Published in final edited form as:

Clin Pharmacol Ther. 2016 March ; 99(3): 298–305. doi:10.1002/cpt.321.

Integrating electronic health record genotype and phenotype datasets to transform patient care

Dan M. Roden^{1,2,3} and Joshua C. Denny^{1,2}

¹Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN

²Department of Biomedical informatics, Vanderbilt University School of Medicine, Nashville, TN

³Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, TN

Abstract

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 mandates the development and implementation of electronic health record (EHR) systems across the country. While a primary goal is to improve the care of individual patients, EHRs are also key enabling resources for a vision of individualized (or personalized or precision) medicine: the aggregation of multiple EHRs within or across healthcare systems should allow discovery of patient subsets that have unusual and definable clinical trajectories that deviate importantly from the expected response in a “typical” patient. The spectrum of such personalized care can then extend from prevention to choice of medication to intensity or nature of follow up.

Large datasets allow identification of defined subsets

The development of very large datasets to identify in a robust and reproducible fashion specific patient subsets is a key starting point for development of evidence to implement an approach that treats some patients differently from average.¹ EHRs are one such very large dataset: the EHR includes data generated during routine clinical care and can be used in a stand-alone fashion or be coupled to other data types for discovery.¹ Examples of other data types include information on the sociocultural determinants of health,² systematically acquired patient-reported or mobile device-acquired data, and biobank-derived information including genotype or sequence data as well as other “omic” (transcriptomic, proteomic, metabolomic, etc) data. This review will focus on ways in which coupling EHRs to genomic datasets can be enabling for discovery of genotype-phenotype associations and how these associations can then be implemented in EHRs to start to individualize patient care.

Recent efforts have generated several very large datasets that integrate EHR data of various types to dense genomic information, including the Electronic Medical Records and Genomics (eMERGE) Network,³ the Veterans Administration’s Million Veterans Program,⁴ the Kaiser-Permanente GERA program,⁵ the UK Biobank,⁶ and the Icelandic deCODE resource.⁷ Taken together, these have generated dense genotype information (genome wide

Correspondence: Dan M. Roden, M.D., Professor of Medicine, Pharmacology, and Biomedical Informatics, Director, Oates Institute for Experimental Therapeutics, Assistant Vice-Chancellor for Personalized Medicine, Vanderbilt University School of Medicine, 1285 Medical Research Building IV, Nashville, TN 37232-0575.

association study (GWAS)-level or more) in over a million patients. Importantly, while initial studies in these datasets have demonstrated their value in discovering common genetic loci associated with common human disease through GWAS, more recent work has shown they can be exploited for many other applications, such as identifying rare genetic variants with large effect sizes, pleiotropic effects of common and rare genetic variants, and potential drug targets. While these systems have been expensive to establish, they hold the promise of actually improving efficiencies in both discovery and implementation, since data generated in the course of clinical care is reused for research purposes.⁸ Further, the genetic datasets, once generated, can be reused for multiple analyses. This idea was initially developed by the Wellcome Trust Case Control Consortium,⁹ and has been validated on multiple occasions by individual biobanks⁸ and across the eMERGE Network.¹⁰ One study⁸ in BioVU, the Vanderbilt DNA biobank that now includes DNA samples from >215,000 subjects and is also a participant in eMERGE, compared the costs and time required for traditional NIH-supported pharmacogenetic studies to those in a large pharmacogenetic project in BioVU. The BioVU cohorts were larger (median 1,123 vs 623), less expensive to generate (\$76,674 vs \$1,335,927), and required less time to generate (3 months versus 3 years), recognizing that the costs of developing the EHR itself – a by-product of routine healthcare – are not factored into these calculations.

Phenotyping in the EHR

While the concept of using EHR systems as a tool for discovery in genome science is appealing, a major initial obstacle that this approach had to overcome was whether the phenotypes represented in EHR systems were in fact at all useful for defining important human phenotypes. One of the major challenges has been to understand optimal ways to analyze multiple types of data contained in an EHR to develop algorithms to identify subjects with target diseases (cases) and those who do not have the diseases (controls). Some phenotypes may be relatively “easy” to ascertain. For example, if an investigator is interested in identifying cases of atrial fibrillation, and establishes that a 12-lead electrocardiogram recording the abnormal rhythm is required to establish the subject as a case, all that is required is searching electrocardiograms for instances of atrial fibrillation. Even here, however, algorithms may be imperfect: the electrocardiogram may be misread or the rhythm may be documented only in text notes or in poorly reproduced rhythm strips. While such records might not meet a case definition, they would be inappropriate to include as controls.

Atrial fibrillation is a simple example. For other conditions, combinations of multiple types of data within an EHR, including structured data (such as billing codes) as well as natural language processing approaches to analyze unstructured data (such as clinical notes), may be required to identify “true cases”.^{11,12} The longitudinal follow-up inherent in the EHR may assist in phenotype algorithm development: whether a patient with joint pain has rheumatoid arthritis or another related condition may become clearer over time. Longitudinal follow-up is an especially desirable feature in individualized medicine, where the questions are often which subsets of patients with disease X will go on to develop complication Y, or which subsets of patients exposed to drug A will go on to develop outcome B.

