SUPPLEMENTARY MATERIAL

The Electronic Medical Records and Genomics (eMERGE) Network: Past, Present and Future

Omri Gottesman, MD, Helena Kuivaniemi, MD, PhD, Gerard Tromp, PhD, W. Andrew Faucett, MS, Rongling Li, MD, PhD, Teri A. Manolio, MD, PhD, Saskia C. Sanderson, PhD, Joseph Kannry, MD, Randi Zinberg, MS, CGC, Melissa A. Basford, MBA, Murray Brilliant, PhD, David J. Carey, PhD, Rex L. Chisholm, PhD, Christopher G. Chute, MD, DrPH, John J. Connolly, MD, David Crosslin, PhD, Joshua C. Denny, MD, Carlos J. Gallego, MD, Jonathan L. Haines, PhD, Hakon Hakonarson, MD, PhD, John Harley, MD, PhD, Gail P. Jarvik, MD, PhD, Isaac Kohane, MD, PhD, Iftikhar J. Kullo, MD, Eric B. Larson, MD, MPH, Catherine McCarty, PhD, MPH, Marylyn D. Ritchie, PhD, Dan Roden, MD, Maureen E. Smith, MS, Erwin P. Böttinger, MD, Marc S. Williams, MD and The eMERGE Network

DESCRIPTION OF eMERGE-II SITES

Locations of the nine research groups, their affiliated sites, a coordinating center and the services and support centers constituting eMERGE-II are shown in Figure 1. Outline of the activities of the eMERGE Network is shown in Figure 2 and the organizational structure of the network is represented in Figure 3. In addition to the site-specific information that follows, details of the biorepositories, EMR systems and genotyping projects are summarized in Table 1. The main aims of each site are listed in Table S1 and the primary and secondary phenotypes selected by each site are shown in Table S2.

Cincinnati Children's Hospital Medical Center (CCHMC) and Boston Children's Hospital (BCH)

CCHMC and BCH have combined their efforts to be a single eMERGE site. Their biobank solutions differ. CCHMC now obtains "opt-in" consent at outpatient and inpatient CCHMC patient registration for permission to use in research those biological materials that would otherwise be discarded. At this point there are nearly 15,000 cryopreserved tissues and 15,000 DNA samples being stored. The DNA samples are accumulating at a pace of over 2000 per month. BCH is enrolling 10,000 children and family members into The Gene Partnership with samples being collected and a broad consent to return results, including a web-based interface that presents results to patients and their guardians. This collection at BCH will be participating in many studies of the return of results.

The combination of institutions into one site provides BCH and CCHMC the opportunity to coordinate and solve problems between institutions so that the extension of algorithms to other eMERGE institutions will be simpler. The first goal is to construct the informatic infrastructure that will allow rapid application of algorithms developed at one institution rapidly at the other.

This CCHMC/BCH eMERGE II site is built on a five-year history of collaboration, particularly in patient EMR-related informatics, the basis of much of eMERGE II. All CCHMC/BCH site GWAS and EMR data are linked to a shared data warehouse platform, i2b2,[1-4] and are made available for analyses in a shared Genome-Wide Association Database, GWADB. GWAS data are derived from a variety of sources including clinical service (cytogenetics), normal control, investigator initiated phenotype-specific studies, and the BCH Research Connection – The Gene Partnership (TGP), an initiative focused on understanding the root causes of complex diseases. Under development is a combined EMR view of subjects with defined phenotypes from i2b2 using a modification of the SHRINE[5] distributed query system. This will enable cross-institutional real-time patient selection that scales nationally.

Early childhood severe obesity (2 through 5 years old) has been chosen as a phenotype thinking that early onset disease is more likely to be less complex and therefore more genetically homogeneous than the adult forms of severe obesity the analysis of which has required enormous numbers of cases and the finding of many genes with modest effects. The obesity epidemic is also affecting these children. At CCHMC there are more than 3000 children who appear to satisfy the preliminary criteria for inclusion and are in the top 0.1% for BMI by WHO criteria.

