A statistical approach to utilizing electronic health records

Joseph E. Lucas, Ph. D.
Know your Patients

- Can a health system make use of what it knows about patients to make informed decisions?
  - Lots of data.
  - What questions to ask?
  - How to generate answers?

- Where is my patient going next?

- What will help me know?
  - EHR data, labs, genetics, “quantified self” data, genomics?
Complicated Messy Data

- Electronic Health Records data has
  - Continuous data (labs, age, vitals)
  - Categorical data (gender, race, family history)
  - Written text (nurses’ and physicians’ notes, radiology reports)
  - Images (x-ray, CT, EKG)

- Important information everywhere
  - Example: A diabetic might have any or all of the following
    - Synonym of “diabetes” in a note
    - High lab values (glucose, HbA1C)
    - Relevant medications
    - Billing codes related to treatment of diabetes
    - Predisposing demographics (weight, race, family history)
    - Genetic predisposition (TCF7L2, JAZF1, HHEX, etc)

- We want to incorporate all of this information
- Don’t want to be fooled by mistakes
Approach

- Automated “Patients like me”
- Create groups of homogeneous patients
- This allows:
  - Automated generation of differential diagnosis
  - Novel comparative effectiveness studies
  - Listing of treatment options
  - Identification of Adverse drug events
  - Estimation of disease progression and prognosis
  - Assessment clinical utility of novel lab tests
- Predict probable patient type from other data
- Group patients through time
Model

\[ P(n_i = k \mid X) = f(m_k) \prod_l P_l(n_i = k \mid X) \]

- Product densities from each component
- Each data type is treated independently
- Doesn’t explicitly model correlation across different data types
Topic Model for Text

\[ q_k \sim Dir(\beta) \]
\[ p(w_{i,j} | n_i = k) \sim MN(q_k) \]

\[ P(w_i) \propto \sum_k f(m_k) \prod_j MN(w_{i,j} | q_k) Dir(q_k | \beta) I_{[n_i = k]} \]

Compute the likelihood that \( n_i = k \):

1. Start with likelihood 1.
2. Pick a word from document \( i \).
3. Multiply by the fraction of words in topic \( k \) that are equal to the chosen word.
4. Add the word to the topic and repeat until all words have been added.

- \( i \) – index for document
- \( j \) – index for word
- \( k \) – index for cluster
- \( n_i \) – topic for document \( i \)
- \( q_k \) – latent word frequencies
- \( m_k \) – # documents in cluster \( k \)
- \( w_i \) – indicator vector

Multiply likelihood by \( \frac{2}{5} \).
Topic Models, NLP, Mixed data

Clustering patient visits to identify features of disease processes

**Diagnoses**
- Angina pectoris
- CAD
- Chest pain acute
- Chest pressure
- MI
- Palpitations
- Chest tightness
- Myocardial infarction
- Pericarditis

**Medications**
- Asa/buffers
- Aspirin
- Hctz
- Lantus
- Lipitor
- Lisinopril
- Metoprolol
- Metformin

**Medical (NLP) Concepts**
- No trauma
- No fever
- No diarrhea
- Aspirin
- Back
- Chest
- Dyspnea
- Exercise

**Profile of a patient with heart disease**
- Combined evidence from multiple sources
- Relationships between drugs and patient group
- Metformin is protective
- Lantus: heart failure is side effect

Nice features

- Very fast mixing
- MCMC without ever having to sample parameters
- Can compute MAP

- Model knows when it doesn’t know
- Use to identify trends in other data

Example: Suppose a small percentage of your customers use your clinic. What do they purchase in the rest of the store? Does that inform on other potential clinic users? Can/should you market to them?
Case Study: Data

- 54,000 records from ED
- Contains
  - Vocabularies
    - Notes
    - Orders
    - Patient reported meds
    - Diagnoses
  - Categorical data
    - Chief complaint
    - Gender
    - Disposition
    - Zip code
  - Continuous data
- None is codified
- All data subject to parsing errors

- Age
- Priority
- Vitals
- Weight
Messy data

weakness/aching/heads
weakness/discomfort
weakness/dizziness
weakness/dizziness/recent
weakness/faintness/congestion
weakness/fatigue
weakness/flaccid
weakness/heaviness
weakness/numbness
weakness/pain
weakness/shaking
weakness/tingling
weakness;
weaknessambulatory
weaknesscoughfever
wbswent
weaknesscoughing
weaknessdiahrrea
weaknessdizziness
weaknesses
weaknessfalling
weaknessfatigue
weaknessn\aloc
weaknessper
weaknesss
weaknesssob
weaknesssore
weaknessunstable
weaknessvison
weaknessx
weaknss

Over 50,000 unique “words” with no copy editing. How to clean up mistakes?
Unified Medical Language System (UMLS)

- **Metathesaurus**
  - 10.5 million atoms in thesaurus (1.2 in SNOMED)

- **Semantic Network**
  - 51.4 million relationships (2.9 in SNOMED)

- **Lexicon**
  - 3.3 million spellings, inflections, properties, modifiers, abbreviations

- **Together: 23 gigabytes of plaintext**
  - At 60 words per minute this is about 7 years of non-stop typing.
UMLS Challenges

- Incredible number of arbitrary decisions
  - What should be the semantic types?
    - Why is Mammal a semantic type but not primate
  - What words to include?
  - What relationships?

- Overly inclusive
  - 10.5 million atoms in thesaurus (1.2 in SNOMED)
  - Orders of magnitude more words in the thesaurus than are unique words in the 50k records
  - How does one curate something this large?

