Tutorial 6:
Applications of Natural Language Processing in Clinical Research and Practice

Yanshan Wang¹, Ahmad P. Tafti¹, Sunghwan Sohn¹, Rui Zhang²

¹Department of Health Sciences Research, Mayo Clinic
²Department of Pharmaceutical Care & Health Systems, University of Minnesota
Overview of Natural Language Processing in Clinical Domain
Data capture and sharing

Advanced clinical processes

Improved Outcome

HITECH Act law 2009
Motivation for Clinical NLP

20% Structured Data
- Demographics, Lab results, Medication, Diagnosis...

80% Unstructured Data
- Clinical notes
- Patient provided information
- Family history
- Social history
- Radiology reports
- Pathology reports
- ...
The Nature of EHR Data

• Primary function is to record clinical events and facilitate billing and the communication among the care team.

• Significant dependence on narrative text, which is often the gold standard for clinical findings.

• Using administrative/billing structured data as a surrogate for clinical data is problematic
  • Variations in coding, miscoding, incomprehensive
  • Misleading
Speakers and Topics
Big Data Infrastructure for Large-scale Clinical NLP

Ahmad P. Tafti is a Research Associate at Mayo Clinic, with a deep passion for improving health informatics using diverse medical data sources combined with advanced computational methods. Dr. Tafti's major interests are AI, machine learning, and computational health informatics. Dr. Tafti has published over 20 first-author peer-reviewed publications in prestigious journals and conferences (e.g., CVPR, AMIA, ISVC, JMI, PLOS, IEEE Big Data), addressing medical text and medical image analysis and understanding using advanced computational strategies.

- Big Data Social Media
- Harnessing Social Media; What and Why?
- Quantitative and Qualitative Analysis of Social Media
- How We Can Draw Demographic-Specific Disparities Using Social Media
  - Gender-Specific
  - Age-Specific
  - Ethnicity-Specific
- Case Study: Gender Disparity in Side Effects Reporting of Chronic Pain Medications
Advances of NLP in Clinical Research

Rui Zhang is an Associate Professor and KcKnight Presidential Fellow in the College of Pharmacy and the Institute for Health Informatics (IHI), and also graduate faculty in Data Science at the University of Minnesota (UMN). He is the Leader of NLP Services in Clinical and Transnational Science Institution (CTSI) at the UMN. His work has been recognized on a national scale including Journal of Biomedical Informatics Editor’s Choice, nominated for Distinguished paper in AMIA Annual Symposium and Marco Ramoni Distinguished Paper Award for Translational Bioinformatics, as well as highlighted by The Wall Street Journal.

- Background of NLP to Support Clinical Research
- NLP Systems and Tools for Clinical Research
- Use Case 1: NLP to Support Dietary Supplement Safety Research
- Use Case 2: NLP to Support Mental Health Research
Clinical Information Extraction

Sungwhan Sohn is an Associate Professor of Biomedical Informatics at Mayo Clinic. He has expertise in mining large-scale EHRs to unlock unstructured and hidden information using natural language processing and machine learning, thus creating new capacities for clinical research and practice in order to achieve better patient solutions. He has been involved in the development of cTAKES, the most popular NLP tool in the clinical domain. Dr. Sohn’s research facilitates the best use of EHRs to solve clinical problems and improve public health.

- **About EHR and its challenges**
- **Clinical information extraction (IE)**
  - Methodology review (NLP techniques)
    - strength/weakness
- **Clinical documentation variations and their effects on NLP tools**
- **NLP tool portability**
  - Case study of NLP tool portability (asthma ascertainment)
Yanshan Wang is a Research Associate at Mayo Clinic. His current work is centered on developing novel NLP and artificial intelligence (AI) methodologies for facilitating clinical research and solving real-world clinical problems. Dr. Wang has extensive collaborative research experience with physicians, epidemiology researchers, and statisticians. Dr. Wang has published over 40 peer-reviewed articles at referred computational linguistic conferences (e.g., NAACL), and medical informatics journals and conference (e.g., JBI, JAMIA, JMIR and AMIA). He has served on program committees for EMNLP, NAACL, IEEE-ICHI, IEEE-BIBM.

- **Cohort retrieval**
- **Approaches for cohort retrieval**
  - Medical concept embedding
  - Information retrieval
  - Deep patient representation
- **Case studies**
  - Patient cohort retrieval for clinical trials accrual
Big Data Infrastructures for large-scale clinical NLP: Healthcare Social Media Mining

- Big Data Social Media
- Harnessing Social Media in Healthcare; What and Why?
- Quantitative and Qualitative Analysis of Social Media
- How We Can Draw Demographic-Specific Disparities Using Social Media
  - Gender-Specific
  - Age-Specific
  - Ethnicity-Specific

[Case Study: Gender Disparity in Side Effects Reporting of Chronic Pain Medications]

Ahmad Tafti, PhD

Division of Digital Health Sciences
Mayo Clinic
Big Data Social Media

Harnessing Social Media in Healthcare; What and Why? (Contd.)

The **rapid** and **extensive growth** of **social media** persuades an increasing number of patients to use this technology for health related reasons.

It has impacted the **communication style** that **patients** and **physicians** can take to discuss and share health related events, such as **disease diagnosis** and **treatment, symptoms, medications**, and **drug side effects**.
Harnessing Social Media in Healthcare; What and Why? (Contd.)

- **Healthcare-Generated Health Data**
  - EHRs
  - Clinical Notes
  - Radiology Reports
  - Vital Signs
  - Medical Images
  - ...

- **Patient-Generated Health Data**
  - Active/First Symptoms
  - Pain
  - Logistics
  - Drug Reviews
  - ...

Harnessing Social Media in Healthcare; What and Why? (Contd.)

- Social media posts offer a unique opportunity to capture information about patient experiences with health events.

- Social media information can be used to develop patient-centered decision support tools that can be integrated with the EHR to facilitate discussions on treatment choices, risks, and benefits.
Benefits and Advantages

- Faster, Easier Communication
- Professional Networking
- Professional Education
- Organizational promotion
- Boost internal and external visibility
- Customer feedback
- Impress potential customers
- User-generated content
- Patient care
- Patient education
Harnessing Social Media in Healthcare; What and Why? (Contd.)

Challenges

- Unstructured data
- Garbage mixed with gold (information quality issues)
- Damage to Professional Image
- Breaches of Patient Privacy
- Legal and licensing Issues
Harnessing Social Media in Healthcare; What and Why?

Advices and Guidelines on How to Use This Wealth Of Data

• Social Media Guidelines Issued by Health Care Institutions
• IRBs Protocols
• License agreements of the social media
# Qualitative Analysis of Social Media

<table>
<thead>
<tr>
<th>Source</th>
<th>Gender Availability</th>
<th>Age Availability</th>
<th>Ethnicity Availability</th>
<th>Location Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Google+</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Drugs.com</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DailyStrength</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>WebMD</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MedHelp</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Patient.info</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Qualitative Analysis of Social Media

About Me

Age: 61
Gender: Female

I have married to my best friend for 34 yrs. I have 1 son and 1 daughter. I love Bernese Mtn Dogs. I have been a nurse for 30+ yrs. I learned of my diagnosis the day I was scheduled to go Bridal Dress shopping with my daughter.

Pednurse1 06/16/2015

Hi

I had hoped I would lose weight on Levothyroxine but instead I am constantly hungry and have gained 20 additional pounds in 3 years. I am now considering weight loss surgery and wondered if others on this group have found that helpful.
Framework

Text Processing

Training Dataset

- Gender Classification
- Ethnicity Classification
- Sentence Classification

Top Distinctive Terms and Dominant Topics

MedTagger

Scientific Visualization

Web Crawler/APIs

SM-Dataset

- Web Crawler/APIs

- Text Processing

G+ twitter

Facebook

Reddit

WebMD

Patient

Drugs.com

MedHelp
Gender Classification

Gender API: [https://gender-api.com](https://gender-api.com)
gender-guesser: [https://pypi.org/project/gender-guesser](https://pypi.org/project/gender-guesser)
genderize.io: [https://genderize.io](https://genderize.io)
NameAPI: [https://www.nameapi.org](https://www.nameapi.org)
NamSor: [https://www.namsor.com](https://www.namsor.com)
Case Study: Gender Disparity in Side Effects Reporting of Chronic Pain Medications

We are filtering the posts using a list of pain/chronic pain related keywords and health-related ones. The rationale of selecting the keywords is to cover/pull the posts as much as we can. Here is the list based on our best practices:

1) Pain related keywords:
2) Pain Medications:
Filtering the data

3) Twitter Hashtags:

#pain
#methadone
#injury
#arthritis
#osteoarthritis, etc.

4) Disorders:

Asthma
Lupus
Irritable Bowel Syndrome (IBS)
Chronic Fatigue Syndrome
Filtering the data

5) Pharmaceuticals companies:

Novartis
Teva
AstraZeneca
Amgen Inc.
Eli Lilly
Gilead Sciences
Abbott
Bayer AG
AbbVie Inc.
Sanofi S.A.
Pfizer Inc.
Johnson & Johnson
Filtering the data

6) Insurance companies:
   Aetna
   Humana
   HCSC
   Cigna
   Kaiser Permanente
   United Healthcare
   HealthPartners
An Example: Good Reviews

"I feel that drinking 25 mg of methadone has made way more difference in pain compared to 3 Norcos 10milligram. You cannot compare the difference. The methadone helps my pain much better."

10

cat (taken for 1 to 6 months) January 14, 2019

"Methadone has been helping me with my chronic back pain for over six years. This is one of only two medicine I find that help me. I was unemployed for over a year and this medicine not only helps control my back pain, but it is also affordable. This is one of the cheapest cost medicines I have ever had. Yet doctors don't want to let me keep using it. They cannot advise me are give me any alternatives that help with my back pain, but they are trying to wean me off of it. Now my back is hurting me more and they want me to stop taking the only thing that helps me. Why should I take a cheap pain medicine that works when we can take more expensive meds and go thru costly testing. I have nothing but good to say about methadone for pain, yet I am told because of all the worlds addiction and dying from overdoses it will no longer be allowed for me to take. Sad, because after six years of pain management, I will be back where I started with chronic back pain."

9.0

TMT (taken for 5 to 10 years) October 27, 2018
"My experience with methadone has been a life saving one! I am on 80 mg for chronic back pain, and have stayed at this dose for 3 years and have never had the need to go up on my dose which is an amazing thing about this med: once you find the right dose for your pain needs, you can stay there. I did experience some side effects, but for me it's completely worth it because it provides amazing pain relief, and it's not expensive. It's a very effective pain medication!"

8.0

E (taken for 5 to 10 years) March 3, 2018
An Example: Bad Reviews

"When it was first prescribed I'd have given it a 10/10. I've tried almost everything to manage my pain and nothing worked as well with as few side effects as methadone. I took it for years, gradually needing more and more. My doctors assured me it was safe and that it wouldn't be like other opioids. Methadone has ruined my life, even taking it exactly as prescribed. I'm currently withdrawing from it because I refuse to be addicted to it anymore, and its pure HELL. Worse than anything else I've ever come off of in my entire life. Don't use long-term. Please, do yourself a favor and get off now."

AimlessMe January 29, 2018

"I have some bad news for all of you that are sooo happy to be on methadone. Eventually you will be taken off of the drug and when that time comes you will fully understand what an awful medication methadone is to get off of. The withdrawal from it is 3-5 times worse than any other opiate you have ever been on. Imagine severe oxycodone withdrawal or hydromorphone, oxycontin, hydrocodone, opana withdrawal. Methadone withdrawal is absolutely horrid and you will regret ever putting it in your body. The truth hurts and most don't want to hear it but its something the docs don't tell you. Goodluck!!"

