High-Throughput Phenotyping from Electronic Health Records for Clinical and Translational Research

Jyotishman Pathak, PhD
Associate Professor of Biomedical Informatics
Department of Health Sciences Research

August 15th, 2013
Outline

• Clinical phenotyping from electronic health records (EHRs)
  • eMERGE
  • SHARPn

• Standards-based approaches to EHR phenotyping
  • NQF Quality Data Model
  • JBoss® Drools business rules environment
  • PhenotypePortal.org

• On-going and future work
Phenotyping is still a bottleneck…
### Exhibit 8: Changes in Adoption of Basic and Comprehensive EHRs

<table>
<thead>
<tr>
<th>Year</th>
<th>Any EHR</th>
<th>Basic EHR</th>
<th>Comprehensive EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>1.5%</td>
<td>7.2%</td>
<td>8.8%</td>
</tr>
<tr>
<td>2009</td>
<td>2.7%</td>
<td>9.2%</td>
<td>11.9%</td>
</tr>
<tr>
<td>2010</td>
<td>3.6%</td>
<td>11.5%</td>
<td>15.1%</td>
</tr>
<tr>
<td>2011</td>
<td>8.7%</td>
<td>18.0%</td>
<td>26.6%</td>
</tr>
</tbody>
</table>

Electronic health records (EHRs) driven phenotyping

- EHRs are becoming more and more prevalent within the U.S. healthcare system
  - Meaningful Use is one of the major drivers
- Overarching goal
  - To develop high-throughput semi-automated techniques and algorithms that operate on normalized EHR data to identify cohorts of potentially eligible subjects on the basis of disease, symptoms, or related findings both retrospectively and prospectively
EHR-driven Phenotyping Algorithms - I

• Typical components
  • Billing and diagnoses codes
  • Procedure codes
  • Labs
  • Medications
  • Phenotype-specific co-variates (e.g., Demographics, Vitals, Smoking Status, CASI scores)
  • Pathology
  • Radiology

• Organized into inclusion and exclusion criteria
EHR-driven Phenotyping Algorithms - II

1. Diabetes diagnosis (T1 or T2)
   - Yes → 2. DR/ME in Diagnoses or Problem Lists
     - DR/ME ICD9 Code: No → Exclude
     - Yes → 3. Negative Mention of DR/ME
       - No → Exclude
       - Yes → 4. Eye exam within past 2 years
         - No → Exclude
         - Yes → Control

2. DR/ME in Diagnoses or Problem Lists
   - Yes → Data

3. Negative Mention of DR/ME
   - No → Case

Evaluation

Visualization

Phenotype Algorithm

Transform

Transform

Data

Mappings

NLP, SQL
Example: Hypothyroidism Algorithm

No thyroid-altering medications (e.g., Phenytoin, Lithium)

ICD-9s for Hypothyroidism
Abnormal TSH/FT4
Antibodies for TTG or TPO (anti-thyroglobulin, anti-thyroperidase)

Thyroid replace. meds

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

Case 1
Case 2

2+ non-acute visits in 3 yrs

No ICD-9s for Hypothyroidism
No Abnormal TSH/FT4

No thyroid replace. meds

No Antibodies for TTG/TPO

No Hx of myasthenia gravis

Control

[Denny et al., ASHG, 2012; 89:529-542]
Example: Hypothyroidism Algorithm

**Case medications**
- Levothyroxine, levoacetate, levoxyl, thyroxine, liothyronine, synthroid, eltroxin
- Drugs

**Antibody lab tests**
- Anti-thyroglobulin antibodies: HTGA, ThyAB, ATThy- positive
- Anti-thyroperoxidase (TPO), TPO, ATThy- positive
- Anti-thyroid microsomal Ab- positive
- Labs

**ICD-9 codes for hypothyroidism**
- 244, 244.8, 244.9, 245, 245.2, 245.8, 245.9
- Diagnosis

**Abnormal lab values**
- TSH > 5 OR FT4 < 0.5

**Case Definition**
- **Case 1:**
  - ICD-9 code for hypothyroidism + abnormal TSH/FT4
  - Required at least 2 instances of either medication or lab with at least 3 months between the first and last instance of medication and lab

**Case 2:**
- Anti-thyroid, anti-thyroglobulin, OR anti-thyroperoxidase antibodies

**Case Exclusions**
- Exclude if the following information occurs in the record:
  - Secondary causes of hypothyroidism
  - Post surgical or post radiation hypothyroidism
  - Other thyroid diseases
  - Thyroid altering medication

