The success, challenge, and promise of EHR phenotypes for medical and genomic research

Josh Denny, MD, MS
josh.denny@vanderbilt.edu
Vanderbilt University, Nashville, Tennessee, USA

11/15/2014
Disclaimers

I receive funding from:

• NIH: NLM, NHGRI, NIGMS, NCI, NCATS
• Reynolds Foundation (Geriatrics Education)
• National Board of Medical Examiners
Three stories

• Doing “traditional” genomic research with the EHR
• EHRs as a tool to accelerate research and enable new types of research
The vision

"Here's my sequence..."

New Yorker, 2000
How will this vision actually start to be tested and become reality?

"Here's my sequence..."

New Yorker, 2000

- Biomedical research
- Commitment to information technology
- Harnessing the healthcare system for discovery
- Ability to nimbly adapt a healthcare system to evolving evidence
EHR feeds both discovery and implementation

**Discovery**

- De-identified DNA repository
- Vanderbilt BioVU
- Integrated genomic testing for discovery

**Implementation**

- PREDICT
- Integrated genomic testing for clinical care
Genomic achievements since the Human Genome Project

2010: First EMR-based genetic studies

2010: 500th genome-wide association study

2005: First genome-wide association study

2004: Publication of finished human genome sequence

Green et al. Nature. 2011
John Doe

One way hash

~2 million records

The Synthetic Derivative:
updated regularly

A7CCF99DE65732....
One way hash

Extract DNA

BioVU
~190,000 DNAs

Scrubbed

>2 million records

The Synthetic Derivative:
updated regularly
“Scrubbed” Medical Record

- MR# is removed
- Substituted names
- Replaced SSN and phone #
- Shifted Dates
Technology + Policy

• Databank access restricted to Vanderbilt employees

• Must sign Data Use Agreement that prohibits “re-identification”

• Operations Advisory Board and Institutional Review Board oversight

• All data access logged and can be audited per project
Resources for rapid, efficient EMR-based research at VUMC

The Synthetic Derivative
A de-identified and continuously-updated image of the EMR (2 M records)

BioVU
- DNA samples available: ~190,000 samples
- Plasma trial underway

Redeposited genotypes
- Subjects with GWAS data: >18,000
- Subjects with any genotyping: >70,000
The “demonstration project”

• Are genotype-phenotype relations replicated in BioVU?
• Genotype “high-value” SNPs in the first 10,000 samples accrued.
  • 21 established loci (>1 SNP for some)
  • in 5 diseases with known associations:
    Atrial fibrillation
    Crohn’s disease
    Multiple Sclerosis
    Rheumatoid arthritis
    Type II Diabetes
• Develop “electronic phenotype algorithms” to identify cases and controls
Finding cases: Rheumatoid Arthritis

Definite Cases (algorithm-defined)  Possible Cases (require manual review)  Excluded (algorithm-defined)  Controls (algorithm-defined)

255  507  7121  1184

Used for analysis
## RA – Case Definition Evolution

<table>
<thead>
<tr>
<th>#</th>
<th>Definition</th>
<th># Cases</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ICD9 codes for RA + Medications (only in problem list)</td>
<td>371</td>
<td>Found incomplete problem lists</td>
</tr>
<tr>
<td>2</td>
<td>Same as above but searched notes</td>
<td>411</td>
<td>Patients billed as RA but actually other conditions, overlap syndromes, juvenile RA</td>
</tr>
<tr>
<td>3</td>
<td>Above + require “rheumatoid arthritis” and small list of exclusions</td>
<td>358</td>
<td>Overlap syndromes with other autoimmune conditions, conditions in which physicians did not agree</td>
</tr>
<tr>
<td>4</td>
<td>Above + exclusion of other inflammatory arthritides</td>
<td>255</td>
<td>PPV = 97%; a few “possible RA” or family history items remained</td>
</tr>
</tbody>
</table>
Final RA case definition

ICD 9 codes (any of the below)
- 714    Rheumatoid arthritis and other inflammatory polyarthropathies
- 714.0  Rheumatoid arthritis
- 714.1  Felty’s syndrome
- 714.2  Other rheumatoid arthritis with visceral or systemic involvement