Col. Davis (taken for 2 to 5 years) December 20, 2017
# Sentence Classification: ADEs vs Non-ADEs

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Learning method</th>
<th>Number of sentences</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>Area under the ROC</th>
<th>Training time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADEs#1_Combined</td>
<td>bigNN system</td>
<td>7360</td>
<td>88.7</td>
<td>88.5</td>
<td>89.4</td>
<td>0.842</td>
<td>45.7</td>
</tr>
<tr>
<td>ADEs#1_Combined</td>
<td>BoW + SVM</td>
<td>7360</td>
<td>89.4</td>
<td>88.3</td>
<td>88.0</td>
<td>0.841</td>
<td>66.3</td>
</tr>
<tr>
<td>ADEs#1_Combined</td>
<td>BoW + decision tree</td>
<td>7360</td>
<td>84.0</td>
<td>83.7</td>
<td>82.1</td>
<td>0.775</td>
<td>49.5</td>
</tr>
<tr>
<td>ADEs#1_Combined</td>
<td>BoW + naïve Bayes</td>
<td>7360</td>
<td>83.7</td>
<td>82.1</td>
<td>83.5</td>
<td>0.763</td>
<td>48.9</td>
</tr>
<tr>
<td>ADEs#2_Combined</td>
<td>bigNN system</td>
<td>14,017</td>
<td>89.1</td>
<td>88.9</td>
<td>89.3</td>
<td>0.874</td>
<td>69.5</td>
</tr>
<tr>
<td>ADEs#2_Combined</td>
<td>BoW + SVM</td>
<td>14,017</td>
<td>89.5</td>
<td>88.0</td>
<td>89.7</td>
<td>0.875</td>
<td>88.9</td>
</tr>
<tr>
<td>ADEs#2_Combined</td>
<td>BoW + decision tree</td>
<td>14,017</td>
<td>85.5</td>
<td>84.9</td>
<td>84.5</td>
<td>0.861</td>
<td>75.2</td>
</tr>
<tr>
<td>ADEs#2_Combined</td>
<td>BoW + naïve Bayes</td>
<td>14,017</td>
<td>84.3</td>
<td>84.0</td>
<td>85.7</td>
<td>0.855</td>
<td>73.8</td>
</tr>
<tr>
<td>ADEs#3_Combined</td>
<td>bigNN system</td>
<td>21,843</td>
<td>92.7</td>
<td>93.6</td>
<td>93.0</td>
<td>0.905</td>
<td>121.7</td>
</tr>
<tr>
<td>ADEs#3_Combined</td>
<td>BoW + SVM</td>
<td>21,843</td>
<td>92.5</td>
<td>94.0</td>
<td>93.2</td>
<td>0.911</td>
<td>159.5</td>
</tr>
<tr>
<td>ADEs#3_Combined</td>
<td>BoW + decision tree</td>
<td>21,843</td>
<td>88.3</td>
<td>87.5</td>
<td>87.2</td>
<td>0.868</td>
<td>131.5</td>
</tr>
<tr>
<td>ADEs#3_Combined</td>
<td>BoW + naïve Bayes</td>
<td>21,843</td>
<td>87.5</td>
<td>86.2</td>
<td>85.8</td>
<td>0.851</td>
<td>135.3</td>
</tr>
</tbody>
</table>

---

*aADEs: adverse drug events.

*bbigNN: big data neutral network.

*cBoW: bag-of-words.

*dSVM: support vector machine.
Sentence Classification: ADEs vs Non-ADEs
Visualization Results: An Example

Figure 1. (Left) Women were more likely than men to report side effects from gabapentin. Women: 239 posts out of 1407 associated with gabapentin (including side effects, indication, and/or other topics) Men: it was 281 out of 1,973 posts.

(Right) Gender-specific comparative visualization across five different side effects of gabapentin.

**Note:** all side effects identified in this exercise are previously reported side effects of gabapentin.
Figure 2. Percentwise proportion of women and men who did shared their pain-related experiences within three opioid class medications in Twitter, within last 30 days. One can see the number of Oxycodone and Ritalin tweets generated by women is greater than those generated by men. For Methadone, it shows men discussed the medication a little more than women.
Thank You!
Advances of Natural Language Processing in Clinical Research

Rui Zhang, Ph.D.

Associate Professor and McKnight Presidential Fellow,
Department of Pharmaceutical Care & Health Systems,
and Institute for Health Informatics

University of Minnesota, Twin Cities
Outline

• Background of NLP to Support Clinical Research
• NLP Systems and Tools for Clinical Research
• Use Case 1: NLP to Support Dietary Supplement Safety Research
• Use Case 2: NLP to Support Mental Health Research
Clinical Research Informatics (CRI)

- CRI involves the use of informatics in the discovery and management of new knowledge relating to health and disease.
- It includes management of information related to clinical trials and also involves informatics related to secondary research use of clinical data.
- It involves approaches to collect, process, analyze, and display health care and biomedical data for research.
Healthcare Big Data

Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care
# Structured vs. Unstructured Data

## Diagnosis codes

<table>
<thead>
<tr>
<th>Fake ID</th>
<th>ENTRY_DAT</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>34068</td>
<td>5/13/2001</td>
<td>41.85</td>
</tr>
<tr>
<td>37660</td>
<td>8/6/2002</td>
<td>79.99</td>
</tr>
<tr>
<td>140680</td>
<td>8/31/2003</td>
<td>79.99</td>
</tr>
<tr>
<td>23315</td>
<td>5/14/2003</td>
<td>112</td>
</tr>
<tr>
<td>75936</td>
<td>7/9/2004</td>
<td>117.9</td>
</tr>
</tbody>
</table>

## Lab tests

<table>
<thead>
<tr>
<th>Fake ID</th>
<th>TEST</th>
<th>ENTRY_DAT</th>
<th>VALU</th>
</tr>
</thead>
<tbody>
<tr>
<td>3536</td>
<td>pO2</td>
<td>1/23/1996</td>
<td>314</td>
</tr>
<tr>
<td>72921</td>
<td>LDL</td>
<td>2/5/1996</td>
<td>34</td>
</tr>
<tr>
<td>102460</td>
<td>pCO2</td>
<td>1/26/1996</td>
<td>45</td>
</tr>
<tr>
<td>135043</td>
<td>HDL</td>
<td>1/25/1996</td>
<td>35</td>
</tr>
<tr>
<td>135432</td>
<td>MonAb</td>
<td>1/24/1999</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Structured vs. Unstructured Data

80% of all digital data is unstructured

Unstructured data is growing by 60% CAGR

Unstructured data resists utilization and reuse
Leveraging Natural Language Processing (NLP) to Unlock Unstructured Data

- A field of Artificial Intelligence (AI)
- Applications that automatically analyze natural language (English, Chinese)
- Computational linguistics + domain knowledge
- Tasks
  - Word sense disambiguation (WSD)
  - Named entity recognition (NER)
  - Relation extraction (RE)
  - Negation identification (NegIde)
  - Semantic role labelling (SRL)
  - Information extraction
Word Sense Disambiguation (WSD)

• Word sense: a meaning of a word.

• Acronym
  ➢ “The patient underwent a left **BK** amputation.”
    Sense: below knee
  ➢ “**BK** viremia in the past.”
    Sense: BK (virus)

• Abbreviation
  ➢ “CT of head showed old **CVA** on left side.”
    Sense: cerebrovascular accident
  ➢ “Straight with no **CVA** tenderness.”
    Sense: costovertebral angle
Named Entity Recognition (NER)

BRIEF HISTORY: The patient is an (XX)-year-old female with history of previous stroke; hypertension; COPD, stable; renal carcinoma; presenting after a fall and possible syncope. While walking, she accidentally fell to her knees and did hit her head on the ground, near her left eye. Her fall was not observed, but the patient does not profess any loss of consciousness, recalling the entire event. The patient does have a history of previous falls, one of which resulted in a hip fracture. She has had physical therapy and recovered completely from that.

Initial examination showed bruising around the left eye, normal lung examination, normal heart examination, normal neurologic function with a baseline decreased mobility of her left arm. The patient was admitted for evaluation of her fall and to rule out syncope and possible stroke with her positive histories. DIAGONSTIC STUDIES: All x-rays including left foot, right knee, left shoulder and cervical spine showed no acute fractures. The left shoulder did show old healed left humeral head and neck fracture with baseline anterior dislocation. CT of the brain showed no acute changes, left periorbital soft tissue swelling. CT of the maxillofacial area showed no facial bone fracture. Echocardiogram showed normal left ventricular function, ejection fraction estimated greater than 65%.
Determine relationships between entities or events

“We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.” [PMID: 3718110]

Relationship: Hemofiltration-TREATS-Patients

Negation Identification (NegIde)

• Identify pertinent Negatives from narrative clinical reports

- “The chest X-ray showed no infiltrates...”
- “The patient denied experiencing chest pain”
- “no murmurs, rubs or gallops”
- “murmurs, rubs and gallops are absent”
Semantic Role Labeling

- Detect the semantic role played by each noun phrase associated with the verb of a sentence
  - Agent: Noun Phrase (NP) before the verb
  - Patient: NP after the verb
  - Instrument: NP in a Prepositional Phrase (PP)

We dissected a cystic artery with cautery.
Information Extraction

• Automated extraction of family and observation predications from unstructured text
  • Supplied text: "Heart disease on the father side of the family. Mother has arthritis."
  • Extracted elements:
    • Constituent: family {FAMILY HISTORY: FAMMEMB}
    • Constituent: observation {Heart disease: C1576434}
    • Constituent: family {father side of the family: Paternal*}
    • Constituent: family {Mother: MTH}
    • Constituent: observation {arthritis: C1692886}
  • Predications:
    • Family Member{father side of the family}, Observation{Heart disease}, Negated{false}
    • Family Member{Mother}, Observation{arthritis}, Negated{false}
Leveraging NLP in Clinical Research

Clinical Notes

Biomedical Literature

Social Media

NLP (extract, classify, summarize)

- Social history
- Function score
- Medical history
- Smoking status

Biomedical knowledge
(Subj ect – Predicate - Object)

Pharmacovigilance signals
(Drug/supplement - adverse Events)

Clinical researchers
Leveraging Big Data for Pharmacovigilance

https://knowledgent.com/whitepaper/big-data-enabling-better-pharmacovigilance/
Shared NLP Tasks

• Challenges
  - Lack of shared resources and evaluation (de-identification, recognition of medical concepts, semantic modifies, temporal information)

• Shared tasks
  - Informatics for Integrating Biology and the Bedside (i2b2) challenges
  - Conference and Labs of the evaluation Forum (CLEF) eHealth challenges
  - Semantic Evaluation (SemEval) challenges

Springer International Publishing (2016), 10.1007/978-3-319-44564-9_24 pp. 255–266
Computational Linguistics, San Diego, California (2016), pp. 1052-1062
# Open source NLP Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
<th>Institute (PI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedLEE</td>
<td>An expert-based NLP system for unlocking clinical information from narratives</td>
<td>Columbia U (Friedman)</td>
</tr>
<tr>
<td>cTAKES</td>
<td>A UIMA pipeline built around openNLP, Lucene, and LVG for extracting disorders, drugs, anatomical sites, and procedures information from clinical notes</td>
<td>Mayo Clinic (Chute)</td>
</tr>
<tr>
<td>MedEX</td>
<td>A semantic-based medication extraction system designed to extract medication names and prescription information</td>
<td>U Texas Houston (Xu)</td>
</tr>
<tr>
<td>HiTEX</td>
<td>An NLP system distributed through i2b2</td>
<td>Harvard U (Zeng)</td>
</tr>
<tr>
<td>MedTagger</td>
<td>A machine learning based name entity detection system utilizing existing terminologies</td>
<td>Mayo Clinic (Liu)</td>
</tr>
<tr>
<td>BioMedICUS</td>
<td>A UIMA pipeline system designed for researchers for extracting and summarizing information from unstructured text of clinical reports</td>
<td>U Minnesota (Pakhomov)</td>
</tr>
</tbody>
</table>
Chaining NLP Tasks: Pipelines