Figure 1 presents an overview of the approach used in BioVU and across eMERGE for phenotyping. One early lesson, obvious in retrospect, was that engaging clinical users of the EHR is highly desirable in constructing initial algorithms: these individuals are most familiar with how specific diseases may be represented in the EHR, such as the medications, laboratory test results, free text, etc. that may be useful to identify true cases and controls. Once an initial algorithm is constructed, this is deployed across the EHR system until potential cases are identified. These are then manually reviewed and if the positive predictive value (PPV) for the algorithm is less than 95% it is further refined and the process iterated until the PPV exceeds 95%. For rare phenotypes, an algorithm that captures every potential case can be deployed with manual review to establish which, in the initial dataset, are true cases.^{13,14} On the other hand for common diseases, where the datasets will include thousands or tens of thousands of subjects, electronic algorithms are clearly required to extract true cases and controls. Resources such as the eMERGE's Phenotype Knowledgebase (PheKB.org) and i2b2 (informatics for integrating biology and the bedside) present phenotype definitions to investigators using EHRs for discovery research.^{15–18} The development of electronic phenotyping algorithms at one institution has been followed by the evidence that such algorithms perform well across multiple institutions, often with different EHR architectures.¹⁰ Further, phenotype algorithms for complex diseases such as rheumatoid arthritis, type II diabetes, and hypothyroidism that incorporate multiple dimensions of information – billing codes, natural language processing (NLP) of free text, medications, etc – perform better than simple structured data such as billing codes.¹¹

Using EHR-developed phenotypes for discovery

Initial studies to validate EHR-based phenotype approaches sought to replicate known associations between common single nucleotide polymorphisms (SNPs) and common diseases, established in multiple GWAS. An initial study in BioVU genotyped the first 9,483 samples accrued for common variants previously reported to be associated with rheumatoid arthritis, type II diabetes, multiple sclerosis, Crohn's disease, and atrial fibrillation.¹³ Phenotyping algorithms identified 70-698 cases and 808-3,818 controls for the five phenotypes. A total of 21 tests of association were performed on SNPs with previously established odds ratios and each point estimate in BioVU was in the expected direction, and statistically significant associations were detected in all analyses that were adequately powered.

Subsequently, EHR-based biobanks have been used to discover new genetic associations. One notable example was the identification of common variants near *FOXE1*, previously associated with thyroid cancer, in 1,317 cases and 5,053 controls across the eMERGE Network.¹⁰ A notable feature of this study was that the genetic data had been previously generated for other phenotypes at each eMERGE site and the case and control algorithms were successfully deployed across multiple sites. Thus this represents one of the first examples of successful reused of extant EHR/genomic data for discovery. Subsequent efforts in eMERGE have replicated other known associations and have identified new genetic loci associated with phenotypes such as erythrocyte sedimentation rate,¹⁹ varicella zoster virus infection,²⁰ and red and white cell counts.^{21,22}

Drug response phenotypes

Identifying robust genomic associations for drug response phenotypes presents specific challenges within EHR systems. Patients receiving a specific drug must first be identified and then the outcome of interest, within a predefined time window following drug administration, must then be identified. As a consequence, these datasets tend to be smaller and the algorithms more complicated than those used for disease genotype associations. Further, the requirement for ascertainment at multiple time points (baseline, on drug) can introduce bias since not all patients have complete follow up. Differences in a wide range of factors across patients exposed to a drug may account for variable responses and many of these may be inapparent or difficult to measure in the EHR. Whether patients actually take their medication is a source of variability in any drug response study, and EHR-based work is no exception. Seeing a change in a drug response metric (e.g. lowering LDL with a statin) might indicate compliance, but no change could be non-compliance, altered kinetics, or altered response. Measuring drug levels can help, and these data, if obtained, are available in the EHR.

Our initial studies in this area demonstrated that reported genomic predictors of common drug outcomes could be readily replicated in BioVU. These included *CYP2C19**2 and major adverse cardiovascular events (MACE) in patients receiving clopidogrel after coronary stenting,²³ *CYP2C9* and *VKORC1* variants and steady state warfarin dosages,²⁴ and *CYP3A5**3 and tacrolimus concentrations after renal transplant.²⁵ The Pharmacogenomics Research Network (PGRN) supported the development of the PGPpop (Pharmacogenomic discovery and replication in very large patient POPulations) resource for mining EHR systems for pharmacogenomic response phenotypes (Table 1). PGPpop used BioVU, the Marshfield Personalized Medicine Research Project (PMRP), and other datasets for discovery of new genotype-phenotype associations by PGRN investigators. Other drug-response genotypes addressed using EHR resources have included thromboembolism during tamoxifen,²⁶ cough during ACE inhibitor therapy,²⁷ heparin-induced thrombocytopenia,²⁸ creatinine concentrations during vancomycin therapy,²⁹ osteonecrosis during steroid therapy,³⁰ and anthracycline-related changes in ejection fraction.³¹