Medications (any of the below)
- methotrexate [MTX] [amethopterin] sulfasalazine [azulfidine]; Minocycline [minocin][solodyn];
- hydroxychloroquine [Plaquinil]; adalimumab [Humira]; etanercept [Enbrel] infliximab
  [Remicade]; Gold [myochrysine]; azathioprine [Imuran]; rituximab [Rituxan] [MabThera];
- anakinra [Kineret]; abatacept [Orencia]; leflunomide  [Arava]

Keywords (any of the below)
- rheumatoid [rheum] [reumatoid] arthritis [arthritides] [arthritis] [arthristis]
  [arthritis] [arthrtis] [artritis]
Final RA case definition - 2

**AND NOT**

ICD 9 codes (any of the below)

- 714.30 Polyarticular juvenile rheumatoid arthritis, chronic or unspecified
- 714.31 Polyarticular juvenile rheumatoid arthritis, acute
- 714.32 Pauciarticular juvenile rheumatoid arthritis
- 714.33 Monoarticular juvenile rheumatoid arthritis
- 695.4 Lupus erythematosus
- 710.0 Systemic lupus erythematosus
- 373.34 Discoid lupus erythematosus of eyelid
- 710.2 Sjogren's disease
- 710.3 Dermatomyositis
- 710.4 Polymyositis
- 555 Regional enteritis
- 555.0 Regional enteritis of small intestine
- 555.1 Regional enteritis of large intestine
- 555.2 Regional enteritis of small/large intestine
- 555.9 Regional enteritis of unspecified site
- 564.1 Irritable Bowel Syndrome
- 135 Sarcoidosis
- 696 Psoriasis and similar disorders
- 696.0 Psoriatic arthropathy
- 696.1 Other psoriasis and similar disorders excluding psoriatic arthropathy
- 696.8 Other psoriasis and similar disorders
- 099.3 Reiter’s disease
- 716.8 Arthropathy, unspecified
- 274.0 Gouty arthropathy
- 358.0 myasthenia gravis
- 358.00 myasthenia gravis without acute exacerbation
- 358.01 myasthenia gravis with acute exacerbation
- 775.2 neonatal myasthenia gravis
- 719.3 Palindromic rheumatism
- 719.30 Palindromic rheumatism, site unspecified
- 719.31 Palindromic rheumatism involving shoulder region
- 719.32 Palindromic rheumatism involving upper arm
- 719.33 Palindromic rheumatism involving forearm
- 719.34 Palindromic rheumatism involving hand
- 719.35 Palindromic rheumatism involving pelvic region and thigh
- 719.36 Palindromic rheumatism involving lower leg
- 719.37 Palindromic rheumatism involving ankle and foot
- 719.38 Palindromic rheumatism involving other specified sites
- 719.39 Palindromic rheumatism involving multiple sites
- 720 Ankylosing spondylitis and other inflammatory spondylopathies
- 720.0 Ankylosing spondylitis
- 720.8 Other inflammatory spondylopathies
- 720.81 Inflammatory spondylopathies in diseases classified elsewhere
- 720.89 Other inflammatory spondylopathies
- 720.9 Unspecified inflammatory spondylopathy
- 721.2 Thoracic spondylosis without myelopathy
- 721.3 Lumbosacral spondylosis without myelopathy
- 729.0 Rheumatism, unspecified and fibrositis
- 340 Multiple sclerosis
- 341.9 Demyelinating disease of the central nervous system unspecified
- 323.9 transverse myelitis
- 710.1 Systemic sclerosis
- 245.2 Hashimoto’s thyroiditis
- 242.0 Toxic diffuse goiter
- 443.0 Raynaud’s syndrome

**OR**

Keywords (any of the below)