Autism is afflicting American children at an astonishingly increasing frequency. CCHMC and BCH both have large clinics and gifted clinicians attempting to manage and ameliorate the serious problems that these children and their families have. This is a complex phenotype for algorithm development. The complexity is used to apply both heuristic and machine learning approaches are being applied to explore the genetics of this condition.

For secondary phenotypes, the CCHMC/BCH site is preparing to apply the algorithms for asthma and type 2 diabetes, both of which are serious problems in childhood.

1. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19:181-185.
2. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, et al. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res. 2009;19:1675-1681.
3. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12:417-428.
4. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am J Hum Genet. 2011;88:57-69.
5. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16:624-630.

The Children's Hospital of Philadelphia (CHOP)

The Center for Applied Genomics (CAG) at The Children's Hospital of Philadelphia (CHOP) has established one of the world's largest pediatric biorepository and is actively working on over 50 medical disorders affecting children from birth to age 21. Recruitment—maintained by a staff of ten full-time nurses and phlebotomists—is primarily population-based, though a number of disease-specific efforts run concurrently. The CAG biorepository currently totals >70,000 internal samples, including >5,000 complete trios. We maintain an additional 100,000 samples through collaborative projects, including several thousand trios. The biorepository is growing at a rate of 12-15,000 samples per year. All samples/data are encrypted and barcoded at point-of-contact. More than 85% of participants consent/assent to re-contact to participate in further studies. All (internal) participants consent/assent to retrospective and prospective access to longitudinal EMRs. EMRs are handled through EpicCare®—in place in CHOP since 2001—and the mean length of EMRs for repository participants is >5.5 years. EMRs are enriched by data from a range of internal projects, including a follow-up study of 10,000 biorepository participants who completed a 3-hour neurobehavioral battery, and 1,500 of whom completed a neuroimaging study. In addition all participants complete an enrollment questionnaire that includes self-reported demographic, health, and family-history data. Samples are stored and managed by a robotic-based system from REMP (Oberdiessbach, Switzerland), which facilitates rapid data share and dramatically reduces the potential for placement/calling errors. More than 99% of samples are available for re-genotyping under CLIA/CAP certification or next-generation sequencing. The Center's analytical capabilities are designed to identify genetic variants across a spectrum of medical disorders and we have active projects on >50 disease areas. The biorepository is notable for its large proportion (43%) of African American samples.

The primary objectives of the CHOP eMERGE project are 1) to build upon existing eMERGE initiatives to adapt phenotype algorithms relevant to pediatric cohorts (lipids being the primary phenotype in this regard), 2) to define novel phenotypes with relevance to adult and pediatric cohorts (asthma, attention deficit hyperactivity disorder, atopic dermatitis, and gastroesophageal reflux disease are primary phenotypes in this regard), 3) to conduct GWAS with minimal risks to patient privacy from sharing of EMR data, 4) to develop consent and community consultation procedures for conducting research, and 5) to begin incorporating genomic research results into clinical care.

Geisinger Health System

Geisinger Health System (GHS) is an integrated, comprehensive health care delivery system that serves a large, stable, mainly rural population in north central and northeastern Pennsylvania. Geisinger has a fully functional and integrated EMR system, and is a recognized leader in the use of EMR and health information technology. To leverage the health system's assets for genomic medicine in 2006 Geisinger launched a biobanking program, the MyCode® project, a central repository of patient samples (blood, DNA, serum and tissue) that are linkable to EMR for broad research use in a manner that protects confidentiality of patient information.[1]

The goals of the Geisinger eGenomic Medicine (GeM) Program are to: 1) Identify genetic variants associated with abdominal aortic aneurysm (AAA),[2] extreme obesity,[3] and related conditions; 2) Develop approaches to incorporate genomic data into clinical care, using EMR-based clinical decision support tools; and 3) Identify sociocultural concerns of patients residing in rural areas regarding return of genetic findings, and provide patient and physician education. The strategy for integration of genetic data into EMR is modeled on processes developed by Geisinger's Clinical Transformation Team, a dedicated department devoted to creating IT-enabled improvements in patient care. A "Care Gaps" strategy is used, in which clinical areas are evaluated to

identify IT processes that can be incorporated into the EMR to improve patient care. Geisinger's ELSI Advisory Committee will develop policies and guidelines in conjunction with the eMERGE Network and NHGRI for return of results to participants. It will also create educational materials for participants and clinicians.

