Joint Analysis of Multiple Data Types in 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?

Do they want to?

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

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

- Product of densities
- Model for each data type
- No nuisance parameters in MCMC

- Probabilities for cluster membership
- E-M algorithm for MAP
- ID cluster features in new data

Normal – Gamma Clustering

Multinomial – Dirichlet Clustering

Multiply likelihood by \( \frac{2}{5} \).
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
<table>
<thead>
<tr>
<th>Weakness/Aching/Headaches</th>
<th>Weakness/Shaking</th>
<th>Weakness/Numbness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weakness/Discomfort</td>
<td>Weakness/Tingling</td>
<td>Weakness/Pain</td>
</tr>
<tr>
<td>Weakness/Dizziness</td>
<td>Weakness;</td>
<td>Weakness/fatigue</td>
</tr>
<tr>
<td>Weakness/Dizziness/Recent</td>
<td>Weaknessambulatory</td>
<td>Weakness/Flaccid</td>
</tr>
<tr>
<td>Weakness/Faintness/Congestion</td>
<td>Weaknesscoughfever</td>
<td>Weakness/Sore</td>
</tr>
<tr>
<td>Weakness/Fatigue</td>
<td>Weaknesscoughing</td>
<td>Weakness/unstable</td>
</tr>
<tr>
<td>Weakness/Flaccid</td>
<td>Weaknessdiarrhea</td>
<td>Weaknessx</td>
</tr>
<tr>
<td>Weakness/Heaviness</td>
<td>Weaknessdizziness</td>
<td>Weaknss</td>
</tr>
<tr>
<td>Weakness/Numbness</td>
<td>Weaknesses</td>
<td></td>
</tr>
<tr>
<td>Weakness/Pain</td>
<td>Weaknessfalling</td>
<td></td>
</tr>
</tbody>
</table>

Over 50,000 unique “words” with no copy editing. How to clean up mistakes?
Metamap Results Example

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
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

angina pectoris not otherwise
classified
chest pain acute
chest pain musculoskeletal
chest pain other
chest pain unspecified
chest pressure
cardiac pain
perd gastro-oesophageal reflux disease
musculoskeletal chest pain
palpitations
unstable engine
chest tightness
shortness of breath
acute chest pain other
myocardial infarction unspecified
pericarditis acute
congestive heart disease

Orders

Abc automated blood count
Aspirin chewable tab aspirin
Basic metabolic panel
Brp pro-brain natriuretic peptide
Cardiac monitoring
Chest x-ray
Chest x-ray portable
Ck-creative kinase
Ck-mb-creative kinase
Ck-mb band
Consult Cardiology admit
Copepoe orders have been entered
Creatine kinase
Creatine kinase mb isoenzyme
Ct(chest x-ray)
Ct(chest x-ray portable)
Ct(heart x-ray)
Ct(heart x-ray portable)
Dose
Ecg standard
Fragment sk-dimer

Diagnosis

Medications

asabufers
aspirin
aspirin
aspirin
aspirin
aspirin
aspirin
aspirin
aspirin
labilir
labinol
labinol
labinol
metoprolol
nitroglycerin
ntk
plane
toprol-x
lax
imidar
tentri
topiramate
tentri
coreg
asac
omepazole
metformin
romana
simvastatin

meta Notes

-diminished-
-dysuria-
-fever-
-nausea-
-palpitations-
-vomiting-
-trauma-
-histor-
-emotional-
-appetite-
-back-
-chest-
-dyspnea-
-emotions-
-exercise-
-fevers-
-hearts-
-jaws-
-nausea-
-pain-

age

0.04
0.03
0.02
0.01
-50
0
50
100
150

Base distribution
Cluster 45
Drug – Patient Association: Diabetes patient cluster

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>
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
Just the Beginning

“Real World” Data Sources
- EMR
- Claims
- Quantified self
- Social media
- High-throughput molecular data
- Job performance

Interested Industries
- Patients
- Care providers
- Pharma
- Clinical research
- Insurance
- Regulators
- Your employer
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