- juvenile [juv] rheumatoid [rheum] [reumatoid] [rhumatoid] arthritis [arthritides] [arthritis] [arthrises] [arthritus] [arthritus] [arthritis]
- juvenile [juv] arthritis [arthritides] [arthritis] [arthrises] [arthritus] [arthritus] [arthritis]
- juvenile chronic arthritis [arthritides] [arthritis] [arthrises] [arthritus] [arthritus] [arthritis]
- juvenile [juv] RA, JRA
- Inflammatory [inflammatory] [inflam] osteoarthritis [osteoarthrosis] [OA]
- Reactive [psoriatic] arthritis [arthropathy], [arthritides] [arthritis] [arthrises] [arthritus] [arthritus] [arthritis]
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 BioVU
5. Genetic association tests

Billing codes
ICD9 & CPT

Clinical Notes
(NLP - natural language processing)

Medications
ePrescribing & NLP

Labs & test results
NLP

True cases

< 95%

≥ 95%
Natural language processing

Clinical Notes, test reports, etc

Find Biomedical Concepts and Qualifiers (KnowledgeMap and SecTag)

CC: SOB
HPI: This is a 65yo w/ h/o CHF, ... no dm2... on atenolol 50mg daily... Mother had RA.

Medication Extraction (MedEx)

Structured Output
DrugName: atenolol
Strength: 50 mg
Frequency: daily

Structured Output
certainty (positive, negated)
Who experienced it? (patient or family member?)

chief_complaint:
  C0392680: Shortness of Breath
history_present_illness:
  Congestive Heart Failure
  Type 2 diabetes, negated
mother_medical_history:
  rheumatoid arthritis
### Validating EMR phenotype algorithms
(Using first 10,000 patients in BioVU)

<table>
<thead>
<tr>
<th>Disease</th>
<th>Methods</th>
<th>Definite Cases</th>
<th>Controls</th>
<th>Case PPV</th>
<th>Control PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial fibrillation</td>
<td>NLP of ECG impressions ICD9 codes CPT codes</td>
<td>168</td>
<td>1695</td>
<td>98%</td>
<td>100%</td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>ICD9 codes Medications (NLP)</td>
<td></td>
<td></td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>ICD9 codes Medications (NLP) NLP exclusions Labs</td>
<td></td>
<td></td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>ICD9 codes or text diagnosis</td>
<td>66</td>
<td>1857</td>
<td>87%</td>
<td>100%</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>ICD9 codes Medications (NLP) NLP exclusions</td>
<td>170</td>
<td>701</td>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Common themes:**
Billing codes – 5/5
NLP – 5/5
Meds – 4/5
Labs – 2/5

NLP = Natural language processing
Results

<table>
<thead>
<tr>
<th>Disease</th>
<th>Marker 1</th>
<th>Gene / Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial fibrillation</td>
<td>rs2200733</td>
<td>Chr. 4q25</td>
</tr>
<tr>
<td></td>
<td>rs10033464</td>
<td>Chr. 4q25</td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>rs11805303</td>
<td>IL23R</td>
</tr>
<tr>
<td></td>
<td>rs17234657</td>
<td>Chr. 5</td>
</tr>
<tr>
<td></td>
<td>rs1000113</td>
<td>Chr. 5</td>
</tr>
<tr>
<td></td>
<td>rs17221417</td>
<td>NOD2</td>
</tr>
<tr>
<td></td>
<td>rs2542151</td>
<td>PTPN22</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>rs3135388</td>
<td>DRB1*1501</td>
</tr>
<tr>
<td></td>
<td>rs2104286</td>
<td>IL2RA</td>
</tr>
<tr>
<td></td>
<td>rs6897932</td>
<td>IL7RA</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>rs6457617</td>
<td>Chr. 6</td>
</tr>
<tr>
<td></td>
<td>rs6679677</td>
<td>RSBN1</td>
</tr>
<tr>
<td></td>
<td>rs2476601</td>
<td>PTPN22</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>rs4506565</td>
<td>TCF7L2</td>
</tr>
<tr>
<td></td>
<td>rs12255372</td>
<td>TCF7L2</td>
</tr>
<tr>
<td></td>
<td>rs12243326</td>
<td>TCF7L2</td>
</tr>
<tr>
<td></td>
<td>rs10811661</td>
<td>CDKN2B</td>
</tr>
<tr>
<td></td>
<td>rs8050136</td>
<td>FTO</td>
</tr>
<tr>
<td></td>
<td>rs5219</td>
<td>KCNJ11</td>
</tr>
<tr>
<td></td>
<td>rs5215</td>
<td>KCNJ11</td>
</tr>
<tr>
<td></td>
<td>rs4402960</td>
<td>IGF2BP2</td>
</tr>
</tbody>
</table>