• Any practical NLP task must perform sub-tasks (low-level tasks must execute sequentially)
• Pipelined system enables applications to be decomposed into components
• Each component does the actual work of analyzing the unstructured information
• Unstructured information management architecture (UIMA)
Evaluation of Clinical Research and NLP

• The goal of clinical research (trial, cohort study):
  - To assess the association between a risk factor or intervention with an clinical outcome
    • Internal validation: measured on the original study sample
    • External validation: measured on a different sample

• The goal of NLP method development
  - To produce computational solutions to special problem
    • Intrinsic: measuring on attaining its immediate objective
    • Extrinsic: evaluating its usefulness in an overarching goal where NLP is part of a more complex process
Evaluation of Clinical Research and NLP

- NLP development mainly focuses on intrinsic evaluation
  - Document (patient status, report type)
  - Documents section (current med, past med history, discharge summary)
  - Named entities and concepts (diagnosis, symptoms, treatments)
  - Semantic attributes (negation, severity, temporality)

- Intrinsic evaluation may not be informative when they apply on higher level problem (patient level) or new data
  - In clinical practice, any >0% error rate (the misclassification of a drug or a history of severe allergy) is unacceptable
  - True negative are rarely considered in NLP evaluation, but is key factor in clinical research (medical screening)

- It is unclear how best to incorporate and interpret NLP performance when using outputs from NLP approaches in clinical research.
NLP-PIER (Patient Information Extraction for Research)
NLP-PIER

• A clinical notes processing platform including an NLP query and search engine for clinical and translational researchers

• System is secured by an authentication and authorization layer

• System was designed to give clinical researchers access to NLP capabilities for searching clinical notes in an environment that is compliant for accessing protected health information (PHI)

http://athena.ahc.umn.edu/nlppier/
NLP-PIER

- Users only have access to sets of notes that are defined externally and configured in the Elasticsearch engine
- Access granted through CTSI-BPIC
- Provide researchers direct access to patient data in free text of **170 million** clinical notes for **>2.9 million** patients (as of May 2019)
NLP-PIER Search Capabilities

• Keyword searching
• Advanced query syntax
  ➢ NOT, AND, OR
  ➢ Grouping
  ➢ Distance syntax
• Identified UMLS concepts
  ➢ Historical, negated modifiers
• Word vector - based query expansion
  ➢ Misspellings
  ➢ Contextually related terms
Other Capabilities

- Personalized filters
- Set export settings
- Save and share queries
- Query expansion
- Patient/encounter counts
Query Help Options

Query string syntax

NLP-PIER leverages elastic search for indexing unstructured EMR notes. Queries utilize the Lucene query syntax, a powerful and flexible syntax for finding that surgical needle in the clinical notes haystack. Commonly used query types are listed below. Examples can be pasted as templates into the search box by clicking on the example itself. Modify as necessary according to your needs.

Example Queries (NLP-PIER defaults to logical AND queries; case of logical operator matters)

heart failure
  Find notes containing both heart and failure; relative position does not matter

"heart failure"
  Find where the terms heart and failure occur next to one another in the same note

heart AND failure
  Logical AND query: same as heart failure, default search behavior

heart OR failure
  Logical OR query: find notes containing either heart or failure

heart NOT failure
  Logical NOT query: find notes containing heart and missing (does not contain) failure

"heart failure" + female
  Find notes containing the phrase "heart failure" and the term female; + and AND are equivalent operators

"heart irregular" - 10
  Find notes containing the terms heart and irregular within 10 terms of each other

mrn:xxxxxxx
  Restrict results to the specified MRN. Can be used in combination with other terms, e.g., keywords and/or service date(s)

service_date [2018-07-07 TO 2018-07-14]
  Restrict results to those with a service date within a range. Ranges using [ ] are inclusive; use () for exclusive ranges. These can be used in combination. Wildcards can be used as upper or lower bounds, e.g., service_date [ * TO 2018-12-31]. Single service date values are also permitted, e.g., service_date 2012-06-02

cus:C0033213
  Find notes tagged with UMLS CUIs (Concept Unique Identifier). Can be combined using logical AND / OR operators, e.g., cus:C0039706 OR cus:C2137071

Query syntax pointers from elasticsearch. Or consult the Lucene reference query syntax documentation directly from the Lucene API site.
## Settings Menu

These options control which filters are displayed along the left side of the search results and how many filter options per field are displayed. Changes persist across logins.

<table>
<thead>
<tr>
<th>Epic Categories</th>
<th>Filter category</th>
<th>Enabled</th>
<th>Filter values displayed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department Id</td>
<td></td>
<td></td>
<td></td>
<td>CDR department identifier</td>
</tr>
<tr>
<td>Encounter Center</td>
<td></td>
<td></td>
<td></td>
<td>Encounter center name in Epic</td>
</tr>
<tr>
<td>Encounter Center Type</td>
<td></td>
<td></td>
<td></td>
<td>Encounter center type in Epic</td>
</tr>
<tr>
<td>Encounter Clinic Type</td>
<td></td>
<td></td>
<td></td>
<td>Encounter Clinic type in Epic</td>
</tr>
<tr>
<td>Encounter Department</td>
<td></td>
<td></td>
<td></td>
<td>Encounter department in Epic</td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
<td></td>
<td></td>
<td>Specialty name in Epic</td>
</tr>
<tr>
<td>Encounter Id</td>
<td></td>
<td></td>
<td></td>
<td>Epic visit number</td>
</tr>
<tr>
<td>Filling Date</td>
<td></td>
<td></td>
<td></td>
<td>Date note was filled</td>
</tr>
<tr>
<td>Filling Datetime</td>
<td></td>
<td></td>
<td></td>
<td>Date/time note was filled</td>
</tr>
<tr>
<td>Mnr</td>
<td></td>
<td></td>
<td></td>
<td>Epic patient identifier</td>
</tr>
<tr>
<td>Patient Id</td>
<td></td>
<td></td>
<td></td>
<td>CDR Patient ID</td>
</tr>
<tr>
<td>Prov Id</td>
<td></td>
<td></td>
<td></td>
<td>Provider ID in Epic</td>
</tr>
<tr>
<td>Prov Name</td>
<td></td>
<td></td>
<td></td>
<td>Provider name in Epic</td>
</tr>
<tr>
<td>Prov Type</td>
<td></td>
<td></td>
<td></td>
<td>Provider type in Epic name</td>
</tr>
<tr>
<td>Provider Id</td>
<td></td>
<td></td>
<td></td>
<td>CDR department identifier</td>
</tr>
<tr>
<td>Service Date</td>
<td></td>
<td></td>
<td></td>
<td>Date of Service</td>
</tr>
<tr>
<td>Service Id</td>
<td></td>
<td></td>
<td></td>
<td>CDR encounter identifier</td>
</tr>
<tr>
<td>Text Source Format</td>
<td></td>
<td></td>
<td></td>
<td>Plain text, rich text, format of analyzed note</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HL7 LOINC</th>
<th>Filter category</th>
<th>Enabled</th>
<th>Filter values displayed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kod</td>
<td></td>
<td></td>
<td></td>
<td>Kind of Document axis in HL7-LOINC DO</td>
</tr>
<tr>
<td>Role</td>
<td></td>
<td></td>
<td></td>
<td>Role axis in HL7-LOINC DO</td>
</tr>
<tr>
<td>Setting</td>
<td></td>
<td></td>
<td></td>
<td>Setting axis in HL7-LOINC DO</td>
</tr>
<tr>
<td>Smd</td>
<td></td>
<td></td>
<td></td>
<td>Subject Matter Domain in HL7-LOINC DO</td>
</tr>
<tr>
<td>Tos</td>
<td></td>
<td></td>
<td></td>
<td>Subject Matter Domain in HL7-LOINC DO</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NLP Annotations</th>
<th>Filter category</th>
<th>Enabled</th>
<th>Filter values displayed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Confidence Medical Concepts</td>
<td></td>
<td></td>
<td>10</td>
<td>UMLS CUIs identified by BioMediCUS NLP pipeline, lower confidence detection</td>
</tr>
<tr>
<td>Medical Concepts</td>
<td></td>
<td></td>
<td>20</td>
<td>UMLS CUIs identified by BioMediCUS NLP pipeline</td>
</tr>
</tbody>
</table>

- **Change number of displayed filters here**
- **Select type of filters that are enabled here**
Query Expansion Word Vectors

- Related misspellings
- Semantically related terms

Find Common Misspellings Here
Find Related Terms Here
EMERSE (Electronic Medical Record Search Engine)

- Enables users to search clinical notes (dictated or typed) from our electronic medical record (CareWeb and MiChart) for terms.
- EMERSE aids in cohort identification, eligibility determination and data abstraction in a variety of research, clinical, and operational settings.
- Similar to PIER search engine
- Expert curated Synonyms

Search with Synonyms
Visualization with NER
24,123 patients matched the search criteria

24,123 patients matched the search criteria

24,123 patients matched the search criteria

Cohort Identification

Cohort Identification

Cohort Identification
Use Case 1: NLP to Support Dietary Supplement Safety Research

- Expanding Supplement Terminology from Clinical Notes
- Detecting Supplement Use Status
- Detecting Safety Signals about Supplements in Clinical Notes
- Mining biomedical Literature to Discover DSIs
- Active Learning to Reduce Annotation Costs

1R01AT009457 (PI: Rui Zhang)
Introduction to Dietary Supplements

• Dietary supplements
  – Herbs, vitamins, minerals, probiotics, amino acids, others.

• Use of supplements increasing
  – More than half of U.S. adults take dietary supplements (Center for Disease Control and Prevention)
  – One in six U.S. adults takes a supplement simultaneously with prescription medications
  – Sales over $6 billion per year in U.S. (American Botanical Council, 2014)

https://nccih.nih.gov/health/supplements
Safety of Dietary Supplements

• Doctors often poorly informed about supplements
  ➢ 75.5% of 1,157 clinicians

• Supplements are NOT always safe
  ➢ Averagely 23,000 annual emergency visits for supplements adverse events
  ➢ Drug-supplement interactions (DSIs)
    • Concomitant administration of supplements and drugs increases risks of DSIs
    • Example: Docetaxel & St John’s Wort (hyperforin component induces docetaxel metabolism via P450 3A4)

Regulation for Dietary Supplements

• Regulated by Dietary Supplement Health and Education Act of 1994 (DSHEA)
  – Different regulatory framework from prescription and over-the-counter drugs
  – Safety testing and FDA approval NOT required before marketing
  – Postmarketing reporting only required for serious adverse events (hospitalization, significant disability or death)
Limited Supplements Research

• Supplement safety research is limited
  – Not required for clinical trials
  – Not found until new supplement is on the market
  – Voluntary adverse events reporting underestimates the safety issues
  – Pharmacy studies only focuses on specific supplements
  – DSI documentation is limited due to less rigorous regulatory rules on supplements
  – No existing standard supplement terminology
Limited Supplements Research

• Limited knowledge on supplements $^{1,2}$
  – Safety (adverse effects, interactions, precautions, etc.)
  – Efficacy
  – Mechanism of action
  – Bioavailability/dosing
  – Metabolism/excretion
  – Other essential data elements (naming, type, source, origin, etc.)

