

The eMERGE Network: Challenges and Lessons Learned

Lin Gyi¹, Cathy A. McCarty², Rex L. Chisholm³, Chris G. Chute⁴, Paul K. Crane⁵, Gail Jarvik⁵, Iftikhar Kullo⁴, Eric Larson⁵, Daniel R. Masys⁶, Dan M. Roden⁶, Rongling Li¹ for the eMERGE Network

¹National Human Genome Research Institute, Bethesda, MD, ²Marshfield Clinic Research Foundation, Marshfield, WI, ³Northwestern University, Chicago, IL, ⁴Mayo Clinic, Rochester, MN,

⁵Group Health Cooperative/University of Washington, Seattle, WA, ⁶Vanderbilt University, Nashville, TN

The eMERGE Network
electronic Medical Records & Genomics



Abstract

Widespread adoption of the electronic medical record (EMR), though expensive and logistically challenging, can potentially establish new frontiers in personalized medicine. The electronic Medical Records and Genomics (eMERGE) Network (www.gwas.net) is a consortium of five participating sites (Group Health Seattle, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University) funded by the NHGRI to investigate synergies between EMR and genomic research. The goal of eMERGE is to conduct genome-wide association studies in approximately 19,000 individuals using EMR-derived phenotypes and DNA from linked biorepositories. While eMERGE is still underway, dissemination of important challenges and lessons learned from the network can benefit the scientific community. Challenges faced by eMERGE include development of EMR-based algorithms requiring expertise in clinical care, genetics/genomics, and biomedical informatics; implementation and validation of algorithms among different types of EMR data; informed consent or re-consent of patients for genomic research; and the return of incidental findings. The lessons learned include improving model consent language to better inform patients participating in genomic research, and designing EMR-based algorithms to be transportable to different institutions with varying data structures. However, a standardized EMR system would be more efficient for cost-effective research. The key strengths of eMERGE include its collaborative nature, potential for external Network initiatives, the ability to rapidly extract phenotypes from the EMR, transportability of algorithms, and cost-effectiveness for longitudinal clinical research. The eMERGE Network is uniquely poised to develop novel strategies for leveraging the EMRs in genomic research and thereby facilitate personalized medicine.

Introduction

The electronic Medical Records and Genomics (eMERGE) Network is a five-member site national consortium formed to develop, disseminate, and apply approaches to research that combines DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research to identify genetic risk factors for clinical disease.

Specific Aims

- Develop and validate electronic phenotyping algorithms for phenotype classification in genomic research
- Identify genetic variants related to complex traits through genome-wide association (GWA) analyses
- Develop, implement and evaluate the process of consent and community consultation for genomic research
- Develop best practices to protect patients to maximize data sharing, and to benefit society