Ritchie et al., AJHG 2010
eMERGE goals

• To perform GWAS using EMR-derived phenotypes
• To initiate implementation of actionable variants into the EMR
Hypothyroidism: A “no genotyping” GWAS?

Domain experts define phenotype (VU)

Create initial EMR-based algorithm (VU)

Evaluate & refine

Share algorithm

Case PPV=92.4%
Control PPV=98.5%

Denny et al., AJHG 2011
Hypothyroidism: “No-Genotyping” GWAS

Denny et al., AJHG 2011
### eMERGE GWAS completed

<table>
<thead>
<tr>
<th>Site</th>
<th>Primary phenotype</th>
<th>Secondary Phenotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health</td>
<td>Dementia</td>
<td>white blood cell counts, monocyte count, herpes zoster</td>
</tr>
<tr>
<td>Marshfield</td>
<td>Cataracts</td>
<td>diabetic retinopathy</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>Peripheral Arterial Disease</td>
<td>red blood cell counts, ESR levels, Platelet levels</td>
</tr>
<tr>
<td>Northwestern</td>
<td>Type 2 Diabetes</td>
<td>HDL/LDL, height</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>PR Duration, QRS Duration</td>
<td>PheWAS</td>
</tr>
</tbody>
</table>

**Network Phenotypes**

- Autoimmune Hypothyroidism
- Resistant hypertension

**bold**=GWAS completed with significant results

**NEW**=first description or new findings

... another ~30 in progress as part of eMERGE II...
Sharing algorithms: PheKB.org

>70 phenotypes, 21 public; >100 implementations with 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>
<tbody>
<tr>
<td>PPV (primary site)</td>
<td></td>
</tr>
<tr>
<td>PPV (secondary sites)</td>
<td></td>
</tr>
<tr>
<td>PPV (primary site)</td>
<td></td>
</tr>
<tr>
<td>PPV (secondary sites)</td>
<td></td>
</tr>
</tbody>
</table>

Case

Control

VANDERBILT UNIVERSITY MEDICAL CENTER
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><strong>14%</strong>#</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%* (for a subset)</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
Why did Resistant HTN perform poorly at some sites?

1. First pharmacogenetic algorithm in eMERGE
2. The algorithm was complex:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Boolean Operators</th>
<th>Max Depth</th>
<th>Temporal Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic retinopathy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Controls</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Height</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Serum lipid level</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Low HDL cholesterol level</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Peripheral arterial disease</td>
<td>17</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>QRS duration</td>
<td>28</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Resistant hypertension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>172</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Controls</td>
<td>26</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>15</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Cataract</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>9</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Controls</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Using Machine Learning for Phenotyping

Synthetic Derivative

NLP Concepts → Meds → ICD9

Filter to RA related items

Refined SVM → Naïve SVM

AUC

Data Set
- Refined
- Naïve

Number of training records

Carroll et al, AMIA 2011
Blood pressure comparison over time

Who has hypertension?

No hypertension

Has hypertension

Systolic | Diastolic | Systolic HTN Threshold | Diastolic HTN Threshold
HTN Cases Distribution with Random Forest Probability Prediction (AUC=0.956)
Phenotype Definitions in 2014
Challenge: How do we automate this?

Word processing documents that contain lists of standardized lists of billing codes, lab measurements, UMLS concepts, etc.
Efforts to standardize these

• Observational Medical Outcomes Partnership (OMOP)
• Observational Health Data Sciences and Informatics (OHDSI)
• Quality Data Model/HQMF and Measure Authoring Tool (MAT)
• SHARPn and phenotypeportal.org
• Phenotype execution and modeling architecture (PhEMA)
Tools to enable end user access - Informatics for Integrating Biology and the Bedside (i2b2)
Use of the VU Synthetic Derivative by other investigators

- Total Number of SD Users
  - New interface with real-time querying released

- Number of sets created by SD Users
  - New interface released

- However, SD team programming extractions have risen too...

- Service charges
  - July, 2011
  - July, 2012
  - July, 2013
  - Mar, 2014
Computational Phenotyping

- Principal components analysis
- Latent Dirichlet allocation
- Nonnegative tensor factorization
Extreme Phenotypes: LDL values in BioVU

Analysis of ~32,000 BioVU subjects with existing exome chip data for LDL

<table>
<thead>
<tr>
<th>SNP</th>
<th>Phenotype</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCSK9 R46L</td>
<td>mean LDL</td>
<td>1.8 x 10^{-18}</td>
</tr>
<tr>
<td></td>
<td>minimum LDL</td>
<td>6.1 x 10^{-13}</td>
</tr>
<tr>
<td></td>
<td>LDL &lt; 50</td>
<td>1.9 x 10^{-6}</td>
</tr>
</tbody>
</table>
The genome-wide association study

PheWAS requirement: A large cohort of patients with genotype data and many diagnoses
Original PheWAS

Table 2. Diseases previously associated with the five SNP studied and current PheWAS ORs

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene/region</th>
<th>Disease</th>
<th>Cases</th>
<th>Previous OR</th>
<th>PheWAS P-value</th>
<th>PheWAS OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3135388</td>
<td>DRB1*1501</td>
<td>MS</td>
<td>89</td>
<td>1.99</td>
<td>2.77 x 10^-6</td>
<td>2.24 (1.56-3.16)</td>
</tr>
<tr>
<td>rs17234657</td>
<td>Chr. 5</td>
<td>SLE</td>
<td>141</td>
<td>2.06</td>
<td>0.51</td>
<td>1.13 (0.79-1.58)</td>
</tr>
<tr>
<td>rs2200733</td>
<td>Chr. 4q25</td>
<td>CD</td>
<td>200</td>
<td>1.54</td>
<td>0.00080</td>
<td>1.57 (1.19-2.04)</td>
</tr>
<tr>
<td>rs1333049</td>
<td>Chr. 9p21</td>
<td>AF and flutter</td>
<td>606</td>
<td>1.75</td>
<td>0.14</td>
<td>1.15 (0.95-1.39)</td>
</tr>
<tr>
<td>rs6457620</td>
<td>Chr. 6</td>
<td>RA</td>
<td>461</td>
<td>1.30</td>
<td>0.02</td>
<td>1.18 (1.06-1.32)</td>
</tr>
</tbody>
</table>

N = 6,005

Denny et al., Bioinformatics 2010
PheWAS of “all” NHGRI GWAS Catalog SNPs

- 3,144 SNPs with prior GWAS-discovered associations
- 674 SNPs with 86 phenotypes
- 751 SNP-phenotype associations

**Replication Arm**

- Test for replication of 751 associations using PheWAS

**Discovery Arm**

- Replication of novel associations

Denny et al, Nat Biotech 2013

NHGRI GWA Catalog
www.genome.gov/GWASTudies
PheWAS Population

- 13,835 European-Ancestry individuals from 5 eMERGE sites with available GWAS data
- 2,080,550 unique dates of interaction with the EMR
- Mean follow-up of 15.7 ± 10.3 years
Replications of NHGRI GWAS associations via PheWAS

P-value for replication:
- All - 210/751: $2 \times 10^{-98}$
- Powered - 51/77: $3 \times 10^{-47}$

Denny et al, Nat Biotech 2013
Q-Q plot

ROC analysis

AUC=0.83

- T2D vs. T1D were most common FPs
- Several “FPs” actually real associations not in GWAS Catalog

