Mining the EHR to understand diseases, drugs, and adverse events

Nigam Shah
nigam@stanford.edu
1. Clinical questions
2. Insights from the data
3. Predictive models
## Structured vs. Unstructured Data

<table>
<thead>
<tr>
<th>row ID</th>
<th>patient ID</th>
<th>hospital ID</th>
<th>order</th>
<th>ICD9 code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1968</td>
<td>156</td>
<td>199280</td>
<td>12</td>
<td>2930</td>
</tr>
<tr>
<td>2627</td>
<td>214</td>
<td>197273</td>
<td>7</td>
<td>2930</td>
</tr>
<tr>
<td>2764</td>
<td>223</td>
<td>105694</td>
<td>17</td>
<td>2930</td>
</tr>
<tr>
<td>5902</td>
<td>902</td>
<td>116367</td>
<td>15</td>
<td>2930</td>
</tr>
<tr>
<td>5957</td>
<td>505</td>
<td>117919</td>
<td>1</td>
<td>2930</td>
</tr>
<tr>
<td>6463</td>
<td>546</td>
<td>127873</td>
<td>14</td>
<td>2930</td>
</tr>
<tr>
<td>6807</td>
<td>588</td>
<td>178452</td>
<td>16</td>
<td>2930</td>
</tr>
<tr>
<td>8475</td>
<td>735</td>
<td>149547</td>
<td>5</td>
<td>2930</td>
</tr>
<tr>
<td>9379</td>
<td>807</td>
<td>121760</td>
<td>13</td>
<td>2930</td>
</tr>
<tr>
<td>10811</td>
<td>923</td>
<td>151107</td>
<td>5</td>
<td>2930</td>
</tr>
<tr>
<td>12122</td>
<td>1040</td>
<td>118695</td>
<td>6</td>
<td>2930</td>
</tr>
<tr>
<td>12135</td>
<td>1041</td>
<td>159277</td>
<td>13</td>
<td>2930</td>
</tr>
<tr>
<td>12140</td>
<td>1042</td>
<td>138732</td>
<td>5</td>
<td>2930</td>
</tr>
<tr>
<td>15121</td>
<td>1332</td>
<td>161256</td>
<td>10</td>
<td>2930</td>
</tr>
<tr>
<td>14487</td>
<td>1269</td>
<td>172465</td>
<td>6</td>
<td>2930</td>
</tr>
<tr>
<td>14526</td>
<td>1271</td>
<td>151184</td>
<td>13</td>
<td>2930</td>
</tr>
<tr>
<td>13630</td>
<td>1178</td>
<td>132240</td>
<td>7</td>
<td>2930</td>
</tr>
<tr>
<td>14976</td>
<td>1321</td>
<td>130634</td>
<td>3</td>
<td>2930</td>
</tr>
<tr>
<td>16311</td>
<td>1439</td>
<td>126530</td>
<td>13</td>
<td>2930</td>
</tr>
<tr>
<td>16414</td>
<td>1446</td>
<td>150830</td>
<td>11</td>
<td>2930</td>
</tr>
<tr>
<td>15526</td>
<td>1354</td>
<td>144830</td>
<td>5</td>
<td>2930</td>
</tr>
<tr>
<td>17424</td>
<td>1551</td>
<td>131267</td>
<td>7</td>
<td>2930</td>
</tr>
<tr>
<td>17478</td>
<td>1559</td>
<td>147179</td>
<td>5</td>
<td>2930</td>
</tr>
<tr>
<td>29983</td>
<td>2658</td>
<td>126865</td>
<td>5</td>
<td>2930</td>
</tr>
<tr>
<td>30536</td>
<td>2758</td>
<td>120236</td>
<td>9</td>
<td>2930</td>
</tr>
<tr>
<td>31006</td>
<td>2759</td>
<td>107939</td>
<td>4</td>
<td>2930</td>
</tr>
<tr>
<td>31982</td>
<td>2842</td>
<td>124867</td>
<td>25</td>
<td>2930</td>
</tr>
<tr>
<td>30087</td>
<td>2667</td>
<td>149534</td>
<td>10</td>
<td>2930</td>
</tr>
<tr>
<td>18125</td>
<td>1610</td>
<td>145005</td>
<td>16</td>
<td>2930</td>
</tr>
<tr>
<td>19805</td>
<td>1701</td>
<td>167723</td>
<td>9</td>
<td>2930</td>
</tr>
<tr>
<td>18272</td>
<td>1621</td>
<td>102458</td>
<td>22</td>
<td>2930</td>
</tr>
<tr>
<td>33498</td>
<td>3002</td>
<td>185653</td>
<td>28</td>
<td>2930</td>
</tr>
<tr>
<td>33699</td>
<td>3024</td>
<td>139229</td>
<td>14</td>
<td>2930</td>
</tr>
</tbody>
</table>