Informatics to Support Supplements Research

• Online resources
  – Provides DS knowledge across various resources
  – Need informatics method to standard and integrate knowledge

• Biomedical literature
  – Contains pharmacokinetics and pharmacodynamics knowledge
  – Discover undefined pathways for DSIs
  – Find potential DSIs by linking information
  – Limited studies to discover DSIs

• Electronic health records
  – EHR provides patient data for supplement use
  – Detailed supplements usage information documented in clinical notes
  – No studies investigating the supplements in clinical notes
Challenges

• Lexical variations of supplements in clinical notes
• Detailed usage information related to supplements
• No standardized and consistent DS knowledge representation
1.1. Expanding Supplement Terminology in Clinical Notes using Word Embeddings

• Thesaurus-based method (e.g., MeSH, SNOMED-CT)

• Distributional semantics
  • Word similarity is estimated based on the distribution of the words in the corpus
    • Traditional methods
      • Vector models (high dimensional; sparsity issue)
    • Word embeddings
      • Reveal hidden relationship between words (similarity and relatedness)
      • More efficient; can be trained a large amount of unannotated data
Objective

• To apply word embedding models to expand the terminology of DS from clinical notes: semantic variants, brand names, misspellings
  • Corpus size
  • Compare two word embedding models
    • Word2vec, GloVe

<table>
<thead>
<tr>
<th>calcium</th>
<th>chamomile</th>
<th>cranberry</th>
<th>dandelion</th>
<th>flaxseed</th>
<th>garlic</th>
<th>ginger</th>
</tr>
</thead>
<tbody>
<tr>
<td>ginkgo</td>
<td>ginseng</td>
<td>glucosamine</td>
<td>lavender</td>
<td>melatonin</td>
<td>turmeric</td>
<td>valerian</td>
</tr>
</tbody>
</table>
Method Overview

Figure 1. The overview and workflow of the method. EHR: electronic health record.
Model Training

- Corpus size

<table>
<thead>
<tr>
<th>Time span of clinical notes for 7 corpora</th>
<th>3 months</th>
<th>6 months</th>
<th>9 months</th>
<th>12 months</th>
<th>15 months</th>
<th>18 months</th>
<th>21 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size</td>
<td>214 948</td>
<td>312 557</td>
<td>388 891</td>
<td>454 459</td>
<td>520 127</td>
<td>577 362</td>
<td>635 176</td>
</tr>
<tr>
<td>Semantic variants</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Brand names</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Misspellings</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>31</td>
<td>31</td>
<td>36</td>
<td>30</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>MAP</td>
<td>0.313</td>
<td>0.294</td>
<td>0.356</td>
<td>0.247</td>
<td>0.242</td>
<td>0.280</td>
<td>0.263</td>
</tr>
</tbody>
</table>

MAP: mean average precision; DS: dietary supplements.

- Hyperparameter tuning
  - Window size (i.e., 4, 6, 8, 10, and 12)
  - Vector size (i.e., 100, 150, 200, 250)

- Glove trained on the same corpus
  - Window size and vector size

- Optimal parameters were chosen based on human annotation (intrinsic evaluation)
Figure 1. The number of semantic similar terms identified by human experts based on 40 top-ranked terms by word2vec for each DS from 7 corpora
## Results: Query Expansion Examples

<table>
<thead>
<tr>
<th>Initial Query</th>
<th>word2vec Expanded Query</th>
<th>Expanded Examples</th>
</tr>
</thead>
</table>
| **Black cohosh**   | **Misspelling:** black kohosh, black kohash; **Brand name:** remifemin Estroven Estrovan estraven icool amberen amberin Estrovera EstroFactor | • Please try black cohosh or Estroven for hot flashes.  
• Pt has discontinued Remifemin but still has symptoms.  
• Recommend Estroven trial for symptoms of menopause. |
| **Turmeric**       | **Misspelling:** tumeric                                                                  | • Pt emailed wondering about taking Tumeric  
• Patient states that she sometimes takes the supplements Tumeric |
| **Folic acid**     | **Brand name:** Folgard, Folbic  
**Other name:** Folate                                                                       | • Patient is willing to try Folgard if ok with provider.  
• Patient is on folate and does not smoke. |
| **Valerian**       | **Misspelling:** velarian  
**Brand name:** myocalm pm, somnapure                                                       | • Taking Velarian root and benadryl as well  
• I would recommend moving to 6mg dose first, then trying somnapure if still not helping. |
| **Melatonin**      | **Misspelling:** Melantoin, melotonin  
**Brand name:** alteril, neuro sleep                                                           | • Can try melantoin for sleep aid.  
• Try alteril - it is over the counter sleep aid Let me know if this is not better over the next few weeks |
### Results: Comparison of Base and Expanded Queries

<table>
<thead>
<tr>
<th>Dietary supplements</th>
<th>Number of expanded terms</th>
<th>Number of clinical notes</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base query</td>
<td>Expanded query</td>
<td>Additional records found</td>
</tr>
<tr>
<td>Black cohosh</td>
<td>13,782</td>
<td>23,641</td>
<td>9,859</td>
</tr>
<tr>
<td>Calcium</td>
<td>7,024,626</td>
<td>7,053,856</td>
<td>29,230</td>
</tr>
<tr>
<td>Cranberry</td>
<td>187,586</td>
<td>189,239</td>
<td>1,653</td>
</tr>
<tr>
<td>Dandelion</td>
<td>4,316</td>
<td>4,375</td>
<td>59</td>
</tr>
<tr>
<td>Fish oil</td>
<td>1,305,996</td>
<td>1,311,777</td>
<td>5,781</td>
</tr>
<tr>
<td>Folic acid</td>
<td>839,710</td>
<td>1,058,627</td>
<td>218,917</td>
</tr>
<tr>
<td>Garlic</td>
<td>91,342</td>
<td>92,481</td>
<td>1,139</td>
</tr>
<tr>
<td>Ginger</td>
<td>88,870</td>
<td>88,870</td>
<td>0</td>
</tr>
<tr>
<td>Ginkgo</td>
<td>19,020</td>
<td>27,502</td>
<td>8,482</td>
</tr>
<tr>
<td>Ginseng</td>
<td>9,663</td>
<td>10,748</td>
<td>1,085</td>
</tr>
<tr>
<td>Glucosamine</td>
<td>468,774</td>
<td>469,925</td>
<td>1,151</td>
</tr>
<tr>
<td>Green tea</td>
<td>29,810</td>
<td>29,816</td>
<td>6</td>
</tr>
<tr>
<td>Melatonin</td>
<td>647,389</td>
<td>647,601</td>
<td>212</td>
</tr>
<tr>
<td>Milk thistle</td>
<td>18,930</td>
<td>19,298</td>
<td>368</td>
</tr>
<tr>
<td>Saw palmetto</td>
<td>38,934</td>
<td>38,947</td>
<td>13</td>
</tr>
<tr>
<td>Turmeric</td>
<td>25,172</td>
<td>37,959</td>
<td>12,787</td>
</tr>
<tr>
<td>Valerian</td>
<td>15,023</td>
<td>15,330</td>
<td>307</td>
</tr>
<tr>
<td>Vitamin E</td>
<td>384,072</td>
<td>384,072</td>
<td>0</td>
</tr>
</tbody>
</table>
# Results: Comparison with External Source

Comparison between word embedding expanded queries and external source expanded queries (task 2) for 14 DS (selected examples)

<table>
<thead>
<tr>
<th>Dietary supplements</th>
<th>Number of external source terms</th>
<th>Number of word embedding expanded terms</th>
<th>Number of clinical notes</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[ <strong>Number of clinical notes</strong> ]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>External source query</td>
<td>Word embedding query</td>
<td>Additional records found</td>
</tr>
<tr>
<td>Calcium</td>
<td>15</td>
<td>12</td>
<td>7453873</td>
<td>7543569</td>
</tr>
<tr>
<td>Cranberry</td>
<td>21</td>
<td>3</td>
<td>196944</td>
<td>198625</td>
</tr>
<tr>
<td>Flaxseed</td>
<td>10</td>
<td>2</td>
<td>169349</td>
<td>169343</td>
</tr>
<tr>
<td>Ginkgo</td>
<td>6</td>
<td>3</td>
<td>20275</td>
<td>28093</td>
</tr>
<tr>
<td>Turmeric</td>
<td>18</td>
<td>3</td>
<td>35719</td>
<td>48749</td>
</tr>
</tbody>
</table>

External sources: Natural Medicines Comprehensive Database (NMCD), Dietary Supplement Label Database (DSLD)
1.2. Extracting Supplements’ Usage Information in Clinical Notes

To classify the use status of the supplements in clinical notes into four categories: Continuing, Discontinued, Started, and Unclassified.

Fan et al., AMIA Joint Summit 2017:493-501. (The 2nd place in Student Paper Competition)
## Results: Performance comparison

<table>
<thead>
<tr>
<th>Type</th>
<th>Features</th>
<th>Decision tree</th>
<th>Random forest</th>
<th>Naïve Bayes</th>
<th>SVM</th>
<th>Maximum Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Type 1</td>
<td>raw uni\textsuperscript{a}</td>
<td>0.819</td>
<td>0.817</td>
<td>0.816</td>
<td>0.858</td>
<td>0.853</td>
</tr>
<tr>
<td>Type 2</td>
<td>uni</td>
<td>0.846</td>
<td>0.845</td>
<td>0.844</td>
<td>0.878</td>
<td>0.876</td>
</tr>
<tr>
<td>Type 3</td>
<td>tf-idf</td>
<td>0.862</td>
<td>0.857</td>
<td>0.857</td>
<td>0.862</td>
<td>0.857</td>
</tr>
<tr>
<td>Type 4</td>
<td>bi\textsuperscript{a}</td>
<td>0.760</td>
<td>0.720</td>
<td>0.716</td>
<td>0.760</td>
<td>0.720</td>
</tr>
<tr>
<td>Type 5</td>
<td>uni + bi</td>
<td>0.872</td>
<td>0.864</td>
<td>0.863</td>
<td>0.872</td>
<td>0.864</td>
</tr>
<tr>
<td>Type 6</td>
<td>uni + bi+tri\textsuperscript{a}</td>
<td>0.863</td>
<td>0.852</td>
<td>0.850</td>
<td>0.863</td>
<td>0.852</td>
</tr>
<tr>
<td>Type 7</td>
<td>indi\textsuperscript{a} only</td>
<td>0.848</td>
<td>0.847</td>
<td>0.846</td>
<td>0.861</td>
<td>0.860</td>
</tr>
<tr>
<td>Type 8</td>
<td>uni + bi+indi</td>
<td>0.860</td>
<td>0.860</td>
<td>0.860</td>
<td>0.875</td>
<td>0.865</td>
</tr>
<tr>
<td>Type 9</td>
<td>uni + bi+tri + indi</td>
<td>0.860</td>
<td>0.857</td>
<td>0.857</td>
<td>0.872</td>
<td>0.861</td>
</tr>
</tbody>
</table>

\textsuperscript{a}uni: unigrams; bi: bigrams; tri: trigrams; indi: indicators

*P: precision, R: recall, F: F-measure

70% training, 30% testing

**Type 1**: raw unigrams without normalization; **Type 2**: unigrams (normalized);

**Type 3**: TF-IDF (term frequency – inversed document frequency) for unigrams;

**Type 4**: bigrams; **Type 5**: unigrams + bigrams; **Type 6**: unigrams + bigrams + trigrams; **Type 7**: indicator words only;

**Type 8**: unigrams + bigrams + indicator words with distance (window size);

**Type 9**: unigrams + bigrams + trigrams + indicator words with distance

Performance comparison

The Performance of Maximum Entropy with Type 8 in Test Set

<table>
<thead>
<tr>
<th>Status</th>
<th>Number</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuing</td>
<td>233</td>
<td>0.86</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>Discontinued</td>
<td>166</td>
<td>0.94</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>Started</td>
<td>178</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Unclassified</td>
<td>173</td>
<td>0.92</td>
<td>0.84</td>
<td>0.88</td>
</tr>
<tr>
<td>Total (weighted)</td>
<td>750</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