Adverse drug reactions present a specific problem in ascertainment, may require manual case and control curation, and highlight the need for very large datasets to identify what are often rare events. Metrics of drug elimination, such as estimated glomerular filtration rate, are often available and can be used to assess contributions of renal function to variable outcomes. Similarly, data can be obtained on effects of co-administration of a study drug and known inhibitors of specific pathways of drug elimination. Temporal features are especially important for most drug response algorithms. For example, temporal elements, including those extracted with NLP, were important to identify methotrexate-induced liver injury in an algorithm that performed with a PPV of 59%.¹⁴ For such complex, rare phenotypes, an approach that captures all possible cases and then subjects them to manual curation and review by expert clinicians may be required. Kawai et al ascertained 250 cases of bleeding during long term warfarin therapy and 250 controls receiving long term warfarin therapy without bleeding in BioVU.³² A candidate gene analysis (*CYP2C9*, *CYP4F2*, and *VKORC1*) identified *CYP2C9**3 as a risk factor for bleeding with an odds ratio for major

bleeding of 2.05 (95% confidence interval (CI) 1.04–4.04). The association between bleeding and *CYP4F2* variants seen in an administrative database study³³ was not replicated, and larger numbers may be required. Importantly, major bleeding was not a common outcome in the randomized controlled trials (RCTs), discussed further below, comparing genotype-guided therapy to conventional therapy for warfarin,^{34,35} although a trend toward less bleeding was seen in the largest of these trials (10/501 cases vs 4/514 cases). Another key point raised in particular by studies of warfarin pharmacogenomics, including the RCTs, is the need for ancestral diversity. Genetically-informed algorithms performed worse than conventional care in African-American (AA) subjects,³⁴ likely reflecting the fact that different variants contribute to variable dose requirements in AAs.^{24,36–38}

Curating the EHR phenome enables PheWAS

Prospectively-acquired datasets such as the UK Biobank or community cohorts such as Framingham or Jackson Heart and others acquire data in a predefined and comprehensive fashion. By contrast, the EHR represents the range of conditions and follow-up for which patients seek medical attention. Unlike prospective cohorts, the EHR does not provide unambiguous cases and controls. Rather, as outlined above, algorithms using a variety of different data sources need to be developed and validated.

It is increasingly clear that the development of tools to interrogate and curate the EHR “phenome” opens the way to novel forms of new knowledge development. One approach to generating a curated structured dataset from the EHR is illustrated in Figure 2. For example, a conventional genetic association study starts with a phenotype of interest and searches for associated genetic loci or variants. The phenome-wide association study (PheWAS) reverses this paradigm, using the phenome as the interrogation target. The initial implementations sought associations between genotypes at single nucleotide polymorphism (SNP) sites and the phenome,^{10,39,40} generated by algorithmically-defined diagnoses for cases (representing about 1500 different diseases and traits) and corresponding controls. For example, in the eMERGE hypothyroidism project, the odds ratios for association between hypothyroidism and *FOXE1* variants were near-identical with GWAS and PheWAS; PheWAS also identified specific associated thyroid disorders as well as an association with atrial arrhythmias.¹⁰ The top association in a GWAS of normal cardiac conduction (assessed by variability in QRS durations on normal ECGs) was in a sodium channel gene *SCN10A*, and PheWAS demonstrated that the top SNP was associated with atrial fibrillation, developing over years after the initial normal ECG.³⁹ After these initial studies demonstrated feasibility of the approach, 3,144 PheWAS using GWAS data across eMERGE sites analyzed SNPs previously identified as potential mediators of disease susceptibility or other physiologic traits in 1,358 EHR derived phenotypes across 13,835 individuals.⁴¹ This study replicated 51 out of 77 sufficiently powered prior GWAS associations, showing PheWAS can be used as a replication tool to complement other genetic association studies. Others have used EHR-based PheWAS approaches to replicate new associations,⁴² to identify functional variants,⁴² highlight potential drug targets,⁴³ and to identify signals in multiethnic populations.⁴⁴

Other applications of phenome scanning

PheWAS highlights the value of a curated EHR-based phenome and opens the possibility of using other approaches to interrogate this dataset. For example, the large-scale eMERGE study identified 63 phenotypes at a P value $< 4.6 \times 10^{-6}$ with previously-unreported associations with specific SNPs, complementing the *SCN10A*-atrial fibrillation and the *FOXE1*-atrial arrhythmia examples and highlighting the potential for PheWAS as a tool for discovery of genetic pleiotropy.

