Analytic Approaches to Phenotypic Complexity

Jeremy Warner M.D., M.S.
Joshua Denny M.D., M.S.
David Kreda
Gil Alterovitz Ph.D.
Visual Analytic Approaches to Phenotypic Complexity

Jeremy Warner M.D., M.S.
Joshua Denny M.D., M.S.
David Kreda
Gil Alterovitz Ph.D.
Goal: Drilling into Phenome

Phenotypes and Phenomes

Phenotype Algorithms

Introducing Phenome Drilling

Inspiration: the SMART Genomics Advisor

Phenome Drilling: the SMART Phenomics Advisor

Q&A
**phenotype**

composite of an organism's observable characteristics or traits

**disease phenotype**

emerging manifestation of a defined underlying pathophysiology

**human disease phenome**

span of definable human disease phenotypes
Classified By ICD

ICD is the standard diagnostic tool for epidemiology, health management and clinical purposes.

World Health Organization (WHO)

1893 Bertillon Classification of Causes of Death
1900 International List of Causes of Death
1949 International Statistical Classification of Diseases
1979 ICD-9
1999 ICD-10
20xx ICD-11
Complex Disease Phenotypes

Common or rare disease lacking specific structured definition, often with 1-to-many relationships:

1 disease $\rightarrow$ many genotypes  
  e.g. ALL

1 genotype $\rightarrow$ many diseases  
  e.g. *BRAF* V600E

1 disease $\rightarrow$ *new* induced complex phenotypes

+ a wrinkle $\rightarrow$ co-existing & unrelated diseases
Phenotype Algorithms Today
Phenotype Algorithms

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 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

PheKB.org
Primary Hypothyroidism Algorithm

No thyroid-altering medications e.g., phenytoin, lithium

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

Thyroid replacement meds

No secondary causes e.g., pregnancy, ablation

Phenotype Case

2+ non-acute visits in 3 years

- No ICD-9s for hypothyroidism
- No abnormal TSH/FT4

No thyroid replacement meds

No antibodies for TTG/TPO

No history of myasthenia gravis

Control

Denny et al. AJHG 2011
## Primary Hypothyroidism Validation

<table>
<thead>
<tr>
<th>Site</th>
<th>Case PPV (%)</th>
<th>Control PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Marshfield</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>82</td>
<td>96</td>
</tr>
<tr>
<td>Northwestern</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>All sites weighted</td>
<td>92.4</td>
<td>98.5</td>
</tr>
</tbody>
</table>
PheKB: Algorithm Performance

Cases

Controls

PPV

Site Implementations

Median
Sharing Algorithms at PheKB.org

To Date:

- **64 total phenotypes**
- **20 public algorithms** (eMERGE, PGRN, VESPA)
- < 0.5% of 17,000 ICD-9-CM’s
- < 2% of 1600 PheWAS Codes
Do these efforts scale ???
Estimated Frequency of Specific Genotypes in Childhood Leukemias

Pui C et al. JCO 2011;29:551-565
Proposed Precursor B-lymphoblastic neoplasms

B lymphoblastic leukaemia/lymphoma, not otherwise specified
B lymphoblastic leukaemia/lymphoma with t(9:22) (q34;q11.2); BCR/ABL
B lymphoblastic leukaemia/lymphoma with t(v;11q23); MLL rearranged
B lymphoblastic leukaemia/lymphoma with t(12;21) (p13;q22); TEL/AML1 (ETV6-RUNX1)
B lymphoblastic leukaemia/lymphoma with hyperdiploidy
B lymphoblastic leukaemia/lymphoma with hypodiploidy (Hypodiploid ALL)
B lymphoblastic leukaemia/lymphoma with t(5;14)(q31;q32)(IL3-IGH)
B lymphoblastic leukaemia/lymphoma with t(1;19)(Q23;P13.3); (E2A-PBX1; TCF3/PBX1)

~66% of currently known mutations are covered

apps.who.int/classifications/icd11/browse
What About Induced Phenotypes

disease-related complications

caused primarily or secondarily by underlying disease process

treatment-related complications

caused by treatment of underlying disease process

syndromes

associated, not clearly caused by, a unifying disease process
Induced Phenotypes & Syndromes

- HIV/AIDS
  - PJP Pneumonia
  - Lymphoma
  - Dyslipidemia
    - Type 2 Diabetes
      - Coronary Arterial Disease
      - Neuropathy
      - Lactic Acidosis

