Pancreatic ductal adenocarcinoma (PDAC) is the 4th leading cause of cancer deaths in the US and the only cancer predicted to rise in the next decade. PDAC is most often found in stage III and IV with a survival measured in months. Our goal is to identify asymptomatic early stage PDAC from the Duke EMR data using a supervised topic model to and to follow high risk patients prospectively.

Duke Electronic Medical Record (EMR)
- Spans 2004-2013
- 210,140 patients primarily from Durham County
- 11,550 unique ICD9 diagnosis codes
- 15,293 patients with diabetes
- 11,234 patients over 50 years with diabetes
- 5,712 over 50 with diabetes >3 years “new-onset diabetes”

The timeline shows median times of diagnosis codes prior to PDAC and their respective frequency.
- Diagnoses found closer to a PDAC diagnosis pose greater risk.
- While individually, each code may not warrant concern, in conjunction they may be predictive of risk.
Supervised Latent Dirichlet Allocation Approach

**Theory of sLDA**

- $\gamma$: Label set
- $\theta$: Topic proportion in each patient
- $Z_\omega$: Topic assignment for each diagnosis code
- $\beta$: Codes contribution for each topic
- $\omega$: Observed diagnosis code
- $\alpha$: Dirichlet topic prior parameter
- $\eta$: Dirichlet word prior parameter
- $\mu$: Label prior for topic

**Group partition**

- 189 PDAC patients
- ~2000 general patients

**Regression Coefficient**

- Get regression coefficients from the topic model to make a predictive list.
- Positive topics are those with regression coefficients above 0.

**Word Cloud**

- 105 groups of regression coefficients, and over 200 positive topics, are the sum of loadings in all positive topics for each diagnosis code. The top 50 diagnosis codes are shown.

Label pancreatic cancer patients with 0; label all general patients with 1.

Randomly partition subset of general population into groups of about 2000 patients. Apply the topic model to each of the partitioned groups separately.
Predictive Performance and Future Work

### Predictive Performance

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>patient_label</td>
<td>prediction_score</td>
</tr>
<tr>
<td>2</td>
<td>1.06027</td>
<td>1.0292589</td>
</tr>
<tr>
<td>3</td>
<td>1.04522</td>
<td>1.0643422</td>
</tr>
<tr>
<td>4</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>5</td>
<td>1.06694</td>
<td>1.0292589</td>
</tr>
<tr>
<td>6</td>
<td>1.02318</td>
<td>1.0292589</td>
</tr>
<tr>
<td>7</td>
<td>1.02318</td>
<td>1.0292589</td>
</tr>
<tr>
<td>8</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>9</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>10</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>11</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>12</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>13</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
<tr>
<td>14</td>
<td>1.40835</td>
<td>1.0433043</td>
</tr>
</tbody>
</table>

### Results

- Patients with a label of 0, without a pancreatic cancer 157.* diagnosis codes for PDAC, and a prediction score greater than 0, which predicts PDAC (yellow) are candidate high risk.

- 500 high risk patients were found in common across multiple trials of different seeds in the general patient population.

### Future Work

- Apply the topic model to see the predictive performance in other cancers or neurodegenerative diseases that also develop silently.

- Provide clinicians a list of high risk patients for a prospective study.

- Validate externally with EMR data from other health systems.