Panel: Phenotyping Using the Electronic Medical Record – Experiences from the eMERGE Network

Peggy L. Peissig, MBA[a], Joshua C. Denny, MD, MS[b], Christopher G. Chute, MD, DrPH[c], Abel Kho MD, MS[d], David Carrell, PhD[e]

[a]Biomedical Informatics Research Center, Marshfield Clinic, Marshfield, WI, USA
[b]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
[c]Health Sciences Research, Mayo Clinic, Rochester, MN
[d]General Internal Medicine, Northwestern University, Chicago, IL
[e]Information Technology, Group Health Center for Health Studies, Seattle, WA

Abstract

With the wide spread adoption of Electronic Medical Records, large amounts of clinical data are becoming available for research. The Electronic Medical Records & Genomics Network (eMERGE) consists of five leading institutions involved with electronic medical record (EMR) phenotyping and biobanking. Each eMERGE institution has a different population and approach to phenotyping. These institutions include Group Health, Marshfield Clinic, Mayo Clinic, Northwestern University and Vanderbilt University. Each site has used electronic algorithms to identify both site-specific phenotypes as well as a network-wide phenotype (applied at all five sites) for genome analysis. Specifically, the panel will compare and contrast the biorepository populations, available EMR data sources, phenotyping approaches, algorithms and validation results. The panel will also discuss the limitations of each approach and present the lessons learned from this effort. By assembling representatives from each of these institutions into a single "expert" panel, others will be able to evaluate multiple approaches and identify trade-offs associated with each of the EMR phenotyping strategies with the eventual goal of developing "best practices" for their environment.

Keywords: Electronic Medical Record Phenotyping

Introduction

With the completion of the Human Genome Project in 2004 and the recent advances in high throughput genotyping, large amounts of genetic data have become available for Genome-wide Association Studies (GWAS). These studies help us learn how genetic alterations influence human disease and understand the variability in response to diagnostic or therapeutic interventions.

As in all research that attempts to identify and quantify relationships between exposures and outcomes, rigorous characterization of the phenotype is an essential component of any GWAS. Careful phenotyping is both critical to the eventual results discovered in a study and a great challenge due to the tremendous variety of phenotyping options that can be employed with the data. With the implementation of Electronic Medical Records (EMRs), large amounts of clinical and environmental data are available for the development and validation of phenotyping algorithms. Because EMR data is rarely collected with research in mind, deriving valid phenotypic information from EMR data is frequently less than straightforward. The overall utility and reliability of EMR data for phenotyping has yet to be validated for use in large GWA studies.

eMERGE: the Electronic Medical Records & Genomics Network (eMERGE) is consortium formed in response to a Request For Application (RFA) by the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences[1]. The goal of the RFA was 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 phenotyping experts from the eMERGE consortium who will share their experiences in developing EMR phenotypes. Specifically each panelist will present an overview of their institutions biorepository population, EMR data sources, phenotyping approach, and validation

results. This will be followed by a structured discussion by all panelists on the following topics:

- Comparison of the biorepositories focusing on populations and available EMR data sources;
- Comparison of EMR phenotyping an validation approaches highlighting limitations including EMR-derived challenges to accurate phenotype identification such as purposeful and accidental billing errors;
- Discussion of the roles of different methodologies (e.g., NLP, billing codes, lab values) in defining high-quality phenotypes;
- Discussion of experiences with cross-institution phenotyping and local adaptations required for the network-wide phenotype;

Table 1: eMERGE Biorepository Characteristics

Institution	Biorepository Overview	Recruitment Model	Repository Size	EMR Summary	Phenotype	Phenotyping Methods*
Group Health (Seattle, WA)	**GHC Biobank:** Alzheimer's Disease Patient Registry and Adult Changes in Thought Study	Disease specific	**~4000** >96% Caucasian	20+ years pharmacy data 15+ years ICD9 data Comprehensive EMR since 2004	**Alzheimer's Disease (AD) & Dementia**	Coded data extraction,* Mining free-text via regular expressioins, Manual chart review
Marshfield Clinic (Marshfield, WI)	**Personaized Medicine Research Project** Marshfield Clinic, an integrated regional health system	Geographic	**20,000 - 21,000** 98% Caucasian	Comprehensive EMR since 1985 75% participants have 20+ years medical history	**Cataracts and Low HDL**	Coded data extraction, NLP, Intellegent Character Recognition
Mayo Clinic (Rochester, MN)	**Case-control:** Mayo Clinic Non-Invasive Vascular Laboratory & Exercise Stress Testing Lab	Disease specific	**3,500** Includes Controls >96% Caucasian	Comprehensive EMR since 1995 40 years history of data extraction	**Peripheral Arterial Disease (PAD)**	Coded data extraction, NLP
Northwestern University (Chicago, IL)	**Nugene Project:** Northwestern affiliated hospitals and outpatient clinics	Clinic & Hospital	**8,500 / 20,000** 12% AA 8% Hispanic	20+ years ICD9 data Comprehensive EMR since 2000	**Type 2 Diabetes**	Coded data extraction, text searches
Vanderbilt University (Nashville, TN)	**BioVU:** Vanderbilt Clinic, diverse outpatient clinic	Outpatient lab draws	**50,000 / 200,000** 11% AA	35+ years medical history data Comprehensive EMR since 2000	**QRS Duration**	Coded data extraction, NLP

*NLP=Natural Language Processing. Coded data extraction refers to retrieving data in the EMR that has been coded (e.g. ICD9, Labs, CPTs, etc.)