Syndromes:
- Polyarthritis
- Alopecia
- Malar Rash
- Systemic Lupus Erythematosus

Primary disease process
Disease-related phenotypes
Treatment-related phenotypes
Induced Phenotype
GWAS of Normal Cardiac Conduction

Find individuals with \textbf{normal} cardiac conduction

THEN

Find genetic variants associated with QRS duration

Hypothetical Record

\textit{“Normal” ECG}  
\begin{itemize}
  \item No heart disease
  \item No Na-blocking drugs
  \item No abnormal K, Ca, Mg
\end{itemize}

Myocardial infarction  
Atrial fibrillation

Long QRS

Myocardial Infarction??  
Atrial Fibrillation??

Ritchie, Denny et al. Circulation 2013
GWAS of QRS Duration

SCN5A/SCN10A

$n=5,272$

Ritchie, Denny et al. Circulation 2013
PheWAS of rs6795970 (SCN10A) (longer QRS duration in normal hearts)

Ritchie et al. Circulation 2013

n=13,617

cardiac arrhythmias \(\rightarrow\) \(\leftarrow\) atrial fibrillation

- \(\log_{10}(p\text{-value})\)

1 disease codes
What happens to heart healthy population?

Followed \(n=5,272\) heart healthy population for developing atrial fibrillation based upon genotype.

Atrial fibrillation-free survival

HR=1.49 per G allele
\[p=0.001\]

Ritchie, Denny et al. Circulation 2013
Co-Existing & Unrelated Diseases

Phenome Drilling
Phenotype Neighborhood

All diagnoses co-occurring at least once with primary diagnosis in a given cohort

Size (outer) = number occurring
Size (inner) = number co-occurring
Color = ICD chapter

PheWAS does not do secondary association (gray lines)
Phenome Neighborhood Drilling

Goal
Isolate induced & syndromic phenotypes from unrelated diseases

Technique
Co-occurrence proportionality implies associative strength

Low co-occurrence
Common diagnosis

High co-occurrence
Uncommon diagnosis
Multiple Myeloma Phenome Drilling

Multiple myeloma is a plasma cell neoplasm characterized by bony destruction, infections, kidney damage, and anemia

Effective drugs prolong survival but introduce complications
### Dichotomous Features

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Multiple myeloma defined as:</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRQ CCS Group 40 multiple myeloma or</td>
<td></td>
</tr>
<tr>
<td>ICD-9-CM 203.1* plasma cell leukemia or</td>
<td></td>
</tr>
<tr>
<td>ICD-9-CM 238.6 neoplasm of uncertain behavior of plasma cells</td>
<td></td>
</tr>
</tbody>
</table>

**AHRQ single-level CCS diagnosis categories evaluated**

| Significance | One-sided exact binomial test, Bonferroni correction |

| Phenome | MIMIC II |

---

**Abbreviations**

AHRQ: Agency for Healthcare Research Quality  
CCS: Clinical Classifications Software
Multiple Myeloma Phenome Mapping

Warner, Alterovitz, et al. JAMIA 2013
# Partial List of MM Associations

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>O:E Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCS 207</td>
<td>Pathological fracture</td>
<td>30.7</td>
<td>6.02 x 10^{-17}</td>
</tr>
<tr>
<td>V42.81</td>
<td>Bone marrow replaced by transplant</td>
<td>53.4</td>
<td>1.20 x 10^{-14}</td>
</tr>
<tr>
<td>285.22</td>
<td>Anemia in neoplastic disease</td>
<td>24.8</td>
<td>2.10 x 10^{-14}</td>
</tr>
<tr>
<td>CCS 259</td>
<td>Residual codes; unclassified</td>
<td>5.8</td>
<td>4.82 x 10^{-13}</td>
</tr>
<tr>
<td>733.13</td>
<td>Pathologic fracture of vertebrae</td>
<td>23.4</td>
<td>3.90 x 10^{-12}</td>
</tr>
<tr>
<td>CCS 2617</td>
<td>E Codes: Adverse effects of medical drugs</td>
<td>8.2</td>
<td>5.67 x 10^{-10}</td>
</tr>
<tr>
<td>275.42</td>
<td>Hypercalcemia</td>
<td>13.9</td>
<td>3.70 x 10^{-05}</td>
</tr>
<tr>
<td>288</td>
<td>Neutropenia</td>
<td>21.4</td>
<td>4.40 x 10^{-05}</td>
</tr>
</tbody>
</table>