While the input function for initial PheWAS has been individual SNPs, other possible inputs include all SNPs across a gene, or all variants in a pathway. Further, the input function need not be genetic: an investigator might ask what diseases are overrepresented after exposure to a drug, or search for pair-wise overlaps of genetically-defined heritability across all other conditions in the phenome.⁴⁵ A topology-based network approach⁴⁶ in 11,210 subjects with diabetes identified clinical characteristics and associated genotypes for three subsets: one with diabetic nephropathy and retinopathy; a second with cancer and cardiovascular diseases; and the third associated with cardiovascular disease, neurological disease, allergy, and HIV infection. These data support the long held clinical impression that “type II diabetes” is a label for a disease that has a range identifiable subsets and point to how high dimensional EHR – coupled eventually to genotype data to further refine subset definitions – may assist in identifying such subsets for individualized therapy.

Phenome interrogation may contribute to drug development and drug repurposing

A study of >110,000 adult patients with cancer at Vanderbilt and the Mayo Clinic over 15 years demonstrated that in the subset with type II diabetes, metformin was associated with improved cancer survival and a 22% lower overall mortality compared to other oral hypoglycemics or insulin.⁴⁷ These data support previous studies suggesting that metformin could produce a beneficial effect in a range of cancers. Phenome scanning in follow-up to a very large GWAS of rheumatoid arthritis (RA) identified pleiotropic disease or biomarker associations for approximately two-thirds of 98 RA-associated genes.⁴⁸ These data then suggest that drugs currently marketed for such “pleiotropic” disease associations might be also effective in rheumatoid arthritis or in subsets of patients with rheumatoid arthritis. Some of these drug targets (anti-TNF agents, rituximab) are already used in the disease while others have been developed for other conditions, notably malignancies and have not yet entered human trials for rheumatoid arthritis.⁴³

Finding rare nonsense variants associated with a “protective” phenotype suggests drugs inhibiting the function of the gene product may be useful therapeutics. One example that has brought this concept to fruition was the association reported in 2006 between nonsense variants in *PCSK9* and lower low-density lipoprotein (LDL) as well as protection against coronary artery disease.⁴⁹ These data suggested inhibition of PCSK9 might be a useful therapeutic strategy in coronary artery disease, and two new drugs with this mechanism of action were approved in 2015.⁵⁰ Similarly, in an experiment that studied >113,000 subjects drawn from both EHRs and community and prospective cohorts, very rare *NPC1L1* loss-of-

function variants were found to be less prevalent in cases of coronary artery disease than among controls, supporting suppression of the gene product (an effect of the lipid-lowering agent ezetemibe) as a valid therapeutic strategy.⁵¹ A third example is the identification of rare nonsense variants in *SLC30A8* as protective against diabetes in the deCODE database as well as community and prospective cohorts.

Implementation

A common vision is that large genotype datasets (up to whole genomes) will be routinely acquired in large numbers of patients and embedded in EHRs to be used by knowledgeable clinicians, supported by sophisticated clinical decision support systems.⁵² As described below, this idea is now being tested across multiple EHR systems for pharmacogenomic implementation. However, there are multiple major obstacles that need to be addressed as these systems are developed.^{53,54} Some are operational and may interface with policy issues such as storing and retrieving these large datasets, maximizing privacy, and ensuring that genetic data are not used in a discriminatory fashion in individuals (e.g. in non-health insurance) and across populations definable by genetic variation. Another is the general problem of developing and evaluating evidence that specific variants are “actionable”, a term taken to mean that their presence would in some way alter routine care. In pharmacogenetic implementation, this could be to avoid a drug or to change a drug dose, while genomic information could be used to intensify screening, undertake specific therapies, or to acquire long-term care insurance. Once a decision is made to implement genomic information in an EHR context, it must be coupled to clinical decision support to assist the provider by supplying background information and advice on genotype-specific actions to be taken: reduce a drug dose, change drugs, start more frequent cancer screening, etc. Systems that implement such approaches also need to develop methods to track and assess outcomes.

Pharmacogenomic implementation

To date, routine implementation of genomic data in healthcare has been largely confined to variants affecting drug response, using either a point-of-care or a preemptive approach.^{53,54} The point-of-care approach relies on genotyping at the time of drug prescription and very rapid turnaround of genomic variant data: programs using this approach have been implemented for clopidogrel⁵⁵ and for warfarin.^{34,35} The preemptive strategy relies on identifying subjects at increased likelihood of receiving either a single drug⁵⁶ or a range of drugs^{57,58} that have been associated with variable responses due to known pharmacogenetic factors.

