Engineering a Learning Healthcare System: The Role of Data Science

Jyotishman Pathak, PhD
Director, Clinical Informatics Program and Services
Kern Center for the Science of Health Care Delivery
Associate Professor of Biomedical Informatics
Department of Health Sciences Research

October 8th, 2014
Learning Health Systems Seminar
University of Utah, Salt Lake City
DISCLOSURES

Relevant Financial Relationship(s)
None

Off Label Usage
None
Our health care system: Today

Less than 50% of elderly patients are up to date on clinical preventive services

Elderly patients with co-morbidities require up to 19 medication doses daily

Every year the average elderly patient sees 7 doctors across 4 practices

Average surgery patient is seen by 27 different health care providers

1 out of 5 elderly patients are readmitted within 30 days

Less than half of non-surgical patients follow-up with their primary care provider after discharge

Preventive

Science

Insights poorly managed

Self Management

Evidence

Evidence poorly used

Outpatient Care

Care

Experience poorly captured

Hospital

Follow-Up

Patient Experience

Missed Opportunities, Waste, and Harm
Information Challenges in the U.S. Health Care System

• A patient's vital medical information is scattered across medical records kept by many different caregivers in many different locations – and all of the patient’s medical information is unavailable at the time of care

• Clinicians do not always have ready access to complete information about their patients; do not know how other doctors are treating their same patients; or how other healthcare providers around the country treat patients with the same condition

[Bush, 2004]
Information Challenges in the U.S. and World Healthcare System

• Clinicians do not always have the best **information** to select the best treatments for their patients, resulting in an unacceptable lag time before new scientific advances are used in patient care.

• Consumers do not always have access to useful, credible health **information** about treatment alternatives, which hospitals and physicians are best for their needs, or their own health status.

[Bush, 2004]
[your institution’s name]

Health Care Value Equation

\[
\text{QUALITY} + \text{SATISFACTION} + \text{EFFICIENCY} = \text{COST}
\]
Our Health Care System: Tomorrow

- Real-time access to knowledge
- Digital capture of care experience
- Engaged, empowered patients
- Incentives aligned for value
- Leadership-instilled culture of learning

Continuous Learning, Best Care, Lower Cost
10 Recommendations

Foundational elements
1. The digital infrastructure – Improve the capacity to capture clinical, delivery process, and financial data for better care, system improvement, and creating new knowledge.
2. The data utility – Streamline and revise research regulations to improve care, promote the capture of clinical data, and generate knowledge.

Care improvement targets
3. Clinical decision support
4. Patient-centered care
5. Community links
6. Care continuity
7. Optimized operations
8. Supportive policy environment
9. Financial incentives
10. Performance transparency
11. Broad leadership

1. The digital infrastructure – Improve the capacity to capture clinical, delivery process, and financial data for better care, system improvement, and creating new knowledge.
2. The data utility – Streamline and revise research regulations to improve care, promote the capture of clinical data, and generate knowledge.
This requires…

- Understanding information modeling
- Understanding health IT standards
- Understanding electronic health records
- Understanding advanced analytics
- Understanding clinical decision support
- Understanding data visualization

Themes and research areas in Biomedical Informatics and Data Science
This requires…

• Understanding information modeling
• Understanding health IT standards
• Understanding electronic health records
• Understanding advanced analytics
• Understanding clinical decision support
• Understanding data visualization

My focus for today
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
   - Exclude

2. DR/ME in Diagnoses or Problem Lists
   - DR/ME ICD9 Code
   - Yes
   - Case
   - No
   - DR/ME Problem List
   - No
   - Exclude
   - Yes

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

4. Eye exam within past 2 years
   - Yes
   - Control
   - No
   - Exclude

Phenotype Algorithm

Transform

Transform

Data

Mappings

NLP, SQL

Evaluation

Visualization

[eMERGE Network]
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)

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 Definition**

**Case 1:**
- ICD-9 code for hypothyroidism OR abnormal TSH/FT4
- Thyroid replacement medication use
- Require 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-thyeroxidase 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

**Antibody lab tests**
- Anti-thyroglobulin antibodies: H-TGA, ThyAB, ATHyg - positive
- Anti-thyroxidase: H-MPO, TPO, ATHyP - positive
- Anti-thyroid antibodies: ThyAb – positive

**Pregnancy exclusion ICD 9 codes**
(if present with abnormal TSH or FT4 within six months before pregnancy to one year after pregnancy cannot be a case)

**Conway et al. AMIA 2011: 274-83**
Example: Hypothyroidism Algorithm

Case Definitions:
- Case 1:
  - ICD-9 code for hypothyroidism
  - Abnormal lab values: TSH > 5 OR FT4 < 0.5
- Case 2:
  - Anti-thyroid, anti-thyroglobulin, OR anti-thyroperoxidase antibodies

Case Exclusions:
- 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

Diagnosis:
- Hypothyroidism
- Secondary causes of hypothyroidism
- Post surgical or post radiation hypothyroidism
- Other thyroid diseases
- Thyroid altering medication

Labs:
- Anti-thyroglobulin antibodies: H-TGA, ThyAB, ATHyg - positive
- Anti-thyroperoxidase: TPO, TPO, ATHyg - positive

Drugs:
- Levotiroxine, levoxil
- Synthroid, levothyroxine
- Thyroxine, liothyronine

Procedures:
- Operation, biopsy, thyroidectomy
- Radiation therapy

NLP:
- Natural Language Processing
- Text mining

Vitals:
- Blood pressure, heart rate, respiratory rate

[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>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>
</tbody>
</table>

[eMERGE Network]
“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>FPs, 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>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial fibrillation</td>
<td>rs2200733</td>
<td>Chr. 4q25</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>rs10033464</td>
<td>Chr. 4q25</td>
<td>5.0</td>
</tr>
<tr>
<td>rs11805303</td>
<td>IL23R</td>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td>rs17234657</td>
<td>Chr. 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1000113</td>
<td>Chr. 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs17221417</td>
<td>NOD2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2542151</td>
<td>PTPN22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>rs3135388</td>
<td>DRB1*1501</td>
<td></td>
</tr>
<tr>
<td>rs2104286</td>
<td>IL2RA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6897932</td>
<td>IL7RA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>rs6457617</td>
<td>Chr. 6</td>
<td></td>
</tr>
<tr>
<td>rs6679677</td>
<td>RSBN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2476601</td>
<td>PTPN22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>rs4506565</td>
<td>TCF7L2</td>
<td></td>
</tr>
<tr>
<td>rs12255372</td>
<td>TCF7L2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12243326</td>
<td>TCF7L2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10811661</td>
<td>CDKN2B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs8050136</td>
<td>FTO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs5219</td>
<td>KCNJ11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs5215</td>
<td>KCNJ11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs4402960</td>
<td>IGF2BP2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>rs3135388</td>
<td>DRB1*1501</td>
<td></td>
</tr>
<tr>
<td>rs2104286</td>
<td>IL2RA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6897932</td>
<td>IL7RA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[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

Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium

Algorithm Development Process - Modified

- Standardized and structured representation of phenotype definition criteria
- Use the NQF Quality Data Model (QDM)
- Conversion of structured phenotype criteria into executable queries
  - Use JBoss® Drools (DRLs)
- Standardized representation of clinical data
  - Create new and re-use existing clinical element models (CEMs)

Rules

Mappings
NLP, SQL
Data

Phenotype Algorithm

Semi-Automatic Execution

Visualization

Evaluation

[Welch et al., JBI 2012; 45(4):763-71]
[Pathak et al., JAMIA 2013; 20(e2): e341-8]
The SHARPn “phenotyping funnel”

Phenotype specific patient cohorts

Mayo Clinic EHR data → QDMs → CEMs → DRLs → Phenotype specific patient cohorts → Intermountain EHR data

[Welch et al., JBI 2012; 45(4):763-71]
[Pathak et al., JAMIA 2013; 20(e2): e341-8]
Clinical Element Models
Higher-Order Structured Representations

BloodPressurePanel
key BloodPressure
items
SystolicBloodPressure SystolicBP
  data 120 mmHg
DiastolicBloodPressure DiastolicBP
  data 80 mmHg
quals
BodyPostion BodyPosition
  data Sitting

[Stan Huff, IHC]
CEMs available for patient demographics, medications, lab measurements, procedures etc.

**Medication CEM template**
- associatedCode
- Change_status
- Conditional
- Dosage
- Duration
- End_date
- Form
- Frequency
- Generic
- Negation_indicator
- Route
- Start_date
- Strength
- Subject
- Uncertainty_indicator

**Sign/Symptom CEM template**
- Alleviating_factor
- associatedCode
- Body_laterality
- Body_location
- Body_side
- Conditional
- Course
- Duration
- End_time
- Exacerbating_factor
- Generic
- Negation_indicator
- Relative_temporal_context
- Severity
- Start_time
- Subject
- Uncertainty_indicator

**Disease/Disorder CEM template**
- Alleviating_factor
- Associated_sign_or_symptom
- associatedCode
- Body_laterality
- Body_location
- Body_side
- Conditional
- Course
- Duration
- End_time
- Exacerbating_factor
- Generic
- Negation_indicator
- Relative_temporal_context
- Severity
- Start_time
- Subject
- Uncertainty_indicator

**Procedure CEM template**
- associatedCode
- Body_laterality
- Body_location
- Body_side
- Conditional
- Device
- End_date
- Generic
- Method
- Negation_indicator
- Relative_temporal_context
- Start_date
- Subject
- Uncertainty_indicator

**Lab CEM template**
- Abnormal_interpretation
- associatedCode
- Conditional
- Delta_flag
- Estimated_flag
- Generic
- Lab_value
- Negation_indicator
- Ordinal_interpretation
- Reference_range_narrative
- Subject
- Uncertainty_indicator

**Anatomical Site CEM template**
- associatedCode
- Body_laterality
- Body_site
- Conditional
- Generic
- Negation_indicator
- Subject
- Uncertainty_indicator
SHARPn data normalization pipeline
SHARPN Data Normalization Architecture

Normalization pipeline

1. Syntactic Parsing
2. HL7/CCD/CDA to Object form
3. Mapping resources
   - Resource configuration
   - TypeSystem configuration
   - Syntactic and semantic mapping configuration
4. Map to UIMA types
5. Generate CEM
6. CTAKEs (NLP)
7. Semantic Normalization

Unstructured Information Management Architecture

mirth™

Data transmitted via NwHIN

Optional Component

Healthcare system

Relational CEM DB

Document DB

EMPI

Batch CEM Processing

EMPI Patient Correlation

MySQL

[Welch et al., JBI 2012; 45(4):763-71]
[Pathak et al., JAMIA 2013; 20(e2): e341-8]
Algorithm Development Process - Modified

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

**Rules**

**Semi-Automatic Execution**

**Evaluation**

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

**Visualize**

**Algorithm Development Process**

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

**XML**

**Data**

**Mappings**

**NLP, SQL**

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

[Pathak et al., JAMIA 2013; 20(e2): e341-8]
Example algorithm: Hypothyroidism

### EMERGE, Network Supplemental Genotyping Project – Phenotype Definition

**Chronic Autoimmune Hypothyroidism (presumptive Hashimoto’s hypothyroidism)**

**Project Outline:** Selection of all Caucasian patients with hypothyroidism without a secondary cause of surgical removal or radiological ablation. The search is designed to eliminate subclinical hypothyroidism (by requiring that patients be on a replacement medication), medication-induced hypothyroidism (e.g., PTU, lithium, or history of amiodarone), and transient causes (e.g., pregnancy or radioactive iodine thyroiditis).

**Phenotype Description:** Patients with presumptive autoimmune hypothyroidism (Hashimoto’s hypothyroidism), requiring replacement therapy.

**Case definition**

**Case inclusion criteria:** (all three conditions required):
- ICD9 code for hypothyroidism OR abnormal TSH/FT4
- Thyroid replacement medication use
- Require at least 2 instances of either medication or lab (a combination is acceptable) with at least 3 months between the first and last instance of medication or lab

**Case exclusions (if occurring at any time in the record):**
- Secondary causes of hypothyroidism: ICD9 codes 244.0, 244.1, 244.2, or 244.3
- Post-surgical or post-radiation hypothyroidism (by ICD9 codes or CPT codes for the procedures)
- Other thyroid diseases (Graves, thyroid cancer, MEN syndromes, etc) by ICD9 codes, CPT codes, or text word diagnoses (can be limited to problem lists)
- Any thyroid-altering medication (see below)

**Time-dependent case exclusions:**
- Recent pregnancy TSH/FT4 (any pregnancy billing code or lab test if all Case Definition codes, labs, or medications fall within 6 months before pregnancy to one year after pregnancy)
- Recent contrast exposure (all abnormal lab or medication references occurring within 6 weeks following a contrast study)

