MAYO CLINIC

Tutorial 6:

Applications of Natural Language Processing in Clinical Research and Practice

Yanshan Wang[1], Ahmad P. Tafti[1], Sunghwan Sohn[1], Rui Zhang[2]

[1]Department of Health Sciences Research, Mayo Clinic

[2]Department of Pharmaceutical Care & Health Systems, University of Minnesota

MAYO CLINIC

Overview of Natural Language Processing in Clinical Domain

HITECH Act law 2009

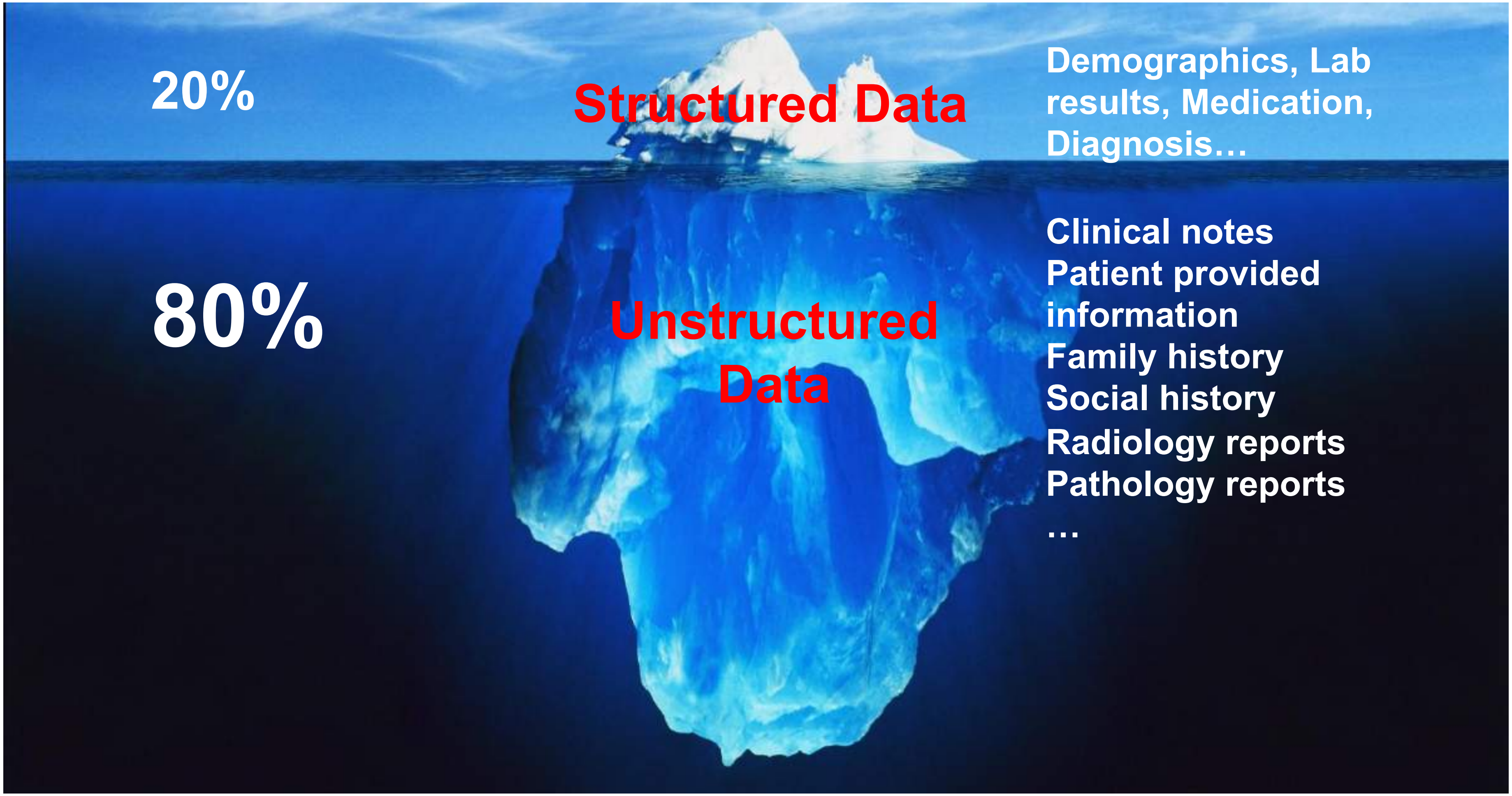
Data capture and sharing

Advanced clinical processes

Improved Outcome

MAYO CLINIC

Motivation for Clinical NLP



MAYO CLINIC

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

MAYO CLINIC

Speakers and Topics

MAYO CLINIC

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, JMIR, 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**

MAYO CLINIC

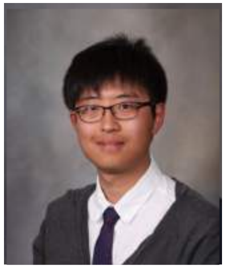
Clinical Information Extraction

Sunghwan 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)