1. Gerhard GS, Carey DJ, Steele GD, Jr. Electronic Health Records in Genomic Medicine. In: Willard HF, Ginsburg GS, eds. *Genomic and Personalized Medicine.* Vol 1. 2 ed. Waltham, MA: Academic Press; 2013:287-294.
2. Gretarsdottir S, Baas AF, Thorleifsson G, et al. Genome-wide association study identifies a sequence variant within the *DAB2IP* gene conferring susceptibility to abdominal aortic aneurysm. Nat Genet. 2010;42:692-697.
3. Wood GC, Chu X, Manney C, Strodel W, Petrick A, Gabrielsen J, Seiler J, Carey D, Argyropoulos G, Benotti P, Still CD, Gerhard GS. An electronic health record-enabled obesity database. BMC Med Inform Decis Mak 2012;12:45.

Group Health Cooperative, University of Washington and the Fred Hutchinson Cancer Research Center

The Group Health (GH) biobank for eMERGE II utilizes patients enrolled at Group Health Cooperative, a large, regional, integrated delivery system based in Seattle, Washington, serving more than 620,000 persons. The GH eMERGE biobank includes subjects aged > 50 years from two ongoing biorepositories at GH: 1) the Alzheimer's disease Patient Registry/Adult Changes in Thought (ADPR/ACT) study; and 2) the Northwest Institute of Genetic Medicine (NWIGM) prospective biorepository. The current eMERGE II cohort consists of 5,299 subjects from these biorepositories, including 3,561 that have been genotyped and whose data form the current analytic sample for the eMERGE-II project. An additional 1,738 subjects have exome chip, but not GWAS data. All subjects have long-term GH EMR.

Seattle, with other Network sites, will continue to conduct GWAS analyses for multiple new phenotypes identified through the EMR. Phenotypes planned to be explored by this site include those related to infectious disease susceptibility (BuGWAS), specifically susceptibility to Clostridium difficile, reactivation of the varicella zoster virus, and fungal nail infection (onchomychosis). Additionally, we are studying the relationship of chromosomal abnormalities identified by GWAS with bone marrow disorders. Finally, Seattle investigators will generate and study implementation of pharmacogenetic data. Using one-on-one interviews with medical system leaders, focus groups with physicians and patients, and prototype development and testing, Seattle investigators are assessing the challenges of integrating genomic information into clinical care. GH, the University of Washington and the Fred Hutchinson Cancer Research Center will collaborate within and extend eMERGE through exchange of knowledge, technology, and best practices for generating EMR-based GWAS and integrating with clinical care. We have specific collaborations with the Alzheimer's Disease Genetics Consortium and with external genetic and genomic studies.

Marshfield Clinic Foundation, Essentia Institute of Rural Health and Pennsylvania State University