The program at St. Jude Children’s Research Hospital focuses on a group of patients seeking specialized care for relapsed acute childhood leukemia, a situation in which drug treatment regimens are complex and defined.⁵⁸ The Vanderbilt PREDICT program uses algorithms derived from EHR data to identify patients “at risk” of receiving clopidogrel, warfarin, or simvastatin within the next three years.⁵⁹ Each program generates genotypes at multiple known variant sites across multiple pharmacogenes, embeds variant data in the EHR, and develops clinical decision support that fires when a drug is prescribed to a patient with a

variant. The Vanderbilt program reported that 91% of the first 10,000 subjects studied carried a potentially actionable variant for one of 5 drug-gene pairs, and that high-risk genotypes were present in 4.8% of subjects;⁶⁰ these data argue for a preemptive multiplexed approach since even with only 5 drug-genes pairs, most subjects have some actionable variant for some drug.

Identifying actionable pharmacogenetic variants

One of the issues facing those wishing to implement these approaches is identifying variants for which a change from routine dosing would be recommended for specific drugs. Early adopter programs generated lists of actionable variants by internal review of extant data, but the task has now been made easier by the Clinical Pharmacogenomics Implementation Consortium (CPIC) that publishes specific drug-genotype guidelines on which genotypes, or in some cases diplotypes, alter protein function and thus merit changes from routine drug dosing.⁶¹ CPIC specifically avoids the question of whether patients should be genotyped but rather examines the question that if a genotype result is available, what actions should be considered. Implementation recommendations are made based on multiple types of evidence, such as whether a drug is metabolized exclusively by a single pathway, what the clinical consequences of unusually high or low concentrations of parent drug or metabolite might be, and whether alternate treatment approaches are available.

Addressing Actionability

The first study to show a role for CYP2C19 in bioactivating clopidogrel reported that subjects heterozygous for *CYP2C19* loss of function *2 allele generate on average less anti-platelet effects than those with wild-type *1 alleles, but there was overlap between those with the *1/*1 and those with the *1/*2 genotypes.⁶² Subsequently, a meta-analysis of outcomes after clopidogrel for coronary stenting showed higher rates of in-stent thrombosis and MACE in subjects with *CYP2C19* variants.⁶³ The FDA then changed the label for clopidogrel to include a black box warning that if genotype data were available and indicate a CYP2C19 loss of function variant, alternate anti-platelet therapies should be prescribed. As a result of the clinical studies and the black box warning, some centers have adopted routine *CYP2C19* screening in subjects scheduled for coronary angiography. Most recently, the University of Florida group reported that switching from clopidogrel to alternate antiplatelet therapies in subjects with *CYP2C19* variants reduced MACE risk after stenting.⁶⁴

Advocates for pharmacogenomic implementation argue that what is known about clopidogrel bioactivation and the meta-analysis are sufficient to justify this approach now. Others argue that an RCT is required. One difficulty with conducting an RCT is that the effect size is likely to be greatest in homozygotes (rare group; e.g. <3% of subjects in PREDICT⁶⁰) while that in heterozygotes is smaller. Further, randomizing homozygotes may be equivalent to administering placebo, and the resultant lack of clinical equipoise for some investigators may make randomization of homozygotes difficult. Nevertheless an RCT is underway but over the course of the study, newer stent technologies and newer antiplatelet drugs are becoming available.

An RCT of genotype-guided versus conventional therapy showed that HLA-B*5701 genotyping could eliminate serious skin reactions during initiation of abacavir.⁶⁵ On the other hand, RCTs of genotype-guided versus conventional warfarin therapy have shown either no difference³⁴ or a small difference³⁵ in time in therapeutic range (TTR) during the first 30–90 days of treatment. Whether the TTR is the appropriate metric for such a trial, the role of ancestry might play in determining these results, and the question of genetics and long-term bleeding risk discussed above are still open issues.

Penetrance in Mendelian disease

A looming problem in implementing genomic medicine in the EHR context is that for the vast majority of non-synonymous genetic sequence variants, currently available tools to predict *in vitro* or *in vivo* function are imperfect at best. The development of large databases such as ClinGen that aspire to present not only variant frequency data but also associated phenotypes developed from a range of sources including crowd-sourcing, will help tackle this problem.⁶⁶ The availability of Mendelian disease gene genotype or sequence data in large numbers of unselected subjects with EHRs may allow estimates of penetrance across populations. The eMERGE Network took this approach in a study of the hemochromatosis-related genotypes (homozygote C282Y/HFE homozygotes or C282Y/H63E compound heterozygotes) in 538 carriers in a set of 39,000 genotyped subjects.⁶⁷ The homozygotes were more likely to carry the clinical diagnosis of hereditary hemochromatosis than the compound heterozygotes and as in previous reports men were more likely to carry the diagnosis than women: 24.4% of homozygous men and 14.0% of homozygous women carried the diagnosis. However, the prevalence of liver disease and diabetes were both higher than the prevalence of the diagnosis of hemochromatosis and at one site, 17% of homozygotes were receiving iron (generally contraindicated in this iron overload disease) and none had a diagnosis of hereditary hemochromatosis. These data suggest that in individuals with existing genotype data, opportunistic screening is relatively common at risk genotypes should be undertaken to identify individuals at risk for the disease.

