A review of approaches to identifying patient phenotype cohorts using electronic health records
Shivade, Raghavan, Fosler-Lussier, Embi, Elhadad, Johnson, Lai

Chaitanya Shivade

JAMIA Journal Club
March 6, 2014
Introduction

- EHR based Phenotyping
  - Not a well defined term in literature

- Our scope
  - Identification of patient cohorts using EHR

- Why?
  - Clinical trial recruitment, survival analysis, etc.
Challenges

Different data sources

• Structured
  • Lab results, medications, diagnoses, etc.
• Unstructured
  • Progress notes, radiology reports, etc.
• Other
  • Images, genetic information, etc.

Consistency problems

• Data standards, mismatch, mappings.

Administrative roadblocks

• Privacy
Motivation

1. Review fragmented efforts
2. Summarize phenotypes
3. Data sources
4. Study Techniques
5. Large scale systems
6. Find opportunities
Literature Search

Limitations of standard approaches

- Missing articles
- Conference proceedings

Manual review from 2010-2012

- Journals
  - Journal of American Medical Informatics Association
  - Journal of Biomedical Informatics
- Conference Proceedings
  - Clinical Research Informatics Conference
  - AMIA Annual Symposium
Method

<table>
<thead>
<tr>
<th>JAMIA 460</th>
<th>JBI 365</th>
<th>CRI 84</th>
<th>AMIA 381</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusion after examining title and abstract</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Included articles: 76

Relevant references: 53

After removal of duplicates: 129

Exclusion after reading full text

Finally included: 97
Criteria

**Inclusion**

- Identifying diagnosis
- Clinical Trial recruitment solutions
- New methods, data-sources

**Exclusion**

- No automation
- Non EHR studies
- Distant application to cohort identification
Top Phenotypes

- Cancer
- Peripheral Arterial Disease
- Hypertension
- Asthma
- Pneumonia
- Drug side effect
- Cataract
- Rheumatoid Arthritis
- Heart Failure
- Diabetes
Phenotype details

- **Cancer**
  - Breast, lung, colon, pancreatic, brain, etc.

- **Diabetes**
  - Type-2, hypoglycemia, diabetic retinopathy.

- **Heart failure**
  - Congestive heart failure, heart failure.

- **Hypertension**
  - Hypertension, resistant hypertension.

- **Infectious diseases**
  - Pneumonia, hepatitis, pertussis, etc.
# Data Sources

<table>
<thead>
<tr>
<th>Category</th>
<th>Cancer</th>
<th>Diabetes</th>
<th>Heart Failure</th>
<th>Rheumatoid Arthritis</th>
<th>Cataract</th>
<th>Drug Side Effect</th>
<th>Pneumonia</th>
<th>Asthma</th>
<th>Peripheral Arterial Disease</th>
<th>Hypertension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Medications</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Lab</td>
<td>6</td>
<td>14</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Vitals</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Clinical</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>10</td>
<td>14</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Treatment</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Notes</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Genomic</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Data Characteristics

- Model development
  - Train and Test
  - Cross validation
  - Agreement statistic
- Gold standard
  - Manual chart review
  - ICD-9 codes
  - Systematic calculation using other variables
- Size
  - Units: Number of patients, samples, documents
Methods

1. Rule Based
2. Natural Language Processing
3. Machine Learning and Statistics
4. Hybrid Approaches
Rule Based

- Simple, reliable, effective
- Not generalizable

Rule derivation
- Rules based on clinical judgment
  - Kho et al. [2012] ADA for Type 2 Diabetes
  - Klompas et al. [2008] CDC for Hepatitis
- Refinement of previous rules
  - Wright et al. [2011] Novel data sources
- Automatically generated rules
  - Li et al. [2012] JBoss Drools
Natural Language Processing

- Notes contain valuable information
- Abbreviations, misspellings, ambiguity, etc.

**Approaches**

- Named entity extraction [UMLS terms]
  - HITex, MetaMap, GATE
- Use of keywords
- Use of semantic web technologies
  - Cui et al. [2012] EpiDEA

- Difficult to develop.
Machine Learning & Statistics

- Generalize

- Approaches
  - Comparison of popular algorithms
  - Decision tree based
    - Explanatory models.
  - Other algorithms
  - Statistical methods
    - Chi-squared tests, survival analysis, etc.

- Large datasets.
Hybrid approaches

- Rules and Machine learning
  - Smoking Status using cTAKES

- Multi-modal approach
  - Pessig et al. [2012] Data mining, NLP, OCR

- NLP and machine learning
  - Xu et al. [2011] Comparison. Rule Based depends on NLP

- NLP and rules
  - Coden et al. [2009] IBM MedTAS
Cohort Identification Systems

Multi-site
- eMERGE initiative
- Successful
- 14 EHR-oriented phenotyping algorithms
- CICTR project
  - Lacked acceptance

Single-site
- DISCERN at Duke
- STRIDE at Stanford

Lessons
- Standard terminologies
- Privacy
Implications for future research

- Terminologies
  - Broad vs Narrow Coverage

- Causality
  - Variables used for prediction

- Visual models
  - Clinical vs Informatics

- Techniques
  - Rule Based: Automatic rule mining
  - NLP: End to end usage
  - Machine learning: Explanatory models

- Widely accepted tools
  - Compare in-house and open tools
Summary

- Fragmented efforts for same task
  - Lack of standard solutions

- Varied data sources
  - Comprehensive usage

- Varied approaches
  - Aggregation is missing

- Some best practices
  - Standard terminologies
  - Multiple data sources
Thank you!

shivade@cse.ohio-state.edu

Chaitanya
Shivade