**Case Exclusions Temporarily sensitive exclusions**
- Recent pregnancy TSH/FT4
- Recent contrast exposure

**Vitals**

**Labs**
- CPT codes for post radiation hypothyroidism
  - 77261, 77287, 77288, 77330, 77380

**Procedures**
- Post surgical or post radiation hypothyroidism
- Other thyroid diseases
- Thyroid altering medication

**Drugs**
- Levotiroxine, levoacetate, levoxyl, thyroxine, liothyronine, synthroid, eltroxin

**NLP**
- Temporarily sensitive exclusions
- Recent pregnancy TSH/FT4
- Recent contrast exposure

[Conway et al. AMIA 2011: 274-83]
<table>
<thead>
<tr>
<th>Phenotyping Algorithms</th>
<th>Data Categories used to define EHR-driven Phenotyping Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clinical gold standard</td>
</tr>
<tr>
<td></td>
<td>EHR-derived phenotype</td>
</tr>
<tr>
<td></td>
<td>Validation (PPV/NPV)</td>
</tr>
<tr>
<td></td>
<td>Sensitivity (Case/Cntrl)</td>
</tr>
<tr>
<td>Alzheimer’s Dementia</td>
<td>Demographics, clinical examination of mental status, histopathologic examination</td>
</tr>
<tr>
<td>Cataracts</td>
<td>Clinical exam finding (Ophthalmologic examination)</td>
</tr>
<tr>
<td>Peripheral Arterial Disease</td>
<td>Clinical exam finding (ankle-brachial index or arteriography)</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Laboratory Tests</td>
</tr>
<tr>
<td>Cardiac Conduction</td>
<td>ECG measurements</td>
</tr>
<tr>
<td></td>
<td>[eMERGE Network]</td>
</tr>
</tbody>
</table>
“EHR Depth” plays an important role

<table>
<thead>
<tr>
<th>Time frame of EMR data</th>
<th>Patients with ≥2 visits</th>
<th>Identified subjects (TP + FP), No.</th>
<th>TPs, No.</th>
<th>FP, No.</th>
<th>TNs, No.</th>
<th>FNs, No.</th>
<th>PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007 (1 year)</td>
<td>74,212</td>
<td>2970</td>
<td>2089</td>
<td>881</td>
<td>50,632</td>
<td>681</td>
<td>70%</td>
</tr>
<tr>
<td>2006–2007 (2 years)</td>
<td>82,679</td>
<td>3280</td>
<td>2366</td>
<td>914</td>
<td>50,599</td>
<td>404</td>
<td>72%</td>
</tr>
<tr>
<td>2005–2007 (3 years)</td>
<td>83,792</td>
<td>3374</td>
<td>2466</td>
<td>908</td>
<td>50,605</td>
<td>304</td>
<td>73%</td>
</tr>
<tr>
<td>2004–2007 (4 years)</td>
<td>84,326</td>
<td>3273</td>
<td>2550</td>
<td>723</td>
<td>50,790</td>
<td>220</td>
<td>78%</td>
</tr>
<tr>
<td>2003–2007 (5 years)</td>
<td>84,617</td>
<td>3051</td>
<td>2616</td>
<td>435</td>
<td>51,078</td>
<td>154</td>
<td>86%</td>
</tr>
<tr>
<td>2002–2007 (6 years)</td>
<td>84,788</td>
<td>2967</td>
<td>2650</td>
<td>317</td>
<td>51,196</td>
<td>120</td>
<td>89%</td>
</tr>
<tr>
<td>2001–2007 (7 years)</td>
<td>84,903</td>
<td>2936</td>
<td>2692</td>
<td>244</td>
<td>51,269</td>
<td>78</td>
<td>92%</td>
</tr>
<tr>
<td>2000–2007 (8 years)</td>
<td>84,993</td>
<td>2919</td>
<td>2721</td>
<td>198</td>
<td>51,315</td>
<td>49</td>
<td>93%</td>
</tr>
<tr>
<td>1999–2007 (9 years)</td>
<td>85,072</td>
<td>2858</td>
<td>2743</td>
<td>115</td>
<td>51,398</td>
<td>27</td>
<td>96%</td>
</tr>
<tr>
<td>1998–2007 (10 years)</td>
<td>85,125</td>
<td>2768</td>
<td>2755</td>
<td>13</td>
<td>51,500</td>
<td>15</td>
<td>99.5%</td>
</tr>
<tr>
<td>1997–2007 (gold standard)</td>
<td>85,172</td>
<td>2770</td>
<td>2770</td>
<td>0</td>
<td>51,513</td>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