The Performance of Classifier on 25 dietary supplements

<table>
<thead>
<tr>
<th>Dietary Supplement</th>
<th>Number</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfalfa</td>
<td>30</td>
<td>0.904</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Biotin</td>
<td>30</td>
<td>0.927</td>
<td>0.900</td>
<td>0.904</td>
</tr>
<tr>
<td>Black cohosh</td>
<td>30</td>
<td>0.937</td>
<td>0.933</td>
<td>0.933</td>
</tr>
<tr>
<td>Coenzyme Q10</td>
<td>30</td>
<td>0.809</td>
<td>0.800</td>
<td>0.799</td>
</tr>
<tr>
<td>Cranberry</td>
<td>30</td>
<td>0.945</td>
<td>0.933</td>
<td>0.934</td>
</tr>
<tr>
<td>Dandelion</td>
<td>30</td>
<td>0.939</td>
<td>0.933</td>
<td>0.926</td>
</tr>
<tr>
<td>Echinacea</td>
<td>30</td>
<td>0.913</td>
<td>0.900</td>
<td>0.902</td>
</tr>
<tr>
<td>Fish oil</td>
<td>30</td>
<td>0.938</td>
<td>0.933</td>
<td>0.933</td>
</tr>
<tr>
<td>Flax seed</td>
<td>30</td>
<td>0.900</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Folic acid</td>
<td>30</td>
<td>0.911</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Garlic</td>
<td>30</td>
<td>0.919</td>
<td>0.900</td>
<td>0.903</td>
</tr>
<tr>
<td>Ginger</td>
<td>30</td>
<td>0.893</td>
<td>0.867</td>
<td>0.861</td>
</tr>
<tr>
<td>Ginkgo</td>
<td>30</td>
<td>0.943</td>
<td>0.933</td>
<td>0.932</td>
</tr>
<tr>
<td>Ginseng</td>
<td>30</td>
<td>0.947</td>
<td>0.933</td>
<td>0.935</td>
</tr>
<tr>
<td>Glucosamine</td>
<td>30</td>
<td>0.936</td>
<td>0.933</td>
<td>0.933</td>
</tr>
<tr>
<td>Glutamine</td>
<td>30</td>
<td>0.938</td>
<td>0.933</td>
<td>0.934</td>
</tr>
<tr>
<td>Kava kava</td>
<td>30</td>
<td>0.913</td>
<td>0.900</td>
<td>0.902</td>
</tr>
<tr>
<td>Lecithin</td>
<td>30</td>
<td>0.939</td>
<td>0.933</td>
<td>0.934</td>
</tr>
<tr>
<td>Melatonin</td>
<td>30</td>
<td>0.806</td>
<td>0.800</td>
<td>0.801</td>
</tr>
<tr>
<td>Milk thistle</td>
<td>30</td>
<td>0.787</td>
<td>0.767</td>
<td>0.751</td>
</tr>
<tr>
<td>Saw palmetto</td>
<td>30</td>
<td>0.907</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>St. John’s Wort</td>
<td>30</td>
<td>0.910</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Turmeric</td>
<td>30</td>
<td>0.927</td>
<td>0.900</td>
<td>0.886</td>
</tr>
<tr>
<td>Valerian</td>
<td>30</td>
<td>0.944</td>
<td>0.933</td>
<td>0.928</td>
</tr>
<tr>
<td>Vitamin E</td>
<td>30</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

# Deep Learning Text Classification Methods

- **Word-level CNN**
  - Filter size: [1, 2, 3, 4, 5, 6], number of filters: 256

- **Bi-LSTM**
  - Hidden units: 256

- **Stacked Bi-LSTM**
  - Layers: 2, Hidden units: 128

<table>
<thead>
<tr>
<th>Models</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure (weighted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-CNN</td>
<td>0.808</td>
<td>0.804</td>
<td>0.803</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>0.882</td>
<td>0.880</td>
<td>0.879</td>
</tr>
<tr>
<td>Stacked Bi-LSTM</td>
<td>0.918</td>
<td>0.916</td>
<td>0.916</td>
</tr>
</tbody>
</table>
1.3. Mining Biomedical Literature to Discovery Drug-Supplement Interactions (DSIs)

Researchers at the University of Minnesota in Minneapolis are exploring interactions between cancer drugs and dietary supplements, based on data extracted from 23 million scientific publications, according to lead author Rui Zhang, a clinical assistant professor in health informatics. In a study published last year by a conference of the American Medical Informatics Association, he says, they identified some that were previously unknown.

http://www.wsj.com/articles/what-you-should-know-about-how-your-supplements-interact-with-prescription-drugs-1456777548
http://www.foxnews.com/health/2016/03/01/what-should-know-about-how-supplements-interact-with-prescription-drugs.html
We have shown that ECHINACEA preparations and some common alkylamides weakly inhibit several cytochrome P450 (CYP) isoforms, with considerable variation in potency. (19790031)

Tamoxifen and toremifene are metabolised by the cytochrome p450 enzyme system, and raloxifene is metabolised by glucuronide conjugation. (12648026)
# Results: Selected Interactions

<table>
<thead>
<tr>
<th>Drug/Supplement</th>
<th>Predicate</th>
<th>Gene/Gene Class</th>
<th>Predicate</th>
<th>Supplement/Drug</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Echinacea</td>
<td>INH</td>
<td>CYP450</td>
<td>INT</td>
<td>Docetaxel</td>
<td>Y</td>
</tr>
<tr>
<td>Echinacea</td>
<td>INH</td>
<td>CYP450</td>
<td>INT</td>
<td>Toremifene</td>
<td>N</td>
</tr>
<tr>
<td>Echinacea</td>
<td>STI</td>
<td>CYP1A1</td>
<td>INT</td>
<td>Exemestane</td>
<td>N</td>
</tr>
<tr>
<td>Grape seed extract</td>
<td>INH</td>
<td>CYP3A4</td>
<td>INT</td>
<td>Docetaxel</td>
<td>N</td>
</tr>
<tr>
<td>Kava preparation</td>
<td>STI</td>
<td>CYP3A4</td>
<td>INT</td>
<td>Docetaxel</td>
<td>Y</td>
</tr>
</tbody>
</table>

INH, INHIBITS; STI, STIMULATES; INT, INTERACTS_WITH

Echinacea: fights the common cold and viral infections
Grape seed extract: cardiac conditions
Kava: treat sleep problems, relieve anxiety and stress
## Results: Selected Predications

<table>
<thead>
<tr>
<th>Semantic predication</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Echinacea INHIBITS CYP450</td>
<td>We have shown that ECHINACEA preparations and some common alkylamides weakly inhibit several cytochrome P450 (CYP) isoforms, with considerable variation in potency. (19790031)</td>
</tr>
<tr>
<td>Grape seed extract INHIBITS CYP3A4</td>
<td>Four brands of GSE had no effect, while another five produced mild to moderate but variable inhibition of CYP3A4, ranging from 6.4% by Country Life GSE to 26.8% by Loma Linda Market brand. (19353999)</td>
</tr>
<tr>
<td>Melatonin INHIBITS Cyclooxygenase-2</td>
<td>Moreover, Western blot analysis showed that melatonin inhibited LPS/IFN-gamma-induced expression of COX-2 protein, but not that of constitutive cyclooxygenase. (18078452)</td>
</tr>
<tr>
<td>CYP450 INTERACTS_WITH Toremifene</td>
<td>Tamoxifen and toremifene are metabolised by the cytochrome p450 enzyme system, and raloxifene is metabolised by glucuronide conjugation. (12648026)</td>
</tr>
<tr>
<td>CYP3A INHIBITS Docetaxel</td>
<td>Because docetaxel is inactivated by CYP3A, we studied the effects of the St. John's wort constituent hyperforin on docetaxel metabolism in a human hepatocyte model. (16203790)</td>
</tr>
</tbody>
</table>
1.4. Active Learning to Reduce Annotation Costs for NLP Tasks

- NLP tasks require human annotations
  - Time consuming and labor intensive
- Active learning reduces annotation costs
  - Used in biomedical and clinical texts
  - Effectiveness varies across datasets and tasks

Objectives

• To assess the effectiveness of AL methods on filtering incorrect semantic predication
• To evaluate various query strategies and provide a comparative analysis of AL method through visualization
Method Overview

Query strategies:
- Uncertainty sampling
- Representative sampling
- Combined sampling

Evaluation:
- 10-fold cross validation
- Training = 2700, $L_0=270$
- Testing = 300 using AUC
Datasets and Annotations

• Substance interaction (3,000):
  – INTERACTS_WITH, STIMULATES, or INHIBITS

• Clinical Medicine (3,000):
  – ADMINISTERED_TO, COEXISTS_WITH, COM-Plicates, DIAGNOSES, MANIFESTATION_OF, PRE-Cedes, PREVENTS, PROCESS_OF, PRODUCES, TREATS, or USES

• Inter-rater agreement:
  – Kappa: 0.74 (SI), 0.72 (CM)
  – Percentage agreement: 87% (SI), 91% (CM)
Performance Comparison

When \( L \) is small and \( U \) is large:

- it is unlikely that \( L \) is representative of \( U \)
- given that \( L \) is small and unrepresentative, the prediction model trained on \( L \) is likely to be poor.

\[
\beta = \frac{2|U|}{|L|}
\]

- \( |U| \) is the size of the current unlabeled set
- \( |L| \) is the size of the current labeled set

Table 1. Area under the learning curve (ALC) and number of training examples required to reach target area under the ROC curve (AUC) of the uncertainty, representative, and combined query strategies evaluated on the substance interactions and clinical medicine datasets
Results

Uncertainty Sampling

<table>
<thead>
<tr>
<th>Query Strategy</th>
<th>ALC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive Learning</td>
<td>0.590</td>
</tr>
<tr>
<td>Uncertainty Sampling</td>
<td>0.597 – 0.607</td>
</tr>
</tbody>
</table>
Results

Representative Sampling

<table>
<thead>
<tr>
<th>Query Strategy</th>
<th>ALC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive Learning</td>
<td>0.590</td>
</tr>
<tr>
<td>Uncertainty Sampling</td>
<td>0.597 – 0.607</td>
</tr>
<tr>
<td>Representative Sampling</td>
<td>0.622 – 0.634</td>
</tr>
</tbody>
</table>
Results
Combined Sampling

<table>
<thead>
<tr>
<th>Query Strategy</th>
<th>ALC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive Learning</td>
<td>0.590</td>
</tr>
<tr>
<td>Uncertainty Sampling</td>
<td>0.597 – 0.607</td>
</tr>
<tr>
<td>Representative Sampling</td>
<td>0.622 – 0.634</td>
</tr>
<tr>
<td>ID (manual $\beta$)</td>
<td>0.642</td>
</tr>
</tbody>
</table>

Area Under the ROC Curve (AUC) vs. Labeled set size
Results

Dynamic $\beta$

<table>
<thead>
<tr>
<th>Query Strategy</th>
<th>ALC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive Learning</td>
<td>0.590</td>
</tr>
<tr>
<td>Uncertainty Sampling</td>
<td>0.597 – 0.607</td>
</tr>
<tr>
<td>Representative Sampling</td>
<td>0.622 – 0.634</td>
</tr>
<tr>
<td>ID (manual $\beta$)</td>
<td>0.642</td>
</tr>
<tr>
<td>ID (dynamic $\beta$)</td>
<td>0.641</td>
</tr>
</tbody>
</table>

Area Under the ROC Curve (AUC) vs Labeled set size
Performance Analysis

Uncertainty Sampling (worst performing)

Representative Sampling (best performing)

Use Case 2: NLP in Mental Health Research

- NLP to Extract Symptoms of Severe Mental Illness (SMI) from Clinical Texts
- Deep Neural Network for Phenotyping Youth Depression
Introduction

