The eMERGE Network: A National Consortium of Electronic Health Record-Linked Biobanks Furthering Large-Scale Genomic Research

W.A. Wolf[1], R.L. Chisholm[1], C.G. Chute[2], G. Jarvik[3], E. Larson[3], D.R. Masys[4], C.A. McCarty[5], D.M. Roden[4], J.P. Struewing[6]

Northwestern University, Chicago, IL[1], Mayo Clinic, Rochester, MN[2], Group Health Cooperative/University of Washington, Seattle, WA[3], Vanderbilt University, Nashville, TN[4], Marshfield Clinic Research Foundation, Marshfield, WI[5], National Human Genome Research Institute, Bethesda, MD[6]

Abstract

Medical research institutions are developing biorepositories of genomic DNA coupled to electronic medical record (EMR) information generated by routine clinical care. The goal of the Electronic Medical Records and Genomics (eMERGE) Network (www.gwas.net), organized by NHGRI in late 2007, is to investigate how such resources can be leveraged for genome and informatics science, with extensive ELSI input. Each of the five participating sites will use natural language processing or other tools to identify cases with defined phenotypes and controls for genome wide analyses in ~18,000 subjects, and will share data with the scientific community through NIH's dbGaP. The target phenotypes include type 2 diabetes, cataracts, peripheral arterial disease, normal QRS duration, and late life dementias (Alzheimer's disease). Initial studies underway include testing computational algorithms to identify subjects meeting phenotype criteria, comparing these tools across diverse EMR systems, assessing the potential for combining cases or controls from different sites, and initiation of mechanisms to optimize community consultation. The goal of these studies is to contribute to our understanding of disease and develop recommendations to improve the utility of EMRs for research, setting the stage to integrate genomic and EMR data to achieve the vision and inform best practices for personalized medicine.

Introduction

The eMERGE Network brings together researchers with a wide range of expertise in ethics, genomics, statistics, informatics, and clinical medicine from leading medical research institutions across the country to assess whether EMR data can provide suitable phenotype and environmental exposure data to analyze the genetic and environmental factors contributing to disease susceptibility and therapeutic outcomes. Each center participating in the consortium, organized by the NHGRI with additional funding from the National Institute of General Medical Sciences, has proposed studying the relationship between genetic variation and a common human trait, using genome-wide association studies (GWAS). In addition, the consortium will include a focus on social and ethical issues such as privacy, confidentiality, and interactions with the broader community.

eMERGE Goals

The overall goal of the consortium is to test the ability to leverage EMRs and biorepositories for the conduct of genomic research. Through the collaborative efforts of the consortium and NHGRI, this research will facilitate development of policies and procedures to realize the untapped potential of EMR-linked biorepositories for GWAS to improve the understanding, prevention, and treatment of chronic diseases. To achieve these goals, the consortium will:

1) Evaluate the validity and utility of phenotypic and exposure data from EMRs for use in GWAS

2) Develop and validate electronic algorithms for primary and secondary phenotypes

3) Conduct association studies of genome-wide data with EMR-derived phenotypes and deposit data in dbGaP

4) Assess adequacy of existing consent for genomic technologies, and for sharing data

5) Develop best practices for GWAS in the areas of electronic phenotyping, genomics and analytics, and ELSI topics

Network Structure

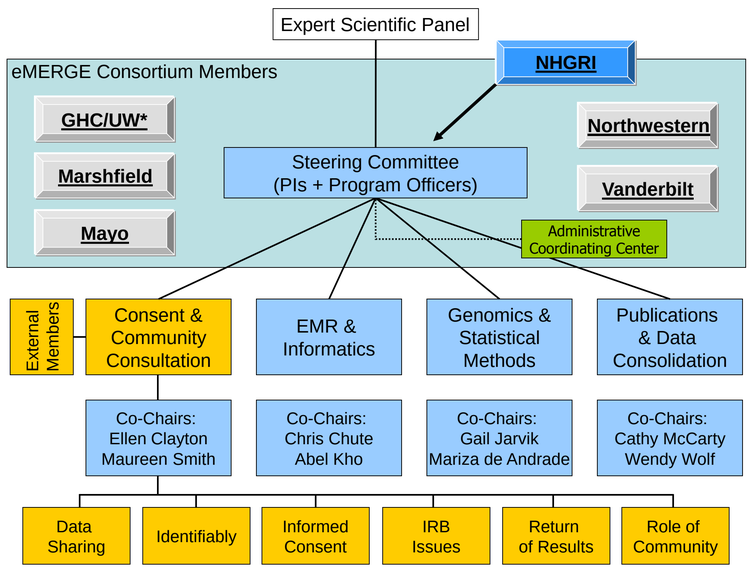
The Steering Committee is the governing body for the consortium and is composed of the Principal Investigators from each institution and the NIH Project Scientist. The main goal of the committee is to identify the best approaches for conducting genome-wide research in biorepositories. Rex Chisholm, PhD chairs the Steering Committee.

