Use of Electronic Medical Records for Genomic Research – Preliminary Results and Lessons from the eMERGE Network

Joshua C. Denny, MD, MS[a], Abel Kho MD, MS[b], Jyoti Pathak, PhD[c], David Carrell, PhD[d], Peggy Peissig, MBA[e], Dan Masys, MD[a]

[a]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[b]Biomedical Informatics Center, Northwestern University, Chicago, IL. USA
[c]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA
[d]Group Health Research Institute, Seattle, WA, USA
[e]Biomedical Informatics Research Center, Marshfield Clinic, Marshfield, WI, USA

Abstract

Wide-spread adoption of electronic medical records (EMRs) containing rich amounts of longitudinal clinical data and the formation of EMR-linked biobanks represent an opportunity to rapidly expand sample sizes and efficiently phenotype subjects for genomic studies. The Electronic Medical Records & Genomics (eMERGE) Network consists of five leading institutions involved in EMR phenotyping with linked DNA biobanks. The goal of eMERGE is to conduct genome-wide association studies (GWAS) in approximately 19,000 individuals using EMR-derived phenotypes and biorepository-derived genome-wide genotypes. These institutions include Group Health Research Institute, Marshfield Clinic, Mayo Clinic, Northwestern University and Vanderbilt University. Each site has used electronic algorithms to identify both site-specific phenotypes and network-wide phenotypes (applied at all five sites) for genomic analysis. The panel will present data from site-specific and network-wide studies illustrating the strengths and limitations of EMRs for genomic studies. Panelists will discuss and compare approaches to developing phenotypic electronic algorithms, challenges in implementing algorithms at each site, and approaches to validation of the algorithms and genomic results. Panelists will also present results from initial studies into performing phenome-wide analyses for genetic associations. Finally, the panel will present lessons learned from these efforts.

Keywords: Electronic Medical Records, genomics, natural language processing, phenotyping

Introduction

The mapping of the human genome has enabled new exploration of how genetic variations contribute to health and disease. Genomic research has been very successful with the results of this work both shedding light on how genetic variants influence susceptibility to common, chronic diseases but also playing an instrumental role in the discovery of new biologic pathways and drug targets – combined with EMRs this is a pivotal first step towards the implementation of personalized medicine in clinical care.

The adoption of EMRs for clinical care has been one of the most important technological advances in healthcare. EMRs represent a robust source of clinical and environmental data that are being increasingly mined for use in clinical research although its utility for genomic research is still being explored. Advances in genetics research are driving a need for ever larger sample sizes and a possible solution has been the establishment of biorepositories linked to EMRs.

eMERGE: the Electronic Medical Records & Genomics Network (eMERGE) is consortium formed by the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences[1]. The goal of the Network is to develop, disseminate and apply research methods that combine EMRs with DNA biorepositories for the conduct of large scale genome-wide association studies. These institutions have developed and validated a variety of phenotypes using the EMR. Table 1 provides a summary of each

institutions biorepository, EMR and phenotyping effort.

Topic

The panel represents data and informatics experts from the eMERGE consortium who will share their experiences in validating the use of EMR for genetic research. Specifically each panelist will present the performances and challenges of implementing locally and externally developed electronic phenotype algorithms, initial results from GWA studies, and their relevance to existing genetic knowledge. Specific challenges to identification of adequate populations for genomic research will also be addressed, including determining ancestry/ethnicity information, achieving adequate sample size, and validating phenotype algorithms across the network. The panel will also include structured discussions on the following topics:

- GWAS and other genomic studies results from each site;
- Comparison of EMR phenotyping and validation approaches and illustrating an approach to cross-institutional algorithm development and implementation, including discussion of approaches to mining similar information from fundamentally different data representations (eg., medication data), and adapting externally-developed algorithms to local data and systems;
- Comparison of using formal chart abstraction method to physician review to validate phenotype algorithms;
- Discussion of experiences with cross-institution implementation of phenotype algorithms and local adaptations required for the network-wide phenotypes;
- Comparison of network-phenotype performance at each site;
- Representation of phenotype data using consolidated health informatics (CHI) standards.

Table 1: eMERGE Biorepository Characteristics

Institution	Biorepository Overview	Recuritment Model	Repository Size	EMR Summary	Phenotype	Phenotyping Methods*
Group Health (Seattle, WA)	Alzheimer's Disease Patient Registry and Adult Changes in Thought Study	Disease specific	**~2,800** Includes controls >96% Caucasian	20+ years pharmacy data 15+ years radiology and pathology reports 15+ years ICD9 data Comprehensive EMR since 2004	**1º: Alzheimer's Disease & Dementia, 2º: White Blood Cell Counts**	Coded data extraction, NLP, Manual chart review, Computer assisted chart abstaction
Marshfield Clinic (Marshfield, WI)	**Personalized Medicine Research Project** Marshfield Clinic, an integrated regional health system	Geographic	**~21,000** 98% Caucasian	Comprehensive EMR since 1985 75% participants have 20+ years medical history	**1º: Cataracts & Low HDL, 2º: Diabetic Retinopathy**	Coded data extration, NLP, Intelligent Character Recogonition
Mayo Clinic (Rochester, MN)	Mayo Clinic Non-Invasisve Vascular Laboratory & Exercise Stress Testing Lab	Disease specific	**~3,300** Includes controls >96% Caucasian	Comprehensive EMR since 1995 40 years of history of data extraction	**1º: Peripheral Arterial Disease (PAD), 2º: Red Blood Cell Counts**	Coded data extraction, NLP
Northwestern University (Chicago, IL)	**Nugene Project:** Northwestern affiliated hospitals and outpatient clinics	Clinic & Hospital	**~10,000** 12% AA 9% Hispanic	20+ years ICD9 data Comprehensive EMR since 2000	**1º: Type 2 Diabetes, 2º: Lipids & Height**	Coded data extraction, NLP, Mining text using regular
Vanderbilt University (Nashville, TN)	**BioVU:** Vanderbilt Clinic, diverse outpatient clinics	Outpatient lab draws	**~80,000/200,000** 11% AA	35+ years medical history data Comprehensive EMR since 2000	**1º: QRS & PR Duration, Other: PheWAS**	Coded data extraction, NLP