**Control definition**

- No billing codes for hypothyroidism, no evidence of thyroid replacement meds
- Must have a normal TSH (and FT4 if checked)
- Must contain at least two Past Medical History sections and Medication lists (could substitute two nonacute clinic visits or requirement for annual physical)

**Control exclusions:**

- Any cause of hypo- or hyperthyroidism

### Variables of Interest

#### Case ICD 9 codes

- 244.2: chronic lymphocytic thyroiditis
- 244.8: chronic thyroiditis NEC/NOS
- 244.9: hypothyroidism NOS
- 245.2: hyperthyroidism NOS

#### Case lab names/values

- Hypothyroidism: TSH > 5 or FT4 < 0.5
- Anti-thyroglobulin antibodies: H-TG, ThyAB, AnTyr- positive
- Anti-thyroxineoxidase: H-TPO, TPO, AnTyr- positive
- Anti-thyroid antibodies: ThyAB – positive

#### Case medications

- Levothyroxine, Synthroid, Levoxyl, Unithroid, armour thyroid, desiccated thyroid, synven,isons, levothyroxine, synthetic triiodothyronine, levox, thyrolar, T3’ and T4’

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

### Control lab names/values

- TSH must be between 0.5 – 5
- FT4 must be between 0.5-1.2 (if checked)

### Case/Control thyroid disease exclusion ICD 9 codes

- Thyroid cancer, all types: 191.9
- Toxic diffuse goiter: 242.0
- Toxic multinodular goiter: 242.2
- Thyrotoxicosis not goiter or other case: 245.2

### Control exclusion ICD9 codes

- 240.2: Simple and unspecified goiter
- 241.2: Nontoxic nodular goiter
- 242.2: Thyrotoxicosis with or without goiter
- 243.2: Congenital hypothyroidism
- 244.2: Acquired hypothyroidism
- 245.2: Thyroiditis
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

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>NQF #</td>
</tr>
<tr>
<td>Title:</td>
</tr>
<tr>
<td>Project Name:</td>
</tr>
<tr>
<td>Status:</td>
</tr>
<tr>
<td>Original Endorsement Date:</td>
</tr>
<tr>
<td>Most Recent Endorsement Date:</td>
</tr>
<tr>
<td>Steward(s):</td>
</tr>
<tr>
<td>Description:</td>
</tr>
</tbody>
</table>
Example: Diabetes & Lipid Mgmt. - II

Population criteria

- Initial Patient Population =
  - AND: "Patient characteristic: birth date" $\geq$ 17 year(s) and $\leq$ 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: $\geq$ 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</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>standard taxonomy</th>
<th>code</th>
<th>descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidase inhibitors</td>
<td>RxNorm</td>
<td>199150</td>
<td>Acarbose 100 MG Oral Tablet</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidase inhibitors</td>
<td>RxNorm</td>
<td>200132</td>
<td>Acarbose 25 MG Oral Tablet</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidase inhibitors</td>
<td>RxNorm</td>
<td>205329</td>
<td>miglitol 25 MG Oral Tablet</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidase inhibitors</td>
<td>RxNorm</td>
<td>205330</td>
<td>miglitol 50 MG Oral Tablet</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidase inhibitors</td>
<td>RxNorm</td>
<td>205331</td>
<td>miglitol 100 MG Oral Tablet</td>
</tr>
<tr>
<td>2.16.840.1.113883.3.464.0001.05</td>
<td>Alph-glucosidase inhibitors</td>
<td>RxNorm</td>
<td>401938</td>
<td>Miglustat 100 MG Oral Capsule</td>
</tr>
</tbody>
</table>
Example: Diabetes & Lipid Mgmt. - V

Population Criteria Section: denominator

-->
<entry typeCode="DRIV">
  <observation classCode="OBS" moodCode="EVN.CRT" isCriterionInd="true">
    <id root="655EBFF4-0530-4D99-A1BB-6F95D24CF2FC"/>
    <code code="ASSERTION" codeSystem="2.16.840.1.113883.5.4"/>
    <value xsi:type="CD" code="DENOM" codeSystem="2.16.840.1.113883.5.1063" codeSystemName="HL7 Observation Value" displayName="Denominator"/>
    <sourceOf typeCode="PRCN">
      <conjunctionCode code="AND"/>
      <observation classCode="OBS" moodCode="EVN" isCriterionInd="true">
        <id root="B76DB255-7D1D-4968-AE98-3E3FA2D58CDF"/>
        <title>Initial Patient Population</title>
      </observation>
    </sourceOf>
  </observation>
</entry>

Computer readable HQMF XML (based on HL7 v3 RIM)
An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms

William K. Thompson, Ph.D.\textsuperscript{1}, Luke V. Rasmussen\textsuperscript{1}, Jennifer A. Pacheco\textsuperscript{1}, Peggy L. Peissig, M.B.A.\textsuperscript{2}, Joshua C. Denny, M.D.\textsuperscript{3}, Abel N. Kho, M.D.\textsuperscript{1}, Aaron Miller, Ph.D.\textsuperscript{2}, Jyotishman Pathak, Ph.D.\textsuperscript{4},
\textsuperscript{1}\textit{Northwestern University, Chicago, IL}; \textsuperscript{2}\textit{Marshfield Clinic, Marshfield, WI}; \textsuperscript{3}\textit{Vanderbilt University, Nashville, TN}; \textsuperscript{4}\textit{Mayo Clinic, Rochester, MN}

<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]
## An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms

William K. Thompson, Ph.D.\(^1\), Luke V. Rasmussen\(^1\), Jennifer A. Pacheco\(^1\), Peggy L. Peissig, M.B.A.\(^2\), Joshua C. Denny, M.D.\(^3\), Abel N. Kho, M.D.\(^1\), Aaron Miller, Ph.D.\(^2\), Jyotishman Pathak, Ph.D.\(^4\).