• Mental illness is a condition that affects a person’s thinking, feeling, and behavior
• There are five major categories of mental illnesses:
  – Anxiety disorders
  – Mood disorders
  – Schizophrenia and psychotic disorders
  – Dementia
  – Eating disorders
Mental Health Records

• Most salient information for research and clinical practice in text filed (70%)
  – Self-reported experience
  – Determining treatment initiation and outcome evaluation
  – 90 documents per patient (South London and Maudsley mental health trust)

• Most clinical researchers and clinicians collect data using standardized instrument
  – Beck Depression Inventory (BDI)
  – the Positive and Negative Syndrome Scale (PANNS)
2.1. Extract Symptoms of Severe Mental Illness (SMI) from Clinical Texts

• Background
  – SMI: schizophrenia, schizoaffective disorder and bipolar disorder
  – Diagnoses (ICD or DSM) form semantically convenient unit
  – Mental disorders have broad symptomatic manifestations
    • Schizophrenia (all, or few of associated symptoms)
  – Symptomatology, compared to diagnoses, offer more objective patient grouping

• Objective:
  – develop NLP models to capture key symptoms of SMI to facilitate the secondary use of mental health data in research

Method

• Data:
  – EHR from a large mental health providing serving 1.2 million residents in UK
  – 3.5 million documents

• NLP task
  – Sentence classification
    • Symptom keywords
    • Clinical relevant modifier terms (product subclassification)
# SMI Keywords and Modifiers

## Table 1: Symptom instance definitions

<table>
<thead>
<tr>
<th>SMI concept</th>
<th>Keyword strings</th>
<th>Modifier strings</th>
<th>Lax or strict modifiers</th>
<th>SNOYMED-CT (SCTID)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression</td>
<td>aggress*</td>
<td></td>
<td></td>
<td>61372001</td>
</tr>
<tr>
<td>Agitation</td>
<td>agitat*</td>
<td></td>
<td></td>
<td>106126000</td>
</tr>
<tr>
<td>Anhedonia</td>
<td>anhedon*</td>
<td></td>
<td></td>
<td>28669007</td>
</tr>
<tr>
<td>Apathy</td>
<td>apath*</td>
<td></td>
<td></td>
<td>20602000</td>
</tr>
<tr>
<td>Arousal</td>
<td>arous*</td>
<td></td>
<td></td>
<td>(none)</td>
</tr>
<tr>
<td>Blunted or flat affect</td>
<td>Affect</td>
<td>blunt*, flat*, restrict*</td>
<td>Optional</td>
<td>6140007/932006/39370001</td>
</tr>
<tr>
<td>Catalepsy</td>
<td>catalep*</td>
<td></td>
<td></td>
<td>247917007</td>
</tr>
<tr>
<td>Catatonic syndrome</td>
<td>catatoni*</td>
<td></td>
<td></td>
<td>247917007</td>
</tr>
<tr>
<td>Circumstantial speech</td>
<td>circumstan*</td>
<td></td>
<td></td>
<td>18343006</td>
</tr>
<tr>
<td>Deficient abstract thinking</td>
<td>Concrete</td>
<td></td>
<td></td>
<td>71573006</td>
</tr>
<tr>
<td>Delusions</td>
<td>delusion*</td>
<td></td>
<td></td>
<td>2073000</td>
</tr>
<tr>
<td>Derailment of speech</td>
<td>derail*</td>
<td></td>
<td></td>
<td>65135009</td>
</tr>
<tr>
<td>Diminished eye contact</td>
<td>eye contact</td>
<td></td>
<td></td>
<td>412786000</td>
</tr>
</tbody>
</table>
Information Extraction

- **TextHunter**
  - Built around ConText algorithm* and GATE framework
    - Matching keywords using regular expression
    - Providing annotation interface
    - Construct SVM model for the concept and evaluate
  - Uses bag-of-word features and knowledge engineering features from ConText

* The ConText algorithm provides context - whether the event occurred (Negation: affirmed or negated), who experienced it (Experiencer: patient or other), and when it occurred (Temporality: historical, recent, not particular) - for a given event from a sentence. [BioNLP Workshop of the Association for Computational Linguistics; June 29, 2007]
Performance Comparison

• Annotated 50 symptoms with 37211 instances (Cohen’s $\kappa$ of 0.83)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Model</th>
<th>P%</th>
<th>R%</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>ConText + ML</td>
<td>83</td>
<td>78</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>ConText</td>
<td>71</td>
<td>97</td>
<td>0.79</td>
</tr>
<tr>
<td>Median</td>
<td>ConText + ML</td>
<td>90</td>
<td>85</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>ConText</td>
<td>84</td>
<td>98</td>
<td>0.91</td>
</tr>
</tbody>
</table>

SMI, severe mental illness.
2.2. Deep Neural Network for Phenotyping Youth Depression

• Background
  – EHR analysis can support recruitment in clinical research
  – Diagnosis codes are frequently missing
  – NLP can detect features in clinical notes and outperformed features by experts
  – NLP outperformed diagnosis for classifying mood state (ROC: 0.85–0.88 vs 0.54–0.55)

• Objective
  – To identify individuals who meet inclusion criteria as well as unsuitable patients who would require exclusion

Evidence-based mental health 2017. DOI: 10.1136/eb-2017-102688
• Phenotype of youth depression
  – Inclusion: Ages 12-18 with DSM defined Major Depressive Disorder or Dysthymic Disorder
  – Exclusion: schizophrenia, bipolar disorder, autism, epilepsy, personality disorder, developmental delay and traumatic brain injury

• Data:
  – 366 patients with 861 physician documents
Dictionary-based Method

• Brute force
  – Positive dictionary (inclusion)
  – Negative dictionary (exclusion)

**Box 1**  Positive dictionary: a dictionary of terms to help identify depression
- Major depressive disorder
- Major depression
- Double depression
- Dysthymic disorder
- Persistent depressive disorder
- Depressive disorder
- Depression
- MDD

**Box 2**  Negative dictionary: terms that would indicate that someone is not suitable
- Bipolar disorder
- Schizophrenia
- Bipolar II
- Bipolar I
- Traumatic brain injury
- Developmental delay
- Personality disorder
- Borderline personality disorder
- Hypomanic
- Autism
- Epilepsy
Deep Neural Network

- Training: 748 docs, 101 suitable and 657 unsuitable pts
- Test: 103 docs, 25 suitable, 78 unsuitable pts
- Implemented in H2O.ai R package
- Two models (DL0 and DL1)
- Construct an aggregate predictor (DL1+0)

Figure 1  The more sensitive DL1 method was initially applied. Following DL1, the more specific DL0 model was then used on the documents selected with DL1. DL, deep learning paradigm.
Performance Comparison

- Demonstrate the potential for this approach for patient recruitment purposes
- A larger sample size is required to build a truly reliable recommendation system
Challenges from NLP Perspective

• Large and labeled datasets are not available for many NLP methods (e.g., neural network)
• Evaluation is still performed based on intrinsic criteria, not for a specific clinical problem
  – Timely detection of suicidal behavior risk
    • Suicidal behavior is relatively rare (low precision)
    • Ensure an appropriate sample to provide interpretable NLP output
• NLP tasks are more complex
  – From simply NER to ascertain novel and complex entities (makers of socioeconomic status, life experience)
  – From single institution to a multi-site application
Questions?

Rui Zhang, Ph.D.
Email: zhan1386@umn.edu
Research Lab: http://ruizhang.umn.edu/
Clinical Information Extraction

Sunghwan Sohn, PhD
Division of Digital Health Sciences, Mayo Clinic
Learning Objects

1) Understand challenges of EHRs

2) Know clinical information extraction
   - Methodology review (high-level)

3) Explore clinical documentation variations and IE-based NLP tool portability
   - Case study of NLP tool portability (asthma ascertainment)
Electronic Health Records

Structured Data
- Pharmacy
- Pathology
- Documentation

Unstructured Data
- ~80% of EHRs
- Often ungrammatical/fragment of text, use of abbreviations
- Clinical notes, radiology reports, operation notes, etc.

Retrieve using query
Extract using NLP
Challenges of EHR

• Volume
  • Much of EHR is free text
  • Requires natural language processing

Clinical IE

• Variability
  • Clinical practice and workflow vary across institutions
  • Clinical language is not homogenous

Clinical documentation variation

• Portability
  • Out-of-box NLP models don’t work well

NLP system portability
NLP can facilitate the extraction and mining of text for structured information and knowledge.

Unstructured Text

Information retrieval
Information extraction
Automated chart review
Knowledge discovery
Clinical decision support

Structured Content

Image courtesy of National Institutes of Health
Methodological Review

Clinical information extraction applications: A literature review

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarrad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng\textsuperscript{1}, Saeed Mehrabi\textsuperscript{2}, Sunghwan Sohn, Hongfang Liu\textsuperscript{*}

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

\section*{ARTICLE INFO}

\textbf{Keywords:} Information extraction, Natural language processing, Application, Clinical notes, Electronic health records

\section*{ABSTRACT}

\textbf{Background:} With the rapid adoption of electronic health records (EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems at the point of care and to enable secondary use of EHRs for clinical and translational research. One critical component used to facilitate the secondary use of EHR data is the information extraction (IE) task, which automatically extracts and encodes clinical information from text.

\textbf{Objectives:} In this literature review, we present a review of recent published research on clinical information extraction (IE) applications.

\textbf{Methods:} A literature search was conducted for articles published from January 2009 to September 2016 based on Ovid MEDLINE In-Process & Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, Web of Science, and ACM Digital Library.

\textbf{Results:} A total of 1917 publications were identified for title and abstract screening. Of these publications, 263 articles were selected and discussed in this review in terms of publication venues and data sources, clinical IE tools, methods, and applications in the areas of disease- and drug-related studies, and clinical workflow optimizations.

\textbf{Conclusions:} Clinical IE has been used for a wide range of applications, however, there is a considerable gap between clinical studies using EHR data and studies using clinical IE. This study enabled us to gain a more concrete understanding of the gap and to provide potential solutions to bridge this gap.
Clinical IE
- methodologies

1) Dictionary lookup
2) Rule-based / expert system
3) Machine learning
4) Deep learning
Dictionary Lookup

- Map medical text to the concepts in dictionary
- Dictionary resources
  - Existing: UMLS, SNOMED-CT, RxNorm, etc.
  - Custom-built
- Predefined concepts (by medical experts)
- Can follow the standard when using (inter)national resources
  - Enable interoperability between computer systems
- Tools: MetaMap, cTAKES, MedTagger, MedXN
Rule-based (Expert System AI)

- Regular expression and rules
  - Flexible handling string and pattern variations
- Suitable to implement existing criteria, expert logics
- Interpretable, customizable
- Labor intensive
- Tools: UIMA Ruta, MedTaggerIE

Chemical threats diagnosis expert system (CTDES) in Advanced Materials Research · November 2012
Machine Learning AI

- Suitable for problems with no explicit criteria
- Require feature engineering
- Not interpretable

- Applications
  - named entity recognition, adverse drug reaction, disease prediction, relation extraction

- Popular techniques:
  - SVM, CRF, decision tree, Naïve Bayes, random forest

IE architecture & BIO tagging

https://www.nltk.org
Deep Learning

- No feature engineering
- Popular techniques
  - RNN (based on LSTM), CNN
- Applications
  - named entity recognition, relation extraction
- DL Information extraction: Words a sequence of tokens - embedding layer in RNN - softmax to classify token’s entity

Effective Use of Bidirectional Language Modeling for Transfer Learning in Biomedical Named Entity Recognition, arXiv:1711.07908
Deep Learning

- Require large training set
- Transfer learning to overcome the burden of large training data
  - use pre-train the model’s weights for the main task

use language modeling (on PubMed abstracts) as a transfer learning approach to pretrain the NER model’s weights.