[Wei et al. IJIM 2012 (Epub ahead of print)]
Genotype-Phenotype Association Results

<table>
<thead>
<tr>
<th>disease</th>
<th>marker</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>

Odds Ratio

[Ritchie et al. AJHG 2010; 86(4):560-72]
Key lessons learned from eMERGE

- Algorithm design and transportability
  - Non-trivial; requires significant expert involvement
  - Highly iterative process
  - Time-consuming manual chart reviews
  - Representation of “phenotype logic” is critical

- Standardized data access and representation
  - Importance of unified vocabularies, data elements, and value sets
  - Questionable reliability of ICD & CPT codes (e.g., billing the wrong code since it is easier to find)
  - Natural Language Processing (NLP) is critical

Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project

Susan Rea\textsuperscript{a,\ast}, Jyotishman Pathak\textsuperscript{b}, Guergana Savova\textsuperscript{c}, Thomas A. Oniki\textsuperscript{d}, Les Westberg\textsuperscript{e}, Calvin E. Beebe\textsuperscript{b}, Cui Tao\textsuperscript{b}, Craig G. Parker\textsuperscript{a}, Peter J. Haug\textsuperscript{a,f}, Stanley M. Huff\textsuperscript{d,f}, Christopher G. Chute\textsuperscript{b}
Algorithm Development Process - Modified

- Standardized and structured representation of phenotype definition criteria
- Use the NQF Quality Data Model (QDM)

Rules

- Conversion of structured phenotype criteria into executable queries
- Use JBoss® Drools (DRLs)

Semi-Automatic Execution

- Standardized representation of clinical data
- Create new and re-use existing clinical element models (CEMs)

[Welch et al., JBI 2012; 45(4):763-71]
The SHARPn “phenotyping funnel”

Phenotype specific patient cohorts

[Welch et al., JBI 2012; 45(4):763-71]
SHARP Data Normalization Architecture
SHARPn data normalization architecture - II

Intermountain Healthcare (IHC) EHR

Mirth Connect

IHC NwHIN Aurion Gateway

SHARP NwHIN Aurion Gateway

Mirth Connect

Mayo Clinic EHR Systems

UIMA Pipeline

CFM Instance Database

Phenotype Rules

CEM CouchDB database with standardized and normalized patient data

[Welch et al., JBI 2012; 45(4):763-71]
Algorithm Development Process - Modified

- Standardized and structured representation of phenotype definition criteria
- Use the NQF Quality Data Model (QDM)

Semi-Automatic Execution

- Standardized representation of clinical data
- Create new and re-use existing clinical element models (CEMs)

[Welch et al., JBI 2012; 45(4):763-71]
Example algorithm: Hypothyroidism

<table>
<thead>
<tr>
<th>Variables of Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case ICD 9 codes</td>
</tr>
<tr>
<td>244 acquired hypothyroidism</td>
</tr>
<tr>
<td>244.8 acquired hypothyroidism NEC</td>
</tr>
<tr>
<td>244.9 hypothyroidism NOS</td>
</tr>
<tr>
<td>245.2 chronic lymphocytic thyroiditis</td>
</tr>
<tr>
<td>245.8 chronic thyroiditis NEC/NOS</td>
</tr>
<tr>
<td>245.9 thyroiditis NOS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case lab names/values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothyroidism: TSH &gt;5 or FT4 &lt;0.5</td>
</tr>
<tr>
<td>Anti-thyroglobulin antibodies: H-TGA, ThyAB, ATDy - positive</td>
</tr>
<tr>
<td>Anti-thyroidperoxidase: H-TPO, TPO, ATP - positive</td>
</tr>
<tr>
<td>Anti-thyroid antibodies: ThyAb - positive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levothyroxine, synthroid, Levoxyl, unitroid, armour thyroid, desiccated thyroid, cymoxanil, tacrolix, levothyroxine, levothyroxine, T3 and T4</td>
</tr>
</tbody>
</table>

*Optional depending on sample size. Will likely require a standard dosage following them to distinguish from lab tests when using NLP to identify.*