In another study, sequencing in 2022 subjects found 121 who carried rare non-synonymous variants in the congenital arrhythmia disease genes *SCN5A* and *KCNH2*.⁶⁸ Three expert annotators disagreed on which ones should be called pathogenic, and review of EHRs of 48 individuals with potentially pathogenic variants identified very few arrhythmia phenotypes. These studies reinforce the ideas that penetrance is incomplete in Mendelian diseases and that current methods to determine pathogenicity need improvement. Notably, embedding sequence data in EHRs may not only enable implementation of “actionable” sequence variants, but also serves as a tool for discovering which variants are actionable: an example of the learning healthcare system.

The Precision Medicine Initiative

Mining of EHR data for meaningful phenotypes will likely continue to be an important feature of genomic research. The coming formation of the US national Precision Medicine Initiative Cohort Program (PMI-CP), whose strategy was outlined in September 2015,⁶⁹ calls for a population of more than 1 million individuals that will share their healthcare data.

This publicly available resource will be available to diverse researchers. It is anticipated the majority of the individuals for the PMI-CP will be derived from healthcare provider organizations, which will share longitudinal EHR data periodically to form the bulk of phenotype data to enable both discovery and implementation approaches described here. Near-term priority scientific opportunities recognized in the PMI-CP report⁶⁹ included analysis of drug-response traits and phenome-wide methods to accelerate drug discovery, both of which are uniquely accelerated through EHR research. The PMI-CP, in combination with other worldwide resources, anticipate an exciting future with huge populations able to be densely studied for many diseases.

Acknowledgments

Grant Support: This work was supported by the awards supporting participation in the NIH's Pharmacogenomics Research Network (U19 HL65962; P50 GM115305) and the electronic Medical Records and Genomics (eMERGE) Network (U01 HG04603; U01 HG008672)

References

1. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015; 372:793–5. [PubMed: 25635347]
2. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. *N Engl J Med*. 2015; 372:698–701. [PubMed: 25693009]
3. Gottesman O, et al. The Electronic Medical Records and Genomics (eMERGE) Network: Past, Present and Future. *Genetics in Medicine*. 2013 in press.
4. Gaziano JM, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*. 2015
5. Banda Y, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*. 2015; 200:1285–95. [PubMed: 26092716]
6. Allen NE, Sudlow C, Peakman T, Collins R. oboU Biobank. UK Biobank Data: Come and Get It. *Science Translational Medicine*. 2014; 6:224ed4.
7. Gulcher JR, Stefansson K. The Icelandic Healthcare Database and informed consent. *N Engl J Med*. 2000; 342:1827–30. [PubMed: 10853008]
8. Bowton E, et al. Biobanks and electronic medical records: enabling cost-effective research. *Sci Transl Med*. 2014; 6:234cm3.
9. The Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–78. [PubMed: 17554300]
10. Denny JC, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet*. 2011; 89:529–42. [PubMed: 21981779]
11. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association : JAMIA*. 2015
12. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome medicine*. 2015; 7:41. [PubMed: 25937834]
13. Ritchie MD, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet*. 2010; 86:560–72. [PubMed: 20362271]
14. Lin C, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association : JAMIA*. 2015; 22:e151–61. [PubMed: 25344930]

15. Carroll RJ, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association : JAMIA*. 2012; 19:e162–9. [PubMed: 22374935]
16. Kho AN, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association*. 2011
17. Mo H, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *Journal of the American Medical Informatics Association : JAMIA*. 2015; 22:1220–30. [PubMed: 26342218]
18. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. *AMIAAnnuSympProc*. 2006:1040.
19. Ding K, et al. Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. *G3 (Bethesda, Md)*. 2013; 3:1061–8.
20. Crosslin DR, et al. Genetic variation in the HLA region is associated with susceptibility to herpes zoster. *Genes and immunity*. 2015; 16:1–7. [PubMed: 25297839]
21. Shameer K, et al. A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Human genetics*. 2014; 133:95–109. [PubMed: 24026423]
22. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. *PloS one*. 2010; 5
23. Delaney JT, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin Pharmacol Ther*. 2012; 91:257–63. [PubMed: 22190063]
24. Ramirez AH, et al. Predicting warfarin dosages in European-American and African-American subjects using DNA samples linked to an electronic health record. *Pharmacogenomics*. 2012 in press.
25. Birdwell KA, et al. The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. *Pharmacogenetics and genomics*. 2012; 22:32–42. [PubMed: 22108237]
26. Onitilo AA, et al. Estrogen receptor genotype is associated with risk of venous thromboembolism during tamoxifen therapy. *Breast cancer research and treatment*. 2009; 115:643–50. [PubMed: 19082882]
27. Onitilo AA, et al. Estrogen receptor genotype is associated with risk of venous thromboembolism during tamoxifen therapy. *Breast Cancer ResTreat*. 2009; 115:643–50.
28. Karnes JH, et al. A genome-wide association study of heparin-induced thrombocytopenia using an electronic medical record. *Thrombosis and haemostasis*. 2015; 113:772–81. [PubMed: 25503805]
29. Van Driest SL, et al. Genome-Wide Association Study of Serum Creatinine Levels during Vancomycin Therapy. *PloS one*. 2015; 10:e0127791. [PubMed: 26030142]
30. Karol SE, et al. Genetics of glucocorticoid-associated osteonecrosis in children with acute lymphoblastic leukemia. *Blood*. 2015; 126:1770–6. [PubMed: 26265699]
31. Wells QS, et al. Genomewide-Association Identifies a Novel Locus for Anthracycline Cardiotoxicity. *Circulation*. 2013; 128:A15509.
32. Kawai VK, et al. Genotype and risk of major bleeding during warfarin treatment. *Pharmacogenomics*. 2014; 15:1973–83. [PubMed: 25521356]
33. Roth JA, et al. Genetic Risk Factors for Major Bleeding in Warfarin Patients in a Community Setting. *Clin Pharmacol Ther*. 2014; 95:636–43. [PubMed: 24503627]
34. Kimmel SE, et al. A Pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. *New England Journal of Medicine*. 2013; 369:2283–93. [PubMed: 24251361]
35. Pirmohamed M, et al. A Randomized Trial of Genotype-Guided Dosing of Warfarin. *New England Journal of Medicine*. 2013; 369:2294–303. [PubMed: 24251363]
36. Perera MA, Cavallari LH, Johnson JA. Warfarin pharmacogenetics: an illustration of the importance of studies in minority populations. *Clin Pharmacol Ther*. 2014; 95:242–4. [PubMed: 24548987]