Marshfield/Essentia/Penn State proposed three aims for their eMERGE II project: 1) develop and validate electronic algorithms for ophthalmic conditions and efficacy of medical therapy for ophthalmic conditions and implement other phenotype algorithms developed across the eMERGE network, 2) leverage GWAS data available for nearly 6000 research subjects aged 50 years and older in Marshfield and an additional 20-25,000 subjects throughout the eMERGE network to undertake genetic discoveries for ophthalmic conditions and ophthalmic pharmacogenetics, and 3) undertake consultation activities with the general community and clinicians related to the incorporation of GWAS results into electronic health records to inform health care decisions. Information about the primary phenotypes can be found in Table S2. In addition to the GWAS analyses using the SNP data, the CNV data are being cleaned and will be used for both the eMERGE 1 (cataract, HDL) and eMERGE II outcomes. Focus group discussions, using the same methodology that was employed in eMERGE-I[1] have been undertaken with general community members who have primary care physicians at Marshfield Clinic, participants of the PMRP biobank,[2] and Marshfield Clinic primary care physicians. The discussions centered around the return of genetic research results and supporting information for patients and providers. The PMRP Community Advisory Group meets several times per year to provide advice on various aspects of eMERGE II and the PMRP cohort is kept informed of research activities through a twice yearly newsletter.[3]

1. McCarty CA, Garber A, Reeser JC, Fost NC. Study newsletters, community and ethics advisory boards, and focus group discussions provide ongoing feedback for a large biobank. *Am J Med Genet* 2011;155:737-741.
2. McCarty CA, Wilke RA, Giampietro PF, Wesbrook S, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large, population-based biobank. *Personalized Medicine* 2005;2:49-79.
3. McCarty CA, Chapman-Stone D, Giampietro PF, Fost NC, PMRP Community Advisory Group. Community consultation and communication for a population-based DNA biobank: the Marshfield Clinic Personalized Medicine Research Project. *Am J Med Genet* 2008;3026-3033.

Mayo Clinic

The major focus at Mayo for phase II of eMERGE is to conduct EMR-based GWAS[1] of multiple cardiovascular phenotypes and develop methods to translate relevant genomic findings to clinical practice with ongoing guidance from the community. The proposed Mayo eMERGE-II cohort (n=6,916) includes the 3,769 eMERGE-I patients and an additional 3,147 individuals, the majority (90%) genotyped on the Illumina 660W-Quad platform. Most of the Mayo patients (>95%) are of European ancestry. The Mayo investigators are in the process of completing EMR-based GWAS to identify common genetic variants that influence: a) inter-individual variation in cardiorespiratory fitness, a powerful marker of adverse outcomes, and b) susceptibility to two common conditions of significant public health importance: venous thromboembolism and heart failure. Validated and transportable phenotyping algorithms for these phenotypes are being developing and made widely available to the scientific community via an open source library (PheKB). The second major goal of Mayo eMERGE phase II is to conduct a randomized-clinical trial to investigate how patients respond to genotype-informed coronary heart disease (CHD) risk. The investigators will enroll 150 patients from the community who are at intermediate 10-y CHD risk, perform CLIA-certified genotyping of multiple CHD susceptibility SNPs, and communicate the results and the implications for CHD risk to patients and their providers using specifically designed risk communication tools. The effectiveness of the communication and the patients' comprehension of risk, response to the information, and planned changes in lifestyle will be assessed by interview and surveys shortly after communicating with CHD risk and at several time points thereafter. Concomitantly, the Mayo team is working on informatics approaches to incorporate genomic information and relevant clinical decision support tools into the EMR.

1. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010;17:568-574. [PMCID: 2995686]
2. Kullo IJ, Jarvik GP, Manolio TA, Williams MS, Roden DM. Leveraging the electronic health record to implement genomic medicine. Genet Med. (in press) PMID: 23018749

