Unsupervised approaches for phenotyping using electronic health record data

July 29, 2020 | Sentinel Innovation Center Seminar

Katherine P. Liao, MD, MPH

Associate Professor of Medicine | Assistant Professor of Biomedical Informatics
Division of Rheumatology, Inflammation, and Immunity, Brigham and Women’s Hospital
Massachusetts Veterans Epidemiology Research and Information Center, VA Boston Healthcare System
Outline

• Rationale for development of phenotyping approaches using EHR

• Brief background of ML for phenotyping
  • Supervised vs unsupervised

• Unsupervised approaches for phenotyping w/ EHR data
  • Strengths and limitations
Who has rheumatoid arthritis (RA) in the EHR?

<table>
<thead>
<tr>
<th>Model</th>
<th>RA by algorithm or criteria, no.</th>
<th>PPV (95% CI), %</th>
<th>Sensitivity (95% CI), %</th>
<th>Difference in PPV (95% CI), %†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narrative and codified (complete)</td>
<td>3,585</td>
<td>94 (91–96)</td>
<td>63 (51–75)</td>
<td>Reference</td>
</tr>
<tr>
<td>Codified only</td>
<td>3,046</td>
<td>88 (84–92)</td>
<td>51 (42–60)</td>
<td>6 (2–9)†</td>
</tr>
<tr>
<td>NLP only</td>
<td>3,341</td>
<td>89 (86–93)</td>
<td>56 (46–66)</td>
<td>5 (1–8)†</td>
</tr>
<tr>
<td>Published administrative codified criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥3 ICD-9 RA codes</td>
<td>7,960</td>
<td>56 (47–64)</td>
<td>80 (72–88)</td>
<td>38 (29–47)†</td>
</tr>
<tr>
<td>≥1 ICD-9 RA codes plus ≥1 DMARD</td>
<td>7,799</td>
<td>45 (37–53)</td>
<td>66 (57–76)</td>
<td>49 (40–57)‡</td>
</tr>
</tbody>
</table>

* The complete classification algorithm was also compared with criteria for RA used in published administrative database studies. RA = rheumatoid arthritis; PPV = positive predictive value; 95% CI = 95% confidence interval; NLP = natural language processing; ICD-9 = International Classification of Diseases, Ninth Revision; DMARD = disease-modifying antirheumatic drug.
† Difference in PPV = PPV of complete algorithm – comparison algorithm or criteria.
‡ Significant difference in PPV compared with the complete algorithm.
Types of EMR data

Natural language processing

Liao, Cai, et al., BMJ 2015
Natural language processing (NLP)

Computational method for text processing based on the rules of linguistics
I saw the girl with the ophthalmoscope.
NLP ≠ “find” command in Word

• Negation
  • The patient has no erosions in the MCPs.

• Inverted syntax
  • Colon, ascending and descending, biopsy

• Relation
  • Tamoxifen is used in the treatment of breast cancer

• Morphologic variations
  • Tobacco, 30 pack years, past smoker, +tob → smoking
## Illustrative dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Dx code</th>
<th>Lab</th>
<th>Dis+</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>22</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>F</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>75</td>
<td>F</td>
<td>1</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>56</td>
<td>M</td>
<td>0</td>
<td>56</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>54</td>
<td>F</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>F</td>
<td>1</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>48</td>
<td>F</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

## Training set
Pattern recognition

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Dx code</th>
<th>Lab</th>
<th>Dis+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>F</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>F</td>
<td>1</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>F</td>
<td>0</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>F</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>82</td>
<td>F</td>
<td>1</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

+200 subjects

Training set

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Dx code</th>
<th>Lab</th>
<th>Dis+</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>22</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>F</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>75</td>
<td>F</td>
<td>1</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>56</td>
<td>M</td>
<td>0</td>
<td>56</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>54</td>
<td>F</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>F</td>
<td>1</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>48</td>
<td>F</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

+1000 features
Pattern recognition

- More potential “features” *may* enable more accurate algorithms
  - Features can also add noise

- Challenge to identify the important features and their patterns

### Training set

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Dx code</th>
<th>Lab</th>
<th>Dis+</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>22</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>F</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>75</td>
<td>F</td>
<td>1</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>M</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>56</td>
<td>M</td>
<td>0</td>
<td>56</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>54</td>
<td>F</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>F</td>
<td>1</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>48</td>
<td>F</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

+200 subjects

+1000 variables
Artificial Intelligence & Machine Learning

• Artificial intelligence (AI)
  • Intelligence demonstrated by machines
    • Contrast to human intelligence

• Machine learning (ML) → subset of AI
  • Requires training set
  • Focus on prediction (vs causality)
    • Does not address why or how to change outcomes
  • Learning structure from data
    • Pattern recognition
  • Examples
    • Least absolute shrinkage and selection operator (LASSO) regression
    • Support vector machine (SVM)
Types of EHR data

Liao, Cai, et al., BMJ 2015
Approach to developing phenotype algorithms using EHR data

• Chart review - not feasible

• Rule-based
  • Relies on human expertise to identify important features
  • Algorithm is a combination of AND, NOT, OR

• Machine learning
  • Data driven method to select features and develop algorithm
Machine learning, NLP, and EHR
Pipeline for phenotyping

Limitations of supervised ML approaches for phenotyping

• Require gold standard labels through manual chart review
  • Notes not always available
  • Time and resource intensive
  • Not scalable

• Inefficient
  • Large amount of unlabeled data contains “noisy labels”
## Comparison of EHR phenotype algorithm approaches

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Supervised or semi-supervised</th>
<th>Unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual chart review for labels</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Feature selection</td>
<td>Manual or automated</td>
<td>Automated</td>
</tr>
<tr>
<td>Rule-based, e.g. 2 ICD + 1 Rx</td>
<td>Option</td>
<td>N</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Option</td>
<td>Y</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Varies</td>
<td>High</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Data available</td>
<td>Needs validation</td>
</tr>
</tbody>
</table>
Unsupervised approaches for phenotyping w/ EHR data
Unsupervised approaches

• Anchor, Halpern et al., 2014
• XPRESS, Agarwal et al., 2016
• APHRODITE, Banda et al., 2017
• PheNorm, Yu et al...Cai, 2017
• MAP, Liao, Sun et al...Cai, 2019
PheNorm: Assumption

Surrogate disease labels $S_i$ (i.e. ICD-9 codes) normalized by a patient’s healthcare utilization $U_i$ (i.e. count of patient notes) are log-normally distributed with mean $\mu_Y$ dependent on the patient’s true disease status $Y_i$

$$\log(S_i) \sim \text{Norm}(\mu_Y + c \log(U_i))$$
Abbreviations

• Main Features
  • $x_{ICD}$: # ICD-9 codes of target phenotype for each patient
  • $x_{NLP}$: # positive NLP mentions only, e.g. not negated, remove mention from family hx, of target phenotype from all notes for a given patient
  • $x_{ICDNLP} = x_{ICD} + x_{NLP}$

• Healthcare utilization: $x_{note} = # \text{notes for each patient}$

• Additional potential features: $x_1 \ldots x_p$
  • Counts of medication, mentions of signs and symptoms in the notes, etc
  • Can be curated through prior knowledge or via data-driven approaches
PheNorm Step 1: Normalization

Step 1

\[ z = \log(1 + x) - \alpha \log(1 + x_{\text{note}}) \]

Normal Mixture Normalization

Accuracy Improvement
PheNorm Step 2: Denoising w/ other features

\[ z = \log(1 + x) - \alpha \log(1 + x_{\text{note}}) \]

Normal Mixture Normalization

Random Corruption

Denoising

\[ y \sim \tilde{X} \]
PheNorm workflow

**Raw Feature**

**Normalized Feature**

\[ z = \log(1 + x) - \alpha \log(1 + x) \]

**Normal Mixture Normalization**

**Random Corruption**

**Denoising**

\[ y \sim \bar{x} \]
Table 1. AUCs of the raw feature $x$, the normalized feature $z$, the PheNorm scores using SAFE feature for denoising with a dropout rate of 0.3, PheNorm$_\text{vote}$, the supervised algorithms trained with SAFE features with $N = 100$, 200, or 300 labels, as well as the XPRESS and Anchor algorithms.