<table>
<thead>
<tr>
<th>Control lab names/values</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSH must be between 0.5 – 5</td>
</tr>
<tr>
<td>FT4 must be between 0.5–1.2 (if checked)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case/Control thyroid disease exclusion ICD 9 codes (if present, cannot be a case or a control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>199* thyroid cancer, all types</td>
</tr>
<tr>
<td>242.0 toxic diffuse goiter</td>
</tr>
<tr>
<td>242.1 toxic multinodular goiter 1/27/2009</td>
</tr>
<tr>
<td>242.2 toxic multinodular goiter</td>
</tr>
<tr>
<td>242.3 toxic nodular goiter, unspecified</td>
</tr>
<tr>
<td>242.4 hyperthyroidism not goiter or other cause</td>
</tr>
<tr>
<td>242.5 hypothyroidism not goiter or other cause</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control exclusion ICD9 codes (if present, cannot be a control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>240.2 Simple and unspecified goiter</td>
</tr>
<tr>
<td>241.2 Nontoxic nodular goiter</td>
</tr>
<tr>
<td>242.1 Thyrotoxicosis with or without goiter</td>
</tr>
<tr>
<td>243.2 Congenital hypothyroidism</td>
</tr>
<tr>
<td>244.2 Acquired hypothyroidism</td>
</tr>
<tr>
<td>245.1 Thyrotoxicosis, not elsewhere classified</td>
</tr>
</tbody>
</table>
NQF Quality Data Model (QDM)

- Standard of the National Quality Forum (NQF)
  - A structure and grammar to represent quality measures and phenotype definitions in a standardized format
- Groups of codes in a code set (ICD-9, etc.)
  - "Diagnosis, Active: steroid induced diabetes" using "steroid induced diabetes Value Set GROUPING (2.16.840.1.113883.3.464.0001.113)"
- Supports temporality & sequences
  - AND: "Procedure, Performed: eye exam" > 1 year(s) starts before or during "Measurement end date"
- Implemented as a set of XML schemas
  - Links to standardized terminologies (ICD-9, ICD-10, SNOMED-CT, CPT-4, LOINC, RxNorm etc.)
Example: Diabetes & Lipid Mgmt. - I

Diabetes Measure Pair: A Lipid management: low density lipoprotein cholesterol (LDL-C) <130, B Lipid management: LDL-C <100

Summary

<table>
<thead>
<tr>
<th>NQF #</th>
<th>0064</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>Diabetes Measure Pair: A Lipid management: low density lipoprotein cholesterol (LDL-C) &lt;130, B Lipid management: LDL-C &lt;100</td>
</tr>
<tr>
<td>Project Name:</td>
<td>National Voluntary Consensus Standards for Ambulatory Care- Part 1 (Phase 3 Cycle 1)</td>
</tr>
<tr>
<td>Status:</td>
<td>Endorsed</td>
</tr>
<tr>
<td>Original Endorsement Date:</td>
<td>AUG 10, 2009</td>
</tr>
<tr>
<td>Most Recent Endorsement Date:</td>
<td>AUG 10, 2009</td>
</tr>
<tr>
<td>Steward(s):</td>
<td>National Committee for Quality Assurance</td>
</tr>
<tr>
<td>Description:</td>
<td>Percentage of adult patients with diabetes aged 18-75 years with most recent (LDL-C) &lt;130 mg/dL B: Percentage of patients 18-75 years of age with diabetes whose most recent LDL-C test result during the measurement year was &lt;100 mg/dL</td>
</tr>
</tbody>
</table>
Example: Diabetes & Lipid Mgmt. - II

Population criteria

- Initial Patient Population =
  - AND: "Patient characteristic: birth date" >= 17 year(s) and <= 74 year(s) starts before start of "Measurement period"

- Denominator =
  - AND: "Initial Patient Population"
  - AND:
    - OR:
      - AND:
        - OR: "Encounter: Encounter acute inpatient or ED"
        - OR:
          - AND: >= 2 count(s) of
            - AND: "Encounter: Encounter non-acute inpatient and outpatient"
            - AND: FIRST:"Encounter: Encounter non-acute inpatient and outpatient" starts before start of SECOND" "Encounter: Encounter non-acute inpatient and outpatient"
          - AND: "Diagnosis active: diabetes"
        - OR:
          - OR: "Medication order: Medications indicative of diabetes"
          - OR: "Medication dispensed: Medications indicative of diabetes"
          - OR: "Medication active: Medications indicative of diabetes"
        - <= 2 year starts before or during "Measurement end date"
Example: Diabetes & Lipid Mgmt. - III