Mount Sinai School of Medicine

The Mount Sinai Hospital and School of Medicine comprise The Mount Sinai Medical Center, a large academic medical center serving nearly 60,000 inpatient and over 600,000 outpatient visits annually. Mount Sinai's patient populations largely reside in urban communities of Northern Manhattan that are socio-economically, ethnically, and religiously diverse. In 2007, Mount Sinai established The Charles Bronfman Institute for Personalized Medicine (IPM) where the Bio*Me*™ Biobank Program was launched in September 2007 with the mission to advance data-driven and gene-based individualization of healthcare. Bio*Me*™ is an ongoing, prospective, hospital-based population study that has enrolled over 23,000 consented participants from Mount Sinai's clinical care facilities.[1] Participants consent to a) using their complete EMR for phenotypic and exposure data, b) future re-contacting for participation in additional IRB approved research, and c) sharing of de-identified phenotypic and genotypic data. The Bio*Me*™ cohort represents the ethnic and racial diversity of Mount Sinai's local communities, with 25% of African Ancestry (AA), 30% of European Ancestry (EA), 36% of Hispanic or Latino ancestry (HL), and 9% of other ancestry. Mount Sinai joined the eMERGE II Network in 2011 with the "Biorepository for Genomic Medicine in Diverse Communities" research program. The Mount Sinai eMERGE II program already provided the eMERGE II Network with high quality GWA genotype datasets for 6,275 of its participants, including 4,363 AA, 1,212 HA, and 700 EA. GWA genotyping (OmniExpress+Exome chips) has been extended to over 15,000 Bio*Me*™ participants by December 2012. Mount Sinai has invested significant resources in the development of a complex information management system led by IPM faculty to enable point-of-care incorporation of genotype information with clinical decision support delivered in Mount Sinai's EpicCare® EHR.

The goals and objectives of the Mount Sinai eMERGE II program are concentrated in the following areas: community-participatory education and research in genomic medicine; development and dissemination among

eMERGE II partner sites of phenotyping validated phenotyping algorithms for chronic kidney disease (CKD) and its progression (CKD progression), and drug-induced liver injury (DILI); support of eMERGE II GWAS projects by providing large sample of genetically diverse patient populations; share experience with development and implementation of the genomic information EHR integration system at Mount Sinai; pilot genomic medicine demonstration projects including eMERGE II Network wide clinical care implementation of pharmacogenetic support and local implementation of APOL1 genomic risk information in management of hypertension and renal care in patients of African ancestry.

1. Tayo BO, Teil M, Tong L, et al. Genetic background of patients from a university medical center in Manhattan: implications for personalized medicine. *PLoS One.* 2011;6:e19166.

Northwestern University

The Northwestern University (NU) eMERGE site is based on the NUgene biorepository through which DNA samples of consented participants are linked to clinical information from both inpatient and ambulatory care settings, and supplemented with data from a participant questionnaire. Participants consent to distribution of their coded DNA samples and data for a broad range of genetic research conducted by third-party investigators. Virtually all of our inpatient and outpatient data are captured electronically. Outpatient records are held in EpiCare®EMR and inpatients records in Cerner Powerchart®.[1]

In eMERGE I, NU led the development of phenotype algorithms for Type II Diabetes,[2] asthma,[3] lipids,[4] and height, in addition to implementing the algorithms developed by other sites. Over 4,950 NUgene participants have been genotyped as part of eMERGE I and as controls for an NICHD study of polycystic ovary syndrome. NU led analysis of the T2D GWAS in both African Americans and in European Americans. Mega analysis across all T2D eMERGE samples successfully replicated TCF7L2, the most consistent T2D risk allele yet identified. In addition, analysis of the African American samples has identified some interesting potential associations (not yet published). NU also led the analysis of LDL cholesterol in African Americans revealing a strong protective variant in APOE 4. NU conducted a three-phased mixed methods approach to consulting with various communities about genetic research and sharing of GWAS and phenotypic data, including NUgene biobank participants, the public, and IRBs.[5,6] NU was the lead eMERGE site in the development of model consent language for biobanks planning to perform high throughput genomic studies and share the data[7] and led the writing of a comparative paper on community engagement efforts.[8]

The eMERGE II project builds on NU's broad experience with using the EHR both for research and in developing clinical decision support tools as well as the genotype data available through eMERGE I activities to integrate genomic data into clinical care.[9,10] The specific goals of the project include 1) expanding the development and implementation of electronic phenotype algorithms and detection of new genomic associations; 2) Consult with physicians and patients regarding the utility of real world examples of clinically relevant genotypes and to inform clinical activities; 3) Develop technical approaches for integrating genetic variant data into our commercial EHR; and 4) Evaluate the impact of key translational elements such as regulatory barriers, EHR decision support tools, and the education of physicians and patients about genomic information and disseminate lessons learned and best practice recommendations.