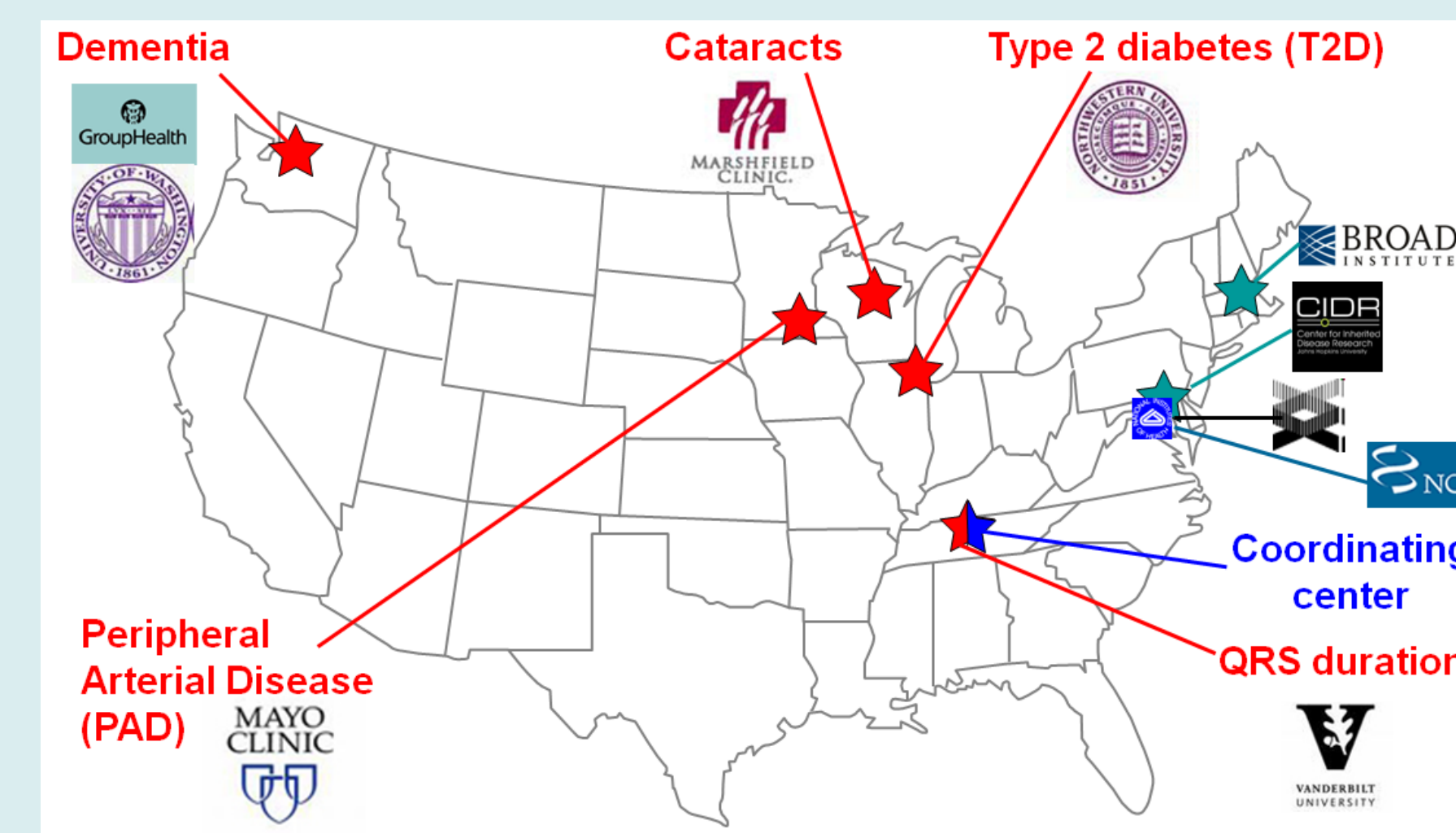
Genotyping

- Genotyping facilities: Broad Institute and Center for Inherited Disease Research (CIDR)
- Platforms: Illumina 1M for individuals of African American ancestry and Illumina 660W Quad for individuals of European ancestry and other race/ethnicity
- Quality control (QC): genotyping QC and centralized data cleaning QC