Bidirectional Recurrent Neural Networks for Medical Event Detection in Electronic Health Records, arXiv:1606.07953
Right approach?

Not about selecting fancy technology, but about understanding the strengths / weaknesses and the nature of your project.

- **Rule based**
  - Input
  - Rule
  - Output

- **ML based**
  - Input
  - Output
  - Rule
Right approach? Cont.

Data perspective
- Variance (pattern)
- Dict. lookup
- Rules
- Machine learning
- Deep learning

User perspective
- Interpretable (outcome)
- Deep learning
- Machine learning
- Rules
- Dict. lookup

Customizable (logic)

Hybrid
Clinical Documentation Variations & NLP System Portability

- The performance of a NLP system often varies across institutions and sources of data.
- Whenever an NLP system developed in one corpus is applied to another corpus, there are questions:
  - How similar are these two corpora?
  - If two corpora differ “how does the difference affect the NLP system portability?”
What was known/not known
- NLP system portability

• Known
  • Validity of portability by comparing the system performance

• Lacked
  • A systematic analysis of the heterogeneous EHR corpus (clinical documentation variations)

• Here
  • Types of clinical documentation variations
  • How they affect NLP system portability
Variations of Clinical Documentation

• Process variation
  • Due to various clinical practice and workflow across institutions
  • Eg) data format, section, note type

• Syntactic (lexical) variation
  • Clinical language is not homogeneous
  • Eg) different words/concepts in cardiology, orthopedics, ophthalmology

• Semantic variation
  • Concept representation
  • Eg) Asthma, destructive airway disease
Similarity Measure

- Create a “vector space model”
- Calculate “cosine similarity”

1) Corpus similarity
2) Medical concept similarity
3) Note type similarity
Corpus Similarity

• The entire corpus of each institution was compared as a whole using
  • tf-idf: each corpus was represented by a normalized $tf-idf$ vector
  • tf-ipf: word distribution at a patient level
• Topic: compare corpora by topic (LDA)
  • The topic $z_k$ for the corpus $C$ is defined as
    \[
    p(z_k|C) = \sum_{d_i \in C} p(z_k|d_i, C)p(d_i|C) = \sum_{d_i \in C} \frac{p(z_k|d_i)}{N}
    \]
Medical Concept Similarity

• A vector representation of medical concepts for each corpus was created using the definition of
  • $cf-idf$ (concept frequency-inverse document frequency)
  • $cf-ipf$ (concept frequency-inverse patient frequency)
Note Type Similarity

• Clinical documents have various note types based on the event
  • e.g., admission, discharge, progression

• Among institutions
  • May have different note types
  • Same note type may contain heterogeneous topics

• Compare topic distributions of note types
  ➢ the topic $z_k$ for the clinical note type $T$ is defined by
  $$p(z_k | T) = \sum_{d_i \in T} \frac{p(z_k | d_i)}{N_T}$$
  where $N_T$ is the number of documents in the note type $T$
Case Study

- Between Mayo Clinic and Sanford Children’s Hospital (SCH)
  - clinical documentation variations
  - performance of the NLP asthma ascertainment system

- EHR
  - Birth cohort
  - Mayo* GE-based vs. SCH EPIC

*changed to EPIC in 2018
Similarities between Mayo and SCH

<table>
<thead>
<tr>
<th>Data source</th>
<th>tf-idf</th>
<th>tf-ipf</th>
<th>topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole corpus</td>
<td>0.669</td>
<td>0.581</td>
<td>0.944</td>
</tr>
<tr>
<td>Asthma-related concepts</td>
<td>0.971</td>
<td>0.855</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Message**

✓ (Word level) Even though clinicians have heterogeneous clinical language that shows up in different EHR systems,
✓ (Concept level) Clinicians share common semantics to describe asthma episodes/events
A heat map of note type similarity (based on topics)

SCH “Telephone Encounter” <-> Mayo “Test MIS,” “Supervisory,” and “Miscellaneous”

SCH “Progress Note” <-> Mayo “Test MIS,” “Supervisory,” and “Limited Exam”

Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J Am Med Inform Assoc. Published online November 30, 2017.
NLP Tool - Asthma Ascertainment

- Implements the predetermined asthma criteria (NLP-PAC)
  - based on presence/absence of asthma-related concepts

Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions, J Am Med Inform Assoc. Published online November 30, 2017. doi:10.1093/jamia/ocx138
NLP-PAC

- Expert rule-based system
- Implemented into the MedTaggerIE
  - open source IE framework built under Apache UIMA

MedTagger

OpenNLP components In cTAKES

MedTagger (Torii et al)
SecTag (Denny et al)
MedTagger (Torii et al)
MedTagger (Torii et al)
ConText (Chapman et al)

Collection Readers
- SentenceAnnotator
- POS Tagger*
- ChunkAnnotator*
- SectionAnnotator
- InformationExtractor
- ContextAnnotator

CAS Consumers

* Indicates optional components

Separates domain-specific NLP knowledge engineering from the generic NLP process.

IE Engine – knowledge driven

Normalization Target
- normDISO
- normDISOdesc
- occlusion~|~occlusion
- occluded~|~occlusion
- oculates~|~occlusion
- occlusive~|~occlusion

Regular Expression
- reDISO
- occlusion
- occlude[ds]
- oculative

Match Rules
- RULENAME="cm_r1a"
- REGEXP="\b(\%reDISOdesc )?\(\%reDISO\)\b"
- LOCATION="SEC:~diagnosis~impression~", NORM="\%DISO:\%normDISOdesc(group(2))_\%normDISO(group(4))"

An information extraction framework for cohort identification using electronic health records, AMIA Summits on Translational Science, 2013
MedTaggerIE
NLP-PAC portability to SCH

1) Prototype NLP-PAC (stage 1)
   - required adjustments to be able to run the Mayo NLP-PAC system on the SCH cohort
   - deal with process variations
     eg) sentence parsing, section segmentation

2) Refined NLP-PAC (stage 2)
   - further reduces process variations and refine the algorithm
     eg) note type to be excluded, assertion adjustment
## NLP-PAC performance for asthma ascertainment

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Mayo (N=497)</th>
<th>SCH stage 1 (prototype, N=298)</th>
<th>SCH stage 2 (refinement, N=298)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sensitivity</td>
<td>0.972</td>
<td>0.840</td>
<td>0.920</td>
</tr>
<tr>
<td>specificity</td>
<td>0.957</td>
<td>0.924</td>
<td>0.964</td>
</tr>
<tr>
<td>PPV</td>
<td>0.905</td>
<td>0.788</td>
<td>0.896</td>
</tr>
<tr>
<td>NPV</td>
<td>0.988</td>
<td>0.945</td>
<td>0.973</td>
</tr>
<tr>
<td>F-score</td>
<td>0.937</td>
<td>0.813</td>
<td>0.908</td>
</tr>
</tbody>
</table>
NLP system portability

✓ Understand clinical documentation variations

✓ Out-of-box system produces considerably lower performance
  ▪ deal with process variations to be technically operable

✓ Further refined system produced comparable performance (eg, negation sublanguage)
Summary

- Right approach of clinical IE (Volume)
  - need to understand strengths/weaknesses and nature of the problem

- EHR data are different among institutions (Variability)
  - exist various types of clinical documentation variations

- Document variations play different roles in assessing the NLP system application (Portability)
  - Need systematic adjustments to deal with the data heterogeneity and improve performance
Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions

Sunghwan Sohn ☑, Yanshan Wang, Chung-Il Wi, Elizabeth A Krusemark, Euijung Ryu, Mir H Ali, Young J Juhn, Hongfang Liu

Patient Cohort Retrieval using Electronic Health Records

Yanshan Wang
Research Associate
Department of Health Sciences Research
Mayo Clinic
Agenda

• Introduction

• Basic Concepts
  • EHR, Phenotyping, Evidence-based Clinical Research, Knowledge Base, Common Data Model

• Patient Cohort Retrieval
  • NLP Approaches for Cohort Retrieval
    • Medical Concept Embedding
    • Information Retrieval
    • Deep Patient Representation
  • Case Study: clinical trials eligibility screening for gastroesophageal reflux disease (GERD)
Introduction

• What do we do?

Computer scientist & Informatician

• How can you contact me?
  • Email: wang.yanshan@mayo.edu
  • LinkedIn, Twitter (yanshan_wang)
Why take this tutorial?

• Patient cohort retrieval is still labor expensive today.

• Most information is embedded in unstructured EHRs.

• Natural language processing is under-utilized for cohort retrieval.
Goal of this tutorial

• To get an understanding of basic concepts about cohort retrieval in clinical domain.

• To connect NLP theory with clinical knowledge.

• To get an introduction into clinical use cases of cohort retrieval.
Suggested reading

• Books

[Images of two books]
Suggested reading

• Papers
  • Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. Miotto et al. 2015.
  • A survey of practices for the use of electronic health records to support research recruitment. Obeid et al. 2017.
  • Clinical information extraction applications: a literature review. Wang et al. 2018
  • Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. Velupillai et al. 2018.
Basic Concepts

• Electronic Health Record
• Phenotyping
• Evidence-based clinical research
• Knowledge bases
• Common Data Model
Basic Concepts

- Electronic Health Record
Basic Concepts

• **Phenotyping**
  • The phenotype (as opposed to genotype, which is the set of genes in our DNA responsible for a particular trait) is the physical expression, or characteristics, of that trait.
  • Phenotyping is the practice of developing algorithms designed to identify specific phenomic traits within an individual[^1].

• **Digital phenotyping using EHRs**
  • Traditionally, clinical studies often use self-report questionnaires or clinical staff to obtain phenotypes from patients. (slow, expensive, could not scale).
  • EHR data come in both structured and unstructured formats, and the use of both types of information can be essential for creating accurate phenotypes[^2].

---

[^1]: eMERGE network.
Basic Concepts

- **Phenotyping**
  - Phenotyping is the practice of developing algorithms designed to identify specific phenomic traits within an individual.

- **Digital Phenotyping using EHRs**
  - EHR data come in both structured and unstructured formats, and the use of both types of information can be essential.

Evidence-based clinical research

• **Observational studies**
  • Types of studies in epidemiology, such as the **cohort study** and the **case-control study**.
  • The investigators retrospectively assess associations between the treatments given to participants and their health status.

• **Randomized control trials**
  • Clinical trials are prospective biomedical or behavioral research studies on **human participants** that are designed to answer specific questions about biomedical or behavioral interventions including new treatments, such as novel vaccines, drugs, and medical devices.
Basic Concepts

• **Cohort/Eligibility Criteria**
  • Inclusion criteria
  • Exclusion criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>clinicaltrials.gov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusion Criteria:</td>
<td></td>
</tr>
<tr>
<td>• Alzheimer’s disease (CDR 0.5, 1, &amp; 2)</td>
<td></td>
</tr>
<tr>
<td>• Active study partner</td>
<td></td>
</tr>
<tr>
<td>• BMI &gt; 21</td>
<td></td>
</tr>
<tr>
<td>• English speaking</td>
<td></td>
</tr>
<tr>
<td>Exclusion Criteria:</td>
<td></td>
</tr>
<tr>
<td>• BMI &lt; 21</td>
<td></td>
</tr>
<tr>
<td>• Consume greater than 14 drinks of alcohol per week</td>
<td></td>
</tr>
<tr>
<td>• Insulin Dependent Diabetes Mellitus</td>
<td></td>
</tr>
<tr>
<td>• Diagnosis of active cancer</td>
<td></td>
</tr>
<tr>
<td>• Myocardial infarction or symptoms of coronary artery disease (e.g. angina) in last year</td>
<td></td>
</tr>
</tbody>
</table>

Basic Concepts

- **Knowledge Bases**
  - UMLS (Unified Medical Language System) (including the Metathesaurus, Semantic Network, the Specialist Lexicon)
    - Used as a knowledge base and resource for a lexicon. Metathesaurus provides the medical concept identifiers. Semantic Network specifies the semantic categories for the medical concepts.
  - SNOMED-CT
    - Standardized vocabulary of clinical terminology.
  - LOINC
    - Standardized vocabulary for identifying health measurements, observations, and documents.
  - MeSH
    - NLM controlled vocabulary thesaurus used for indexing articles for PubMed articles.
  - MedDRA
    - Terminologies specific to adverse event.
  - RxNorm
    - Terminologies specific to medications
Basic Concepts

• **Common Data Model**
  - Common Data Model (CDM) is a specification that describes how data from multiple sources (e.g., multiple EHR systems) can be combined. Many CDMs use a relational database.
  - Observational Medical Outcomes Partnership (OMOP) CDM by Observational Health Data Sciences and Informatics (OHDSI)
Patient Cohort Retrieval for Clinical Trials using NLP
Clinical Trials Eligibility Screening and Recruitment

• Clinical trials recruitment
  • Randomized clinical trials are fundamental to the advancement of medicine. However, patient recruitment for clinical trials remains the biggest barrier to clinical and translational research.

