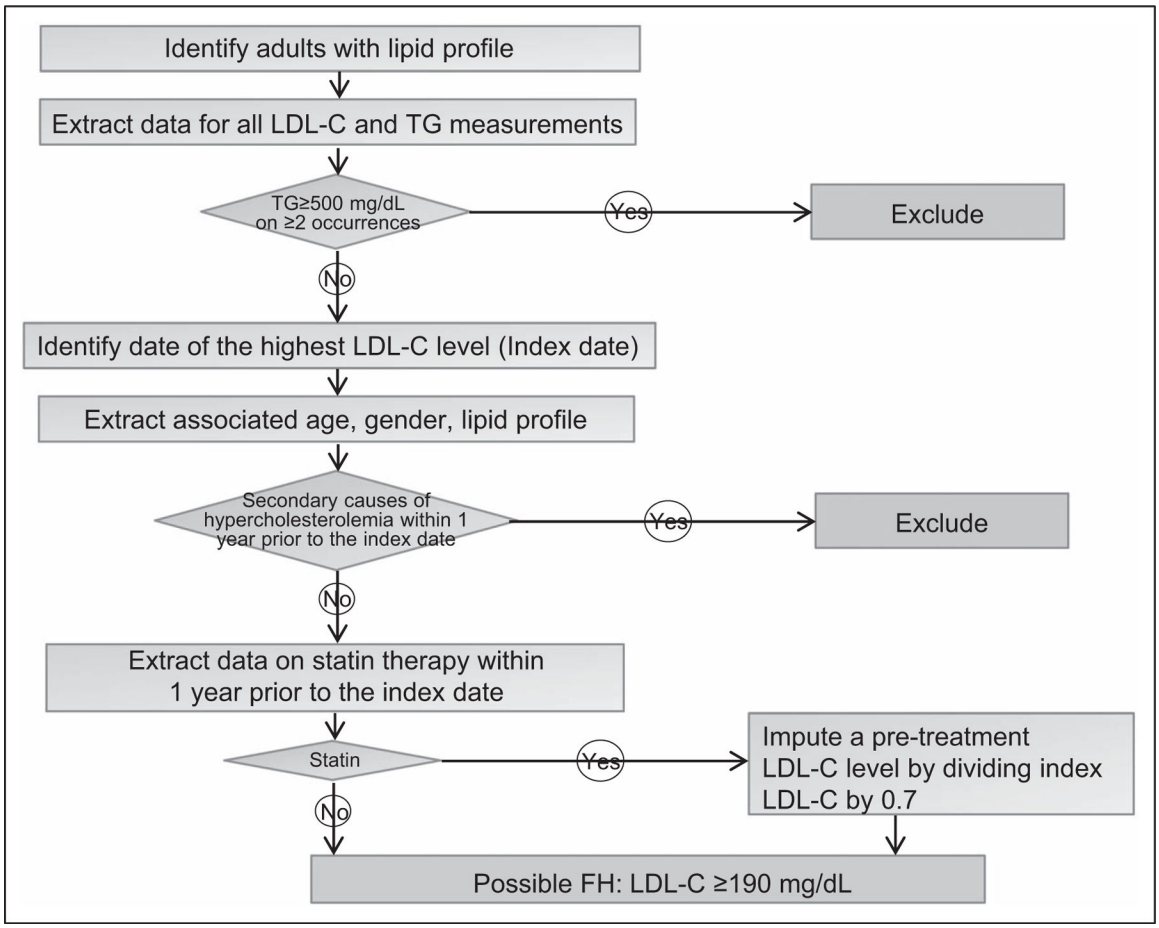

FIGURE 2.

An algorithm to identify cases of primary hypercholesterolemia from the electronic health record. The highest LDL-cholesterol level recorded in the electronic health record is identified. Code-based algorithms are then used to exclude individuals with secondary causes of hypercholesterolemia.