Panel Participants

Peggy Peissig – Marshfield Clinic Research Foundation, Moderator

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 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. Ms. Peissig will present the cataract and low HDL cholesterol phenotyping effort when using the Personalized Medicine Research Project cohort[3]. 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.

Joshua Denny – Vanderbilt University

Dr. Denny is an Assistant Professor in the Department of Biomedical Informatics at Vanderbilt University School of Medicine. Vanderbilt uses a non-consented biobank associated with a de-identified "synthetic derivative model"[2] of the EMR. Vanderbilt's GWAS has defined QRS duration as a phenotype with the goal of detecting genomic variants associated with QRS durations at the extremes of the normal range in both Caucasians and African-American populations (the latter derived in cooperation with Northwestern). Dr. Denny will highlight the coded data retrieval and NLP efforts used to identify and validate the QRS phenotype.

Christopher Chute – Mayo Clinic

Dr. Chute is a Professor of Biomedical Informatics in Health Sciences Research at Mayo Clinic. He is also

the Principal Investigator of Mayo's eMERGE genome-wide association study and Co-chair of the Informatics Section of eMERGE. Mayo Clinic has integrated resources from the electronic medical record, Enterprise Data Trust (warehouse), and the vascular laboratory to characterize patients with Peripheral Arterial Disease for the NHGRI funded genome-wide associations. Dr. Chute will present an overview of Mayo's effort involving a disease specific population, a comprehensive medical record and over 40 years of data abstraction.

Abel Kho – Northwestern University
Dr. Kho is an Associate Professor of Medicine in General Internal Medicine at Northwestern University. Dr. Kho is also Co-chairmen of the Informatics Section of the eMERGE consortium. Northwestern University has developed a Type 2 diabetic algorithm and plans to perform a GWAS on Type 2 diabetes in African-Americans using a population derived from the Northwestern and Vanderbilt biorepositories. A second GWAS, investigating genetic variants associated with asthma, will also be performed in the Northwestern population.

Dr. David Carrell – Group Health
Dr. Carrell is a programmer in Information Technology for the Group Health Center for Health Studies. Group Health (GH) with the University of Washington have phenotyped both research quality Alzheimer's disease and a broader EMR-base definition of dementia from the EMR-linked GH biorepositories to perform a series of GWAS. Dr. Carrell has led all facets of the phenotyping effort including coded data extraction and regular expression free-text mining to identify dementia. Dr. Carrel will present Group Health's phenotyping approach and validation results for dementia and Alzheimer's disease.

Summary

EMRs provide inexpensive and powerful resources for reuse of clinical data for research experiments. This panel session brings together five leaders in the EMR phenotyping domain to present their phenotyping efforts. A comparison of institution populations and EMR data resources will spotlight the challenges associated with developing phenotypes from EMRs and provide a comparison of phenotyping methods.

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:
Peggy L. Peissig
Marshfield Clinic Research Foundation
1000 N. Oak Ave.
Marshfield, WI, 54449
Ph. 715-221-8322
peissig.peggy@marshfieldclinic.org

record, Enterprise Data Trust (warehouse), and the vascular laboratory to characterize patients with Peripheral Arterial Disease for the NHGRI funded genome-wide associations. Dr. Chute will present an overview of Mayo's effort involving a disease specific population, a comprehensive medical record and over 40 years of data abstraction.

Abel Kho – Northwestern University
Dr. Kho is an Associate Professor of Medicine in General Internal Medicine at Northwestern

The eMERGE network has also developed a cross-center phenotype for autoimmune hypothyroidism that will be applied in all biorepositories to identify cases and controls for a GWAS. population. Dr. Kho will present the phenotyping effort associated with defining and validating Type 2 diabetics and asthma phentoypes.

Dr. David Carrell – Group Health
Dr. Carrell is a programmer in Information Technology for the Group Health Center for Health Studies. Group Health (GH) with the University of Washington have phenotyped both research quality Alzheimer's disease and a broader EMR-base definition of dementia from the EMR-linked GH biorepositories to perform a series of GWAS. Dr. Carrell has lead all facets of the phenotyping effort including coded data extraction and regular expression free-text mining to identify dementia. Dr. Carrel will present Group Health's phenotyping approach and validation results for dementia and Alzheimer's disease.

Summary

EMRs provide inexpensive and powerful resources for reuse of clinical data for research experiments. This panel session brings together five leaders in the EMR phenotyping domain to present their phenotyping efforts. A comparison of institution populations and EMR data resources will spotlight the challenges associated with developing phenotypes and provide a comparison of phenotyping methods.

References

4. Genome-Wide Studies in Biorepositories with Electronic Medical Record Data, RFA-HG-07-005.

Univeristy. Dr. Kho is also co-chairmen of the Informatics Section of the eMERGE consortium. Northwestern University has developed a Type 2 diabetic algorithm and plans to perform a GWAS on type 2 diabetes in African-Americans using a population derived from the Northwestern and Vanderbilt biorepositories. A second GWAS, investigating genetic variants associated with asthma, will also be performed in the Northwestern

5. 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.
6. 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

Justin Starren
Marshfield Clinic Research Foundation
1000 N. Oak Ave.
Marshfield, WI, 54449
Ph. 715-221-7299
Starren.justin@mcrf.mfldclin.edu