*Bold* = likely treatment-related complications  
Not bolded = likely disease-related or ambiguous

---

*Warner, Alterovitz, et al. JAMIA 2013*
## Continuous Feature Drilling

<table>
<thead>
<tr>
<th>Continuous Medical Phenomena</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory values</td>
<td>hematocrit</td>
</tr>
<tr>
<td>Time</td>
<td>acute/sub-acute/chronic scaling</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>heart rate</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>oncotype DX</td>
</tr>
</tbody>
</table>

*Warner, et al. JAMIA 2013*
White Blood Count (WBC)

WBC elevation can signify infection, neoplasia, and more

WBC components are also important

[Diagram showing different types of white blood cells and their normal ranges.]

# WBC Phenome Drilling

**Phenotype**: Maximum WBC during hospitalization

**Bounds**: WBC k/µl ≥ 15, 50, and 100

**Significance**: One-sided exact binomial test, Bonferroni correction

**Phenome**: MIMIC II
WBC Phenome Mapping

Significant Diagnoses Shown

ICD-9-CM Code

Patient peak value

75th percentile
50th percentile
25th percentile

Warner, Alterovitz AMIA Ann Symp 2012
**WBC Phenome Mapping**

- **ICD-9-CM Code**
  - **204.10** Chronic lymphoid leukemia (15-100 k/µl)
  - **038.9** Unspecified septicemia (10-45 k/µl)

*Warner, Alterovitz AMIA Ann Symp 2012*
Hospital-acquired complications *likely* to prolong length of stay and increase morbidity and mortality

Large scope of possible complications, *likely* time-dependent

To help determine patterns and characteristics of hospital-acquired complications, use **continuous feature exploration**
**Temporal Phenome Drilling**

<table>
<thead>
<tr>
<th>Length of stay</th>
<th>Direct/surrogate measures of care time</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Buckets”</td>
<td>200 patients (approx.) per interval</td>
</tr>
<tr>
<td>Significance</td>
<td>Fisher’s exact test, false discovery rate adjusted</td>
</tr>
<tr>
<td>Phenome</td>
<td>MIMIC II adults</td>
</tr>
</tbody>
</table>

*Warner, et al. JAMIA 2013*
Temporal Phenome Analysis

ICD-9-CM Code

Warner, et al. JAMIA 2013
Why visualize?

To see things worth exploring for potential phenotype connections!
Moving Forward

http://bit.ly/1ap7MKQ
SMART Genomics API

• Enables integration of genomic data from heterogeneous sources
• Enables integration of clinical and genomic data
• Re-usability eliminates developer barriers
• Promotes the use of genetic data for research

• Tutorial at:
  smartgenomics.wikispaces.com/SMART+Genomics+Tutorial
SMART Genomics Platform

[Diagram of SMART Genomics Platform]

- EHR System: Patient’s Clinical Data
- Personal Genomics System: Patient’s Genomic Data
- Pharmacogenomic Databases
- SMART API
- Cardiac Risk
- Cancer Risk
- BP Centiles
- Diabetes Monograph
- Genomics Advisor App
- Diabetes Monograph
  - Genomics Advisor Module
Genomics Advisor: Detailed Pop-Up
Patients with the AA genotype who are treated with statins may have a decreased, but not absent, risk for adverse cardiovascular events as compared to patients with the AG or GG genotype. Other genetic and clinical factors may also influence a patient's risk for adverse cardiovascular events.
SORRY, SON... THERE'S NO APP FOR THAT
Example App: Dbear SMART EMR

**Problem:** High cost of chronic disease management and issues with long-term patient compliance in diabetes, especially in pediatric cases.

**Solution:** Increase compliance by integrating toy bear from Sproutel with SMART EMR for physicians and to provide parent/child feedback.

**Details:** Integrated bear with the Telecare glucose meter / pump technology to provide clinical data to SMART EMR

* Provides kids with a “fun” way to measure glucose and take insulin. Provide live data for clinicians/patients to view.

* Interactive bear avatar gives feedback to patients.

* Engages patients to increases their likelihood to comply with their diabetes treatment.
Harvard Medical School (cbmi)
DB EMR

First program to integrate genomic, device, EMR/Personal Health Record information.

First program to integrate patient, patient devices (e.g. bear/glucose meter), care giver, and physician data into a unified view to facilitate collaboration on patient care.

