Balancing discovery and implementation in eMERGE

Dan M. Roden, MD
Vanderbilt University School of Medicine
Studying cohorts
• in purpose-generated research datasets
• in the EMR

Using a highly interactive electronic medical record
• to provide real-time clinical advice
• to track outcomes
Discovery science in eMERGE

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>397</td>
<td>1,160</td>
<td>98</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Marshfield</td>
<td>cataracts</td>
<td>4113</td>
<td>514</td>
<td>1,187</td>
<td>91</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>peripheral arterial disease</td>
<td>3043</td>
<td>233</td>
<td>1,884</td>
<td>82</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>Northwestern</td>
<td>type 2 diabetes</td>
<td>1217</td>
<td>92</td>
<td>470</td>
<td>98</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>normal cardiac conduction</td>
<td>2712</td>
<td>81</td>
<td>352</td>
<td>98</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>All sites</td>
<td></td>
<td>13,617</td>
<td>1317</td>
<td>5053</td>
<td>92.4$^a$</td>
<td>98.5$^a$</td>
<td></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.

$^a$ Average weighted for number of samples contributed to the total.

Algorithms can be deployed across multiple EMRs

Analyses can be performed using extant data

Denny et al., 2011
GWAS: Target phenotype

PheWAS: Target genotype

The phenome-wide association study

PheWAS requirement: A large cohort of patients with genotype data and many diagnoses
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
Balancing the discovery and implementation missions

• What can eMERGE contribute to discovery...
  • ...in which others also engaged?
  • ...for which eMERGE is near-uniquely positioned?

• What can eMERGE contribute to implementation...
  • ...in which others also engaged?
  • ...for which eMERGE is near-uniquely positioned?
Discovery versus Implementation
The “easiest” examples

- Some drug responses
- Some cancer susceptibility

Do we really know all there is to know about variable responses to commonly used drugs?
- Rare variants
- Ancestry
Warfarin: not so simple....

Rare variants in VKORC1 associated with high dose requirements

- dose requirement $>20$ mg/day AND serum warfarin $>2.3$

- identified in 8/15 Ashkenazi patients requiring $>11$ mg/day

Harrington et al., 2008
## Warfarin: not so simple....

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Minor Allele Frequency</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2C9*2</td>
<td>rs1799853</td>
<td>12.86%</td>
<td>8.48E-12</td>
</tr>
<tr>
<td>CYP2C9*3</td>
<td>rs1057910</td>
<td>5.72%</td>
<td>3.32E-25</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs2359612</td>
<td>38.47%</td>
<td>6.38E-55</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs9934438</td>
<td>38.11%</td>
<td>1.07E-60</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs9923231</td>
<td>38.14%</td>
<td>3.40E-60</td>
</tr>
</tbody>
</table>
# Warfarin: not so simple....

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Minor Allele Frequency</th>
<th>All</th>
<th>EA</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>Caucasian</td>
<td>AA</td>
<td>n = 1,170</td>
</tr>
<tr>
<td>CYP2C9*2</td>
<td>rs1799853</td>
<td>11.53%</td>
<td>12.86%</td>
<td>2.45%</td>
<td>8.48E-12</td>
</tr>
<tr>
<td>CYP2C9*3</td>
<td>rs1057910</td>
<td>5.22%</td>
<td>5.72%</td>
<td>1.74%</td>
<td>3.32E-25</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs2359612</td>
<td>36.56%</td>
<td>38.47%</td>
<td>23.26%</td>
<td>6.38E-55</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs9934438</td>
<td>34.67%</td>
<td>38.11%</td>
<td>10.76%</td>
<td>1.07E-60</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs9923231</td>
<td>34.69%</td>
<td>38.14%</td>
<td>10.76%</td>
<td>3.40E-60</td>
</tr>
</tbody>
</table>
Multiple gene effect

The warfarin pathway

- 7-OH warfarin (inactive)
- Multiple other metabolites (inactive)
- S-warfarin
- R-warfarin (weak)

CYP2C9

*S2, *3: Coding region variants

VKORC1

- Vitamin K epoxide
- Vitamin K reduced

EPHX1, CALU

GGCX

Common promoter haplotype that correlates with variable liver expression

inactive CYP4F2
Warfarin: not so simple….

