Major US Genomic Medicine Programs: NHGRI’s Electronic Medical Records and Genomics (eMERGE) Network

Dan Roden
Member, National Advisory Council For Human Genome Research Genomic Medicine Working Group
Creating electronic medical records can enable

- Improved care of individual patients
- Identification of specific subsets of patients
- Discovery of new genotype-phenotype and phenotype-genotype associations
- Implementation of Genomic Medicine
• Genomic predictors of disease susceptibility and drug response
• Engaging the Electronic Medical Record (EMR)
Implementing genomic medicine
eMERGE-I goal: to assess utility of DNA collections integrated with electronic medical records (EMRs) as resources for genome science

- Each site identified a phenotype of interest in ~3,000 subjects and conducted a genome-wide association study (GWAS)
- To what extent can identifiers be stripped from EMRs and research utility retained?
- Assess consent for genomic technologies & data sharing
- Develop and promulgate best practices for phenotyping and genomics in EMRs
### eMERGE-I Phenotypes

<table>
<thead>
<tr>
<th>Phenotype</th>
<th># Genotype</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataract</td>
<td>2642</td>
<td>1322</td>
<td></td>
</tr>
<tr>
<td>Dementia</td>
<td>1241</td>
<td>2043</td>
<td></td>
</tr>
<tr>
<td>PAD</td>
<td>1641</td>
<td>1604</td>
<td></td>
</tr>
<tr>
<td>QRS</td>
<td>3034</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>T2 Diabetes</td>
<td>2706</td>
<td>1496</td>
<td></td>
</tr>
</tbody>
</table>

### Data Sharing Memorandum of Understanding
- Each site has final authority regarding their data
- How data may be shared
- Privacy and Confidentiality agreements
- Limitations of Use
## Data Sharing Memorandum of Understanding

- Each site has final authority regarding their data
- How data may be shared
- Privacy and Confidentiality agreements
- Limitations of Use

### eMERGE-I phenotypes

<table>
<thead>
<tr>
<th>Phenotype</th>
<th># Genotype</th>
<th>Data Only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case</td>
<td>Control</td>
</tr>
<tr>
<td>Cataract</td>
<td>2642</td>
<td>1322</td>
</tr>
<tr>
<td>Dementia</td>
<td>1241</td>
<td>2043</td>
</tr>
<tr>
<td>PAD</td>
<td>1641</td>
<td>1604</td>
</tr>
<tr>
<td>QRS</td>
<td>3034</td>
<td>-</td>
</tr>
<tr>
<td>T2 Diabetes</td>
<td>2706</td>
<td>1496</td>
</tr>
</tbody>
</table>
## eMERGE-I phenotypes

<table>
<thead>
<tr>
<th>Phenotype</th>
<th># Genotype</th>
<th>Data Only</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case</td>
<td>Control</td>
<td>Case</td>
<td>Control</td>
</tr>
<tr>
<td>Cataract</td>
<td>2642</td>
<td>1322</td>
<td>1386</td>
<td>1360</td>
</tr>
<tr>
<td>Dementia</td>
<td>1241</td>
<td>2043</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td>PAD</td>
<td>1641</td>
<td>1604</td>
<td>1010</td>
<td>8743</td>
</tr>
<tr>
<td>QRS</td>
<td>3034</td>
<td>-</td>
<td>1019</td>
<td>-</td>
</tr>
<tr>
<td>T2 Diabetes</td>
<td>2706</td>
<td>1496</td>
<td>1101</td>
<td>912</td>
</tr>
<tr>
<td>Platelet indices</td>
<td></td>
<td></td>
<td>13,582</td>
<td></td>
</tr>
<tr>
<td>Red cell indices</td>
<td></td>
<td></td>
<td>16,915</td>
<td></td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td></td>
<td></td>
<td>1306</td>
<td>5013</td>
</tr>
</tbody>
</table>
Approach to electronic phenotyping

Identify phenotype of interest → Case & control algorithm development and refinement → Manual review; assess precision → Deploy at site 1 → Validate at other sites → Genetic association tests; replicate