First mobile app to integrate genomics/sequence information and clinical information.
**Description**

DB Medical Record approach to treatment diabetes compliance involves bringing together not only all of these stakeholders. DB Medical Record links children, parents, caregivers, and physicians together through a common avatar: a toy bear. It also links children to other children with the chronic condition (e.g. diabetes) again via a toy bear. The toy bear itself can educate and provide feedback to the child and caregivers/parents. The bear’s avatar appears is integrated into the medical record. The... More▼
Basic Information

Susie White

Last Known Measurement

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
<th>Date</th>
<th>Value</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>4.56</td>
<td>4/30/2013</td>
<td>8.76</td>
<td>4/22/2013</td>
</tr>
<tr>
<td>Tri</td>
<td>6.08</td>
<td>4/30/2013</td>
<td>7.25</td>
<td>4/22/2013</td>
</tr>
<tr>
<td>HDL</td>
<td>5.7</td>
<td>4/30/2013</td>
<td>8.32</td>
<td>4/22/2013</td>
</tr>
<tr>
<td>LDL</td>
<td>2.3</td>
<td>4/30/2013</td>
<td>2.6</td>
<td>4/22/2013</td>
</tr>
</tbody>
</table>

Condition

- Hyperdicarboxylicaminoaciduria: Past
- Diabetic ophthalmoplegia: Current
- Fracture of tooth: Past
- Headache disorder: Past
- High blood pressure: Current

Medication

- Ibuprofen * 600mg: 3 Tablets/Day

Allergy

- cat fur
- cough
- dog fur
- abdominal pain
- shrimp
- difficulty swallowing

Blood Glucose Plot

Renal diabetes: Grandmother (maternal)
Acute heart disease: Father
Basic Information

Susie White

Last Known Measurement

<table>
<thead>
<tr>
<th></th>
<th>4.56</th>
<th>8.76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol mmol/L</td>
<td>&lt;5.2</td>
<td>4/30/2013</td>
</tr>
<tr>
<td>Tri mmol/L</td>
<td>6.08</td>
<td>7.25</td>
</tr>
<tr>
<td>&lt;1.70</td>
<td>4/30/2013</td>
<td>4/22/2013</td>
</tr>
<tr>
<td>HDL mmol/L</td>
<td>5.7</td>
<td>8.32</td>
</tr>
<tr>
<td>&gt;1.03</td>
<td>4/30/2013</td>
<td>4/22/2013</td>
</tr>
<tr>
<td>LDL mmol/L</td>
<td>2.3</td>
<td>2.6</td>
</tr>
<tr>
<td>&lt;2.60</td>
<td>4/30/2013</td>
<td>4/22/2013</td>
</tr>
</tbody>
</table>

Allergy

cat fur
cough
dog fur
abdominal pain
shrimp
difficulty swallowing

Condition

Hyperdicarboxylicaminoaciduria Past
Diabetic ophthalmoplegia Current
Fracture of tooth Past
Headache disorder Past
High blood pressure Current

Medication

Ibuprofen*600mg 3 Tablets/Day

Blood Glucose Plot

Min: Day: Sun BG:10mmol/L

Family History

Comedocarcinoma Brother
High anal fistula Aunt

Blood Glucose (mmol/L)

1.05
0.67
1.03
0.91

DM1 DM2 HYP CHD
SMART Phenome Advisor

Software tool for patient-centric phenotype exploration

EMR provides patient’s phenotype data

Control and visualization capabilities enables user to

- Define cohorts
- Zoom in/out on phenotype aggregate/detail
- Perform dichotomous PheWAS
- Perform continuous feature phenome mapping
- Explore phenotype neighborhoods
SMART Phenome Advisor

Patient / View Selector

Classification Selector

Feature Selector

Feature Definition Selector

Drill Results Viewer
Phenome Browser Dashboard

Smith, John 50 years

<table>
<thead>
<tr>
<th>Patient Diagnoses</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>038.9 Unspecified septicemia</td>
<td>1899</td>
</tr>
<tr>
<td>263.9 Unspecified protein-calorie malnutrition</td>
<td>579</td>
</tr>
<tr>
<td>410.71 Subendocardial infarction</td>
<td>2009</td>
</tr>
<tr>
<td>425.4 Other primary cardiomyopathies</td>
<td>846</td>
</tr>
<tr>
<td>427.5 Cardiac arrest</td>
<td>753</td>
</tr>
</tbody>
</table>