- What do you do with it?
1. Pt here with c/o N/V and "shakes"
2. decreased po intake x 2 day
3. pt has pain pump which has been out of medication x 1 week
4. pt now taking PO narcotics
5. but presenting with N/V
6. pt with chronic back and neck pain

1. Tremor [Sign or Symptom]
2. Decreased [Quantitative Concept], Oral [Spatial Concept], /day [Temporal Concept]
3. Pain [Sign or Symptom], Pump, device [Medical Device], Drugs [Pharmacologic Substance], week [Temporal Concept]
4. Take [Health Care Activity], Oral [Spatial Concept]
5. Presentation [Idea or Concept], N+ (tumor staging) [Intellectual Product]
6. Chronic [Temporal Concept], Neck pain [Sign or Symptom]
Patients with Chest Pain

angina pectoris not otherwise specified
CAD
chest pain acute
chest pain musculoskeletal
chest pain other
chest pain unspecified
chest pressure
cHEST wall pain
GERD gastro-esophageal reflux disease
MI
musculoskeletal chest pain
palpitations
unstable angina
chest tightness
shortness of breath
acute chest pain other
Myocardial infarction unspecified
pericarditis acute
costochondritis
coronary artery disease
Chest Pain

RN Notes

Diagnosis

Medications

Orders

meta Notes

age

- Base distribution
- Cluster 45
Fever / Febrile Seizure
Pharmacovigilance: Patients with Diabetes

Blood pressure

Dementia

Schizophrenia

Diuretic
Some other associations

<table>
<thead>
<tr>
<th>Drug</th>
<th>Indication</th>
<th>Patient cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tussionex</td>
<td>Opioid</td>
<td>TIA</td>
</tr>
<tr>
<td>Altace</td>
<td>Blood pressure</td>
<td>Chest pain</td>
</tr>
<tr>
<td>Metformin</td>
<td>Blood sugar</td>
<td>Nose bleed</td>
</tr>
<tr>
<td>Tylenol</td>
<td>Analgesic</td>
<td>Rabies</td>
</tr>
<tr>
<td>Buspirone</td>
<td>Anxiolytic</td>
<td>Dog/cat bite</td>
</tr>
<tr>
<td>Cassodex</td>
<td>Chemo</td>
<td>Sickle cell</td>
</tr>
<tr>
<td>Vasotec</td>
<td>Blood pressure</td>
<td>Nose bleed</td>
</tr>
<tr>
<td>Prednisolone</td>
<td>Inflammation</td>
<td>Eye pain</td>
</tr>
<tr>
<td>Wellbutrin</td>
<td>Depression</td>
<td>Nose bleed</td>
</tr>
<tr>
<td>Zomig</td>
<td>Migraine</td>
<td>Chest pain</td>
</tr>
</tbody>
</table>
Models for comparison

- Use UMLS to process RN notes
- Ignore UMLS and just throw out rare “words” from the RN notes
- Cluster with chief complaint
- Cluster with MetaMAP only
Validation of Associations

- Train the model on 20,000 samples selected uniformly at random
  - Test for association between patient clusters and drugs
- Fit the model to the test data
- Model using MetaMap
  - 4% of drug – symptom association tests have p-value<.05 (3021)
  - 56% validate
- Model using terms in RN notes
  - 2% of tests significant (2258)
  - 58% validate
- Model using chief complaint
  - 3% of tests significant (9381)
  - 18% validate
- Model using MetaMap concepts only
  - 3% of tests significant (9319)
  - 22% validate
Differential Diagnosis

- Average correlation (nonparametric)
  - Model with metamap, .61
  - Model with RN notes, .62
  - Chief complaint as classifier, .26
Identify New Patient Populations

1.3% of patients.
Other possible groups to add

- Minor trauma requiring x-ray, cast, splint
  - 13% of patients
- Chronic pain, headache, sickle cell crisis
  - 5.5%
- Cellulitis, local infection
  - 4%
- >50% of patients on either hydroxyurea or procrit presented with sickle cell crisis
## ID Patients by Drug

<table>
<thead>
<tr>
<th>Drug</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ritonavir</td>
<td>metabolic imbalance (kidney disease)</td>
</tr>
<tr>
<td>Hydroxyurea</td>
<td>sickle cell</td>
</tr>
<tr>
<td>Procrit</td>
<td>sickle cell</td>
</tr>
<tr>
<td>Levodopa</td>
<td>Altered mental status</td>
</tr>
</tbody>
</table>
Disease Processes over Time

Learn a parametric model that describes the likelihood of observing particular diseases through time.

- Probability of disease over time

\[ y = \text{disease severity} \]
\[ \log(y) = -a(d^2 + x^2)^{1/2} + bx \]

- ‘a’ and ‘b’ give control of rate of change of disease severity
- ‘d’ gives control of peak width
- Additional parameters allow control of height and temporal shift

Combination of clustering model and temporal disease severity model allows identification and tracking of disease through time.
Disease Trajectory

- New Data set: 3.5 million patients, 7 years
- Why is it important to predict future disease
  - Identification of high risk patients for early intervention
- Early intervention might be:
  - Standard therapy
  - Health coaching (telephone or in person)
  - Enrollment in clinical trial

Trajectory of a patient population with Alzheimer’s

Early signs that patient is developing Alzheimer’s disease
Utilize all types of data to track and predict important features in the evolution of a patient’s disease.
Acknowledgments

Quintiles Payer / Provider
• Brian Kelly
• Jon Morris

Duke University
• Ricardo Henao
• Larry Carin
• Geoff Ginsburg

http://EHRanalytics.blogspot.com

joseph.lucas@quintiles.com
joe@stat.duke.edu