1. 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 Genom 2011;4.
2. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. 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 Ass 2012;19:212-8.
3. Table of Pharmacogenomic Biomarkers in Drug Labels. (Accessed July 13, 2012, at http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm.)
4. Lincoln YSaG, E.G. Naturalistic Inquiry. Beverly Hills: Sage; 1985.
5. Lemke AA, Smith ME, Wolf WA, Trinidad SB. Broad data sharing in genetic research: views of institutional review board professionals. IRB 2011;33:1-5.
6. Lemke AA, Wolf WA, Hebert-Beirne J, Smith ME. Public and Biobank Participant Attitudes toward Genetic Research Participation and Data Sharing. Public Health Genomics 2010;13:368-77.
7. The Electronic Medical Records and Genomics (eMERGE) Network Consent & Community Consultation Workgroup Informed Consent Task Force: MODEL CONSENT LANGUAGE. 2009. http://www.genome.gov/Pages/PolicyEthics/InformedConsent/eMERGEModelLanguage2009-12-15.pdf

8. Lemke AA, Wu JT, Waudby C, Pulley J, Somkin CP, Trinidad SB. Community engagement in biobanking: Experiences from the eMERGE Network. Genomics Soc Policy 2010;6:35-52.
9. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011;3:79re1.
10. Persell SD, Kaiser D, Dolan NC, et al. Changes in performance after implementation of a multifaceted electronic-health-record-based quality improvement system. Med Care 2011;49:117-25.

Vanderbilt University

Vanderbilt proposed four aims for their eMERGE II project: 1) accelerating development of phenotype extraction algorithms through a principled approach to selecting and building phenotype algorithms; 2) to investigate combinations of genotypes highly predictive of disease and disease outcomes; 3) to study patient perspectives of genetic testing to guide drug prescribing; and, 4) to develop tools to maximize the ability to share genomic information while preserving privacy. Pharmacogenetic phenotypes that could potentially alter care were of particular interest, and the primary phenotypes are listed in Table S2. An important part of discovery understanding the range of phenotypes associations with genetic variation. Thus, they are more broadly deploying the EMR-based phenome-wide association study (PheWAS)[1] to broadly investigate pleiotropy with discovered variants and to guide further phenotype investigation. Their eMERGE II project builds upon the PREDICT (Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment) program[2], which seeks to prospectively test patients at risk for receiving medications with known pharmacogenetic influences. The project currently provides computerized decision support advice for clopidogrel, warfarin, and simvastatin based on testing using the Illumina VeraCode ADME platform. More than 10,000 patients have been tested for PREDICT as of December 2012. They have shown that exposure to medications with known pharmacogenetic influences is a common event, with 65% of a normal outpatient population receiving at least one medication with Food and Drug Administration-asserted genetic influences.[3]

1. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–1210.
2. Pulley, J. M. *et al.* Operational implementation of prospective genotyping for personalized medicine: The design of the Vanderbilt PREDICT Project. *Clin Pharmacol Therap*. doi:10.1038/clpt.2011.371
3. Schildcrout, J. S. *et al.* Optimizing drug outcomes through pharmacogenetics: A Case for preemptive genotyping. *Clin Pharmacol & Therap* 2012;92:235–242.