PPV ≥ 95%

PPV < 95%

Table 1. Evaluation of Primary Hypothyroidism Algorithm at the Five eMERGE Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Primary Phenotype</th>
<th>Total Genotyped Subjects</th>
<th>Primary Hypothyroidism</th>
<th>Cases</th>
<th>Controls</th>
<th>Case PPV (%)</th>
<th>Control PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health</td>
<td>dementia</td>
<td>2532</td>
<td></td>
<td>397</td>
<td>1,160</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Marshfield</td>
<td>cataracts</td>
<td>4113</td>
<td></td>
<td>514</td>
<td>1,187</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>peripheral arterial disease</td>
<td>3043</td>
<td></td>
<td>233</td>
<td>1,884</td>
<td>82</td>
<td>96</td>
</tr>
<tr>
<td>Northwestern</td>
<td>type 2 diabetes</td>
<td>1217</td>
<td></td>
<td>92</td>
<td>470</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>normal cardiac conduction</td>
<td>2712</td>
<td></td>
<td>81</td>
<td>352</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>All sites</td>
<td></td>
<td>13,617</td>
<td></td>
<td>1317</td>
<td>5053</td>
<td>92.4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>98.5&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Genotype counts represent all subjects who were found by the hypothyroidism algorithms at each site and who were genotyped. Counts are limited to those classified as “white” in the electronic medical record of each site. PPV = positive predictive value.

<sup>a</sup> Average weighted for number of samples contributed to the total.

Denny et al., 2011
An eMERGE-wide phenotype analyzed with no extra genotyping: hypothyroidism

European Americans (1,306 cases and 5,013 controls)

Denny et al., 2011
The phenome-wide association study

GWAS: Target phenotype → association P value → chromosomal location

PheWAS (ΦWAS): Target genotype → association P value → diagnosis code

PheWAS requirement: A large cohort of patients with genotype data and many diagnoses
PheWAS for rs10759944 near FOXE1

\[ \text{OR}_{\text{GWAS}} = 0.74 \]
\[ \text{OR}_{\text{PheWAS}} = 0.76 \]

N=13617 subjects

Denny et al., 2011
Pleiotropy: PheWAS associations with an IRF4 SNP previously associated with hair and eye color

- All SNPs in the GWAS catalog have now been analyzed by PheWAS
- PheWAS provides a replication tool for conventional GWAS and identifies potential new genetic associations
- All data are publically available at emrphewas.org
What is the Phenotype KnowledgeBase?

The reuse of data from electronic medical records (EMRs) and other clinical data systems holds tremendous promise for improving the efficiency and effectiveness of health research. Clinical data in the EMR is a potential source of rich longitudinal data for research, and the recent government efforts to promote the use of EMRs in the clinical setting may further promote the use of such systems in the US healthcare system. As the use of EMRs expands, the demand for usable data from these systems for research has also expanded.

One such effort by the Electronic Medical Records and Genomics Network (eMERGE) has investigated whether data captured through routine clinical care using EMRs can identify disease phenotypes with sufficient positive and negative predictive values for use in genome-wide association studies (GWAS). Most EMRs captured key information (diagnoses, medications, laboratory tests) used to define phenotypes in a structured format; in addition, natural language processing has also been shown to improve case identification rates.*

PheKB is an outgrowth of that validation effort and provides a collaborative environment of building and

2011-2015: Phase II
eMERGE-II goals

- Expand the electronic phenotyping library and apply to genotyped samples
- Initiate implementation of actionable variants into the EMR
  - Site-specific projects
  - Cross network initiatives
- Define actionability, clinical utility, validity
- Advance methods for integration of genomic information into EMRs, including methods for visualization and Clinical Decision Support
- Evaluate physician and patient attitudes and educational needs
- Continued focus on consent, regulatory, privacy, and security issues; extend to clinical laboratory implementation.

