Outcome Data, Links to Electronic Medical Records

Dan Roden
Vanderbilt University
The eMERGE Network
electronic Medical Records & Genomics

A consortium of biorepositories linked to electronic medical records data for conducting genomic studies
**Type II Diabetes Case Algorithm**

- **Abnormal lab** = Random glucose > 200mg/dl, Fasting glucose > 125 mg/dl, or hemoglobin A1c ≥6.5%.

---

The eMERGE Network  
Electronic Medical Records & Genomics
Type II Diabetes Control Algorithm

1. >=2 in person clinician visits? 
   - Yes: >=1 glucose measure?
     - Yes: Abnormal* glucose or HbA1c?
       - T2DM control
     - No: ICD-9 code for diabetes or related condition?
6. No: Rx diabetes med, inc. insulin or supplies?
5. No: Family Hx of diabetes (type 1 or 2)?
Type II Diabetes Chart Review

Cases
- Blinded clinician review of 100 random charts (50 cases & 50 controls)
  - Case PPV = 98%
  - Control PPV = 98%

Controls
- Absence of data ≠ absence of condition
  - May simply be inadequate data capture in EMR for individual
  - Difference in comprehensiveness of data capture between sites
  - Valid criteria for controls requires same data elements as case
TCF7L2 cases controls
AA 810 873
EA 2413 2392
total 3266 3286
No thyroid-altering medications (e.g., Phenytoin, Lithium)

- ICD-9s for Hypothyroidism
- Abnormal TSH/FT4

Thyroid replacement medication

- No secondary causes (e.g., pregnancy, ablation)

Case

2+ non-acute visits

- No ICD-9s for Hypothyroidism
- Normal TSH

- No thyroid replace. meds
- No hx of myasthenia gravis

Control
# Phenotype Algorithm Validation

<table>
<thead>
<tr>
<th>Site</th>
<th>EMR-based Cases/Controls</th>
<th>Chart Review Cases/Controls</th>
<th>Case PPV (%)</th>
<th>Control PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health</td>
<td>397/1,160</td>
<td>50/50</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Marshfield</td>
<td>514/1,187</td>
<td>50/50</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>233/1,884</td>
<td>100/100</td>
<td>82</td>
<td>96</td>
</tr>
<tr>
<td>Northwestern</td>
<td>92/470</td>
<td>50/50</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>81/352</td>
<td>50/50</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>All sites (weighted)</td>
<td>1,317/5,053</td>
<td>—</td>
<td>92.4</td>
<td>98.5</td>
</tr>
</tbody>
</table>

Denny et al 2011
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 tool 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 validating electronic phenotype algorithms. On this site you can:

- View existing algorithms
- Enter or create new algorithms
- Collaborate with others to create or review algorithms
- View implementation details for existing algorithms

Phenotype algorithms can be viewed by data modalities or methods used:

- CPT codes
- ICD 10 codes
- ICD 9 codes
- Laboratories
An eMERGE-wide phenotype analyzed with no extra genotyping: hypothyroidism

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

Denny et al., 2011
## Phase I Phenotypes

<table>
<thead>
<tr>
<th></th>
<th>GHC/UW</th>
<th>Marshfield</th>
<th>Mayo</th>
<th>Northwestern</th>
<th>Vanderbilt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dementia</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Cataract</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>PAD</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>QRS Duration</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Secondary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Diabetic Retinopathy</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Lipids</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>PheWAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>HDL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Resistant HTN</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
• 203 C282Y homozygotes; n=31,192 Northern Europeans ages 40-69
• 28.4% of men and 1.2% of women had possible “iron-overload related disease”

VUMC BioVU data:
• 41/~5000 C282Y homozygotes
  • 9 diagnosed as having hemochromatosis
  • 32 undiagnosed:
    • 17 possible symptoms
    • 7 receiving iron supplements
PheWAS for rs10759944

\[
\begin{align*}
\text{OR}_{\text{GWAS}} &= 0.74 \\
\text{OR}_{\text{PheWAS}} &= 0.76
\end{align*}
\]

N = 13617 subjects

Denny et al., 2011
Pleiotropy: PheWAS associations with a skin color SNP
The eMERGE Network

electronic Medical Records & Genomics

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

A consortium of biorepositories linked to electronic medical records data for conducting genomic studies.
eMERGE II Goals

- EHR-based Phenotyping
- Establish new Genotype-Phenotype associations
- Clinical Use:
  - Defining actionability/clinical utility/validity
  - Integration into EHR/Visualization/Clinical Decision support
- Physician and Patient attitudes/Education
- Consent/Regulatory
- Privacy/Security/CLIA/CAP
# eMERGE-II Sample Size