Coordinating Center

The eMERGE Network Coordinating Center (CC), located at Vanderbilt University Medical Center with a subcontract to The Pennsylvania State University, is the central communications and administrative hub of eMERGE. The CC coordinates all in-person steering committee meetings and provides project management and administrative support to all six Network workgroups and the External Scientific Panel. The CC has developed and maintains the primary Network web portal (www.gwas.org) and provides organizational dashboards detailing site-specific and Network progress against goals. The CC also provides leadership in three key areas: advancement of electronic phenotyping, genomic data management, and data privacy. Phenotyping efforts include development of a web-based library for electronic phenotypes (http://PheKB.org) and development and implementation of a high-throughput phenotype-genotype association discovery method combining EMR and genetic data (PheWAS). GWAS data are available from over 18,000 eMERGE-I subjects and an additional 24,000 from eMERGE-II. User-friendly software for de-identification and privacy has been developed and is available for all sites.

Table S1. Summary of eMERGE-II Activities at Each Site

Institution	Aim 1	Aim 2	Aim 3	Aim 4	Aim 5
Cincinnati Children's Hospital and Boston Children's Hospital	Demonstrate real-time execution of phenotypic selection across two distinct pediatric institutions with disparate EMR systems as a model for ensuring phenotypic standardization and for national scalability	Contribute to phenotype selection and perform GWAS and PheWAS	Use return of *COMT* and *CYP2D6* research results to explore parents' responses to and use of their children's results and understand factors that influence their decisions about learning incidental findings	Explore clinicians' perceptions of pharmacogenetic research results after EMR integration	
Children's Hospital of Philadelphia	Adapt eMERGE phenotype algorithms relevant to pediatric cohorts with circulating lipid levels as the primary phenotype	Define novel phenotypes with relevance to adult and pediatric cohorts (asthma, ADHD, atopic dermatitis, and GERD)	Conduct GWAS with minimal risks to patient privacy from sharing of EMR data	Develop consent and community consultation procedures for conducting research	Begin incorporating genomic research results into clinical care
Geisinger Clinic	Identify genetic variants associated with AAA, extreme obesity, and related conditions	Incorporate genomic data into clinical care, using EMR-based CDS	Identify sociocultural concerns of patients in rural areas regarding return of genetic findings. Provide patient and physician education		
Group Health	Study infectious disease susceptibility (BuGWAS), specifically susceptibility to *Clostridium difficile*, reactivation of the varicella zoster virus, and fungal nail infection (onychomycosis)	Investigate the relationship of chromosomal abnormalities (ChroWAS) identified by GWAS with bone marrow disorders	Assess challenges of integrating genomic data into clinical care: interview medical system leaders, focus groups with physicians and patients, and develop and test prototype	Collaborate within and extend eMERGE through exchange of knowledge, technology, and best practices for generating EMR-based GWAS and integrating with clinical care	
Marshfield Clinic Research Foundation	Develop and validate electronic algorithms and efficacy of medical therapy for ophthalmic conditions	Undertake genetic discoveries for ophthalmic conditions and pharmacogenetics	Consultat with the general community and clinicians related to the incorporation of GWAS results into EMR		
Mayo Clinic	Identify genetic variants for inter-individual variation in cardio-respiratory fitness, and susceptibility to VTE and CHF using validated and transportable	Quantify genetic risk of CHD and statin myopathy. Develop risk communication tools with clinical and genetic components to both patients and care	Develop informatics approaches to incorporate genomic information and relevant clinical decision support tools into the EMR	Conduct a randomized-clinical trial to investigate how patients respond to genotype-informed CHD risk	