Low Value: 1: Infectious Diseases
High Value: E: External Causes
Phenome Browser Dashboard

All ICD-9-CM Chapters

Low Value
1: Infectious Diseases

High Value
E: External Causes

Smith, John 50 years

<table>
<thead>
<tr>
<th>Patient Diagnoses</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>38 Septicemia</td>
<td>2084</td>
</tr>
<tr>
<td>260 Protein-calorie malnutrition</td>
<td>592</td>
</tr>
<tr>
<td>411.2 Myocardial infarction</td>
<td>5195</td>
</tr>
<tr>
<td>425.1 Primary/intrinsic cardiomyopathies</td>
<td>972</td>
</tr>
<tr>
<td>427.42 Cardiac arrest</td>
<td>755</td>
</tr>
</tbody>
</table>

ICD PheWAS CCS
Phenome Browser Dashboard

Low Value

1: Infectious Diseases

High Value

E: External Causes

Patient Diagnoses

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>038.9 Unspecified septicemia</td>
<td>1899</td>
</tr>
<tr>
<td>263.9 Unspecified protein-calorie malnutrition</td>
<td>579</td>
</tr>
<tr>
<td>410.71 Subendocardial infarction</td>
<td>2009</td>
</tr>
<tr>
<td>425.4 Other primary cardiomyopathies</td>
<td>846</td>
</tr>
<tr>
<td>427.5 Cardiac arrest</td>
<td>753</td>
</tr>
</tbody>
</table>
Phenotype Neighborhood

Primary Phenotype
- 410.71 Subendocardial infarction
- 427.5 Cardiac arrest

Linked Phenotypes (N=332)
- 414.01 Coronary atherosclerosis of native coronary artery - 55
- 428.0 Congestive heart failure, unspecified - 54
- 401.9 Unspecified essential hypertension - 33
- 997.1 Cardiac complications, NEC - 26
- 427.31 Atrial fibrillation - 25
Phenotype Neighborhood

Primary Phenotype
- 410.71 Subendocardial infarction
- 427.5 Cardiac arrest

Unspecified Septicemia

Linked Phenotypes (N=332)

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>414.01</td>
<td>Coronary atherosclerosis of native coronary artery</td>
<td>55</td>
</tr>
<tr>
<td>428.0</td>
<td>Congestive heart failure, unspecified</td>
<td>54</td>
</tr>
<tr>
<td>401.9</td>
<td>Unspecified essential hypertension</td>
<td>33</td>
</tr>
<tr>
<td>997.1</td>
<td>Cardiac complications, NEC</td>
<td>26</td>
</tr>
<tr>
<td>427.31</td>
<td>Atrial fibrillation</td>
<td>25</td>
</tr>
</tbody>
</table>
Phenotype Neighborhood

Linked Phenotypes
(N=332) # in Cohort

414.01 Coronary atherosclerosis of native coronary artery 55
428.0 Congestive heart failure, unspecified 54
401.9 Unspecified essential hypertension 33
997.1 Cardiac complications, NEC 26
427.31 Atrial fibrillation 25

Primary Phenotype
410.71 Subendocardial infarction
427.5 Cardiac arrest

Primary
N=88

038.9 Unspecified septicemia
Phenotype Neighborhood

Linked Phenotypes
(N=332) # in Cohort
414.01 Coronary atherosclerosis of native coronary artery 55
428.0 Congestive heart failure, unspecified 54
401.9 Unspecified essential hypertension 33
997.1 Cardiac complications, NEC 26
427.31 Atrial fibrillation 25

Primary Phenotype
410.71 Subendocardial infarction
427.5 Cardiac arrest

038.9 Unspecified septicemia

Paroxysmal ventricular tachycardia

Primary N=88
Phenotype Neighborhood

Primary Phenotype

410.71 Subendocardial infarction
427.5 Cardiac arrest

Linked Phenotypes

<table>
<thead>
<tr>
<th>Linked Phenotypes</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>414.01 Coronary atherosclerosis of native coronary artery</td>
<td>55</td>
</tr>
<tr>
<td>428.0 Congestive heart failure, unspecified</td>
<td>54</td>
</tr>
<tr>
<td>401.9 Unspecified essential hypertension</td>
<td>33</td>
</tr>
<tr>
<td>997.1 Cardiac complications, NEC</td>
<td>26</td>
</tr>
<tr>
<td>427.31 Atrial fibrillation</td>
<td>25</td>
</tr>
</tbody>
</table>

