Introduction & Motivation

- Phenotyping algorithms describe how to map patients’ electronic health records to meaningful clinical concepts.
- Current phenotyping algorithms are cumbersome, require significant amounts of human supervision, and are difficult to scale.[1]
- We address these problems by developing:
  - State-of-the-art, automated phenotyping algorithms based upon higher order tensor factorization [2,3]
  - A web service allowing for easy access to phenotyping algorithms.

Data

The study involves a cohort of 7744 patients from an outpatient clinic at the Vanderbilt University Medical Center. Details regarding the data are shown below:

<table>
<thead>
<tr>
<th>Concepts</th>
<th>No. unique</th>
<th>Post processing</th>
<th>Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD codes</td>
<td>11,409</td>
<td>PheWAS mapping:</td>
<td>Binary</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1058 ICD codes</td>
<td></td>
</tr>
<tr>
<td>Medications</td>
<td>5,741</td>
<td>Mapping to classes:</td>
<td>Binary</td>
</tr>
<tr>
<td>Labs</td>
<td>6,841</td>
<td>N/A</td>
<td>Average</td>
</tr>
<tr>
<td>CPT codes</td>
<td>6,841</td>
<td>N/A</td>
<td>Binary</td>
</tr>
</tbody>
</table>

Details regarding clinician-derived phenotypes in this cohort are shown below:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>No. cases</th>
<th>No. controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2 diabetes</td>
<td>1040</td>
<td>3357</td>
</tr>
<tr>
<td>Resistant hypertension</td>
<td>787</td>
<td>2016</td>
</tr>
<tr>
<td>Ocular hypertension</td>
<td>6</td>
<td>82</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>12</td>
<td>52</td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
<td>225</td>
<td>7515</td>
</tr>
<tr>
<td>Major adverse cardiac events</td>
<td>1532</td>
<td>6212</td>
</tr>
</tbody>
</table>

Tensor Construction

Co-occurrences of events within a patient’s history are captured in the tensor as binary values.

Our Approach

- Data
- Medications
- Diagnoses
- Labs
- Procedures
- EHR data

Web Service Uploads

Tensor Construction

Phenotyping Algorithm

Post Processing

Bias Tensor

Signal Tensor

Phenotypes

Sample Phenotypes

Predictive Modeling

We performed classification of patients into case and control status for a series of clinician derived phenotypes. Predictive performance using 150 or more phenotypes can achieve accuracy higher than baseline, providing 10-fold feature reduction. Performance comparison is shown for logistic regression classifier using raw phenotypes, clustered phenotypes, and raw diagnoses and medications as predictive features.

| Predictive feature sets as a function of number of phenotypes. B) AUC comparisons for feature set derived from clusters. For 5 or more clusters, the performance is similar to that of the original set of phenotypes.

Web Services

We developed a scalable web service for running phenotyping algorithms on data uploaded by users. Data is uploaded in the form of JSON files following the HL7 FHIR[4] data model. Different phenotyping algorithms are implemented on the web service, including the algorithm based upon tensor factorization, as well as algorithms from the Phenotype Knowledge Base (PheKB).[5] Results are displayed to the user via interactive visualizations.

References: