Conflict of Interest

Alexander Turchin, MD, MS, FACMI

Has no real or apparent conflicts of interest to report.
Agenda

• Importance of EMR data
• Natural language processing
• Rare events: the NLP approach
• What happened next
• How to explain it
Learning Objectives

• Explain importance of identification of rare clinical events in EMR data

• Define different technologies that can be used to extract information from narrative electronic documents

• Compare efficacy of natural language processing technologies for identification of rare clinical events

• Discuss reasons for differences in performance between natural language processing technologies
Importance of EMR Data
EMR = DATA
EMR = DATA
EMR = DATA

Population management

Research

Quality improvement
What’s under the Hood?
What’s under the Hood?
Caveat Emptor

EMR data are like a box of chocolates…
You never know what you’re gonna get.
Structured vs. Narrative

**Problem List**

Hematology and Oncology
- Prostate cancer

Cardiovascular and Mediastinum
- Hypertension
- Coronary artery disease

Respiratory
- Pneumonia

Endocrine
- Body mass index (BMI) of 24.0-24.9 in adult

Behavioral and Developmental
- Obsessive compulsive disorder

### Medications
- aspirin 325 MG tablet
- lisinopril (PRINIVIL,ZESTRIL) 10 MG tablet
- metoprolol tartrate (LOPRESSOR) 25 MG tablet

<table>
<thead>
<tr>
<th>PatientID</th>
<th>Diagnosis</th>
<th>DurationMonths</th>
<th>AcuteChronic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A12345</td>
<td>Coronary artery disease</td>
<td>32</td>
<td>Chronic</td>
</tr>
<tr>
<td>A12345</td>
<td>Hypertension</td>
<td>74</td>
<td>Chronic</td>
</tr>
<tr>
<td>A12345</td>
<td>Prostate cancer</td>
<td>3</td>
<td>Acute</td>
</tr>
<tr>
<td>B54321</td>
<td>Uterine leiomyoma</td>
<td>12</td>
<td>Chronic</td>
</tr>
<tr>
<td>B54321</td>
<td>Back pain</td>
<td>61</td>
<td>Chronic</td>
</tr>
</tbody>
</table>
Structured vs. Narrative

13,993

9,819

5,627 (30.9% of the total)

Medication intensifications

Structured vs. Narrative

## Structured vs. Narrative

<table>
<thead>
<tr>
<th>My Note</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sally M. Smith</strong></td>
</tr>
<tr>
<td><strong>1234567-8</strong></td>
</tr>
<tr>
<td><strong>11/5/2012</strong></td>
</tr>
</tbody>
</table>

**HISTORY OF PRESENT ILLNESS:** Ms. Smith is a 43-year-old woman with past medical history that includes hypercholesterolemia, diabetes mellitus type 2 and a pilonidal cyst. The cyst was apparently removed when she was 18. Last July she presented with more pain in this area. On exam, it was apparently unclear if there was a recurrence. She was put on a course of Keflex and everything resolved.

- More details
  - Includes context
    - Describes clinical reasoning
    - Non-codable information

{NLP}
Natural Language Processing
Mr. Smith comes today with chief complaint of back pain. Denies history of trauma, urinary retention or weakness.

**Chief complaint**
- back pain

**Negated**
- trauma
- weakness
Natural Language Processing

ABSTRACT

Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable.

I. INTRODUCTION

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification of virtual machines and real-time theory. To what extent can web browsers be constructed to achieve this purpose?
1. **Identify Examples**: collect examples (typically through manual record review) of how the concept being sought is documented in EMR. Hundreds of examples usually necessary for a comprehensive description.

2. **Learn from Examples**: based on the examples, create a *language model* that can recognize the concept being sought. This step can be manual (e.g. in rule-based systems) or automated (in machine-learning-based systems).

3. **Evaluate the Language Model**: test the language model on a new set of examples that were not used to create to determine its accuracy. Several dozen examples typically necessary to have sufficiently narrow confidence intervals.
Rare Events: the NLP Approach
Rare Events

Are not documented persistently in multiple documents, but nevertheless could impact patient care and outcomes

- **Splenectomy**: requires specific immunizations (e.g. pneumococcal) to prevent fatal illness

- **Anaphylaxis to penicillin**: a life-threatening reaction to a common medication

- **Rejection of treatment recommendation by the patient**: can impact both future treatment decisions and long-term outcomes
Rare Events

• Can present a particular challenge for design of NLP tools because it can be difficult to collect enough examples to make the tools sufficiently accurate

• This problem is not unique to NLP
Step 1: Identify Examples
Data Enrichment

• Non-adherence to blood pressure medications
  – Significantly elevated BP (≥ 150/100)
  – No intensification of anti-hypertensive medications

• Blood pressures measured at home
  – Notes with blood pressure ranges
    (e.g. 120-130/70-80)

• Patients declining insulin therapy
  – Elevated blood glucose (HbA1c > 7%)
  – No insulin treatment started
Step 2: Create a Language Model

Manually designed (rule-based) systems

V.S.

Automatically developed (machine-learning-based) systems
Step 2: Create a Language Model

- Can incorporate background knowledge not directly found in the examples
- May not need as many examples

V.S.