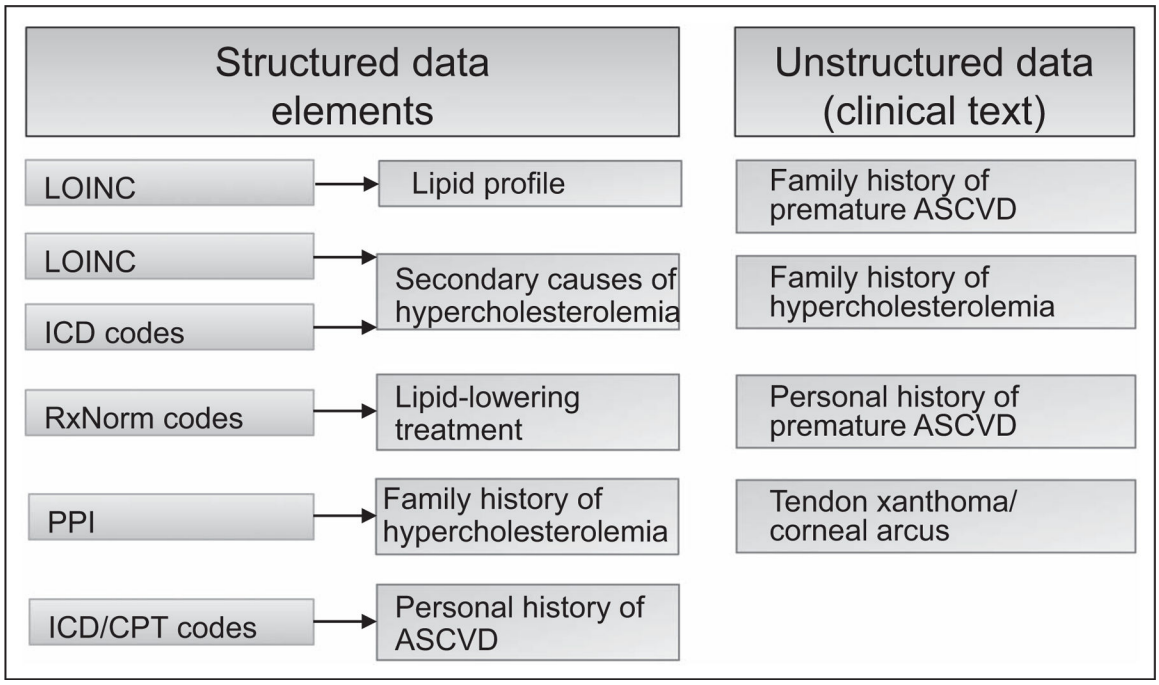

FIGURE 3.

Mining of structured and unstructured data in the electronic health record to ascertain data elements for an electronic phenotyping algorithm for familial hypercholesterolemia [25]. Family history of atherosclerotic cardiovascular disease, premature ASCVD, and hypercholesterolemia were detected using NLP. In this example, clinical data from narrative notes were extracted using NLP implemented using MedTagger (http://ohnlp.org/index.php/MedTagger_Project_Page). ASCVD, atherosclerotic cardiovascular disease; CPT, current procedural terminology codes; EHR, electronic health record; ICD, International Classification of Diseases; LOINC, logical observation identifiers names and codes; NLP, natural language processing; PPI, patient provided information.

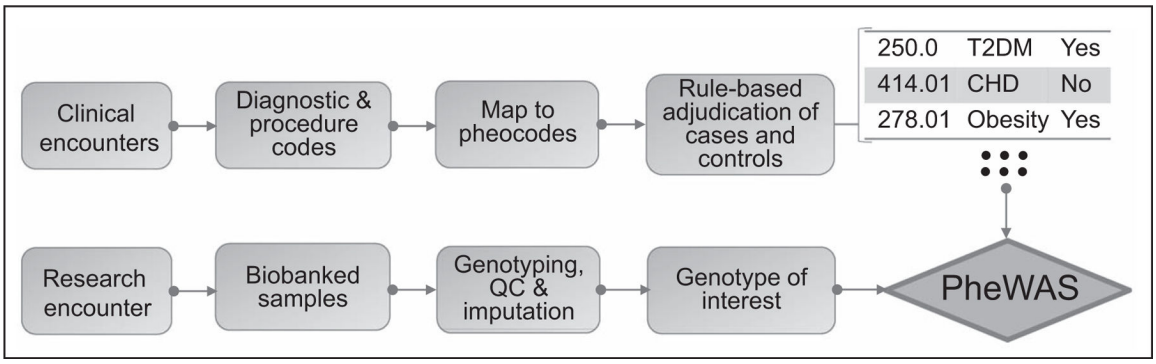

FIGURE 4.

An electronic health record-based phenome-wide association study (PheWAS). Clinical encounters are represented by diagnosis and procedure codes which are transformed to phecodes prior to application of rules to ascertain cases and controls for each phenotype. Linkage of individual phecodes to genotypes allows phenome-wide association study for genetic variants of interest.

Table 1.

Limitations of the electronic health record for genomics research

Missing data	A patient may receive care at more than one healthcare organizations and a single EHR may not adequately capture relevant data
Little information on environmental and lifestyle factors	Information on lifestyle factors such as diet, physical activity, and dietary habits is not adequately captured in EHRs
Family history is poorly documented	Family history is often not documented or documented in insufficient detail
Imprecision in trait ascertainment	Even the most carefully constructed and validated e-phenotyping algorithms will have some degree of imprecision
Bias	Physicians vary in how they document their clinical observations, inferences, diagnosis, and treatment plans
Lack of interoperability	EHR systems differ in technical specifications and functional capabilities and there is variability in data fields and elements (numeric data, structured text, unstructured text, and scanned files and images)

EHR, electronic health record.