\(^1\)Northwestern University, Chicago, IL; \(^2\)Marshfield Clinic, Marshfield, WI; \(^3\)Vanderbilt University, Nashville, TN; \(^4\)Mayo Clinic, Rochester, MN

<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]
An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms
William K. Thompson, Ph.D.¹, Luke V. Rasmussen¹, Jennifer A. Pacheco¹, Peggy L. Peissig, M.B.A.², Joshua C. Denny, M.D.³, Abel N. Kho, M.D.¹, Aaron Miller, Ph.D.², Jayishman Pathak, Ph.D.⁴. ¹Northwestern University, Chicago, IL; ²Marshfield Clinic, Marshfield, WI; ³Vanderbilt University, Nashville, TN; ⁴Mayo Clinic, Rochester, MN

<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>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>

Winner of 2012 AMIA Annual Symposium Distinguished Paper Award

[Thompson et al., AMIA 2012]
What about phenotype algorithm authoring and semi-automated execution?
What about phenotype algorithm authoring?

Design patterns for the development of electronic health record-driven phenotype extraction algorithms


Qualitative evaluation of three phenotype authoring tools to find methotrexate liver injury

Qian Zhu, Huan Mo, MD, MS, Luke V. Rasmussen, Jennifer A. Pacheco, Jie Xu, MS, Richard C. Kiefer, Peter Speltz, William K. Thompson, PhD, Enid Montague, PhD, Andrew Post, MD, PhD, Joshua C. Denny, MD, MS, Jyotishman Pathak, PhD

Evaluation of Existing Phenotype Authoring Tools for Clinical Research

Luke V. Rasmussen, Jie Xu, MS, Ruijue Liu, Qian Zhu, PhD, Jennifer A. Pacheco, Jyotishman Pathak, PhD, William K. Thompson, PhD, Joshua C. Denny, MD, MS, Huan Mo, MD, MS, Richard C. Kiefer, Peter Speltz, Enid Montague, PhD

1Northwestern University, Chicago, IL; 2National University of Singapore, Singapore; 3Mayo Clinic, Rochester, MN; 4NorthShore HealthSystem, Evanston, IL; 5Vanderbilt University, Nashville, TN

[©2014 MFMER | slide-41]
## Author Phenotype Algorithm

### Phenotypes
- My Phenotypes
  - eMERGE Type 2 Diabetes
  - Hypertension
- All Phenotypes
  - Search

### Clinical Terms
- Value Sets
  - Recently Used
  - Search
  - Create New
- Terminologies
  - Browse
    - ICD9
    - ICD10
    - SNOMED-CT
  - Search

### Data Elements
- Care Experience
- Care Goal
- Communication
- Diagnosis
- Device
- Diagnostic Study
- Encounter

### Diagnosis 1

**Drag and drop clinical terms or value sets here, or search for.**

#### Attributes:
- All diagnoses

### Search for value sets:
- Diabetes

### I’m interested in these terminologies:
- ICD-9
- ICD-10
- SNOMED-CT
- LOINC
- RxNorm

### Type 2 Diabetes Dx Codes

<table>
<thead>
<tr>
<th>Name</th>
<th>OID</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2 Diabetes Dx Codes</td>
<td>2.1.1.1.1.1.1.1.1</td>
<td>Group</td>
</tr>
<tr>
<td>Type 1 Diabetes Dx Codes</td>
<td>2.1.1.1.1.1.1.1.2</td>
<td>Group</td>
</tr>
</tbody>
</table>

[http://informatics.mayo.edu/PhEMA]
A Modular Architecture for Electronic Health Record-Driven Phenotyping

Luke V. Rasmussen¹; Richard C. Kiefer², Huan Mo, MD, MS³, Peter Speltz³, William K. Thompson, PhD⁴, Guoqian Jiang, MD, PhD², Jennifer A. Pacheco¹, Jie Xu, MS¹, Qian Zhu, PhD⁵, Joshua C. Denny, MD, MS³, Enid Montague, PhD¹, Jyotishman Pathak, PhD²

¹Northwestern University, Chicago, IL; ²Mayo Clinic, Rochester, MN; ³Vanderbilt University, Nashville, TN; ⁴NorthShore University HealthSystem, Evanston, IL; ⁵University of Maryland Baltimore County, Baltimore, MD
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]
[Pathak et al., JAMIA 2013; 20(e2): e341-8]
Scalable and High-Throughput Execution of Clinical Quality Measures from Electronic Health Records using MapReduce and the JBoss® Drools Engine

Kevin J. Peterson, MS¹,² and Jyotishman Pathak, PhD¹
¹ Dept. of Health Sciences Research, Mayo Clinic, Rochester, MN
² Dept. of Computer Science & Engineering, University of Minnesota, Twin Cities, MN
Example: Initial Patient Population criteria for CMS eMeasure (CMS163V1)

**Initial Patient Population =**
- AND: "Diagnosis, Active: Diabetes" starts before or during "Measurement Period"
- AND: "Patient Characteristic Birthdate: birth date" >= 18 year(s) starts before start of "Measurement Period"
- AND: "Patient Characteristic Birthdate: birth date" <= 75 year(s) starts before start of "Measurement Period"
- AND:
  - OR: "Encounter, Performed: Office Visit"
  - OR: "Encounter, Performed: Face-to-Face Interaction"
  - OR: "Encounter, Performed: Preventive Care Services - Established Office Visit, 18 and Up"
  - OR: "Encounter, Performed: Preventive Care Services-Initial Office Visit, 18 and Up"
  - OR: "Encounter, Performed: Home Healthcare Services"

```mvel
rule "EncounterPerformedOfficeVisit_precondition_21"
dialect "mvel"
no-loop
when
    $measurementPeriod : MeasurementPeriod()
    $p : Patient ()
    $event : edu.mayo.qdm.patient.Encounter(    this during $measurementPeriod )
) from droolsUtil.findMatches("2.16.840.1.113883.3.464.1003.101.12.1001"), $p.getEncounters() then
    insertLogical(new PreconditionResult("EncounterPerformedOfficeVisit_precondition_21", $p, $event ))
end```

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, government agencies, insurers and other stakeholders.

http://phenotypeportal.org

[Endle et al., AMIA 2012]
1. Converts QDM to Drools
2. Rule execution by querying the CEM database
3. Generate summary reports
Scalable and High-Throughput Execution of Clinical Quality Measures from Electronic Health Records using MapReduce and the JBoss® Drools Engine

Kevin J. Peterson, MS¹,² and Jyotishman Pathak, PhD¹
¹ Dept. of Health Sciences Research, Mayo Clinic, Rochester, MN
² Dept. of Computer Science & Engineering, University of Minnesota, Twin Cities, MN

[Peterson AMIA 2014]
## Validation using Project Cypress

<table>
<thead>
<tr>
<th>CMS ID</th>
<th>IPP</th>
<th>NUMER</th>
<th>DENOM</th>
<th>DENEX</th>
<th>DENEXCEP</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS122v1</td>
<td>2/2</td>
<td>2/1</td>
<td>2/2</td>
<td>0/0</td>
<td>0/0</td>
<td>Failure</td>
</tr>
<tr>
<td>CMS126v1</td>
<td>3/3</td>
<td>1/1</td>
<td>3/3</td>
<td>1/1</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS127v1</td>
<td>5/5</td>
<td>1/1</td>
<td>5/5</td>
<td>0/0</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS134v1</td>
<td>2/2</td>
<td>1/1</td>
<td>2/2</td>
<td>0/0</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS145v1</td>
<td>2/2</td>
<td>1/1</td>
<td>2/2</td>
<td>0/0</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS148v1</td>
<td>2/2</td>
<td>1/1</td>
<td>2/2</td>
<td>0/0</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS153v1</td>
<td>4/4</td>
<td>1/1</td>
<td>4/4</td>
<td>1/1</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS159v1</td>
<td>2/2</td>
<td>0/1</td>
<td>2/2</td>
<td>0/0</td>
<td>0/0</td>
<td>Failure</td>
</tr>
<tr>
<td>CMS160v1</td>
<td>3/3</td>
<td>2/2</td>
<td>3/3</td>
<td>0/0</td>
<td>0/0</td>
<td>Success</td>
</tr>
<tr>
<td>CMS163v1</td>
<td>2/2</td>
<td>1/1</td>
<td>2/2</td>
<td>0/0</td>
<td>0/0</td>
<td>Success</td>
</tr>
</tbody>
</table>

[Peterson AMIA 2014]
JSON to Drools

Convert the health-data-standards JSON to JBoss Drools Rules.