	phenotyping algorithms	providers			
Mount Sinai School of Medicine	Develop and implement secure prototype Biobank-EMR genomic data interface and EMR-enabled genomic CDS in clinical care	Develop phenotyping algorithms for chronic kidney disease and its progression	Expand GWAS for cardiovascular and renal phenotypes across minority populations	Implement novel solutions incorporating genomic results with EMR	Explore innovative approaches for community-participatory education and research in genomic medicine
Northwestern University	Expand development of phenotype algorithms and detection of new genomic associations	Consult with physicians and patients regarding the utility of real world examples of clinically relevant genotypes and to inform clinical activities	Develop technical approaches for integrating genetic variant data into the EMR	Evaluate impact of key translational elements (e.g. regulatory barriers, EMR CDS tools, and provider and patient education) and disseminate lessons learned and best practice recommendations	
Vanderbilt University	Accelerate development of phenotype algorithms	Identify combinations of genotypes highly predictive of disease or drug response outcomes	Study patient perspectives of genetic testing to guide drug prescribing	Develop tools to maximize sharing of genomic information while preserving privacy	
Coordinating Center Vanderbilt University with subcontract to the Pennsylvania State University	Facilitate communication between eMERGE sites and coordinate in-person Steering Committee meetings	Provide project management and administrative support for all workgroups and ESP	Provide leadership in advancement of electronic phenotyping, genomic data management, and data privacy	Develop and maintain web portal (www.gwas.org)	Provide organizational dashboards with site-specific and network progress towards goals

AAA, abdominal aortic aneurysm; ADHD, attention deficit hyperactivity disorder; CDS, Clinical Decision Support; CHD, coronary heart disease; CHF, chronic heart failure; EMR, electronic medical records; ESP, external scientific panel; GERD, gastroesophageal reflux disease; GWAS, genome-wide association study; PheWAS, phenome-wide association study; VTE, venous tromboembolism.
For details on biorepositories and EMR, see Table 1 in the main section.

Table S2. eMERGE Phase II Phenotypes

Phenotype	Site	
	Primary	**Secondary**
Abdominal aortic aneurysm (AAA)	Geisinger	Mayo, Vanderbilt
buGWAS: *Clostridium difficile*	Group Health	Northwestern, Vanderbilt
Cardiorespiratory fitness (CRF)	Mayo	Geisinger, Vanderbilt
Diabetes/Hypertension-associated Chronic Kidney Disease (CKD)	Mount Sinai	Marshfield, Northwestern
Lower GI (non-syndromic polyps)	Northwestern	Vanderbilt
Ocular Hypertension and response to medication	Marshfield	Geisinger, Group Health
ACEI cough	Vanderbilt	Northwestern, Group Health
Lipids	CHOP	
Childhood Obesity	CCHMC & BCH	
Extreme Obesity (BMI)	Geisinger	Marshfield
Dose response for lipid lowering agents	Vanderbilt	Mayo, Mount Sinai
Glaucoma	Marshfield	Geisinger, Group Health
Lower GI (diverticulosis and -itis)	Northwestern	Marshfield, Vanderbilt
Drug Induced Liver Injury (DILI)	Mount Sinai	Marshfield, Mayo
Venous Thromboembolism (VTE)	Mayo	Vanderbilt
buGWAS: onychomychosis (*Tinea unguium*)	Group Health	Marshfield
Autism Spectrum Disorder	CCHMC & BCH	
Childhood Asthma	CHOP	
buGWAS: shingles (*Varicella zoster*)	Group Health	Vanderbilt
Heart Failure	Mayo	Group Health
Rapid renal decline in Diabetes/Hypertension-associated Chronic Kidney Disease (CKD)	Mount Sinai	Marshfield, Mayo
Community acquired MRSA	Northwestern	Marshfield, Geisinger
AMD or Dry eye & response to medication	Marshfield	Vanderbilt, Northwestern
Dose Response to lipid lowering agents	Vanderbilt	Mayo, Mount Sinai
Complications of Obesity -- resolution of diabetes after ROUX-EN-Y	Geisinger	Marshfield
Upper GI (Peptic ulcer disease)	Vanderbilt	Northwestern

The phenotyping workgroup of the eMERGE network has as its goal the creation, validation, and execution of phenotype algorithms across eMERGE. Algorithms are adapted and executed at each site to identify cases and controls for each phenotype of interest. These phenotype algorithms typically consist of combinations of billing codes, laboratory and test results, medication exposures, and natural language processing. The algorithms typically take 6-12 months to develop through a process of design and iterative validation. Validation across multiple sites is key to develop a transportable algorithm which can then be used in all eMERGE sites and outside the Network.