<table>
<thead>
<tr>
<th></th>
<th>Participants</th>
<th>Genotyped</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GHC/UW</td>
<td>6,400</td>
<td>3,575</td>
<td></td>
</tr>
<tr>
<td>Marshfield</td>
<td>20,000</td>
<td>4,987</td>
<td></td>
</tr>
<tr>
<td>Mayo</td>
<td>19,000</td>
<td>6,940</td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>10,500</td>
<td>4,962</td>
<td></td>
</tr>
<tr>
<td>VU</td>
<td>140,000</td>
<td>33,228</td>
<td></td>
</tr>
<tr>
<td>Geisinger</td>
<td>19,700</td>
<td>4,191</td>
<td></td>
</tr>
<tr>
<td>Mt. Sinai</td>
<td>21,000</td>
<td>16,000</td>
<td></td>
</tr>
<tr>
<td>CCMC/CHB</td>
<td>40,100</td>
<td>5,586</td>
<td></td>
</tr>
<tr>
<td>CHOP</td>
<td>40,000</td>
<td>8,000</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>316,600</strong></td>
<td><strong>87,469</strong></td>
<td></td>
</tr>
</tbody>
</table>
Collins: Pharmacogenomics will undoubtedly become a very compelling part of medical practice. The limiting factor right now is that oftentimes, if you are ready to write a prescription, you do not want to wait a week to find out the genotype before you decide whether you’ve got the right dose and the right drug. But if everybody’s DNA sequence is already in their medical record and it is simply a click of the mouse to found out all the information you need, then there is going to be a much lower barrier to beginning to incorporate that information into drug prescribing. If you have the evidence, it will be hard, I think, to say that this is not a good thing. And once you’ve got the sequence, it’s not going to be terribly expensive. And it should improve outcomes and reduce adverse events.

Francis Collins, 9/16/2009

“Here’s my sequence…”

New Yorker, 2000
n=58 (germline)
WARNING: DIMINISHED EFFECTIVENESS IN POOR METABOLIZERS

See full prescribing information for complete boxed warning.

• Effectiveness of Plavix depends on activation to an active metabolite by the cytochrome P450 (CYP) system, principally CYP2C19. (5.1)
• Poor metabolizers treated with Plavix at recommended doses exhibit higher cardiovascular event rates following acute coronary syndrome (ACS) or percutaneous coronary intervention (PCI) than patients with normal CYP2C19 function. (12.5)
• Tests are available to identify a patient's CYP2C19 genotype and can be used as an aid in determining therapeutic strategy. (12.5)
• Consider alternative treatment or treatment strategies in patients identified as CYP2C19 poor metabolizers. (2.3, 5.1)

*1/*2 (18.8%) and *2/*2 (2.6%)
Beyond CYP2C19*2...

~7000 European American alleles
~3700 African American alleles

- CYP2C19:
  - 67 missense or nonsense variants
  - 22/67 not previously seen

- CYP2D6
  - 172 missense or nonsense variants
  - 58/172 not previously seen
eMERGE-PGx – Overall Goal

To initiate a multi-site test of the concept that sequence information can be coupled to electronic medical records for use in healthcare
eMERGE-PGx: a PGRN-eMERGE alliance

**PGRN**
- Clinical Pharmacogenomics Implementation Consortium (CPIC)
- Translational Pharmacogenomics Project (TPP)
- PGRN-Seq and other platforms

**eMERGE**
- developing and validating electronic phenotyping algorithms (including for drug responses)
- developing and deploying clinical decision support
Aim 1
Identify target patients
Resequence VIP genes; Identify actionable variants

Aim 2
Develop list of actionable variants (eMERGE, CPIC, ...)
Actionable variants
EMR deposit
- Result display
- Decision support
Outcomes
- Performance metrics
- Healthcare impact

Aim 3
- Create repository of variants of unknown significance
- Initiate studies of function and of genotype-phenotype relationships
What have we learned since 2008?

- Algorithms to accurately identify cases and controls for simple phenotypes can be developed and deployed across multiple EMRs
- In progress: more complex phenotypes
  - Multigene prediction tools
  - Complications of disease
  - Variable drug responses
- The PheWAS experiment is feasible
- Implementation is far from straightforward
Closing thoughts/”answers”

- Populations to be sequenced should include healthcare information
- Advantages of mining in EMRs
  - Real-world
  - Feasibility demonstrated
  - Enabling for implementation (eventually)
  - Rare/extreme phenotypes accessible
  - Potential for coupling to other datasets
- Disadvantages
  - The phenotype is what is in the EMR. Dense exquisite phenotypes need (a lot of) extra work.
Closing thoughts/answers
1. General considerations about power and sample size
2. Relationship of expected disease architecture to study design
3. Strengths and weaknesses of prospective cohort or retrospective case-control designs, family or extremes designs
4. Strengths and weaknesses, especially costs and benefit, of whole genome vs whole exome sequencing
5. Potential differences in analytic approaches for interpretation of WES/WGS sequence data from thousands of samples vs from small numbers of samples
6. Types of validation needed and how this should be factored into initial design
7. Additional data types would one want to gather (RNA, proteomics, etc.) and any impact their availability might have on design of the DNA sequencing component