```java
/* Rule */
rule "PatientCharacteristicBirthdate"
  dialect "mvel"
  no-loop
  when
    $p : Patient{ birthdate != null, toDays(birthdate) < toDays(new Date("01-Jan-1947")) }
  then
    insertLogical(new PreconditionResult("PatientCharacteristicBirthdate", $p))
end
```

http://api.phenotypeportal.org
How is this research applied at Mayo?

- Transfusion related acute lung injury (Kor)
- Pharmacogenomics of depression (Weinshilboum)
- Pharmacogenomics of breast cancer (Wang)
- Pharmacogenomics of heart failure (Pereira)
- Multi-morbidity in depression and heart failure (Pathak)
- Lumbar image reporting with epidemiology (Kallmes)
- Warfarin dosing and heart valve replacements (Pereira)
- Active surveillance for celiac disease (Murray)
- Drug induced liver injury (Talwalkar)
- Drug induced thrombocytopenia and neutropenia (Al-Kali)
How is this research applied at Mayo?

- Transfusion related acute lung injury (Kor)
- Pharmacogenomics of depression (Weinshilboum)
- Pharmacogenomics of breast cancer (Wang)
- Pharmacogenomics of heart failure (Pereira)
- Multi-morbidity in depression and heart failure (Pathak)
- Lumbar image reporting with epidemiology (Kallmes)
- Warfarin dosing and heart valve replacements (Pereira)
- Active surveillance for celiac disease (Murray)
- Drug induced liver injury (Talwalkar)
- Drug induced thrombocytopenia and neutropenia (Al-Kali)
Blood transfusion management

Transfusion-Related Acute Lung Injury (TRALI)
Transfusion-Associated Circulatory Overload (TACO)
Electronic health record surveillance algorithms facilitate the detection of transfusion-related pulmonary complications

Leanne Clifford, Amandeep Singh, Gregory A. Wilson, Pearl Toy, Ognjen Gajic, Michael Malinchoc, Vitaly Herasevich, Jyotishman Pathak, and Daryl J. Kor

32,107 patients > 6 months of age who received blood product transfusion (279,630 products transfused during 111,532 transfusion episodes*)

Continuous electronic surveillance; arterial blood gas analysis with PaO2/FiO2<300

10,184 alerts (8,999 patients)

101,348 episodes (23,108 patients) with no alerts suggesting respiratory worsening after transfusion

Excluded: 10,184

389 alerts reviewed by expert panel for new or worsening bilateral infiltrates

123 TACO

62 Possible TRALI

45 TRALI

35 TACO/TRALI

40 ALI

84 other (bilateral atelectasis, effusions)

Excluded from study

36 TRALI

57 possible TRALI

52 TACO

85 Controls

Excluded: 71

5 patients unable to identify transfusion time of implicated units

2 patients unable to identify transfusion time of implicated units

Excluded: 5

5 patients unable to identify transfusion time of implicated units

Excluded: 7

2 patients also classified as TRALI

5 patients unable to identify transfusion time of implicated units

Excluded: 7

2 patients also classified as TRALI

5 patients unable to identify transfusion time of implicated units

[Clifford et al. Transfusion 2013]
## Missing reported TRALI/TACO cases

<table>
<thead>
<tr>
<th>Transfusion environment</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>PLR (95% CI)</th>
<th>NLR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TRALI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensive Care Unit (n=58)</td>
<td>0.97 (0.84 – 1.00)</td>
<td>0.71 (0.48 – 0.88)</td>
<td>3.41 (1.73 – 6.71)</td>
<td>0.04 (0.01 – 0.27)</td>
</tr>
<tr>
<td>Operating Room (n=67)</td>
<td>0.97 (0.85 – 1.00)</td>
<td>0.97 (0.80 – 1.00)</td>
<td>28.24 (4.11 – 193.87)</td>
<td>0.03 (0.01 – 0.19)</td>
</tr>
<tr>
<td>Other Hospital Wards (n=46)</td>
<td>0.83 (0.58 – 0.96)</td>
<td>0.93 (0.75 – 0.99)</td>
<td>11.67 (3.02 – 45.07)</td>
<td>0.18 (0.06 – 0.51)</td>
</tr>
<tr>
<td><strong>TACO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensive Care Unit (n=34)</td>
<td>0.77 (0.46 – 0.94)</td>
<td>0.81 (0.57 – 0.94)</td>
<td>4.04 (1.59 – 10.24)</td>
<td>0.28 (0.10 – 0.79)</td>
</tr>
<tr>
<td>Operating Room (n=66)</td>
<td>0.92 (0.76 – 0.98)</td>
<td>0.97 (0.81 – 1.00)</td>
<td>27.50 (3.99 – 189.37)</td>
<td>0.09 (0.03 – 0.26)</td>
</tr>
<tr>
<td>Other Hospital Wards (n=30)</td>
<td>1.00 (0.20 – 1.00)</td>
<td>0.96 (0.80 – 1.00)</td>
<td>28.00 (4.09 – 191.88)</td>
<td>0.00 (0.00-0.00)</td>
</tr>
<tr>
<td><strong>Transfusion-Related Pulmonary Complications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensive Care Unit (n=71)</td>
<td>0.94 (0.83 – 0.98)</td>
<td>0.70 (0.46 – 0.87)</td>
<td>3.14 (1.60 – 6.15)</td>
<td>0.08 (0.03 – 0.26)</td>
</tr>
<tr>
<td>Operating Room (n=104)</td>
<td>0.95 (0.86 – 0.98)</td>
<td>0.86 (0.67 – 0.95)</td>
<td>6.86 (2.76 – 17.07)</td>
<td>0.06 (0.02 – 0.16)</td>
</tr>
<tr>
<td>Other Hospital Wards (n=48)</td>
<td>0.95 (0.73 – 1.00)</td>
<td>0.93 (0.75 – 0.99)</td>
<td>13.3 (3.48 – 50.76)</td>
<td>0.05 (0.01 – 0.37)</td>
</tr>
</tbody>
</table>

[Clifford et al. Transfusion 2013]
## Missing reported TRALI/TACO cases

<table>
<thead>
<tr>
<th>Transfusion environment</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PLR</th>
<th>NLR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
</tr>
<tr>