427.1 Paroxysmal ventricular tachycardia
Phenome Browser Dashboard

Low Value 1: Infectious Diseases

High Value E: External Causes

Smith, John 50 years

<table>
<thead>
<tr>
<th>Patient Diagnoses</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>038.9 Unspecified septicemia</td>
<td>1899</td>
</tr>
<tr>
<td>263.9 Unspecified protein-calorie malnutrition</td>
<td>579</td>
</tr>
<tr>
<td>410.71 Subendocardial infarction</td>
<td>2009</td>
</tr>
<tr>
<td>425.4 Other primary cardiomyopathies</td>
<td>846</td>
</tr>
<tr>
<td>427.5 Cardiac arrest</td>
<td>753</td>
</tr>
</tbody>
</table>
Phenome Browser Dashboard

Low Value
- All ICD-9-CM Chapters
- 1: Infectious Diseases

High Value
- E: External Causes

Patient Diagnoses

<table>
<thead>
<tr>
<th>ICD Code</th>
<th>Diagnosis</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>038.9</td>
<td>Unspecified septicemia</td>
<td>1899</td>
</tr>
<tr>
<td>263.9</td>
<td>Unspecified protein-calorie malnutrition</td>
<td>579</td>
</tr>
<tr>
<td>410.71</td>
<td>Subendocardial infarction</td>
<td>2009</td>
</tr>
<tr>
<td>425.4</td>
<td>Other primary cardiomyopathies</td>
<td>846</td>
</tr>
<tr>
<td>427.5</td>
<td>Cardiac arrest</td>
<td>753</td>
</tr>
</tbody>
</table>
Phenome-Wide Association

-\log(p\text{-value})

ICD-9-CM Code

Bonferroni Correction

V Codes

E Codes

427.41

785.51

997.1

414.01

428.0

348.1
Phenome Browser Dashboard

- **Low Value**
  - 1: Infectious Diseases

- **High Value**
  - E: External Causes

**ALL ICD-9 CHAPTERS**

- **CHAPTER 1 (001-139)**
  - Infectious and parasitic diseases

- **CHAPTER 2 (140-239)**
  - Neoplasms

**E CODES (E800-E999)**

- External causes of injury and supplemental classification

- Patient Diagnoses # in Cohort
  - 038.9 Unspecified septicemia 1899
  - 263.9 Unspecified protein-calorie malnutrition 579
  - 410.71 Subendocardial infarction 2009
  - 425.4 Other primary cardiomyopathies 846
  - 427.5 Cardiac arrest 753

**ICD PheWAS CCS**

- Smith, John 50 years
Phenome-Wide Association

-\log(p\text{-value})

ICD-9-CM Code

Points at ICD-9-CM Codes 427.41, 414.01, and 428.0.
Phenome Browser Dashboard

Low Value
- 390: Acute Rheumatic Fever

High Value
- 459: Veins & Lymphatics

Patient Diagnoses

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>038.9 Unspecified septicemia</td>
<td>1899</td>
</tr>
<tr>
<td>263.9 Unspecified protein-calorie malnutrition</td>
<td>579</td>
</tr>
<tr>
<td>410.71 Subendocardial infarction</td>
<td>2009</td>
</tr>
<tr>
<td>425.4 Other primary cardiomyopathies</td>
<td>846</td>
</tr>
<tr>
<td>427.5 Cardiac arrest</td>
<td>753</td>
</tr>
</tbody>
</table>
Phenome Browser Dashboard

4: Circulatory System

Low Value
390: Acute Rheumatic Fever

High Value
459: Veins & Lymphatics

Patient Diagnoses # in Cohort

<table>
<thead>
<tr>
<th>Patient Diagnosis</th>
<th># in Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>038.9 Unspecified septicemia</td>
<td>1899</td>
</tr>
<tr>
<td>263.9 Unspecified protein-calorie malnutrition</td>
<td>579</td>
</tr>
<tr>
<td>410.71 Subendocardial infarction</td>
<td>2009</td>
</tr>
<tr>
<td>425.4 Other primary cardiomyopathies</td>
<td>846</td>
</tr>
<tr>
<td>427.5 Cardiac arrest</td>
<td>753</td>
</tr>
</tbody>
</table>