~25% ~75%

**IMPRESSON (ACC 6075491):**

*addendum beginsexam association only. addendum endsbilateral diagnostic digital mammogram with computer-aided detection 3/31/2011 8:14 am right axillary ultrasound 3/31/2011 8:14 am indication: female, 73 years old, right breast lateral tenderness, no discrete mass. history: postmenopausal patient. comparison: 3/7/2006 (stanford hospital), 7/24/2009 (advanced medicine center) technique: full-field digital mammograms were obtained with computer-aided detection to assist in interpretation of the study, including bilateral craniocaudal and mediolateral oblique views comparison with an additional right lateral view. real-time breast ultrasound was then performed targeted to findings: mammogram: the breast tissue is largely fatty, there is a skin bb marker over a palpable abnormality in the right axillary region. there are no features to suggest malignancy. ultrasound: targeted ultrasound reveals a normal appearing lymph node in the 11 o'clock position 10 cm from the nipple in the right axillary region 9x 6 x 4 mm. otherwise no discrete solid or cystic masses identified. impression: 1. right breast: bi-rads 1, negative. left breast: bi-rads 1, negative. recommend the finding prompting ultrasound should be followed on a clinical basis alone. assuming clinical stability, recommend annual screening mammography.*
The utility of looking into notes

http://repository.edm-forum.org/egems/vol4/iss3/1/

Note: *Negative Documentation refers to patients reporting that they are not suffering from urinary incontinence
The utility of looking into notes

Wei WQ, et al 2015 JAMIA
NLP or Text-mining

Natural language processing (NLP) is a discipline which attempts to understand human (natural) languages using computers.

Text mining is the process of discovering and extracting knowledge from unstructured data.
Phenotyping
Clinical questions
Androgen deprivation & Alzheimer’s risk

www.tinyurl.com/JCO-ADT
Androgen deprivation & Dementia risk

Number at risk

<table>
<thead>
<tr>
<th>No ADT</th>
<th>7446</th>
<th>5448</th>
<th>4587</th>
<th>3889</th>
<th>3231</th>
<th>2708</th>
<th>2233</th>
<th>1833</th>
<th>1518</th>
<th>1214</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADT</td>
<td>1826</td>
<td>1177</td>
<td>872</td>
<td>645</td>
<td>470</td>
<td>367</td>
<td>282</td>
<td>211</td>
<td>160</td>
<td>124</td>
</tr>
</tbody>
</table>
Past Medical/Surgical History: Positive for atrial fibrillation. The patient had an AVR 6 years ago. Peripheral arterial disease with hypertension, peripheral neuropathy, atherosclerosis, hemorrhoids, proctitis, CABG, and cholecystectomy.

Family History: Positive for atherosclerosis, hypertension, autoimmune diseases in the family.

Review of Systems: Weight loss of 25 pounds within the last 6 months, shortness of breath, constipation, bleeding from hemorrhoids, increased frequency of urination, muscle aches, dizziness and faintness, focal weakness and numbness in both legs, knees, and feet.

Laboratory Data and Radiological Results: The patient had a chest x-ray, which showed cardiomegaly with atherosclerotic heart disease, pleural thickening, and small pleural effusion, left costophrenic angle which has not changed when compared to prior examination, COPD pattern. The patient also had a head CT, which showed atrophy with old ischemic changes. No acute intracranial findings.

Discharge Diagnosis: Syncope.

Discharge Medications: The patient was discharged on the following medications: Cardizem 90 mg p.o. thrice daily, digoxin 0.125 mg p.o. once daily, allopurinol 100 mg two times daily, Coumadin 4 mg p.o. q.h.s., and Remeron 15 mg p.o. q.h.s.
Juvenile Idiopathic Arthritis
ICD 9 codes
696.0, 714.0, 714.2, 714.3, 714.9, 720.2, 720.9

Terms:
Juvenile idiopathic arthritis, JIA
Juvenile rheumatoid arthritis, JRA
Psoriatic arthritis
Juvenile spondyloarthropathy, spondyloarthritis, enthesitis related arthritis, sacroiliitis, reactive arthritis

Phenotyping
Insights
Detecting drug-treats-disease relations

- I2b2 2010 Challenge on information extraction from free text
- Goal: find NER, assertions, and relationships between entities.
  - Relationships included <Drug used to treat Indication>
- Best performance: F1 ~ 0.75 for relations
- Problem solved?
Off-label use via machine learning

- We don’t care about note-level accuracy.
- Can we detect useful signals from aggregate statistics based on noisy low level features?
  - For given drug-disorder pair, is drug used to treat the disorder?
  - Off-label if it is not approved.
Training and test sets

13457 Known uses in Medispan

9444 Testable known uses

80:20 Split

7555 Training Known Uses

1889 Test Known Uses

Censor these uses when constructing Medispan and Drugbank features!