Denny et al, Nat Biotech 2013
Factors associated with replication

- Number of prior publications
- Exactness of phenotype match
- SNP location/functional status NOT associated

![Bar chart showing association count and replication rates for different numbers of publications reporting association.](image)
PheWAS replications compared to validated eMERGE algorithms

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>eMERGE Phenotype for GWAS</th>
<th>PheWAS</th>
<th>Replicated SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cases/controls</td>
<td></td>
<td>cases/controls</td>
</tr>
<tr>
<td></td>
<td>replicated SNPs</td>
<td></td>
<td>replicated SNPs</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>cases=2526; controls=5276</td>
<td>15/23</td>
<td>cases=3122; controls=8106</td>
</tr>
<tr>
<td>Dementia</td>
<td>cases=1599; controls=1866</td>
<td>6/18</td>
<td>cases=737; controls=9469</td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td>cases=1317; controls=5053</td>
<td>discovered FOXE1 (OR=0.74)</td>
<td>cases=2051; controls=10106</td>
</tr>
</tbody>
</table>
PheWAS of all GWAS “hits”
Each dot = one phenotype

- GWA catalog association only
- GWA catalog association replicated by PheWAS
- New association found by PheWAS

PheWAS'associations for TERT
Known: glioma

seborrheic keratosis
leukoplakia of oral mucosa

Chromosome
- 6p25.3
- 6p21.32
- 9p21.3
PheWAS of all GWAS “hits”
Each dot=one phenotype

- GWA catalog association only
- GWA catalog association replicated by PheWAS
- New association found by PheWAS

PheWAS associations for IRF4
Known: hair, skin, eye color

- IRF4
- TERT

Known: hair, skin, eye color
Pleiotropy in Thyroid Diseases
Using PheWAS to refine understanding of GWAS: normal cardiac conduction

$SCN5A/SCN10A$

$n=5,272$

Ritchie et al., Circulation 2013
PheWAS of rs6795970 (SCN10A)

N=13617 subjects with EHR data

cardiac arrhythmias
atrial fibrillation

disease codes

Ritchie et al., Circulation 2013
PheWAS results for >3000 SNPs identified in GWAS studies

- search SNPs, phenotypes, genes
- make/save graphs
- export data sets
- R PheWAS package
What happens in the “heart healthy” population?

Examined 5272 “heart healthy” people

Followed for development of atrial fibrillation based on genotype

Ritchie et al., Circulation 2013
PheWAS to define Autism subtypes
ML+NLP to find Autism → PheWAS Comorbidities → Identify Clusters

Based on Doshi-Velez. *Pediatrics* 2013
Lindgren et al, in preparation
Next generation PheWAS Using Natural Language Processing

NLP-to-ICD9 Problem Lists

NLP-CUIs All Notes 1 Code Min.

-log(p)

1.E+00

1.E-01

1.E-02

1.E-03

1.E-04

1.E-05

1.E-06

1.E-07

1.E-08

phenotypes

Multiple sclerosis

Cardiac congenital anomalies

Conduction disorders

Actinic keratosis

Photodermatitis and sunburn

Cardiac disorders

Type 1 diabetes

Relapsing remitting multiple sclerosis

Checking dressing of skin

Proctitis

Generalized multiple sclerosis

Abdominal pain rt upr quad

Type 1 diabetes

Phenotypes

Teixiera et al. AMIA 2014
Phenotyping grouping/collapsing

Disorder of central nervous system
- Demyelinating disease of central nervous system
- Multiple sclerosis
  - Generalized multiple sclerosis
  - Relapsing remitting multiple sclerosis
  - Chronic progressive multiple sclerosis
  - Primary progressive multiple sclerosis
  - etc...

Degenerative disorder
Making our care safer - pharmacogenomics
Finding ADEs in Clinical Text
NLP of ~11 million notes to calculate risk of ADRs on drugs

LePendu et al., *Clin Pharmacol Ther* (2013); 93, 6, 547–555
An example of EMR mining for drug repurposing: Metformin

N=111,673 individuals with cancer at VU and Mayo

Unadjusted

Adjusted for age, gender, tumor stage, smoking, etc.

