Automated Phenotyping of Patient EMR Data: Feature Extraction and Selection

Rolando Acosta Nuñez, William Artman, Cassandra Burdziak
Genetic Data

GATTACACCGTAAAATACATA
GACGTAAGAGGGGAATTCCA
GATTAACAGGATTTGACAGGA
TCAGGATCAGGATACCAGTAA
ACGGATACCTACGATCAAGTT

Phenotype Data

Computational Methods

Gene-Disease Association

Personalized Medicine
Percent of non-federal acute care hospitals with adoption of EHR Systems by level of functionality: 2008-2014

Structured Data
- ICD-9: 714
- BMI: 21.5
- Medication: Methotrexate, 7.5mg PO as a single weekly dose

Unstructured Data
- ….bilateral joint pain and stiffness. Physical exam revealed nodules, swelling, limited range of motion. X-ray showed soft-tissue swelling and early erosions. Blood tests showed presence of rheumatoid factor and elevated SED rate ….

develop disease classification model for patient medical record notes
Septic arthritis, also known as infectious arthritis, may represent a direct invasion of joint space by various microorganisms, most commonly caused by a variety of bacteria. However, viruses, mycobacteria, and fungi have been implicated. Reactive arthritis is a sterile inflammatory process that usually results from an extra-articular infectious process. Bacteria are the most significant pathogens because of their rapidly destructive nature. For this reason, the current discussion concentrates on the bacterial septic arthritides. Failure to recognize and to appropriately treat septic arthritis results in significant rates of morbidity and may even lead to death …

... Streptococcal species, such as Streptococcus viridans, S pneumoniae, and group B streptococci, account for 20% of cases. Aerobic gram-negative rods are involved in 20-25% of cases. Most of these infections occur in people who are very young, who are very old, who are diabetic, who are immunosuppressed, and who abuse intravenous drugs.
Patient-Level CUI Screening:
Spearman’s Rank Correlation, $r > 0$

Note-Level CUI Screening:
non-main CUI mentioned in >5% of notes including main CUI
<table>
<thead>
<tr>
<th>CUI</th>
<th>Patient 1</th>
<th>Patient 2</th>
<th>Patient 3</th>
<th>Patient 4</th>
<th>Patient 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main CUI</td>
<td>92</td>
<td>16</td>
<td>44</td>
<td>368</td>
<td>144</td>
</tr>
<tr>
<td>C1247884</td>
<td>84</td>
<td>20</td>
<td>59</td>
<td>320</td>
<td>152</td>
</tr>
<tr>
<td>C0934556</td>
<td>122</td>
<td>89</td>
<td>153</td>
<td>0</td>
<td>167</td>
</tr>
</tbody>
</table>

\[ r = 1, \ p = 0.017 \]

\[ r = -0.10, \ p = 0.95 \]
Multiple Testing Problem

$\alpha = 0.05$

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>129</td>
</tr>
<tr>
<td>Migraine</td>
<td>71</td>
</tr>
<tr>
<td>Septic Arthritis</td>
<td>100</td>
</tr>
<tr>
<td>Giant Cell Arteritis</td>
<td>93</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>66</td>
</tr>
</tbody>
</table>

Multiple Testing Correction $\alpha_i'$
Bonferroni Correction

\[ \alpha' = \frac{0.05}{\# \text{ of tests}} \]

Benjamini-Hochberg

Permutation Test

<table>
<thead>
<tr>
<th>CUI</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main CUI</td>
<td>92</td>
<td>16</td>
<td>44</td>
<td>368</td>
<td>144</td>
</tr>
<tr>
<td>C1247884</td>
<td>84</td>
<td>20</td>
<td>59</td>
<td>320</td>
<td>152</td>
</tr>
<tr>
<td>C0934556</td>
<td>122</td>
<td>89</td>
<td>153</td>
<td>0</td>
<td>167</td>
</tr>
</tbody>
</table>

\[ R1 \quad R2 \quad R1^* \quad R2^* \]
Giant Cell Arteritis

- Polymyalgia rheumatica: $r = 0.59$
- Temporal artery biopsy: $r = 0.45$

Correlation coefficients:

- Bonferroni: 0.0005319149
- BH: 0.03035306
- Permutation: 0.0005818897

# selected: 77, 92, 78

White: not significant
Giant Cell Arteritis

Highly Correlated Terms by Type

Disease or Syndrome:
- Giant Cell Arteritis

Pharmacologic Substance:
- Corticosteroid
- Methotrexate
- Ibuprofen
- Vitamin D
- Aspirin
- Anti-inflammatory drugs
- Prednisolone

Finding:
- Blood pressure
- Stroke
- Diabetes
- Anemia

Sign or Symptom:
- Swelling
- Lesion
- Fall
- Sweats
- Examined
- Thickened
- Mental status
- Feeling better
- Discomfort
- Head pain
- Flare
- Malaise
- Tenderness
- Tiredness

Diagnostic Procedure:
- Chest X rays
- CT scan
- Palpation
- Magnetic resonance imaging
<table>
<thead>
<tr>
<th>correlation coefficient</th>
<th>Bonferroni</th>
<th>BH</th>
<th>Permutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000694444</td>
<td>5.029342e-08</td>
<td>0.0006302195</td>
<td></td>
</tr>
</tbody>
</table>
Migraine

Highly Correlated Terms by Type

**Finding**
- swelling
- males
- females
- past medical history of
- loss weight
- fever
- lesion
- personal history

**Diagnostic Procedure**
- computerized tomography
- x rays
- mri scan

**Pharmacologic Substance**
- liver
- prednisone
- caffeine
- acetaminophen
- cardiovascular drugs
- blood pressure medicines
- pain relieving drugs

**Sign or Symptom**
- nausea symptoms
- heart problems
- constipation
- headache
- coughing
- tired out
- pain
- diarrhea
- tenderness
# selected:          66                     64                    66

<table>
<thead>
<tr>
<th>Bonferroni</th>
<th>BH</th>
<th>Permutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0007462687</td>
<td>3.527381e-16</td>
<td>0.0005846256</td>
</tr>
</tbody>
</table>

Osteoarthritis
Septic Arthritis
Discussion

- Limitations
  - Relying on quality of sources
  - Diseases with multiple names which correspond to different CUIs in the UMLS
    - Example: Giant Cell Arteritis and Temporal Arteritis

- Conclusion and Future Directions
Cai Lab and Collaborators
• Dr. Tianxi Cai
• Dr. Sheng Yu

Summer Program in Biostatistics and Computational Biology
• Dr. Rebecca Betensky
• Tonia Smith
• Heather Mattie
• Eleanor Murray
• Joshua Barback