*NLP=Natural Language Processing. Coded data extraction refers to retrieving data in the EMR ath has been coded using established standards (e.g., ICD9, CPT).

Panel Participants

Dan Masys – Vanderbilt University, Moderator
Dr. Masys is the Chair of the Department of Biomedical Informatics at Vanderbilt University and also the PI of the Coordinating Center (CC) for the eMERGE network. The CC is ultimately responsible for data coordination and submission to dbGaP, and has worked with other sites to develop data standards, common data dictionaries, and submission protocols. In addition, to ensure network-wide standardization, all initial genetic data is being quality-checked through the CC. Finally, the CC coordinates the network-wide phenotypes.

Joshua Denny – Vanderbilt University
Dr. Denny is an Assistant Professor in the Department of Biomedical Informatics at Vanderbilt University. Vanderbilt uses an opt-out model biobank associated with a de-identified "synthetic derivative representation"[2] of the EMR. Vanderbilt's GWAS investigated genomic variants associated with cardiac conduction in both Caucasians and African-American populations (the latter derived in cooperation with Northwestern). Individuals were selected amongst those without cardiac disease, and analysis with and without evidence of possibly-interfering medications. Our results identified genomic variants associated with atrioventricular conduction, validating use of EMR-based biobanks for genomic studies. Dr. Denny will highlight the coded data retrieval and NLP efforts used to identify and validate the phenotype. Additionally, linkage to the EMR presents the possibility to perform phenome-wide association analyses (PheWAS), scanning the EMR phenome for genetic associations with many diseases not anticipated in the initial GWAS. Dr. Denny will presents results from these initial PheWAS studies.

Abel Kho – Northwestern University
Dr. Kho is an Associate Professor of Medicine in General Internal Medicine at Northwestern University and is Co-Chair of the eMERGE Informatics Working Group. Northwestern's GWAS assessed genomic variants associated with Type 2 Diabetes (T2D) in Caucasian and African American populations derived from the Northwestern and Vanderbilt biorepositories. Results were compared with associations from traditionally defined T2D case and control cohorts. Dr. Kho will address implementation and validation of T2D and lipid algorithms at Northwestern and across other eMERGE sites.

Jyotishman Pathak – Mayo Clinic

Dr. Pathak is an Assistant Professor in Medical Informatics at the Mayo Clinic College of Medicine. He is leading eMERGE's goal for standardized and consistent representation of phenotype data using common data elements and Consolidated Health Informatics (CHI) standards. In addition, he is investigating extraction of medication data from clinical notes and their classification using standard drug vocabularies.

David Carrell – Group Health Research Institute
Dr. Carrell is Multi-Institutional Research Consult at Group Health Research Institute (GHRI) and has expertise in adoption and use of open source natural language processing (NLP) systems in applied research settings. GHRI's primary phenotype is dementia, defined by gold-standard clinical diagnosis as well as an EMR-based algorithm developed at GHRI and applied at other eMERGE sites. GWA studies assessed genomic variants associated with dementia cases based on clinical and EMR definitions. GHRI also implemented phenotype definitions developed at other sites, including cataract type (Marshfield), cardiac conduction (Vanderbilt), and peripheral arterial disease (Mayo Clinic). Dr. Carrell will address the process and challenges of replicating externally-developed EMR-based phenotype algorithms, particularly as they involve NLP-based components.

Peggy Peissig – Marshfield Clinic Research Foundation

Ms. Peissig is the Associate Director of the Biomedical Informatics Research Center, at the Marshfield Clinic Research Foundation. Ms. Peissig has over 18 years of phenotyping experience using Marshfield Clinic's Cattail's EMR. Marshfield's GWAS study uses the Personalized Medicine Research Project[3] cohort and will elucidate combinations of genetic markers that predispose subjects to the development of cataracts and quantify the degree to which low HDL levels contribute to the onset and progression of cataracts. Marshfield used a mixed-mode approach to phenotyping which included extracting "coded" EMR data, natural language processing and intelligent character recognition on unstructured EMR clinical documents and images. Ms. Peissig will present the cataract, low HDL and diabetic retinopathy phenotyping efforts and highlight implementation and validation challenges of these phenotypes across the network.

References

1. Genome-Wide Studies in Biorepositories with Electronic Medical Record Data, RFA-HG-07-005.
2. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep;84(3):362-9.
3. McCarty C, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Med 2005; 2:49-79.

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)

Affirmation
The first author affirms that all panel participants have agreed to participate and have contributed to the preparation of this document.

Address for correspondence:
Joshua Denny, MD, MS
Vanderbilt University Medical Center
2209 Garland Avenue, Rm 442
Nashville, TN 37232
Ph. 615-936-5034
Josh.denny@vanderbilt.edu