---

Xu et al., JAMIA 2014
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Cases</th>
<th>Controls</th>
<th>% Reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clopidogrel in CV disease</td>
<td>225</td>
<td>468</td>
<td>27%</td>
</tr>
<tr>
<td>Warfarin stable dose</td>
<td>1,167</td>
<td>N/A</td>
<td>47%</td>
</tr>
<tr>
<td>Early Repolarization</td>
<td>544</td>
<td>2,609</td>
<td>89%</td>
</tr>
<tr>
<td>Vancomycin stable dose</td>
<td>1,067</td>
<td>N/A</td>
<td>42%</td>
</tr>
<tr>
<td>C. difficile colitis</td>
<td>941</td>
<td>1,710</td>
<td>28%</td>
</tr>
<tr>
<td>Anthracycline cardiomyopathy</td>
<td>528</td>
<td>N/A</td>
<td>39%</td>
</tr>
<tr>
<td>Guillain-Barre Syndrome</td>
<td>97</td>
<td>6,536</td>
<td>99%</td>
</tr>
<tr>
<td>Heart Transplant</td>
<td>181</td>
<td>N/A</td>
<td>42%</td>
</tr>
<tr>
<td>Kidney transplant</td>
<td>1,078</td>
<td>N/A</td>
<td>32%</td>
</tr>
<tr>
<td>Clopidogrel in strokes/TIAs</td>
<td>6</td>
<td>123</td>
<td>22%</td>
</tr>
<tr>
<td>Statin-related myopathy</td>
<td>11</td>
<td>4,342</td>
<td>100%</td>
</tr>
<tr>
<td>Heparin-induced thrombocytopenia</td>
<td>73</td>
<td>2,300</td>
<td>99%</td>
</tr>
<tr>
<td>CV events with COX2 therapy</td>
<td>85</td>
<td>395</td>
<td>34%</td>
</tr>
<tr>
<td>Serious bleeding during warfarin</td>
<td>259</td>
<td>276</td>
<td>43%</td>
</tr>
<tr>
<td>Amiodarone toxicity (lung, thyroid)</td>
<td>97</td>
<td>343</td>
<td>34%</td>
</tr>
<tr>
<td>Chronic inflammatory polyneuropathy</td>
<td>12</td>
<td>14,000*</td>
<td>100%</td>
</tr>
<tr>
<td>Rheumatic Heart Disease</td>
<td>108</td>
<td>3,464</td>
<td>98%</td>
</tr>
<tr>
<td>ACEi cough</td>
<td>1,174</td>
<td>978</td>
<td>52%</td>
</tr>
<tr>
<td>Fluoroquinolones and tenopathy</td>
<td>87</td>
<td>537</td>
<td>90%</td>
</tr>
<tr>
<td>Warfarin stable dose in children</td>
<td>92</td>
<td>N/A</td>
<td>28%</td>
</tr>
<tr>
<td>Metformin efficacy</td>
<td>80</td>
<td>N/A</td>
<td>35%</td>
</tr>
<tr>
<td>Metformin and cancer</td>
<td>619</td>
<td>421</td>
<td>83%</td>
</tr>
<tr>
<td>Bisphosphonates and Atypical Fracture/Jaw Osteonecrosis</td>
<td>16</td>
<td>1,454</td>
<td>99%</td>
</tr>
<tr>
<td>Wolff-Parkinson-White</td>
<td>197</td>
<td>5,551</td>
<td>97%</td>
</tr>
<tr>
<td>Steroid-induced Osteonecrosis</td>
<td>83</td>
<td>352</td>
<td>57%</td>
</tr>
<tr>
<td>Shellfish Anaphylaxis</td>
<td>157</td>
<td>14,000*</td>
<td>99%</td>
</tr>
<tr>
<td>Aspirin Anaphylaxis</td>
<td>101</td>
<td>4,334</td>
<td>98%</td>
</tr>
<tr>
<td>Bell's Palsy*</td>
<td>577</td>
<td>14,000*</td>
<td>97%</td>
</tr>
</tbody>
</table>