<td>TRALI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensive Care Unit</td>
<td>0.77 (0.46–0.94)</td>
<td>0.81 (0.57–0.94)</td>
<td>4.04 (1.59–10.24)</td>
<td>0.28 (0.10–0.79)</td>
</tr>
<tr>
<td>Operating Room</td>
<td>0.92 (0.76–0.98)</td>
<td>0.97 (0.81–1.00)</td>
<td>27.50 (3.99–189.37)</td>
<td>0.09 (0.03–0.26)</td>
</tr>
<tr>
<td>Other Hospital</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TACO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensive Care Unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operating Room</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Hospital</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Of the **88 TRALI cases** correctly identified by the CART algorithm, only **11 (12.5%)** of these were reported to the blood bank by the clinical service.

Of the **45 TACO cases** correctly identified by the CART algorithm, only **5 (11.1%)** were reported to the blood bank by the clinical service.

[Clifford et al. Transfusion 2013]
Active surveillance for TRALI/TACO

TRALI/TACO “sniffers”

Active surveillance for TRALI/TACO

Example of a data-driven learning health care system?

TRALI/TACO “sniffers”
Kohane et al. JAMA 2014

Probabilistic linkage to validate existing data or fill in missing data

Examples of biomedical data
- Pharmacy data
- Health care center (electronic health record) data
- Claims data
- Registry or clinical trial data
- Data outside of health care system

Ability to link data to an individual
- Easier to link to individuals
- Harder to link to individuals
- Only aggregate data exists

Data quantity
- More
- Less
Our focus today
6 CHARACTERS REBOOTING MEDICINE AND HEALTH: O, 1, A, C, G, T

Accelerating Digital Technology:
- WWW
- YouTube
- Data Universe
- Health Information Systems
- Imaging
- Internet
- Social Networking
- Mobile Connectivity
- Wireless Sensors & Devices

Super-Convergence:
- ACGTAGTTAACCGGCTAGCTTAGCTACATCTGAC
- \(\{101101010010001010101010101010101010101010101\}\)
- \(\{1011010100100010101010101010101010101010101\}\)
- \(\{1011010100100010101010101010101010101010101\}\)

Reboot:
- 64M
- 16M
- 4M
- 1M
- 250K

Future focus
- PREVENTION
- PREDICTION
- MANAGEMENT
- DIAGNOSIS
- DISEASE

Based on the work of Dr. Eric Topol, author of “The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care”
The New Medical Data Ecosystem

Medical data is being captured today from many sources. Pulling it together and studying what it means is the next challenge.

- **Environmental Data**
  Sensors can pick up behavioral information. Mapping, location and weather data adds insight into other triggers.

- **Genomic Data**
  Less expensive genome sequencing offers insight into the role genetics may play.

- **Mobile Health Data**
  100,000-plus mobile health apps, plus wearable devices that measure activity and bodily function, offer a constant read on patient health.

- **Electronic Medical Records**
  Digital records include lab and test results, drug prescriptions, and doctors' reports.

- **Public Health Data**
  Insight into community health patterns from federal and state data.

- **Family Health History**

- **Insurance Claims Data**
  Trends in drug and treatment usage.

- **Analytic algorithms and predictive modeling**
  Mine the layers of data for patterns and insight.

- **Patients**
  More precise and personalized diagnosis and care based on a holistic view.

- **Doctors**
  Decision support tools could help quickly evaluate the best treatments.

- **Researchers**
  Detailed information from many patients, along with other data, could lead to new insights into disease and treatment.

---

**Future focus**

- **Prevention**
- **Prediction**
- **Management**
- **Diagnosis**
- **Disease**

**Number of sequenced Human Genomes**
- 2
- 1M
- 4M
- 16M
- 64M

**Destruction of Digital Technology**

**WWW**

- 1970
- 1980
- 1990
- 2000
- 2010
- 2020

**Brought to you by M2FIT and linkd.in/DigitalHealthGroup**

www.misfitwearables.com/references

*MIT Tech. Review, Sep 2014*
Mayo Clinic collaborates with Apple to release updated patient app

Oct. 2, 2014

Beginning Thursday, Oct. 2, patients who use iPhones, iPods and iPads can download an updated Mayo Clinic app for iOS8. Created in collaboration with Apple, the 2.0 version of the app will offer upgraded functionality and new features on iPhones and iPads. An Android update is expected in 2015.

October 2nd, 2014
New sources of data & knowledge:
- Continuous collection of data from personal sensors
- Epigenomics
- Microbiomics
- Different genomes in the body
- Nanomedicine

Need of informatics methods and technologies:
- Big data
- Visualisation
- Ontologies
- Privacy protection
- Interfacing with sensors
- Systems analysis

[Sanchez et al. JAMIA 2014]
Mayo Clinic Data-as-a-Service Architecture

HEALTHCARE APPLICATIONS
- Patient Similarity
- Therapy Response Prediction
- Dynamic Recommendation Engine
- Patient Outcome Prediction
- Adverse Event Prediction
- Wellbeing services

DATA ASSETS

- Clinical
  - Labs
  - Medications
  - Vitals
  - Diagnosis