Data criteria (QDS Data Elements)

- "Diagnosis active: diabetes" using "diabetes Code List GROUPING (2.16.840.1.113883.3.464.0001.37)"
- "Diagnosis active: gestational diabetes" using "gestational diabetes Code List GROUPING (2.16.840.1.113883.3.464.0001.67)"
- "Diagnosis active: polycystic ovaries" using "polycystic ovaries Code List GROUPING (2.16.840.1.113883.3.464.0001.98)"
- "Diagnosis active: steroid induced diabetes" using "steroid induced diabetes Code List GROUPING (2.16.840.1.113883.3.464.0001.113)"
- "Encounter: Encounter acute inpatient or ED" using "Encounter acute inpatient or ED Code List GROUPING (2.16.840.1.113883.3.464.0001.42)"
- "Encounter: Encounter non-acute inpatient and outpatient" using "Encounter non-acute inpatient and outpatient Code List GROUPING (2.16.840.1.113883.3.464.0003.1142)"
- "Laboratory test result: High Density Lipoprotein (HDL)" using "High Density Lipoprotein (HDL) Code List GROUPING (2.16.840.1.113883.3.464.0001.76)"
- "Laboratory test result: LDL test" using "LDL test Code List GROUPING (2.16.840.1.113883.3.464.0001.89)"
- "Laboratory test result: Total Cholesterol" using "Total Cholesterol Code List GROUPING (2.16.840.1.113883.3.464.0001.124)"
- "Laboratory test result: Triglycerides" using "Triglycerides Code List GROUPING (2.16.840.1.113883.3.464.0001.132)"
- "Medication active: Medications indicative of diabetes" using "Medications indicative of diabetes Code List GROUPING (2.16.840.1.113883.3.464.0001.94)"
- "Medication dispensed: Medications indicative of diabetes" using "Medications indicative of diabetes Code List GROUPING (2.16.840.1.113883.3.464.0001.94)"
- "Medication order: Medications indicative of diabetes" using "Medications indicative of diabetes Code List GROUPING (2.16.840.1.113883.3.464.0001.94)"
- "Patient characteristic: birth date" (age) using "birth date HL7 Code List (2.16.840.1.113883.3.464.0001.14)"
## Example: Diabetes & Lipid Mgmt. - IV

<table>
<thead>
<tr>
<th>standard OID</th>
<th>standard concept</th>
<th>standard taxonomy</th>
<th>code</th>
<th>descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.16.840.1.113883.3.464.0001.94</td>
<td>Medications indicative of diabetes</td>
<td>GROUPING</td>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>&quot;Medications indicative of diabetes&quot; RxNorm code list</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.94</td>
<td>Medications indicative of diabetes</td>
<td>GROUPING</td>
<td>2.16.840.1.113883.3.464.0001.06</td>
<td>&quot;Medications indicative of diabetes&quot; RxNorm code list</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.94</td>
<td>Medications indicative of diabetes</td>
<td>GROUPING</td>
<td>2.16.840.1.113883.3.464.0001.07</td>
<td>&quot;Medications indicative of diabetes&quot; RxNorm code list</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>standard OID</th>
<th>standard concept</th>
<th>code</th>
<th>descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidas inhibitors</td>
<td>RxNorm</td>
<td>199150</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidas inhibitors</td>
<td>RxNorm</td>
<td>200132</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidas inhibitors</td>
<td>RxNorm</td>
<td>205329</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidas inhibitors</td>
<td>RxNorm</td>
<td>205330</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidas inhibitors</td>
<td>RxNorm</td>
<td>205331</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidas inhibitors</td>
<td>RxNorm</td>
<td>401938</td>
</tr>
</tbody>
</table>
Population Criteria Section: denominator

```
--> <entry typeCode="DRIV">
  <observation classCode="OBS" moodCode="EVN.CRT" isCriterionInd="true">
    <id root="655E8FF4-0530-4D39-A188-6F95D24CF2FC"/>
    <code code="ASSERTION" codeSystem="2.16.840.1.113883.5.4"/>
    <value xsi:type="CD" codeSystem="2.16.840.1.113883.5.1063" codeSystemName="HL7 Observation Value" display="Denominator"/>
    <sourceOf typeCode="PRCN">
      <conjunctionCode code="AND"/>
      <observation classCode="OBS" moodCode="EVN" isCriterionInd="true">
        <id root="876DB255-7D1D-4968-AE98-3E3FA2D58CD9"/>
        <title>Initial Patient Population</title>
      </observation>
    </sourceOf>
    <sourceOf typeCode="PRCN">
      <conjunctionCode code="AND"/>
      <act classCode="ACT" moodCode="EVN" isCriterionInd="true">
        <sourceOf typeCode="PRCN">
          <conjunctionCode code="OR"/>
          <act classCode="ACT" moodCode="EVN" isCriterionInd="true">
            <sourceOf typeCode="PRCN">
              <conjunctionCode code="AND"/>
              <act classCode="ACT" moodCode="EVN" isCriterionInd="true">
                <sourceOf typeCode="PRCN">
                  <conjunctionCode code="OR"/>
                  <act classCode="ACT" moodCode="EVN" isCriterionInd="true">
                    <sourceOf typeCode="PRCN">
                      <conjunctionCode code="OR"/>
                      <act classCode="ACT" moodCode="EVN" isCriterionInd="true">
                        <sourceOf typeCode="PRCN">
                          <conjunctionCode code="OR"/>
                          <act classCode="ACT" moodCode="EVN" isCriterionInd="true">
                            <!-- encounter pattern -->
                            <templateId root="2.16.840.1.113883.3.560.1.4"/>
                            <id root="673A2145-D199-4601-9F90-765EC420166A"/>
                          </act>
                        </sourceOf>
                      </conjunctionCode>
                    </act>
                  </act>
                </sourceOf>
              </act>
            </sourceOf>
          </act>
        </sourceOf>
      </act>
    </sourceOf>
  </observation>
</entry>
```
# Algorithm Table