The Expert Scientific Panel (ESP) evaluates the progress of the program and provides recommendations to the National Advisory Council for Human Genome Research about the progress and scientific direction of all components of the program. The ESP is composed of five senior scientists with expertise in key disciplines for the network:

- Chair: Howard McLeod, PharmD, Institute for Pharmacogenomics and Individualized Therapy, University of North Carolina
- Gerardo Heiss, MD, School of Public Health, University of North Carolina
- Stan Huff, MD, Intermountain Healthcare, University of Utah
- Jeff Murray, MD, Department of Pediatrics, University of Iowa
- Lisa S. Parker, PhD, Center for Bioethics and Health Law, University of Pittsburgh

Genome-wide genotyping for the project will be performed by the Broad Institute (PI Stacey Gabriel) or Johns Hopkins University (PI David Valle).

Vanderbilt University serves as the Administrative Coordinating Center (ACC), which facilitates consortium planning, communication, and data sharing. The ACC will also provide a data privacy consultation service for institutional datasets prior to dbGaP data submission.



*Group Health Cooperative includes a partnership with University of Washington and the Fred Hutchinson Cancer Research Center in Seattle.

Biorepository & Phenotype Summary

	Biorepository Details						
Institution	**Biorepository**	**Current/ Expected Size**	**Age Range/ Mean Age**	**% Female**	**EMR Summary**	**Supplemental Survey Data**	**Primary Phenotypes for GWAS**
Group Health Cooperative (Seattle, WA)	**GHC Biobank**: Alzheimer's Disease Patient Registry & Adult Changes in Thought Study	**~4,000** >96% Caucasian	40 – 90+ (over 65)	58.2	• 20+ yrs pharmacy data • 15+ yrs ICD9 data • Comprehensive EMR since 2002	Y	**Alzheimer's Disease (AD) & Dementia** • Research quality diagnoses of all-cause dementia and AD based on complete physical, neuropsychological, laboratory, and neuroimaging studies, with results reviewed at consensus clinical conferences
Marshfield Clinic (Marshfield, WI)	**Personalized Medicine Research Project**: Marshfield Clinic, an integrated regional health system	**20,000/ 21,000** 98% Caucasian	18 – 90+ (48)	57.1	• Comprehensive EMR since 1990 • 75% participants have 20+ yrs medical history	Y	**Cataracts** • Case criteria includes senile cataract diagnosis and surgery; age & chronic steroid use considered • Identifying cataract type by NLP & image recognition • Control criteria considers diagnosis, surgery, recent eye exam, and age
Mayo Clinic (Rochester, MN)	**Case-control:** Mayo Clinic Non-Invasive Vascular Laboratory & Exercise Stress Testing Lab	**1500 cases & 2000 controls** >96% Caucasian	21 - 80 67.4	36.0 (cases) 39.0 (controls)	• Comprehensive EMR since 1995 • 40 yr history of data extraction	Y	**Peripheral Arterial Disease (PAD)** • Cases are patients identified from the non-invasive vascular lab with ankle-brachial index <0.9 • Controls are patients from the exercise stress lab who have a normal stress test and normal ABI
Northwestern University (Chicago, IL)	**NUgene Project**: Northwestern affiliated hospitals and outpatient clinics	**8,500/ 20,000** 12% AA 8% Hispanic	18 – 90+ (50)	58.7	• 20+ yrs ICD9 data • Comprehensive EMR since 2000	Y	**Type 2 Diabetes** • Susceptibility loci already identified – proof of principle • Caucasian GWAS and pooled African American GWAS in collaboration with Vanderbilt
Vanderbilt University (Nashville, TN)	**BioVU**: Vanderbilt Clinic, diverse outpatient clinic	**50,000/ 200,000** 11% AA	18 – 90+ (52)	57.0	• 35+ yrs medical history data • Comprehensive EMR since 2000	N	**QRS duration** • Potential surrogate for arrhythmia susceptibility • Examine subjects at extremes of normal range and validate associations in prospective clinical trial subjects

eMERGE Biorepositories

Group Health - Cohort from 2 study samples

- Alzheimer's Disease (AD) Patient Registry (ADPR) – new cases of AD from GH
- Adult Changes in Thought (ACT) – bulk of GH cohort
  - Prospective study of GH members >65 without dementia at baseline
  - Study visits include assessments of cognitive functioning; those with lower performance are further evaluated to detect cases of dementia and AD using research criteria for diagnosis

Marshfield – Population-based biobank

- Personalized Medicine Research Project (PMRP)
  - Participants recruited from Marshfield Epidemiologic Study Area in central WI
  - 96% of healthcare events captured
  - 55% belong to health plan

Mayo – Case and control study samples

- Peripheral arterial disease (PAD) cases
  - Patients referred to Non-invasive Vascular Lab for lower extremity arterial evaluation who have an ABI ≤0.9
- PAD controls
  - Subjects referred for screening exercise stress ECG in the Cardiovascular Health Clinic who have no ischemia on the test and have a normal ABI