- Fast
- Can model non-linear relationships
Classification Methods

- Typically employ bag-of-words approach (i.e. do not analyze spatial relationships between words in a sentence).
- Naïve Bayes
- Logistic Regression
- Support Vector Machines (SVMs)
Naïve Bayes
Logistic Regression

Inputs: X1, X2, X3 || Weights: θ1, θ2, θ3 || Outputs: Happy or Sad
Support Vector Machines
Sequence Labeling Methods

- These methods are “aware” of the sequence of words in a sentence and use it to inform classification.
- Conditional Random Fields (CRFs)
- Recurrent Neural Networks (RNNs)
Conditional Random Fields

Patient \rightarrow \text{refused} \rightarrow \text{insulin}

Sentence

Labels

Example sentence: "Patient refused insulin."
Neural Networks
Deep Learning

Input Layer

Hidden Layers

Output Layer
Recurrent Neural Network

Hidden Layers

Input Layer

Output Layer
Canary

- A GUI-based platform allowing users without computer science background to create [rule-based] NLP tools to identify concepts of interest in narrative electronic data

- Supports advanced NLP features:
  a) Concept-value extraction (e.g. ejection fraction)
  b) Identification of concepts across sentence boundaries
  c) Parallel processing
  d) Portability of language models between Canary installations
  e) Can analyze text in multiple languages

- Freely available at http://canary.bwh.harvard.edu
Canary

Language models are created using word classes (semantic groupings) and phrase structures (rules defining how a concept can be documented in the text).
Step 3: Evaluate

All true positives

True positives identified by NLP

Sensitivity (Recall)

VS.
Step 3: Evaluate

True positives identified by NLP

Positive Predictive Value (Precision)

All concepts identified by NLP
Step 3: Evaluate

\[ F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]
Why F1?

Reality: Rare Event

Model: Common Event

Arithmetic mean: 52.5%

Harmonic mean: 9.1%

Sensitivity: 100%

PPV: 5%
What Happened Next
Nitty-Gritty: Data

• Training dataset: 50,046 documents (2,660,475 sentences).

• Evaluation dataset: 1,503 documents (86,487 sentences).

• Prevalence of insulin decline by patients: 0.02% in both sets (at the sentence level)
• All machine learning models were trained using both words and lemmas-based methods.

• Naïve Bayes and Logistic Regression never reached F1 of 0.5 on the training set and were not further evaluated.

• Regularization parameter was optimized in the SVM model using cross-validation.

• We also used Synthetic Minority Oversampling Technique (SMOTE) to compensate for the low prevalence of true positives using SVMs.

• Regularization parameter and decision threshold were optimized for the CRF and RNN models using cross-validation.
Nitty-Gritty: Canary

- Canary language model was designed by a clinician with no formal computer science training who had access to the same training set used by the machine learning models.

- The final model contained 148 word classes and 284 phrase structures

- Using this language model, Canary processed text at 1 MB (c. 200 documents) per CPU core per minute
## Accuracy

<table>
<thead>
<tr>
<th>System</th>
<th>Sensitivity</th>
<th>PPV</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.714</td>
<td>0.526</td>
<td>0.606</td>
</tr>
<tr>
<td>CRF</td>
<td>0.563</td>
<td>0.474</td>
<td>0.514</td>
</tr>
<tr>
<td>RNN</td>
<td>0.706</td>
<td>0.632</td>
<td>0.667</td>
</tr>
<tr>
<td>Canary</td>
<td>0.955</td>
<td>1.000</td>
<td>0.977</td>
</tr>
</tbody>
</table>
How to Explain It
Machine Learning – What Worked

• Non-linear boundaries (SVM vs. Logistic Regression).
• Oversampling (SMOTE for SVM)
• Taking context into account (RNN)
Man vs. Machine

Natural food is better!
Man vs. Machine

Natural medicine is better!
Man vs. Machine

Natural intelligence is worse?
Man vs. Machine

- Walking on two feet
- Stitching a T-shirt
- Translating to another language
- Understanding and communicating emotions
- Etc.

- Mathematical calculations
- Chess / Go
Man vs. Machine

- Background knowledge of English
- Background knowledge of subject matter
- Insufficient number of examples
- Imbalance between positive and negative examples

Ability to generalize
Man vs. Machine

- Canary could integrate information far apart in the sentence
- Insufficient number of examples
- Imbalance between positive and negative examples

Ability to take context into account
How could we do better?

- One way to improve machine-learning NLP is to create large publicly available repositories of marked-up text (corpora).

- More likely to be helpful for basics (e.g. named entities) and less likely for complex concepts representing clinical workflow.
Rare in Text ≠ Rare in Life

- 0.02% prevalence at the *sentence* level
- 0.9% prevalence at the *document* level
- 30% prevalence at the *patient* level
Conclusions

• Rare events are an important category of EMR data that may require special approach to identification

• At the current state of technology human-designed NLP tools can achieve significantly higher accuracy than machine learning methods, though they can take time to develop

• Several techniques can improve performance of machine learning methods, but further improvements are needed
Thank you

- Nicholas Alexander
- Lee-Shing Chang
- Wendong Ge
- Matt Goldberg
- Peter Goldberg
- Naoshi Hosomura

- Victor J. Lei
- Shervin Malmasi
- Stephen Skentzos
- Alex Solomonoff
- Dmitriy Timerman
- Huabing Zhang

Funding: Sanofi
Don’t forget to complete online session evaluation!

Questions?

Alexander Turchin
aturchin@bwh.harvard.edu