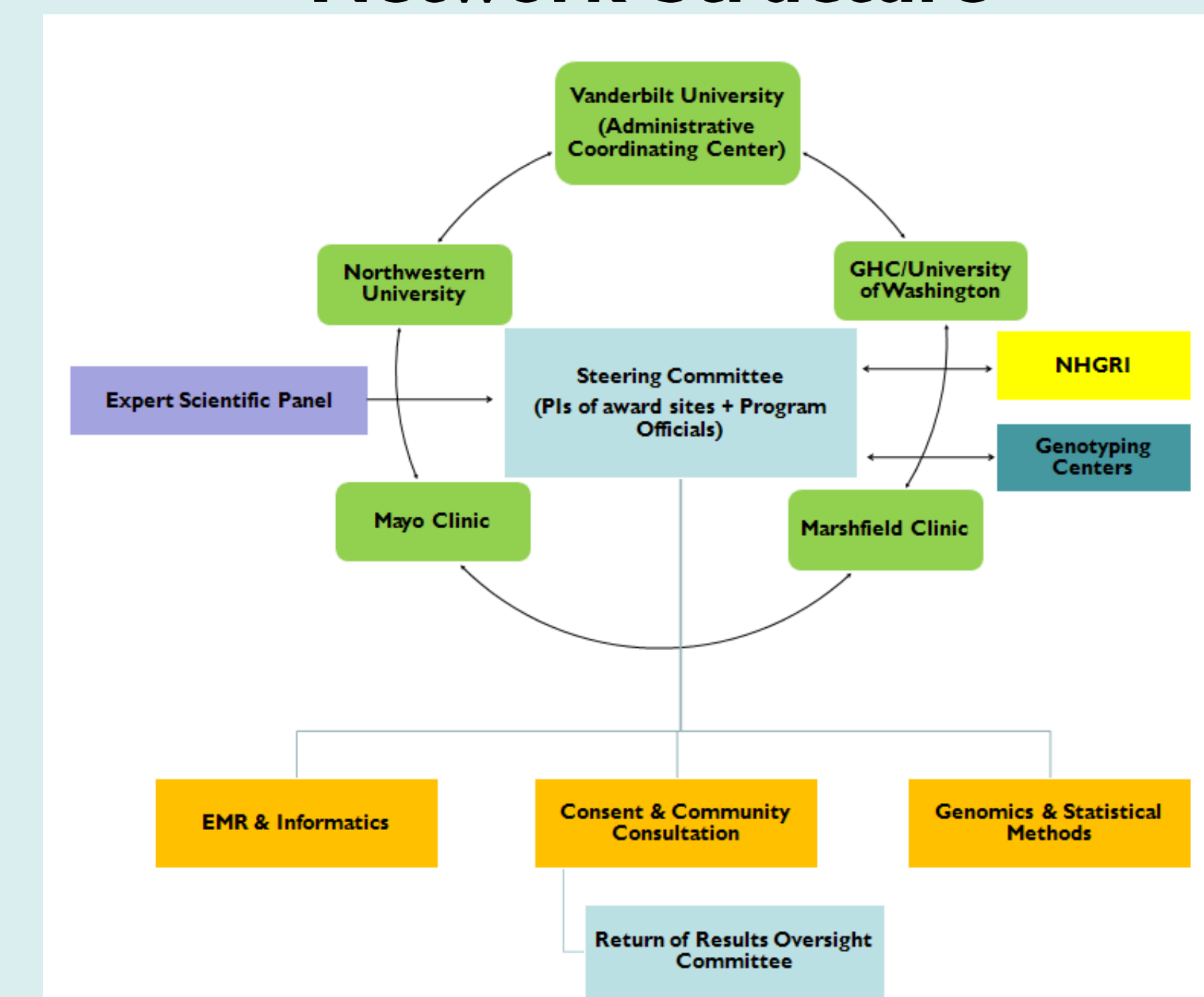
Phenotyping

- Identification: electronic algorithms for structured data extraction (i.e. ICD-9 code), free-text data mining, and/or natural language processing (NLP)
- Validation: manual chart review
- Phenotypes: 6 primary site-specific phenotypes, 2 network phenotypes, 5 secondary network phenotypes, and approximately 776 network phenotypes identified by ICD-9 codes using a PheWAS approach

eMERGE Network



Network Structure



Network Structure

- The **Steering Committee** is the governing body for the consortium and is composed of the Principal Investigators from each institution and the NIH Project Scientist
- An **Expert Scientific Panel** provides input to the NHGRI director about the progress and direction of the Network
- Vanderbilt University Medical Center is the site of the **Administrative Coordinating Center (ACC)** which provides support to the network including coordination, organization of committees, and support for the Expert Scientific Panel
- The **Informatics group** works to determine the validity, reliability, and comprehensiveness of EMR data for Genome-Wide Association Studies
- The **Consent and Community Consultation (C&CC) group** addresses 6 focus areas: consent, data sharing, community engagement, return of results, IRB issues, and identifiability, by leveraging expertise and community engagement activities at each site
- The **Genomics group** facilitates the GWAs timeline and sample quality control for the network
- The **Return of Results Oversight Committee** is one of the focus areas of the C&CC working group. The committee focuses on evaluating circumstances for return of results and works primarily with local investigators and IRBs

Phenotypes and Sample Sizes

Institution	Primary Phenotype	Network Phenotype*	Repository Size	GWA Study Size	EMR Description	Phenotyping Methods
Group Health, University of Washington (GHC Biobank)	Alzheimer's Disease and Dementia	WBC (White Blood Cell Count)	~4000; >96% EA	3,370; 97% EA	Vendor-based EMR since 2004; 20+ yrs pharmacy 15+ yrs ICD9	Structured data extraction, Mining free-text via regular expressions, Manual chart review
Marshfield Clinic (Personalized Medicine Research Project)	Cataracts and HDL-Cholesterol	Diabetic Retinopathy	~20,000; 98% EA	3,968; 99% EA	Internally developed EMR since 1985; 75% pts have 20+ yrs medical history	Structured data extraction, NLP, Intelligent Character Recognition
Mayo Clinic	Peripheral Arterial Disease	Red Blood Cell (RBC) indices	15,000; >96% EA	3,412; 99% EA	Internally developed EMR since 1995; 40 yrs data extraction	Structured data extraction, NLP
Northwestern University, (NUgene Project)	Type 2 Diabetes	Lipids & Height	~10,000; 12% AA 8% Hispanic	3,564; 52% AA	Vendor based EMR since 2000; 20+ yrs ICD9	Structured data extraction, NLP
Vanderbilt University (BioVU)	QRS Duration	PheWAS** (Phenome-Wide Association Study)	100,000; 11% AA	3,061; 16% AA	Internally developed EMR since 2000; 35+ yrs medical history	Structured data extraction, NLP
eMERGE Network		Hypothyroidism and Resistant Hypertension		~20,000	The cross network phenotypes were chosen based on the importance of the scientific question, whether GWAS had been performed for the trait before, and the effort required to develop accurate electronic phenotyping algorithms	