Large scale GWAS for drug response discovery: the VESPA project
Vanderbilt Electronic Systems for Pharmacogenomic Assessment

Bowton et al., *Sci Trans Med.* 2014
## Predicting Warfarin Dose

Trained from the stable doses in the EMR

<table>
<thead>
<tr>
<th>SNP (Gene)</th>
<th>Beta</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1057910 (CYP2C9*3)</td>
<td>0.83</td>
<td>2.70x10^{-26}</td>
</tr>
<tr>
<td>rs9934438 (VKORC1)</td>
<td>0.87</td>
<td>4.48x10^{-61}</td>
</tr>
</tbody>
</table>

Ramirez et al. *Pharmacogenomics*. 2012
Two in-progress GWAS of Drug-ADEs from the EMR

ACEI-cough
(NLP of allergy sections, automated)

Heparin-induced thrombocytopenia
(automated+manual review)
eMERGE-PGx – Overall Goal

A multi-site test of targeted next-gen sequencing of 84 genes, validation, and EMR decision support to guide care in ~9,000 eMERGE patients
Preliminary PGRN-Seq Results

*SCN5A* and *KCNH2* in 2,200 Patients

- 83 rare (MAF < 1%) in *SCN5A*, 45 in *KCNH2*
- 121/128 MAF < 0.5%, 92 singletons
- Three labs assessed known/likely pathogenicity

![Venn Diagram]

- Lab 1: 16/121
- Lab 2: 24/121
- Lab 3: 17/121
- 4
A paradox, and an opportunity…

Large numbers of patients, of diverse ancestries, are required to develop evidence to “personalize” medicine.

EHR linked DNA samples

- eMERGE: 360k
- Kaiser-Permanente: 300k?
- Million Veterans program: 200k; will be 1 million
- UK Biobank: 500k
- China Kadoorie: 500k

Total: >2.5 million

Current GWAS imputed set: 51,038
Summary

• EHR-linked DNA biobanks can be used for genomic and pharmacogenomic discovery. They can be cost efficient and fast. Big populations are (will be) needed for genomic discovery, deciphering rare variants, and drug-drug interactions.

• Tools to provide access to data, algorithms, and results (Research repositories, PheKB.org, phewascatalog.org).

• Phenotype algorithms are typically portable across EHR systems, healthcare settings, NLP systems, etc.

• Challenges remain in high-throughput methods, algorithm sharing, computational methods for complex phenotypes, sharability
A real case: What personalizing medicine really means

57yo with DM2, FHx heart disease, ↑chol admitted for chest pain, receives stent

January

clopidogrel started

April

Recath, stent
“Plavix x 1 year minimum. ASA life long.”

In-stent thrombosis, restent

In-stent thrombosis, restent

Cath, more stents

December

9th admission, 5th intervention, 9th stent

CYP2C19*2/*2 clopidogrel poor metabolizer

Switched to prasugrel

vanderbilt university medical center
Acknowledgements

Informatics
• Lisa Bastarache
• Hua Xu
• Josh Peterson
• Brad Malin
• Dan Masys
• Robert Carroll
• Wei-Qi Wei
• Jeremy Warner

BioVU/SD
• Melissa Basford
• Jill Pulley
• Erica Bowton
• Jay Cowan
• Sunny Wang
• Jenny Madison
• Sue Bradeen

Medicine and Genetics
• Dan Roden
• Ellen Clayton
• Jessica Delaney
• Sara Van Driest
• Jonathan Mosley
• Andrea Ramirez
• Peter Weeke
• Todd Edwards

Biostatistics
• Jonathan Schildcrout
• Yaping Shi

eMERGE Network
• Children’s hospital of Philadelphia
• Boston Children’s/Cincinnati Children’s Hospitals
• Northwestern
• Marshfield Clinic
• Mayo Clinic
• Group Health/UW
• Mount Sinai
• Geisinger

Funding
• VICTR/NCATS
• NHGRI
• NLM
• NIGMS
• NCI