<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>ORS 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>

[Thompson et al., AMIA 2012; (Epub ahead of print)]
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Uses NLP</th>
<th>Term Negation</th>
<th>No Evidence Of</th>
<th>Document Section Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic retinopathy</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Height</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Serum lipid level</td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Low HDL cholesterol level</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Peripheral arterial disease</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>QRS duration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Resistant hypertension</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cataract</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

[Thompson et al., AMIA 2012; (Epub ahead of print)]
Algorithm Development Process - Modified

- Standardized and structured representation of phenotype definition criteria
- Use the NQF Quality Data Model (QDM)

Rules

- Conversion of structured phenotype criteria into executable queries
  - Use JBoss® Drools (DRLs)

Semi-Automatic Execution

- Standardized representation of clinical data
  - Create new and re-use existing clinical element models (CEMs)

Phenotype Algorithm

- Use the NQF Quality Data Model (QDM)

Mappings

- Conversion of structured phenotype criteria into executable queries
  - Use JBoss® Drools (DRLs)

Data

- Standardized representation of clinical data
  - Create new and re-use existing clinical element models (CEMs)

NLP, SQL

- Conversion of structured phenotype criteria into executable queries
  - Use JBoss® Drools (DRLs)

[ Welch et al., JBI 2012; 45(4):763-71]
JBoss® Drools rules management system

- Represents knowledge with declarative production rules
  - Origins in artificial intelligence expert systems
- Simple when <pattern> then <action> rules specified in text files
- Separation of data and logic into separate components
- Forward chaining inference model (Rete algorithm)
- Domain specific languages (DSL)
Example Drools rule

rule “Glucose <= 40, Insulin On”

when

$msg : GlucoseMsg(glucoseFinding <= 40, currentInsulinDrip > 0 )

then

glucoseProtocolResult.setInstruction(GlucoseInstructions.GLUCOSE_LESS_THAN_40_INSULIN_ON_MSG);

end
Modeling and Executing Electronic Health Records Driven Phenotyping Algorithms using the NQF Quality Data Model and JBoss® Drools Engine

Dingcheng Li, PhD¹, Gopu Shrestha, MS¹, Sahana Murthy, MS¹ Davide Sottara, PhD² Stanley M. Huff, MD³ Christopher G. Chute, MD, DrPH¹ Jyotishman Pathak, PhD¹
¹Mayo Clinic, Rochester, MN ²University of Bologna, Italy ³Intermountain Healthcare, Salt Lake City, UT

[Li et al., AMIA 2012; (Epub ahead of print)]
The “executable” Drools workflow

[Li et al., AMIA 2012; (Epub ahead of print)]
What is the Phenotype Portal?