ICD9 = Ninth International Classification of Diseases; NLP = Natural Language Processing; EA = European Americans; AA = African Americans

Structured data extraction = retrieving data that have been stored in a predefined format

*Network phenotyping is not limited to the repository size of the parent institution

**Phenome-wide association study: using prevalent ICD9 codes to identify a significant amount of clinical phenotypes that may associated with select risk genetic markers

Sex Chromosome Anomalies

Sex Chromosome Anomaly	Site A	Site B	Site C	Site D	Site E*	Total
XX/XO mosaic	3	2	0	4	-	9
XO (Turner Syndrome)	0	1	0	0	-	1
XXY/XY mosaic	1	0	0	0	-	1
XXY (Klinefelter Syndrome)	1	1	4	1	-	7
XX, large LOH blocks on X	1	9	0	1	-	11
XXX (normal phenotype)	0	1	0	0	-	1
YYY (not reportable)	0	0	0	1	-	1
LOH (Loss of heterozygosity)	2	0	2	12	-	16

*Site E's data are not yet available.

**Sex chromosome anomaly events recorded out of ~ 13,800 available genotyped samples

The genotyping facilities collected sex chromosome anomaly data from the eMERGE Network sites; the Return of Results Oversight committee is working to determine which of these genotypes are reportable and how they should be discussed.

Network Challenges

- Phenotyping: development of EMR-based algorithms requiring expertise in clinical care, genetics/genomics, and biomedical informatics; implementation and validation of algorithms among different types of EMR data
- Genotyping: addressing quality control of pre-genotyping, genotyping and post-genotyping, data cleaning for site-specific data and combined network data
- Protecting human subjects: ensuring adequate human subject protections and addressing patients' concerns regarding such research
- Return of incident findings: requirements for CLIA certification, re-consent, IRB approvals,

Lessons Learned:

- Improved model consent language to better inform patients participating in genomic research based on EMRs; eMERGE's model consent language has been posted on NHGRI's Informed Consent website at <http://www.genome.gov/27526660>
- Developed methods to extract potentially identifiable clinical characteristics and modify them (by grouping or suppression) to minimize threats to the confidentiality of a patient's genomic information, while maximizing the EMR information preserved
- Designed and improved EMR-based algorithms to be transportable to different institutions with varying data structures
- Realized the key strengths of the eMERGE Network including its collaborative nature, potential for external Network initiatives, the ability to rapidly extract phenotypes from the EMR, and cost-effectiveness for longitudinal research

Acknowledgements

The eMERGE Network was initiated and funded by NHGRI, in conjunction with additional funding from NIGMS through the following grants: U01-HG-004610 (Group Health Cooperative/University of Washington, PI: Eric Larson); U01-HG-004608 (Marshfield Clinic, PI: Cathy McCarty); U01-HG-004599 (Mayo Clinic, PI: Chris Chute); U01-HG-004609 (Northwestern University, PI: Rex Chisholm); U01-HG-004603 (Vanderbilt University, PI: Dan Roden); also serving as the Administrative Coordinating Center, PI: Daniel Masys).

Genome-wide genotyping for the project is performed by the Broad Institute (U01-HG-004424 PI: Stacey Gabriel) or the Center for Inherited Disease Research (CIDR) at Johns Hopkins University (U01-HG-004438 PI: David Valle).

Members of the Expert Scientific Panel include: Gerardo Heiss, PhD (University of North Carolina), Stan Huff, PhD (University of Utah) Howard McLeod, PhD (University of North Carolina, Chair), Jeff Murray, PhD (University of Iowa), and Lisa Parker, PhD (University of Pittsburgh).

We would also like to acknowledge all of the eMERGE participants.