Northwestern – Hospital/clinic-based biobank

- NUgene Project
  - Participants recruited from NU affiliated hospitals and outpatient clinics
  - Population representative of medical center patients and metropolitan Chicago
  - Over 50% participants enrolled through primary care clinics

Vanderbilt – Hospital/clinic-based biobank

- BioVU
  - Repository combining DNA extracted from discarded blood samples with de-identified clinical information from the Vanderbilt University Medical Center electronic medical record system
  - Matched genotype/phenotype data are accumulating rapidly: 500-1000 samples/week, over 40,000 samples to date

Phenotypes for GWAS

The successful identification of genetic variants influencing disease susceptibility depends on a number of different factors, but few are as important as the phenotype definition and accurate identification of cases and controls. Each eMERGE site has proposed its own phenotype for downstream GWAS in ~3,500 subjects (see above table). The sites are developing and validating electronic phenotyping algorithms, as well as evaluating the validity and utility of EMR data for genomic research. Collectively, eMERGE members will develop and disseminate best practices for electronic phenotyping using EMR data to the broader scientific community.

Network-wide Phenotype

In addition to individual site phenotypes, the consortium is working to define a phenotype that would require a common definition applied to all 5 sites, with subjects drawn from each for additional genotyping.

Network-wide Analysis

One of the key goals of the consortium is to assess how effectively EMR data can be pooled across sites for network-wide analysis. Using data from samples genotyped for individual site phenotypes, the network plans to perform network-wide analysis of these phenotypes. Work to identify and define these phenotypes is currently underway.

Data Harmonization & dbGaP Deposit

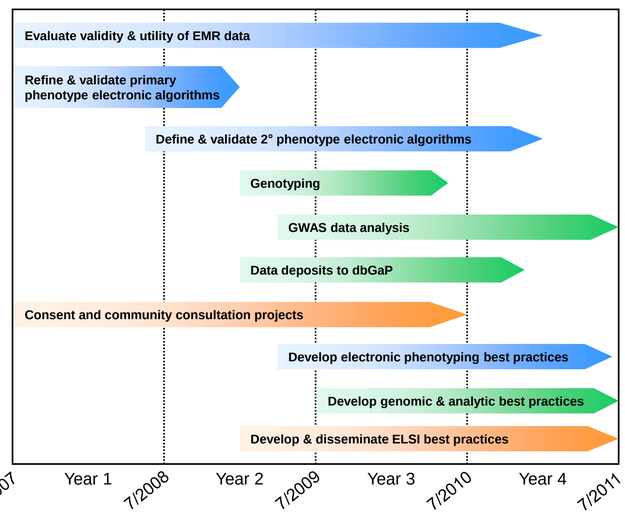
EMR-linked biorepositories include an incredible breadth of data that could be deposited into dbGaP and leveraged for future studies. In concert with NHGRI, the consortium has prioritized classes of data for submission. The first round data submission will include variables used to establish the primary phenotypes, in addition to demographic data and limited covariates. The network is identifying additional data elements of common interest for deposit to dbGaP, including other phenotypes, medications, lab results, reports and vital signs. The consortium is working to harmonize definitions for these phenotypes and variables.

Consent & Community Consultation

Biorepositories for genetic research pose a variety of ethical and procedural challenges, particularly as technology and policy evolve. As part of the eMERGE Network, the sites are addressing issues such as the adequacy of consent, data sharing, and return of individual genetic results in the context of each institution's biobank, with the ultimate goal of developing best practices for biobank operations.

Site	Overview Projects
Group Health Cooperative	• Assess beliefs, attitudes, opinions & experiences of consumers (research subjects and non-subjects) about: 1) informed consent, 2) data sharing & 3) return of results • Convene consensus panel to develop best practices for informed consent and other policies
Marshfield Clinic	• Review consent form with respect to "non-identifiable" data and other issues • Review non-disclosure of personal genetic results • Develop and test computer based program for consent process
Mayo Clinic	• Develop community-based best practices for biobanks using Deliberative Democracy methods • Study research participant understanding and concerns, refine consent tools & procedures, and study effectiveness of recommended approaches
Northwestern University	• Study stakeholders' views of informed consent & data sharing for GWAS, including the NIH GWAS Policy though 1) focus groups of NUgene participants & the public, and 2) survey of IRB professionals • Convene consensus meetings with key professional stakeholders to develop best practices for ethical conduct of GWAS
Vanderbilt University	• Assess ethical, scientific and societal advantages/disadvantages of BioVU model & determine best practices for oversight, community involvement and communication as the resource grows • Assess understanding & acceptability of model with patients, community members, and faculty/staff using various methods

eMERGE Timeline



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); U01-HG-004608 (Marshfield Clinic); U01-HG-04599 (Mayo Clinic); U01-HG-004609 (Northwestern University); U01-HG-04603 (Vanderbilt University, also serving as the Administrative Coordinating Center). We would like to thank the ACC for their assistance with the poster. We also want to thank external members of the Consent & Community Consultation workgroup for their contributions to the consortium.

The eMERGE Network — electronic Medical Records & Genomics — A consortium of biorepositories linked to electronic medical records data for conducting genomic studies