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>RA</th>
<th>CD</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{ICD}$</td>
<td>0.844</td>
<td>0.868</td>
<td>0.824</td>
<td>0.812</td>
</tr>
<tr>
<td>$z_{ICD}$</td>
<td>0.875$^{0.031\ast}_{0.010}$</td>
<td>0.901$^{0.033\ast}_{0.008}$</td>
<td>0.877$^{0.053\ast}_{0.013}$</td>
<td>0.859$^{0.047\ast}_{0.012}$</td>
</tr>
<tr>
<td>PheNorm$_{ICD}$</td>
<td>0.899$^{0.024\ast}_{0.004}$</td>
<td>0.925$^{0.028\ast}_{0.009}$</td>
<td>0.911$^{0.033\ast}_{0.005}$</td>
<td>0.900$^{0.041\ast}_{0.005}$</td>
</tr>
<tr>
<td>$x_{NLP}$</td>
<td>0.840</td>
<td>0.898</td>
<td>0.906</td>
<td>0.904</td>
</tr>
<tr>
<td>$z_{NLP}$</td>
<td>0.864$^{0.025\ast}_{0.011}$</td>
<td>0.923$^{0.025\ast}_{0.011}$</td>
<td>0.947$^{0.041\ast}_{0.007}$</td>
<td>0.931$^{0.026\ast}_{0.006}$</td>
</tr>
<tr>
<td>PheNorm$_{NLP}$</td>
<td>0.884$^{0.019\ast}_{0.003}$</td>
<td>0.934$^{0.014\ast}_{0.005}$</td>
<td>0.946$^{0.001}_{0.004}$</td>
<td>0.935$^{0.004\ast}_{0.002}$</td>
</tr>
<tr>
<td>$x_{ICD_NLP}$</td>
<td>0.865</td>
<td>0.903</td>
<td>0.902</td>
<td>0.901</td>
</tr>
<tr>
<td>$z_{ICD_NLP}$</td>
<td>0.895$^{0.030\ast}_{0.008}$</td>
<td>0.936$^{0.032\ast}_{0.009}$</td>
<td>0.944$^{0.042\ast}_{0.008}$</td>
<td>0.933$^{0.032\ast}_{0.007}$</td>
</tr>
<tr>
<td>PheNorm$_{ICD_NLP}$</td>
<td>0.899$^{0.004\ast}_{0.002}$</td>
<td>0.936$^{0.001}_{0.002}$</td>
<td>0.945$^{0.001}_{0.002}$</td>
<td>0.935$^{0.002}_{0.002}$</td>
</tr>
<tr>
<td>PheNorm$_{\text{vote}}$</td>
<td>0.899</td>
<td>0.937</td>
<td>0.945</td>
<td>0.933</td>
</tr>
</tbody>
</table>

Comparison is with the previous step; asterisk indicates positive increment at the significance level of 0.05.
MAP: a refinement of PheNorm

• Limitations of PheNorm
  • Output linear score vs predicted probability of disease
  • Does not identify threshold value for classifying subjects as cases

• MAP (multi-modal automated phenotyping)
  • Fit a sequence of mixture models $\rightarrow$ predicted probabilities for all patients & estimates of disease prevalence from each fitting
  • Synthesize information via model averaging
  • Classifying as a case if predicted probabilities exceed threshold
Step 1: Assemble NLP & ICD data for each PheWAS group

- Mappings
  - ICD9 codes in a Phecode group → UMLS CUIs
  - ICD9 code → UMLS CUI
  - ICD9 string → UMLS CUI
  - PheWAS string → UMLS CUI