37. Perera MA, et al. Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. *The Lancet*. 2013; 382:790–6.
38. Limdi NA, et al. Race influences warfarin dose changes associated with genetic factors. *Blood*. 2015; 126:539–45. [PubMed: 26024874]
39. Ritchie MD, et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation*. 2013; 127:1377–85. [PubMed: 23463857]
40. Denny JC, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010; 26:1205–10. [PubMed: 20335276]
41. Denny JC, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*. 2013; 31:1102–11.
42. Hebbbring SJ, Schrodri SJ, Ye Z, Zhou Z, Page D, Brilliant MH. A PheWAS approach in studying HLA-DRB1*1501. *Genes and immunity*. 2013; 14:187–91. [PubMed: 23392276]
43. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. *Nature biotechnology*. 2015; 33:342–5.
44. Hall MA, et al. Detection of pleiotropy through a Phenome-wide association study (PheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLoS genetics*. 2014; 10:e1004678. [PubMed: 25474351]
45. Finucane HK, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015; 47:1228–35. [PubMed: 26414678]
46. Li L, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*. 2015; 7:311ra174–311ra174.
47. Xu H, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association : JAMIA*. 2014
48. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014; 506:376–81. [PubMed: 24390342]
49. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *The New England Journal of Medicine*. 2006; 354:1264–72. [PubMed: 16554528]
50. Desai NR, Sabatine MS. PCSK9 inhibition in patients with hypercholesterolemia. *Trends in cardiovascular medicine*. 2015; 25:567–74. [PubMed: 25771732]
51. Stitzel NO, et al. Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N Engl J Med*. 2014; 371:2072–82. [PubMed: 25390462]
52. Collins F. Opportunities and challenges for the NIH--an interview with Francis Collins. Interview by Robert Steinbrook. *N Engl J Med*. 2009; 361:1321–3. [PubMed: 19759378]
53. Shuldiner AR, et al. The Pharmacogenomics Research Network Translational Pharmacogenetics Program: overcoming challenges of real-world implementation. *Clin Pharmacol Ther*. 2013; 94:207–10. [PubMed: 23588301]
54. Manolio TA, et al. Implementing genomic medicine in the clinic: the future is here. *Genet Med*. 2013; 15:258–67. [PubMed: 23306799]
55. Stimpfle F, et al. Impact of point-of-care testing for CYP2C19 on platelet inhibition in patients with acute coronary syndrome and early dual antiplatelet therapy in the emergency setting. *Thrombosis research*. 2014; 134:105–10. [PubMed: 24856643]
56. Weitzel KW, et al. Clinical pharmacogenetics implementation: approaches, successes, and challenges. *American journal of medical genetics Part C, Seminars in medical genetics*. 2014; 166C:56–67.
57. Pulley JM, et al. Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project. *Clin Pharmacol Ther*. 2012; 92:87–95. [PubMed: 22588608]
58. Bell GC, et al. Development and use of active clinical decision support for preemptive pharmacogenomics. *Journal of the American Medical Informatics Association : JAMIA*. 2014; 21:e93–9. [PubMed: 23978487]