  - Demo
  - Procedures
- Omics
  - Genomics
  - Proteomics
  - Metabolomics
- Environmental
- Social network
- Medical device
- Sensors

DATA PREPARATION

- Extract
- Load
- Transform

REAL-TIME ANALYTICS

- Descriptive
  - Statistics
  - Visualization
- Diagnostic
  - Survival Analysis
  - Regression
- Predictive
  - Classification & Clustering
  - Inference Analysis
- Prescriptive
  - Treatment Plan
  - Decision Support

HIGH-THROUGHPUT SCALABLE PHENOTYPING

1. DEFINE
2. ASSESS
3. STRATIFY

Batch
Map Reduce
Script
Pig
SQL
Hive
Online
HBase
Real-Time
Storm
In-memory
Spark

Metadata Management: HCatalag

Multi Processing: YARN

HDFS Storage

Operations (Ambari)

Security

[Adapted from Hortonworks.com]
Some experiences and opportunities for big data in translational research

Christopher G. Chute, MD, DrPH\textsuperscript{1}, Mollie Ullman-Cullere, MS, MSE\textsuperscript{2}, Grant M. Wood, BS\textsuperscript{3}, Simon M. Lin, MD\textsuperscript{4}, Min He, PhD\textsuperscript{4} and Jyotishman Pathak, PhD\textsuperscript{1}

Performance of HDP (Hortonworks Data Platform) on a simple query

![Graph showing performance comparison between SQL, HDFS, and HDP](image)

- 60\% Speed up comparing to SQL
- 4x Speed up comparing to SQL

[Chute et al. GIM 2013]
Mining >800 million search logs from MayoClinic.com

1. Data Collection
   - Online Health Information Seeker
   - Web Search Engine (Google, Bing)
   - MayoClinic.com

2. Dataset Creation
   - Define Data Requirements
   - Data Request
   - Web Analytics Tool
   - Search Query Data
   - Data Preprocessing
   - Dataset for Experiments

3. Semantic Annotation
   - UMLS MetaMap
   - UMLS Metathesaurus
   - Dataset with UMLS Semantic Types and Concepts

4. Health Categorization
   - Define Health Categories
   - Develop Rule Based Classifier
   - Assign UMLS Semantic Types and Concepts to Selected Health Categories
   - Evaluate Rule Based Approach
   - Categorize Search Queries into Health Categories

[Jadhav et al. JMIR 2014]
The user experience for online health information searching may differ by device usage of smart devices for health information seeking varies with the device used, especially between Personal Computers (PCs) and Smart Devices (SDs).

For PCs and SDs, we analyzed and compared:

- **Data collection tool:** IBM Watson Analytics tool.
- **Structural properties of the queries (length of query operators and special characters)**
- **Linguistic characteristics of the queries**
  - Types of health search queries:
  - Length of the health search queries:
  - Misspellings in the health search queries
  - Frequently searched health categories

**Structural Analysis**

- **Number of Operators**
  - Number of Operators: 0 (PC 97.35%, SD 96.53%)
  - Number of Operators: >0 (PC 2.65%, SD 3.47%)
- **Query Operators Usage**
  - AND: (PC 86.53%, SD 85.05%)
  - OR: (PC 4.37%, SD 3.08%)
  - &: (PC 5.20%, SD 6.78%)
  - Other: (PC 1.42%, SD 1.28%)
- **Special Characters**
  - >0 (PC 95.66%, SD 96.72%)
  - >1 (PC 4.34%, SD 3.29%)
- **Spelling Mistake**
  - >0 (PC 87.47%, SD 87.88%)

**Linguistic Analysis**

- **Part-of-Speech**
  - Nouns:
    - 0 (PC 3.19%, SD 1.67%)
    - 1 (PC 28.17%, SD 26.93%)
    - 2 (PC 46.87%, SD 47.38%)
    - 3 (PC 17.75%, SD 19.79%)
    - >3 (PC 4.01%, SD 4.23%)
  - Verbs:
    - 0 (PC 83.34%, SD 78.96%)
    - >0 (PC 16.66%, SD 21.05%)
  - Adverbs:
    - 0 (PC 95.56%, SD 95.38%)
    - >0 (PC 4.45%, SD 4.62%)
  - Adjectives:
    - 0 (PC 69.71%, SD 66.14%)
    - >0 (PC 30.29%, SD 33.86%)

Insights from this study can be leveraged by healthcare-centric application developers and users in their quest for health information, and promote participation in learning and managing their health. Increasingly, individuals are taking active participation in learning and managing their health by leveraging online resources.
### Top 10 Search Queries (2011-2013)

<table>
<thead>
<tr>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mediterranean diet</td>
</tr>
<tr>
<td>Kidney infection</td>
</tr>
<tr>
<td>Lupus</td>
</tr>
<tr>
<td>Lupus symptoms</td>
</tr>
<tr>
<td>Ovarian cancer symptoms</td>
</tr>
<tr>
<td>Mayo Clinic symptom checker</td>
</tr>
<tr>
<td>Heart attack symptoms</td>
</tr>
<tr>
<td>Hypothyroidism</td>
</tr>
<tr>
<td>Celiac disease</td>
</tr>
<tr>
<td>Lyme disease symptoms</td>
</tr>
</tbody>
</table>

[Jadhav et al. JMIR 2014]
Online Information Seeking for Cardiovascular Diseases: A Case Study from Mayo Clinic

Ashutosh JADHAV<sup>a,1</sup>, Stephen WU<sup>b</sup>, Amit SHETH<sup>a</sup> and Jyotishman PATHAK<sup>b</sup>

<sup>a</sup>Knoesis Center, Wright State University, Dayton, OH, USA
<sup>b</sup>Mayo Clinic, Rochester, MN, USA

<table>
<thead>
<tr>
<th>Top 1-10 Queries</th>
<th>Top 11-20 Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart attack symptom</td>
<td>Cardiomyopathy</td>
</tr>
<tr>
<td>Blood pressure chart</td>
<td>Heart palpitations</td>
</tr>
<tr>
<td>How to lower blood pressure</td>
<td>Blood pressure medication</td>
</tr>
<tr>
<td>Heart rate</td>
<td>Symptoms of stroke</td>