UMLS= Unified Medical Language System
CUI= concept unique identifier
<table>
<thead>
<tr>
<th>Code String</th>
<th>Code</th>
<th>ICD_9</th>
<th>ICD9_Str</th>
<th>CUI ICD9</th>
<th>CUI ICD9_String</th>
<th>CUI Code_String</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rheumatoid arthritis</td>
<td>714.0</td>
<td>714.0</td>
<td>rheumatoid arthritis</td>
<td>C0003873</td>
<td>C0003873</td>
<td>C0003873</td>
</tr>
<tr>
<td></td>
<td>714.1</td>
<td>714.1</td>
<td>Felty’s syndrome</td>
<td>C0015773</td>
<td>C0015773</td>
<td>C0003873</td>
</tr>
<tr>
<td></td>
<td>714.2</td>
<td>714.2</td>
<td>Other rheumatoid arthritis with visceral or systemic involvement</td>
<td>C0157914</td>
<td>C0157914</td>
<td>C0003873</td>
</tr>
<tr>
<td>Rheumatoid arthritis and other inflammatory polyarthropathies</td>
<td>714.4</td>
<td>714.4</td>
<td>Chronic postrheumatic arthropathy</td>
<td>C0152084</td>
<td>C0152084</td>
<td>C0157913</td>
</tr>
<tr>
<td></td>
<td>714.8</td>
<td>714.8</td>
<td>Other specified inflammatory polyarthropathies</td>
<td>C0157919</td>
<td>C0157919</td>
<td>C0157913</td>
</tr>
<tr>
<td></td>
<td>714.89</td>
<td>714.89</td>
<td>Other specified inflammatory polyarthropathies</td>
<td>C0157919</td>
<td>C0157919</td>
<td>C0157913</td>
</tr>
<tr>
<td></td>
<td>714</td>
<td>714</td>
<td>Rheumatoid arthritis and other inflammatory polyarthropathies</td>
<td>C0157913</td>
<td>C0157913</td>
<td>C0157913</td>
</tr>
</tbody>
</table>
Step 2: Joint Analysis of NLP & ICD

- Fit multiple Poisson and log-normal mixture models to \{NLP, ICD\} counts → probabilities of phenotype(+)

- Adjust for healthcare utilization
Step 3: Synthesize information from all model fittings

- Each fitted model provides a predicted probability of phenotype for each patient
- The final predicted probability of phenotype(+) is the average predicted probabilities from all fitted models

Step 4: Cut-off estimate based on population prevalence p

- Fitted mixture models → estimated phenotype prevalence
- Classify p% patients with highest predicted probabilities as phenotype(+) (as opposed to the standard method based on ICD code thresholding)
Performance of phenotype algorithms across conditions

[Graph showing AUC for various conditions]
Applications: Phenomics Library

• Veterans Affairs Health Centers
  • ~22 million veterans nationwide
    • Million Veteran Program (MVP)
  • Ported and validated supervised and unsupervised approaches
EHR research platform for translational studies

VA EHR data
Summary

• Phenotyping approaches designed for prevalent conditions
• Optimized for EHR data
• Robust and portable
• Supervised vs unsupervised based on downstream use
  • Cohort creation
  • Phenotype screens, e.g. PheWAS
  • Association studies

• Future directions
  • Algorithms for incident or recurrent conditions
  • Can existing algorithms catch incident conditions within a time window?
Thank you

BWH
Selena Huang
Brittany Weber
Austin Cai
Zeling He
Nicholas Link (VA)
Kumar Dahal
Dana Weisenfeld
Charlotte Golnik
Thany Seyok
Andrew Cagan
Jackie Stratton

DBMI, Harvard Medical School
Tianxi Cai
Chuan Hong
Hajime Uno (DFCI)
Junwei Liu (HSPH)
Amanda King
Isaac Kohane
Susanne Churchill

Boston VA Healthcare System/MAVERIC
J. Michael Gaziano
Christopher O’Donnell
Kelly Cho

Paul Monach
Anne Ho
Jennifer Huffman
Lauren Costa
Petra Schubert
Laura Tarko
Ashley Galloway

National Heart Lung and Blood Institute
NIAMS