2011-2015: Phase II
+ pediatric sites
Network-wide patient survey: biobanking consent

Questions:
• Do participants view specific consent to be a requirement for sharing biosamples and data for future research?
• Which biospecimen and biobanking-related research practices are likely to have the greatest impact on willingness to participate under broad consent?

Plan
• Survey 100,000 participants and patients across the eMERGE institutions to elicit a wide cross-section of patient perspectives.

Outcome
• Recommendations to inform future policy for the ethical conduct of human subject research
Network-wide return of results project: hemochromatosis

<table>
<thead>
<tr>
<th>Site</th>
<th>C282Y/C282Y</th>
<th>C828Y/H63D</th>
<th>H63D/H63D</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geisinger</td>
<td>12</td>
<td>67</td>
<td>110</td>
<td>189</td>
</tr>
<tr>
<td>GHC/Seattle</td>
<td>17</td>
<td>60</td>
<td>72</td>
<td>149</td>
</tr>
<tr>
<td>Marshfield</td>
<td>15</td>
<td>52</td>
<td>87</td>
<td>154</td>
</tr>
<tr>
<td>Mayo</td>
<td>44</td>
<td>179</td>
<td>206</td>
<td>4</td>
</tr>
<tr>
<td>Mt. Sinai</td>
<td>1</td>
<td>12</td>
<td>29</td>
<td>42</td>
</tr>
<tr>
<td>Northwestern</td>
<td>19</td>
<td>64</td>
<td>81</td>
<td>164</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>39</td>
<td>152</td>
<td>141</td>
<td>332</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>147</strong></td>
<td><strong>586</strong></td>
<td><strong>726</strong></td>
<td><strong>1459</strong></td>
</tr>
</tbody>
</table>

- Do these patients carry the clinical diagnosis?
- Do they have clinical phenotypes?
Site-specific Genomic Medicine Implementation Pilot Projects

• Developing genetic risk scores and evaluating their potential clinical impact
  • Marshfield: Age-related Macular Degeneration (7 SNPs)
  • Mayo: Coronary Artery Disease (28 SNPs)

• Genotyping specific variants and evaluating impact on physicians and patients of returning results:
  • Mount Sinai: ApoL1 variants and development of renal dysfunction in hypertensives
  • Northwestern: Impact of genotyping for HFE and FVL variants in an Internal Medicine clinic
Site-specific Genomic Medicine Implementation Pilot Projects

• **Whole genome sequencing**
  • Geisinger: WGS for undiagnosed disease in trios

• **Pharmacogenomics focus**
  • Geisinger: Preemptive genotyping for IL28B in patients with hepatitis C
  • Cincinnati Children’s/Boston Children’s: Assay CYP2D6 and provide results to parents and providers
  • Children’s Hospital of Philadelphia: Response to beta-adrenergic agonists in children with asthma
  • Vanderbilt: Multiplexed preemptive pharmacogenomic testing
eMERGE-PGRN Partnership

PGx capabilities:
• CPIC guidelines
• Resequencing platform for 84 Very Important Pharmacogenes
• CLIA & QC standards

EMR-informatics capabilities
• Privacy
• Electronic phenotyping
• Large populations
• Decision support
Identify target patients

Develop list of actionable variants

Resequence VIP genes; Identify actionable variants

Aim 1

Target: 9000 subjects

Drug-Genome pairs study

CYP2C19-Clopidogrel

VKORC1/CYP2C9-Warfarin*

SLCO1B1-Simvastatin

* BCH DGI only VKORC1/CYP2C9-Warfarin

* Geisinger and M/E/PSU also have CYP4F2-Warfarin
Identifying target patients

Using a Predictive Algorithm in Recruitment

Participants = Newly Recruited Subjects*

* CCHMC and CHOP have a hybrid approach with a new subject cohort and an existing subject cohort being reconsented.
Identify target patients

Resequence VIP genes; Identify actionable variants

Develop list of actionable variants

Aim 1

Aim 2

Actionable variants

EMR deposit
- Result display
- Decision support
- Outcomes
- Performance metrics
- Healthcare impact
Identify target patients