</tr>
<tr>
<td>Broken heart syndrome</td>
<td>Heat stroke</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>Echocardiogram</td>
</tr>
<tr>
<td>Low blood pressure</td>
<td>Heart disease</td>
</tr>
<tr>
<td>Stroke symptoms</td>
<td>Orthostatic hypotension</td>
</tr>
<tr>
<td>Normal blood pressure</td>
<td>Heart healthy recipes</td>
</tr>
<tr>
<td>High blood pressure symptoms</td>
<td>Heart arrhythmia</td>
</tr>
</tbody>
</table>
Is there any correlation between Search Queries and future Healthcare Utilization?

Pursuing Insights about Healthcare Utilization via Geocoded Search Queries

Shuang-Hong Yang*
Twitter Inc.
San Francisco, CA
syang@twitter.com

Ryen W. White
Microsoft Research
Redmond, WA
ryenw@microsoft.com

Eric Horvitz
Microsoft Research
Redmond, WA
horvitz@microsoft.com

Yang et al. SIGIR 2013; White and Horvitz, JAMIA 2013

From health search to healthcare: explorations of intention and utilization via query logs and user surveys

Ryen W White, Eric Horvitz
Research opportunities in Data and Information Sciences

- Scientific infrastructure
  - Streaming and autonomic computing
  - New programming models for computation

- Data and information management
  - Quality and provenance control
  - Link and graph mining
  - Dimensionality reduction

- Mining and analytics
  - Relational and structured learning
  - Parallel and distributed machine learning
  - Large-scale recommendation systems

- Security and privacy
  - Risk of re-identification
  - Anomaly and threat detection
### NIH Big Data-to-Knowledge (BD2K) Training Opportunity: Mayo Clinic & UMN

<table>
<thead>
<tr>
<th>Bootcamp Day</th>
<th>Coursework faculty</th>
<th>Proposed course and hands-on training module</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1 (part 1)</td>
<td>Jyotishman Pathak, PhD (Mayo)</td>
<td>Big Data methods and tools for computational medicine</td>
</tr>
<tr>
<td>Day 1 (part 2)</td>
<td>Christopher G. Chute, MD, DrPH (Mayo)</td>
<td>Standardization and normalization of Big Data</td>
</tr>
<tr>
<td>Day 2 (part 1)</td>
<td>Hongfang Liu, PhD (Mayo)</td>
<td>Natural language processing using Big Data</td>
</tr>
<tr>
<td>Day 2 (part 2)</td>
<td>Claudia Neuhauser, PhD (UMN)</td>
<td>Quantitative techniques for analysis of Big Data</td>
</tr>
<tr>
<td>Day 3 (part 1)</td>
<td>Vipin Kumar, PhD (UMN)</td>
<td>Knowledge discovery and data mining from Big Data</td>
</tr>
<tr>
<td>Day 3 (part 2)</td>
<td>Gyorgy Simon, PhD (UMN)</td>
<td>Machine learning and predictive modeling from Big Data</td>
</tr>
<tr>
<td>Day 4 (part 1)</td>
<td>David Pieczkiewicz, PhD (UMN)</td>
<td>Visualization analytics using Big Data</td>
</tr>
<tr>
<td>Day 4 (part 2)</td>
<td>Susan M. Wolf, JD (UMN)</td>
<td>Ethical and legal issues in handling Big Data</td>
</tr>
<tr>
<td>Day 5 (part 1)</td>
<td>Nilay D. Shah, PhD (Mayo)</td>
<td>The promise of Big Data for comparative effectiveness research</td>
</tr>
<tr>
<td>Day 5 (part 2)</td>
<td>Lila J. Rutten, PhD, MPH (Mayo)</td>
<td>Big Data resources and applications in population health research and improvement</td>
</tr>
<tr>
<td>Day 6* (part 1)</td>
<td>Optum Labs, IBM Research (Watson), Cray Supercomputer Inc.</td>
<td></td>
</tr>
<tr>
<td>Day 6* (part 2)</td>
<td>Trainee mentorship and career advancement opportunities</td>
<td></td>
</tr>
<tr>
<td>Day 7**</td>
<td>BDC4CM program committee meetings</td>
<td></td>
</tr>
</tbody>
</table>

**Big Data Coursework for Computational Medicine (BDC4CM):** [http://bdc4cm.org](http://bdc4cm.org)
Concluding remarks

• Biomedical big data will continue to grow exponentially

• An important focus should be on optimal data “re-use”, adoption of standards, and develop applications impacting patient care

• Exciting time to be in the field of biomedical informatics and data science
It takes a village...
Acknowledgment

- Mayo Clinic eMERGE Phenotyping team
  - Christopher Chute, MD, DrPH
  - Suzette Bielinski, PhD
  - Mariza de Andrade, PhD
  - John Heit, MD
  - Hayan Jouni, MD
  - Adnan Khan, MBBS
  - Sunghwan Sohn, PhD
  - Kevin Bruce
  - Sean Murphy

- Mayo Clinic SHARPn Phenotyping team
  - Christopher Chute, MD, DrPH
  - Dingcheng Li, PhD
  - Cui Tao, PhD (now @ UTexas)
  - Gyorgy Simon, PhD (now @ UMN)
  - Craig Stancl
  - Cory Endle
  - Sahana Murthy
  - Dale Suesse
  - Kevin Peterson

- PhEMA and DEPTH collaborators
  - Joshua Denny, MD
  - William Thompson, PhD
  - Guoqian Jiang, MD, PhD
  - Enid Montague, PhD
  - Luke Rasmussen, MS
  - Richard Kiefer
  - Peter Speltz
  - Jennifer Pachecho
  - Huan Mo

- Mayo Clinic Clinical Informatics Program
  - Daryl Kor, MD
  - Maryam Panahiazhar, PhD
  - Che Ngufor, PhD
  - Dennis Murphree, PhD
  - Sudhi Upadhyaya

Funding

- NIH R01 GM105688 (PI – Pathak): PheMA
- AHRQ R01 HS23077 (PI – Pathak): DEPTH
- NIH R01 GM103859 (PI – Pathak): PGx
- NIH R25 EB020381 (PI – Pathak): BDC4CM
- ONC 90TR002 (PI – Chute): SHARPn
- NIH U01 HG006379 (PI – Chute): eMERGE
- Mayo Clinic Center for the Science of Healthcare Delivery
Thank You!

Pathak.Jyotishman@mayo.edu
http://jyotishman.info