20%
Cancer patients are eligible¹

<5%
Cancer patients are participate¹

85%
Clinical trials fail to retain enough patients²

NLP for Clinical Trials Eligibility Screening

All patients

Clinical trials criteria

Natural Language Processing

EHR

Eligible patients

Recruit
NLP Approaches for Cohort Retrieval

• Medical Concept Embedding
• Information Retrieval
• Deep Patient Representation
Medical Concept Embedding

Cohort Criteria

EHRs

NLP Engine

Concept
Concept
Concept

Concept
Concept
Concept
Concept

Concept
Concept
Concept

Medical Concept Embedding

- Cohort Criteria
- EHRs
- NLP Engine
- Knowledge-based

Concepts

Concept
Concept
Concept
Concept

Concept
Concept
Concept
Medical Concept Extraction and Representation


Semantic Lexicon Extraction

- UMLS-based lexicon discovery from text
  - Retrieved 10,000 eligibility criteria sentences from clinicaltrials.gov.

Semantic Lexicon Extraction

- **UMLS-based lexicon discovery from text**
  - Processed the corpus to identify UMLS-recognizable semantic units that matched the medical concepts in the MRCONSO table of the Metathesaurus of UMLS.
  - Used the Stanford tagger and the Penn Treebank tag set for part-of-speech (POS) tagging. All words tagged as nouns, verbs, adjectives or adverbs were considered content words, which potentially had semantic types in the UMLS Semantic Network.
  - Used MRSTY table of the Metathesaurus and rules to find semantic terms.

Semantic Lexicon Extraction

- UMLS-based lexicon discovery from text
  - Investigated the coverage of the sample corpus provided by annotation procedure, using the Metathesaurus, Semantic Network, and preference rules.

Eligibility criteria extraction and representation using CDM

Limitations of Concept Embedding

- Accuracy of medical concept extraction
- Extensive annotation efforts
- Generalizability/Portability
Deep Patient Representation


Deep Patient: Overall Framework

EHRs are extracted from the clinical data warehouse and are aggregated by patient

Unsupervised deep feature learning to derive the patient representations

Predict patient future events from the deep representations
Deep Patient: Learning

- Multi-layer neural network
  - each layer of the network produces a higher-level representation of the observed patterns, based on the data it receives as input from the layer below, by optimizing a local unsupervised criterion

- Hierarchically combine the clinical descriptors into a more compact, non-redundant and unified representation through a sequence of non-linear transformations
Deep Patient: Data Processing

- Patients data available in the data warehouse

  - **Structured**
    - Diagnoses
    - Lab Tests
    - Medications
    - Procedures
  - **Unstructured**
    - Clinical Notes
  - **Demography**
    - Gender
    - Age
    - Race

- Normalize the clinically relevant phenotypes
  - group together the similar concepts in the same clinical category to reduce information dispersion

- Aggregate data by patients in a vector form
  - bag of phenotypes
Deep Patient: Architecture

- The first layer receives as input the EHR bag of phenotypes.
- Every intermediate level is fed with the output of the previous layer.
- The last layer outputs the Deep Patient representations.
Deep Patient Representation for Cohort Retrieval

Information Retrieval for Cohort Retrieval

IR Tool for Cohort Retrieval

User Interface
(visualization, analytics, reporting, etc.)

Collating Results

Unstructured data

Structured data
IR for Cohort Retrieval Prototype: CREATE

Cohort Retrieval Enhanced Analysis by Text from EHR (CREATE)

Structured EHR → Structured Data
Unstructured EHR → Clinical Texts

ICD9/10, CPT, SNOMED CT...

Transform

Structured Index → Filtered Cohort
Unstructured Index

Index

Structured Query
Unstructured Query

Machine Learning

End User

Full-text Query

“Adults with inflammatory bowel disease (ulcerative colitis or Crohn’s disease)”

Structured Data Flow
Unstructured Data Flow

Case Study: clinical trials eligibility screening for gastroesophageal reflux disease (GERD)

Identify a cohort of patients with and without chronic reflux using the definitions spelled out below. We wish to test people with and without chronic reflux as our working hypothesis is that the prevalence of Barrett's esophagus is comparable between those with and without chronic reflux.

**Inclusion criteria:**

1. Age greater than 50 years.
2. Gastroesophageal reflux disease. This can be defined using ICD 9 or ICD 10 codes. Additional criteria which could be used to define GERD broadly are chronic (> 3 mo) use of a proton pump inhibitor (drug names include omeprazole, esomeprazole, pantoprazole, rabeprazole, dexlansoprazole, lansoprazole) or a H2 receptor blocker (ranitidine, famotidine, cimetidine). Prior endoscopic diagnosis of erosive esophagitis can also be used to make a diagnosis of GERD.
3. Male gender
4. Obesity defined as body mass index greater than equal to 30. This is a surrogate marker for central obesity.
5. Current or previous history of smoking
6. Family history of esophageal adenocarcinoma/cancer or Barrett's esophagus

**Exclusion criteria**

1. Previous history of esophageal adenocarcinoma/cancer or Barrett's esophagus, previous history of endoscopic ablation for Barrett's esophagus.
2. Previous history of esophageal squamous cancer or squamous dysplasia.
3. Treatment with oral anticoagulation including warfarin/Coumadin.
4. History of cirrhosis or esophageal varices
5. History of Barrett's esophagus: this can be defined with ICD 9/10 codes.
6. History of endoscopy (will need to use a procedure code for EGD) in the last 5 years.
**i2b2 (informatics for integrating biology and bedside)**

- Using ontology knowledge bases to represent EHR data.
- Standardized EHR data designed for multi-site research and population health research.
<table>
<thead>
<tr>
<th>Criteria</th>
<th>ICD 9</th>
<th>ICD 10</th>
<th>CPT 4</th>
<th>Medication</th>
<th>Addressed by I2B2</th>
<th>Addressed by ACE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Age greater than 50 years.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2. Gastroesophageal reflux disease (any of 2.1, 2.2, 2.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Gastroesophageal reflux disease defined by Dx</td>
<td>530.81</td>
<td>K21.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2 Gastroesophageal reflux disease defined by drug, duration of use &gt;= 3 months over the last 5 years</td>
<td>530.19</td>
<td>K21.0</td>
<td></td>
<td>omeprazole, esomeprazole, pantoprazole, rabeprazole, dexlansoprazole, lansoprazole, ranitidine, famotidine, cimetidine</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2.3 Gastroesophageal reflux disease defined by prior endoscopic diagnosis of erosive esophagitis</td>
<td>530.19</td>
<td>K21.0</td>
<td></td>
<td>Not able to find specific code for esophagitis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Male gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>4. Obesity defined as body mass index greater than equal to 30.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>5. Current or previous history of smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
<td>Partially</td>
</tr>
<tr>
<td>6. Family history of esophageal adenocarcinoma/cancer or Barrett’s esophagus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
<td>Partially</td>
</tr>
<tr>
<td>7. Caucasian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td><strong>Exclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Previous history of esophageal adenocarcinoma/cancer</td>
<td>150.9</td>
<td>C15.9</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2. Previous history of endoscopic ablation for Barrett’s esophagus.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>3. Previous history of esophageal squamous carcinoma (included in 1)</td>
<td>150.9</td>
<td>C15.9</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>4. Previous history of esophageal squamous dysplasia</td>
<td>622.10</td>
<td>N87.9</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>5. Current Treatment (drug) with oral anticoagulation - warfarin</td>
<td></td>
<td></td>
<td></td>
<td>warfarin</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>6. Current Treatment (drug) with oral anticoagulation - Coumadin. (included in 5)</td>
<td></td>
<td></td>
<td></td>
<td>Coumadin</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>7. History of cirrhosis</td>
<td>571.5</td>
<td>K74.60</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>8. History of esophageal varices</td>
<td>456.20</td>
<td>I85.00</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>9. History of Barrett’s esophagus</td>
<td>530.85</td>
<td>K22.7</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>10. History of endoscopy in the last 5 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43235-43270</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Screening patients by Inclusion criteria 1, 3, 4, 7 and all Exclusion criteria using i2b2. Get patient set A (n=31749)

From patient set A, screening patients by Inclusion criteria 2.1 using i2b2. Get patient set B (n=8667)

From patient set A, screening patients by Inclusion criteria 2.2 using i2b2. Get patient set C (n=1577)

From patient set A, screening patients by Inclusion criteria 2.3, 5, 6 using ACE and CREATE. Get patient set D (n=230)

Union patient sets B, C, and D. Get patient set E (n=9080)
Thank you!
Clinical NLP: Challenges and Opportunities
Data!!
Data!!

Sorry
No Data Available

HIPPA
PHI
HIPAA: the Health Insurance Portability & Accountability Act of 1996 public law

• To ensure the privacy of Americans’ personal health records by protecting the security and confidentiality of health care information – an Individual’s Protected Health Information (PHI).
Data!!

**PHI**

- Name
- Postal addresses
- All elements of dates except year
- Telephone number
- Fax number
- Email address
- URL address
- IP addresses
- Social security number
- Account numbers
- License numbers

- Medical record number
- Health plan beneficiary number
- Device identifiers and their serial numbers
- Vehicle identifiers and serial number
- Biometric identifiers (finger and voice prints)
- Full face photos and other comparable images
- Any other unique identifying number, code, or characteristic
Is it true?
Public Corpora

• **Challenges Data**
  • i2b2 NLP Challenges data
  • OHNLP Challenge data
  • TREC 2011 and 2012 Medical Records track

• **MIMIC II**
  • > 40,000 de-identified intensive care unit stays

• **Mtsamples**
  • publicly available transcribed medical reports

• **THYME corpus**
  • de-identified clinical, pathology, and radiology records

Public Corpora

• **Challenges Data**
  - i2b2 NLP Challenges data
  - TREC 2011 and 2012 Medical Records track

• **MIMIC II**
  - > 40,000 de-identified intensive care unit stays

• **Mtsamples**
  - publicly available transcribed medical reports

• **THYME corpus**
  - de-identified clinical, pathology, and radiology records

Interpretation of Machine/Deep Learning

Why false positive??
Why false negative??

Input → BLACK BOX → Output
NLP is Still Under-utilized

The number of natural language processing (NLP)-related articles compared to the number of electronic health record (EHR) PubMed articles from 2002 through 2015.

Thank you!

- Yanshan Wang
  - wang.yanshan@mayo.edu

- Ahmad P. Tafti
  - tafti.ahmad@mayo.edu

- Sunghwan Sohn
  - Sohn.sughwan@mayo.edu

- Rui Zhang
  - zhan1386@umn.edu