"known" use

9444 Testable known uses

Pick a drug-disease pair

Pick a drug with similar freq

Known Use?

Yes

No

"not used"

"probably not used..."
Off-label drug use

8,861 positive examples
34,973 negative examples

EMR Prior knowledge

Clinical Notes Medi-span, Drugbank

Example features:
- Co-mentions
- Drug-first fraction

Example features:
- Similarity with on-label uses
- Fraction of uses that are approved

Prior knowledge

7,112 positive and 27,938 negative training examples

1,749 positive and 7,035 negative test examples

Evaluate in test set

PPV 0.952
Spec 0.990
Recall 0.758
F1 0.844

Classifier for “used-to-treat” relationship

Example features:
- Co-mentions
- Drug-first fraction

Prioritize 407 well-supported usages

EMR Prior knowledge

2,362,950 drug-indication pairs

Predicted
- Confident
- Novel
- Supported

Risk
- Low
- High

Low High
32 35
17 27

Cost

Data sources:
- EMR
- Prior knowledge
- Clinical Notes
- Medi-span, Drugbank
- FAERS
- MEDLINE

Risk
- High
- Low

Cost
- Low
- High

Evaluate in test set

PPV 0.952
Spec 0.990
Recall 0.758
F1 0.844
‘Just enough’ text mining
Trade-off: simple or advanced [text-processing]

(A) Benchmarking Task

- Annotator
- REVEAL
- cTAKES

- 1249 Notes from i2b2 2008 Challenge

(B) Research Tasks

- Classifier for “used-to-treat” relation
- Detection of drug-drug interactions
- Safety profile of Cilostazol in PAD patients

- 9 million Textual notes from STRIDE
- Annotator
- REVEAL

- 1 day
- 3 months

- Patient – Feature Matrix Representation
A note on evaluating your “NLP”

1. **String matching**: can you grep, can you grep with typos, can you find the right term, span etc.
2. **Knowledge graph handling**: can you use a knowledge graph to infer that Simvastatin is a type of statin
3. **Context and negation**: can you differentiate mentions that are about patient vs. other, negated, historical vs present
4. **Intuitive**: can you infer things that are not mentioned
   - e.g. 5 feet tall, 200 lbs —> obesity
5. **Phenotype**: can you recognize [known] phenotypes correctly
   - e.g. exposure to drug + ALP >= 2x ULN + normal lab measurements prior to exposure to drug --> drug induced liver injury
6. **Functional**: how accurately can the output of the processing be used to accomplish a research task, such as detect adverse drug events
7. **Utility**: if the method was used to generate results, would it change practice
   - e.g. we give pneumovax to a 100 more patients, because text-mining told us that they had a splenectomy
Ask: about the cost-utility trade-off

- EHR mining is a process
- Text is one of the many sources
- Time needs special handling

- Machine learning is used in many places
  - Sorting the documents that contain the text of interest
  - In the processing of the text to extract features and facts ("NLP")
  - In the processing of time to extract features
  - Finding associations among the extracted features
Acknowledgements

Group Members:
- **Scientists**: Ken Jung, Alison Callahan, Juan Banda
- **Fellows**: Rohit Vashisht, Azadeh Nikferjam, Katie Quinn
- **Engineer**: Vladimir Polony
- **BMI Students**: Sarah Poole, Alejandro Schuler, Vibhu Agarwal
- **Med Students**: Mehr Kashyap, Jassi Pannu

**Alums**: Anna Bauer-Mehren (Roche), Srini Iyer (Facebook), Amogh Vasekar (Citrix), Sandy Huang (Berkeley), Paea LePendu (Lexigram), Rave Harpaz (Oracle), Tyler Cole (Barrow Inst.), Sam Finlayson (Harvard), Will Chen (Yale), Yen Low (Netflix), Elsie Gyang (Fellowship in Surgery), Suzanne Tamang (Instructor)

**Collaborators**: Purvesh Khatri, Tina Hernandez-Boussard, Winn Haynes, Kevin Nead, Nick Leeper

**Funding:**
- NIH – NLM, NIGMS, NHGRI, NINDS, NCI, FDA
- Stanford Internal – Dept. of Medicine, Population Health Sciences, Clinical Excellence Research Center, Dean’s office
- Fellowships – Med Scholars, Siebel Scholars Foundation, Stanford Graduate Fellowship
- Industry – Apixio, CollabRx, Healogics, Janssen R&D, Oracle, Baidu USA, Amgen

IT: Alex Skrenchuk, SCCI team