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Minor Allele Frequency</th>
<th></th>
<th></th>
<th></th>
<th>All n = 1,170</th>
<th>EA n = 1,025</th>
<th>AA n = 145</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>Caucasian</td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2C9*2</td>
<td>rs1799853</td>
<td>11.53%</td>
<td>12.86%</td>
<td>2.45%</td>
<td></td>
<td>8.48E-12</td>
<td>1.45E-11</td>
<td>0.5047</td>
</tr>
<tr>
<td>CYP2C9*3</td>
<td>rs1057910</td>
<td>5.22%</td>
<td>5.72%</td>
<td>1.74%</td>
<td></td>
<td>3.32E-25</td>
<td>9.06E-24</td>
<td>0.01556</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs2359612</td>
<td>36.56%</td>
<td>38.47%</td>
<td>23.26%</td>
<td></td>
<td>6.38E-55</td>
<td>1.30E-58</td>
<td>0.3112</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs9934438</td>
<td>34.67%</td>
<td>38.11%</td>
<td>10.76%</td>
<td></td>
<td>1.07E-60</td>
<td>1.50E-58</td>
<td>0.002842</td>
</tr>
<tr>
<td>VKORC1</td>
<td>rs9923231</td>
<td>34.69%</td>
<td>38.14%</td>
<td>10.76%</td>
<td></td>
<td>3.40E-60</td>
<td>4.80E-58</td>
<td>0.002842</td>
</tr>
<tr>
<td>CYP4F2</td>
<td>rs2108622</td>
<td>28.10%</td>
<td>30.53%</td>
<td>10.84%</td>
<td></td>
<td>9.00E-07</td>
<td>1.85E-06</td>
<td>0.3671</td>
</tr>
<tr>
<td>EPHX1</td>
<td>rs2292566</td>
<td>14.22%</td>
<td>14.09%</td>
<td>15.14%</td>
<td></td>
<td>0.9372</td>
<td>0.5237</td>
<td>0.132</td>
</tr>
<tr>
<td>GGCX</td>
<td>rs11676382</td>
<td>9.04%</td>
<td>9.97%</td>
<td>2.45%</td>
<td></td>
<td>0.2755</td>
<td>0.3374</td>
<td>0.5976</td>
</tr>
<tr>
<td>GGCX</td>
<td>rs699664</td>
<td>37.81%</td>
<td>34.35%</td>
<td>37.93%</td>
<td></td>
<td>0.04851</td>
<td>0.05031</td>
<td>0.7907</td>
</tr>
<tr>
<td>CALU</td>
<td>rs339097</td>
<td>1.34%</td>
<td>0.05%</td>
<td>10.42%</td>
<td></td>
<td>0.06144</td>
<td>NA</td>
<td>0.04574</td>
</tr>
<tr>
<td>CYP2C9*6</td>
<td>rs9332131</td>
<td>0.31%</td>
<td>0.10%</td>
<td>1.74%</td>
<td></td>
<td>0.0008942</td>
<td>NA</td>
<td>0.001348</td>
</tr>
<tr>
<td>CYP2C9*8</td>
<td>rs7900194</td>
<td>NA</td>
<td>NA</td>
<td>6.94%</td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>0.00701</td>
</tr>
<tr>
<td>CYP2C9*11</td>
<td>rs28371685</td>
<td>0.48%</td>
<td>0.25%</td>
<td>2.08%</td>
<td></td>
<td>0.6528</td>
<td>NA</td>
<td>0.427</td>
</tr>
</tbody>
</table>
Discovery versus Implementation
Some other “easy” examples

- Factor V Leiden
- HFE
- APOL1

The poster children:
Are these the only ones?
Deploy? How? How to measure impact?
Discovery versus Implementation

Getting harder

- Complex combinations of markers (e.g. risk scores): genomic and other

- Development and validation
- How to deploy
- How to measure impact and outcome
Discovery science that 346,000 DNA samples coupled to EMRs can enable

- PheWAS
- Complex outcomes:
  - Longitudinal over time
  - Disease x drug x response
  - Variable outcomes by disease subtypes
Discovery science that 346,000 DNA samples coupled to EMRs can enable

• PheWAS

• Complex outcomes:
  • Gene x Longitudinal over time
  • Gene x Disease x drug x response
  • Gene x Variable outcomes by disease subtypes
Discovery science that 346,000 DNA samples coupled to EMRs can enable:

- PheWAS
- Complex outcomes:
  - Gene x Longitudinal over time
  - Gene x Disease x drug x response
  - Gene x Variable outcomes by disease subtypes
- Consideration of ancestry issues
- To what extent can data be deidentified and retain discovery value?
Implementation science that 346,000 DNA samples coupled to EMRs can enable

- What? What evidence matters?
- How?
- In who?
- Educating providers and patients
- Decision support
- Outcomes
Studying cohorts
- in purpose-generated research datasets
- in the EMR

Using a highly interactive electronic medical record
- to provide real-time clinical advice
- to track outcomes