Clinical Phenotyping in the EMR: Challenges and Opportunities

Josh Denny, MD, MS
Jan 22, 2014
Goals of Phenotyping Workgroup

1. Develop, validate, and implement ~27 EHR phenotypes for genomic study across eMERGE sites
   – For each phenotype: Lead site develops, validates
     → one to two other sites deploy, validate, revise
     → deploy across network
   – Use existing genotyped records
   – Preserve privacy and promote data/algorithm reuse

2. Improve the process of EHR phenotyping
## Network Phenotyping Progress

<table>
<thead>
<tr>
<th></th>
<th>1st Round</th>
<th>2nd Round</th>
<th>3rd Round</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCHMC / BCH</td>
<td>Childhood Obesity</td>
<td>Autism</td>
<td>GERD/ Appendicitus/ Epilepsy/ or Pulm HTN</td>
</tr>
<tr>
<td>CHOP</td>
<td>Asthma</td>
<td>Atopic Dermatitits</td>
<td>Lipids or ADHD</td>
</tr>
<tr>
<td>Geisinger</td>
<td>Abdominal Aortic Aneurysm</td>
<td>Extreme Obesity</td>
<td>Remission of Diabetes after ROUX-EN-Y</td>
</tr>
<tr>
<td>GHC</td>
<td>Clostridium difficile (Cdiff)</td>
<td>Zoster</td>
<td>CADD as Quantitative Measure</td>
</tr>
<tr>
<td>Mayo</td>
<td>Venous thromboembolism</td>
<td>Cardio Respiratory Fitness</td>
<td>Heart Failure</td>
</tr>
<tr>
<td>MC/EIRH / PSU</td>
<td>Occular HTN</td>
<td>Glaucoma</td>
<td>Age Macular Degeneration</td>
</tr>
<tr>
<td>MS</td>
<td>Diabetic Hypertensive CKD</td>
<td>Rapid renal Decline in Diabetic HTN Nephropathy</td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td>Diverticulosis</td>
<td>Colon Polyps</td>
<td>caMRSA</td>
</tr>
<tr>
<td>VU</td>
<td>ACE-1 Cough</td>
<td>Statins for MACE</td>
<td>Upper GI/PUD</td>
</tr>
</tbody>
</table>

- **Complete**
- **Due 1st Qtr 2014**
- **In Development**
What we learned - Finding phenotypes in the EMR

Algorithm Development and Implementation

1. Identify phenotype of interest
2. Case & control algorithm development and refinement
3. Manual review; assess precision
4. Deploy in EMR Biobanks
5. Genetic association tests

≥95%
Sharing algorithms: PheKB.org

66 phenotypes, 20 public; 73 implementations; PPVs; social networking features; versioning; etc.
Algorithm Performance across PheKB

Drug-induced liver injury

<table>
<thead>
<tr>
<th>Site Implementations</th>
<th>Median</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>PPV (primary site)</th>
<th>PPV (secondary sites)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>Control</td>
</tr>
</tbody>
</table>

20%  40%  60%  80%  100%
But not everything is transportable…
An Algorithm for **Resistant Hypertension**

<table>
<thead>
<tr>
<th>Site</th>
<th>Case 1 PPV</th>
<th>Case 2 PPV</th>
<th>Control 1 PPV</th>
<th>Control 2 PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>96%</td>
<td>84%</td>
<td>14%#</td>
<td>91%</td>
</tr>
<tr>
<td>Site 2</td>
<td>100%</td>
<td></td>
<td></td>
<td>97%</td>
</tr>
<tr>
<td>Site 3</td>
<td></td>
<td>95%-&gt;46%*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site 4</td>
<td>84%</td>
<td></td>
<td>94%-&gt;3%*</td>
<td></td>
</tr>
<tr>
<td>Site 5</td>
<td>96%</td>
<td>88%</td>
<td>84%</td>
<td>84%</td>
</tr>
</tbody>
</table>

*Due to algorithm implementation issues; now manually curated
#Due to difficulty extracting the necessary components from the EMR
phewascatalog.org
PheWAS results for >3000 SNPs identified in GWAS studies