Resequence VIP genes; Identify actionable variants

Develop list of actionable variants

Actionable variants

EMR deposit
- Result display
- Decision support
- Outcomes
- Performance metrics
- Healthcare impact

Aim 1

Aim 2

Aim 3

- Create repository of variants of unknown significance
- Initiate studies of function and of genotype-phenotype relationships

SPhINX
A resource of the eMERGE Network

eMERGE-PGx project
Outcomes

Process outcomes
• Recruitment
• PGRN-Seq Sequencing metrics
• Comparison to Validation Genotyping
• EMR Integration and Clinical decision support
• Returned Results
• Education: clinicians, patients

Healthcare outcomes
• Statins: Myopathy, Drug Switch
• Clopidogrel: Stent or ACS event? Within 30 days?
• Warfarin: Time to steady state? Time out of range? Bleeding? Thrombosis?
• Thiopurines: Blood counts, (disease outcome), …
• Return of results project: 6 ACMG “actionable” genes
A paradox, and an opportunity…
Large numbers of patients, of diverse ancestries, are required to develop evidence to “personalize” medicine.

Current GWAS imputed set: 51,038
Anonymizing records while enabling research

- At least $k$ subjects with specified code groups ($k=2$ in this example)
- Test this scheme by setting $k=5$ and examining 192 phenotype-genotype associations in
  - 5,944 $k$-anonymized records
  - 5,944 records drawn from 104,904 $k$-anonymized records (biobank)
  - 5,944 records drawn from 1,366,786 $k$-anonymized records (entire EMR)
Anonymizing records while enabling research

- Everyone with a medical record (1.5M patients)
- Everyone in Biorepository (100K patients)
- Specific Cohort (5000 patients)

Heatherly, et al. 2013

rs2200733
Creation of the eMERGE Genomics dataset

- Creation of QC Pipeline – High throughput and high quality
- Generating a merged set across multiple genotyping platforms → Imputation
- eMERGE-I (5 sites)
  - 2 platforms: Illumina 660 & 1M
- eMERGE-II (10 sites)
EMR-linked biobanks in eMERGE-II

<table>
<thead>
<tr>
<th>Site</th>
<th>#of subjects with DNA samples linked to EMRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health Seattle</td>
<td>7,000</td>
</tr>
<tr>
<td>Marshfield</td>
<td>20,000</td>
</tr>
<tr>
<td>Mayo</td>
<td>19,000</td>
</tr>
<tr>
<td>Northwestern</td>
<td>11,000</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>175,000</td>
</tr>
<tr>
<td>Geisinger</td>
<td>22,000</td>
</tr>
<tr>
<td>Mt. Sinai</td>
<td>22,000</td>
</tr>
<tr>
<td>CHOP</td>
<td>60,000</td>
</tr>
<tr>
<td>Cincinnati/Boston</td>
<td>10,000</td>
</tr>
<tr>
<td>TOTAL</td>
<td>346,000</td>
</tr>
</tbody>
</table>

A paradox, and an opportunity...
Large numbers of patients, of diverse ancestries, are required to develop evidence to “personalize” medicine.
## EMR-linked biobanks in eMERGE-II

<table>
<thead>
<tr>
<th>Site</th>
<th># of subjects with DNA samples linked to EMRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health Seattle</td>
<td>7,000</td>
</tr>
<tr>
<td>Marshfield</td>
<td>20,000</td>
</tr>
<tr>
<td>Mayo</td>
<td>19,000</td>
</tr>
<tr>
<td>Northwestern</td>
<td>11,000</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>175,000</td>
</tr>
<tr>
<td>Geisinger</td>
<td>22,000</td>
</tr>
<tr>
<td>Mt. Sinai</td>
<td>22,000</td>
</tr>
<tr>
<td>CHOP</td>
<td>60,000</td>
</tr>
<tr>
<td>Cincinnati/Boston</td>
<td>10,000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>346,000</strong></td>
</tr>
</tbody>
</table>