Smith, John 50 years

ICD  PheWAS  CCS
**Phenotype Neighborhood**

**Linked Phenotypes**

(N=332) # in Cohort

- 414.01 Coronary atherosclerosis of native coronary artery: 55
- 428.0 Congestive heart failure, unspecified: 54
- 401.9 Unspecified essential hypertension: 33
- 427.31 Atrial fibrillation: 25
- 427.41 Ventricular fibrillation: 19

**Primary Phenotype**

- 410.71 Subendocardial infarction
- 427.5 Cardiac arrest
Continuous Feature Phenome Dashboard

**White blood cell count**

- **Low Value**
  - Percentile: 10%
  - Numeric: 10 k/µl

- **High Value**
  - Percentile: 90%
  - Numeric: 50 k/µl

**Parameters Trends**

- MIN
- MEAN
- MEDIAN
- MAX

- MIN
- MEAN
- MEDIAN
- MAX

**Trends**

- Flat
- Increasing
- Decreasing
Continuous Feature Phenome Dashboard

**White blood cell count**

- **Low Value**:
  - Percentile: 0%
  - Numeric: 0 k/µl
- **High Value**:
  - Percentile: 100%
  - Numeric: 12,500 k/µl

**Parameters Trends**

- MIN
- MEAN
- MEDIAN
- MAX

- Flat trend
- Increasing trend
- Decreasing trend
- Upward trend
- Downward trend
Continuous Feature Phenome Dashboard

White blood cell count

Percentile

Low Value: 0 %
High Value: 95 %

Numeric

Low Value: 0 k/µl
High Value: 100 k/µl

Parameters

MIN MEAN MEDIAN MAX

Trends

80
Continuous Feature Phenome Dashboard

### HEMATOLOGY
- White blood cell count
- Hematocrit
- Platelet count

### BLOOD CHEMISTRY
- Sodium
- Potassium
- Chloride

### URINE CHEMISTRY

**White blood cell count**

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Value</td>
<td>0%</td>
</tr>
<tr>
<td>High Value</td>
<td>100%</td>
</tr>
</tbody>
</table>
Continuous Feature Phenome Dashboard

White blood cell count

Percentile
Low Value 0 %
High Value 100 %

Numeric
Low Value 0 k/µl
High Value 12 500 k/µl

Parameters
- MIN
- MEAN
- MEDIAN
- MAX

Trends
- MIN
- MEAN
- MEDIAN
- MAX
Acknowledgments

People
Leo Celi (MIMIC II)
Daniel Scott (MIMIC II)
Peter Szolovits
Isaac Kohane
Kenneth Mandl
Dan Roden
Lisa Bastarache
Amin Zollanvari
Robin Joyce
Kelly Bodio

Graham Snyder
Quan Ding
Peijin Zhang
Federico Cismondi
Tom Lasko
Peter Yang

Grants
NIH 5R21DA025168-02 (GA)
NIH 1R01HG004836-01 (GA)
NIH 4R00LM009826-03 (GA)
ONC SHARP III
Questions???
Computational Considerations

<table>
<thead>
<tr>
<th>Task</th>
<th>CPU seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pull all ICD-9 counts from MIMIC II</td>
<td>3.460</td>
</tr>
<tr>
<td>Pull ICD-9 counts for primary phenotype</td>
<td>0.242</td>
</tr>
<tr>
<td>Create &amp; show phenotype neighborhood</td>
<td>1.274</td>
</tr>
<tr>
<td>Calculate &amp; show secondary adjacencies for 038.9</td>
<td>2.237</td>
</tr>
<tr>
<td>Calculate &amp; show secondary adjacencies for 427.1</td>
<td>2.227</td>
</tr>
<tr>
<td>Create &amp; show full PheWAS</td>
<td>0.263</td>
</tr>
<tr>
<td>Calculate full two-feature analysis</td>
<td>91.832</td>
</tr>
<tr>
<td>Display full two-feature analysis</td>
<td>28.642</td>
</tr>
</tbody>
</table>
# Computational Considerations

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>VirtualBox VM</td>
<td>OS: Ubuntu 12.04 LTS</td>
</tr>
<tr>
<td>4 Intel i5-2500k @ 3.30 GHz CPUs</td>
<td>R v. 2.15.3</td>
</tr>
<tr>
<td>4GB RAM</td>
<td>Rstudio v. 0.97.248</td>
</tr>
</tbody>
</table>