Phenotyping is the process of identifying a cohort of patients based on certain diseases, symptoms or clinical findings. The Phenotype Portal is a tool funded by the SHARPn Project from the Office of the National Coordinator (ONC). It will enable clinicians and investigators to identify patient cohorts using electronic health record (EHR) data by leveraging informatics-based phenotyping processes. In turn, these cohorts will facilitate clinical trial enrollment, outcomes research, and inform clinical decision support. Currently, the field has various barriers in technological research and tool development, and Phenotype Portal is the first such platform for generating and executing Meaningful Use standards-based phenotyping algorithms that can be shared across multiple institutions and investigators.

Traditionally, a patient's medical information is stored inconsistently and in multiple locations, both electronically and non-electronically. The Phenotype Portal will work towards creating a unified framework for normalizing and standardizing clinical data, which will allow for the exchange of patient information among care providers.
PhenotypePortal Architecture
PhenotypePortal Demo

• http://107.20.139.191:8080/htp2014/
On-going/Future planned activities: 2013-14

- Measure Authoring Toolkit (MAT) integration
  - Seamless authoring capabilities
- API decoupling/additional REST interfaces
  - Invocation of services w/o UI client
- Improve scalability and performance
  - No need to wait for long batch processing as the execution can be done in real time. As a result, multiple translators can be distributed to handle large patient sets.
  - Batch processing can be done and code has been written to integrate with Hadoop to parallelize execution across large patient sets.
  - Additional optimizations are still possible, such as caching, etc.
Sustainability plan

- NIH funded R01 (July 2013—June 2017)
  - National Infrastructure for EHR-driven phenotyping algorithm modeling and execution (PI: Pathak)
  - Expand HTP infrastructure to additional workflow engines (KNIME), repositories (i2b2), and analytical platforms (R)
- Mayo Clinical Informatics Program (Jan 2013-)
  - Library of Cohort Definitions for use in Mayo Clinic practice and research
  - Real-time phenotype/cohort dashboard
Concluding remarks

• EHRs contain a wealth of phenotypes for clinical and translational research
• EHRs represent real-world data, and hence has challenges with interpretation, wrong diagnoses, and compliance with medications
  • Handling referral patients even more so
• Standardization and normalization of clinical data and phenotype definitions is critical
• Phenotyping algorithms are often transportable between multiple EHR settings
  • Validation is an important component
Acknowledgment

- Mayo HTP team
- Intermountain HTP team
- Erin Martin
Thank You!
Back-Up
1. Converts QDM to Drools
2. Rule execution by querying the CEM database
3. Generate summary reports
Chronic stable coronary artery disease: lipid control

Date range for the algorithm:
From: Aug 15 1995
To: May 15 2012

Execute

File Info Criteria Summary Demographics Workflow

Graph Option:
○ Pie Chart
○ Column Chart

Summary Chart

Numerator Population: 1559
Denominator Population: 1921
Initial Patient Population: 1921
Exception Population: 1872

Phenotypes

 Algorithms

- Create Phenotype
- Upload Phenotype

Phenotypes

- Disease of the skin and subcutaneous tissue
- Diseases of the blood and blood forming organs
- Diseases of the circulatory system (3)
  - Acute rheumatic fever
  - Cerebrovascular disease
  - Chronic rheumatic heart disease
- Diseases of arteries, arterioles, and capillaries
- Diseases of pulmonary circulation (1)
- Diseases of veins and lymphatics
- Hypertensive disease (1)
- Ischemic heart disease (1)
  - Acute myocardial infarction
  - Angina pectoris
- Other forms of ischemic heart disease (1)
- Chronic stable coronary artery disease: lipid control
  - Other diseases of circulatory system
  - Diseases of the digestive system
  - Diseases of the genitourinary system
  - Diseases of the musculoskeletal system
  - Diseases of the nervous system
  - Diseases of the respiratory system (1)
  - Endocrine, nutritional and metabolic diseases (2)
  - Infectious and parasitic diseases
  - Mental and behavioral disorders
  - Neoplasms
Translator

• Distinguishing Features
  • Operates on real-time patient set and/or batch processing
  • Drools rules could be reused and combined
  • Rules can give immediate feedback as a new information is known as new lab results or patient visits are added to the data.
  • Easier integration into clinical setting because the translator can work on observed changes in data instead of only scheduling batch processes. Clients would not have to trigger algorithm execution as this execution would be triggered when new data is encountered. This would minimize points of integration.
  • Open source license, built on established projects such as PopHealth and Cypress.