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP</th>
<th>PheWAS Phenotype</th>
<th>Cases</th>
<th>P-value</th>
<th>OR</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>50087459</td>
<td>Alzheimer's disease</td>
<td>737</td>
<td>5.237e-28</td>
<td>2.41</td>
<td>TOMM40</td>
</tr>
<tr>
<td>19</td>
<td>50087459</td>
<td>Dementias</td>
<td>1170</td>
<td>2.408e-28</td>
<td>2.11</td>
<td>TOMM40</td>
</tr>
<tr>
<td>6</td>
<td>341321</td>
<td>Actinic keratosis</td>
<td>2505</td>
<td>4.141e-25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>26201120</td>
<td>Iron metabolism disorder</td>
<td>40</td>
<td>3.409e-25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>50087459</td>
<td>Delirium dementia and amnestic disorders</td>
<td>1566</td>
<td>8.027e-24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>154985403</td>
<td>Age-related macular degeneration</td>
<td>749</td>
<td>7.157e-20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>341321</td>
<td>Non-melanoma skin cancer</td>
<td>1931</td>
<td>3.818e-17</td>
<td>1.5</td>
<td>IRF4</td>
</tr>
<tr>
<td>6</td>
<td>25929749</td>
<td>Iron metabolism disorder</td>
<td>40</td>
<td>5.306e-17</td>
<td>6.84</td>
<td>SLC17A1</td>
</tr>
</tbody>
</table>

- search SNPs, phenotypes, genes
- make/save graphs
- export data sets
eMERGE Record Counter and SPHINX

Result Set Total: 25

This number may be rounded up or down. It may not perfectly match individual counts.

Gender Distribution:
- Female: 12
- Male: 13
- Unknown: 0

Included Criteria:
- Contains Medication '32968 - clopidogrel'
- Contains ICD code in group 410-Acute myocardial infarction
- OR Contains CPT code 33534

Group Count: 121

Included Criteria:
- Contains ICD code in group 390-Rheumatic fever without mention of heart disease
- Contains ICD code in group 391-Rheumatic fever with heart involvement
- Generates total count: 47
Key Questions for eMERGE 3

• What types of phenotypes to explore?
• How to make the process faster/better?
• How to improve accuracy and reproducibility?
• How can we best leverage the unique nature of the EMR?
New Phenotypes for Discovery

- Move beyond just disease-gene to more detailed phenotypes:
  - less common or rare phenotypes
  - pharmacogenomics (follow-up on eMERGE-PGx)
  - disease subtypes
  - longitudinal phenotypes

- Phenotypes for clinical implementation (e.g., for CDS)

- Bigger sample sizes are needed, but eMERGE has 350k+ lives covered!

- These may be harder than other phenotypes... may need to decide that less is more
Rare phenotypes

- Adverse drug events (Steven-Johnson Syndrome)
- Rare diseases (e.g., Wegener’s granulomatosis)

Rationale:
- May have stronger genetic signals
- Clinical impact
- EMR may be best way to capture

Problem:
- Would likely need new genotyping/sequencing
- GWAS may not be detailed enough (rare variants)
New Methods for eMERGE 3

• Expand work on common infrastructures for phenotyping
• Use phenotyping within Clinical Decision Support frameworks
• High throughput phenotyping with machine learning, active learning, etc.
• Phenomic methods (PheWAS, DrugWAS, etc.) to investigate pleiotropy and comorbidity
• Refine phenotype algorithms to include all patient statuses
  – “gray areas” such as probable and suspect cases
Central Resources

• Expand record counter functionality
  – Options of implementing federated queries and automated processes
  – Virtual data warehouses
• Structured data dictionaries and data validation tools
• Sites would contribute to these efforts, but one standard should be set to cooperative development
Key Questions for eMERGE 3

• What types of phenotypes to explore?
  – common, rare, pharmacogenomic, subtypes

• How to make the process faster/better?
  – new methods, standards, federated search, CDS

• How to improve accuracy and reproducibility?
  – standards, extensible methods

• How can we best leverage the unique nature of the EMR?
  – phenomic approaches, longitudinal phenotypes