59. Schildcrout JS, et al. Optimizing Drug Outcomes Through Pharmacogenetics: A Case for Preemptive Genotyping. *Clin Pharmacol Ther.* 2012; 92:235–42. [PubMed: 22739144]
60. Van Driest SL, et al. Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. *Clin Pharmacol Ther.* 2013
61. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther.* 2011; 89:464–7. [PubMed: 21270786]
62. Hulot JS, et al. Cytochrome P450 2C19 loss-of-function polymorphism is a major determinant of clopidogrel responsiveness in healthy subjects. *Blood.* 2006; 108:2244–7. [PubMed: 16772608]
63. Mega JL, et al. Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *JAMA.* 2010; 304:1821–30. [PubMed: 20978260]
64. Cavallari LH, et al. Abstract 11802: Clinical Implementation Of CYP2C19-genotype Guided Antiplatelet Therapy Reduces Cardiovascular Events After PCI. *Circulation.* 2015; 132:A11802.
65. Mallal S, et al. HLA-B*5701 Screening for Hypersensitivity to Abacavir. *The New England Journal of Medicine.* 2008; 358:568–79. [PubMed: 18256392]
66. Rehm HL, et al. ClinGen — The Clinical Genome Resource. *New England Journal of Medicine.* 2015; 372:2235–42. [PubMed: 26014595]
67. Gallego, Carlos J., et al. Penetrance of Hemochromatosis in HFE Genotypes Resulting in p.Cys282Tyr and p.[Cys282Tyr];[His63Asp] in the eMERGE Network. *The American Journal of Human Genetics.* 2015; 97:512–20. [PubMed: 26365338]
68. Van Driest S, et al. Rare potentially pathogenic variants in the congenital arrhythmia syndrome disease genes SCN5A and KCNH2 are detected frequently but rarely associated with arrhythmia phenotypes in electronic health records. Accepted for presentation, American Society for Human Genetics 2014 Scientific Sessions. 2014
69. [Accessed November 23, 2015] 2015. <https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf>
70. Wei WQ, et al. Characterization of statin dose response in electronic medical records. *Clin Pharmacol Ther.* 2014; 95:331–8. [PubMed: 24096969]
71. McGeachie MJ, et al. CTNNA3 and SEMA3D: Promising loci for asthma exacerbation identified through multiple genome-wide association studies. *The Journal of allergy and clinical immunology.* 2015
72. Naidoo D, et al. A polymorphism in HLA-G modifies statin benefit in asthma. *The pharmacogenomics journal.* 2015; 15:272–7. [PubMed: 25266681]
73. Hoh BL, et al. CYP2C19 and CES1 polymorphisms and efficacy of clopidogrel and aspirin dual antiplatelet therapy in patients with symptomatic intracranial atherosclerotic disease. *Journal of neurosurgery.* 2015:1–6.

Figure 1.

General approach to EHR-based phenotyping adopted in the electronic Medical Records and Genomics (eMERGE) network. PPV: Positive predictive value

Figure 2.

Creation of a structured deidentified dataset from the EHR. Identifiers are removed, biomedical concepts are identified and structured, medications are extracted and structured, and custom classifiers (such as case or control definitions, or algorithms to define smoking status) are applied. Over time, the structured dataset can thus become more useful as a research discovery tool than to original EHR.

Table 1

PGPop studies

Project title	Sites	Cases	Controls	Outcome/comments
Statin dose-response	PMRP BioVU	4278 subjects with multiple statin doses		Association of <i>PRDM16</i> variants with 50% effective dose ⁷⁰
Asthma response to inhaled corticosteroids in asthma	Harvard Pilgrim Harvard Crimson PMRP BioVU	393	1286	Identification of SNPs in <i>CTNNA3</i> and <i>SEMA3D</i> associated with asthma exacerbation. <i>SEMA3D</i> replicated in PGPop. ⁷¹
Pharmacogenetic effects of statins in asthma	PMRP BioVU	447	1206	Statins may be effective in asthma by upregulating miRNA-152 family members to modulate expression of HLA-G. PGPop replicated an association between rs1063320 (in the 3'UTR of <i>HLA-G</i>) and asthma exacerbations. ⁷²
Major adverse cardiovascular events during statin therapy	BioVU PMRP	2679	5000	Ongoing
Pharmacogenetics of methotrexate-induced liver injury	Harvard Partners BioVU Kaiser Permanente	344	4872	Algorithm development for rare adverse events may yield low positive predictive values. ¹⁴
Association of CES1 G143E with increased clopidogrel responsiveness and/or increased bleeding events	BioVU	1926	1812	Ongoing
Association of CYP2C19 with increased risk of stroke in intracranial atherosclerotic disease with clopidogrel	BioVU PMRP	84	97	CYP2C19 variants were unexpectedly associated with fewer clopidogrel failures. ⁷³
Genome-wide Meta-Analysis of Allopurinol Response in Patients with Gout or Hyperuricemia	BioVU PMRP	2610 allopurinol users		Ongoing

PMRP: Personalized Medicine Research Program at the Marshfield Clinic