MAYO CLINIC

Patient Cohort Retrieval using EHRs



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

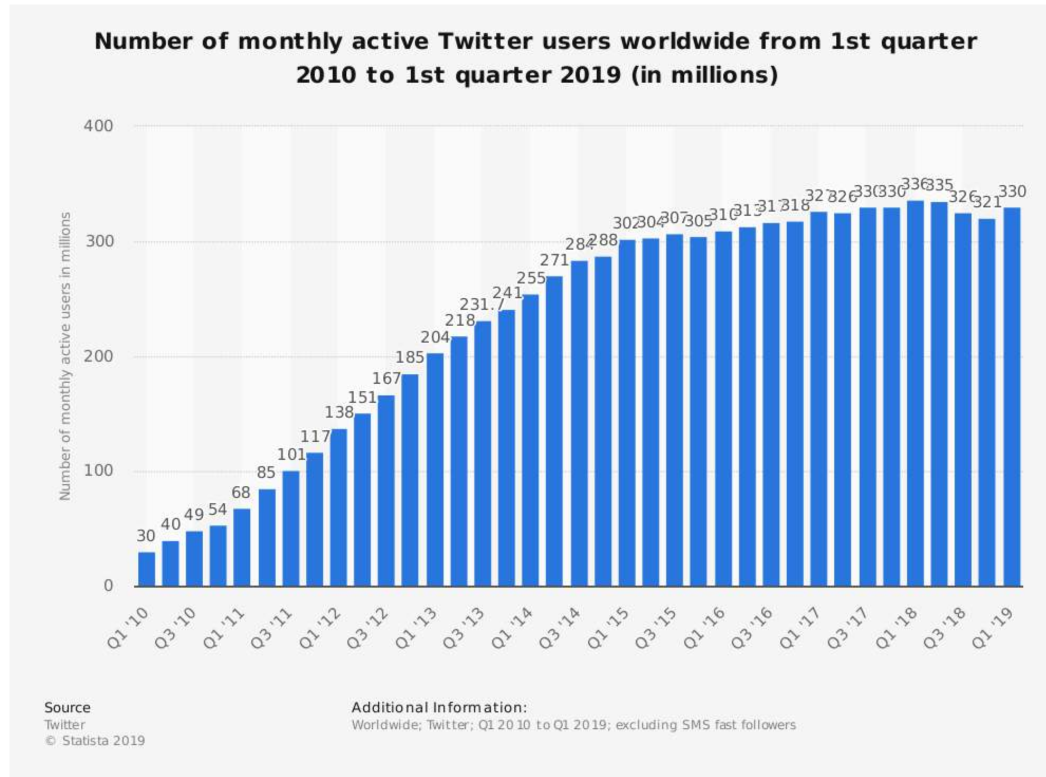
[**Case Study:** Gender Disparity in Side Effects Reporting of Chronic Pain Medications]

Ahmad Tafti, PhD

Division of Digital Health Sciences
Mayo Clinic

MAYO CLINIC

Big Data Social Media

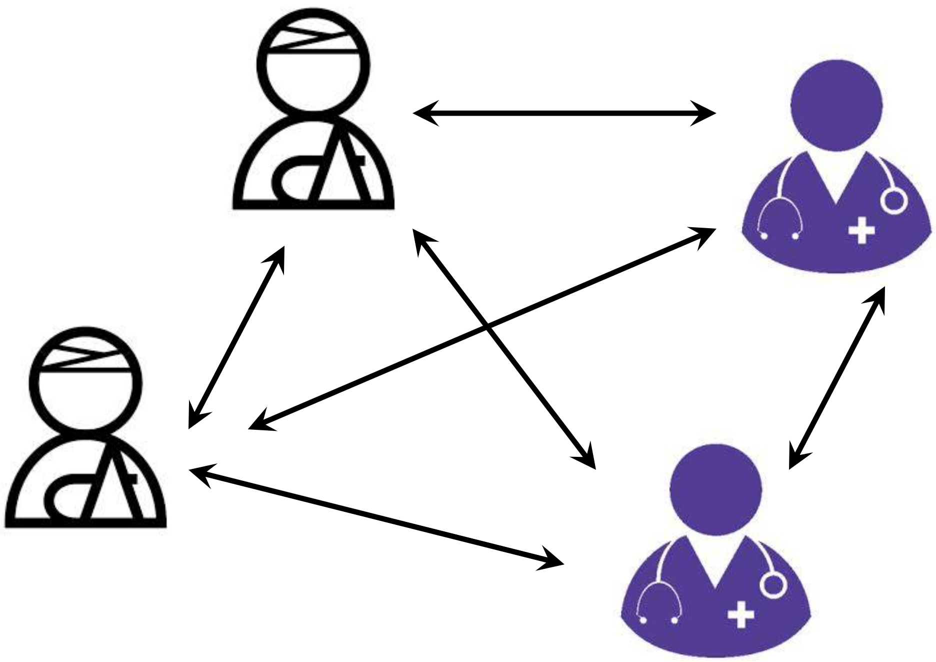


https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

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.

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

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?

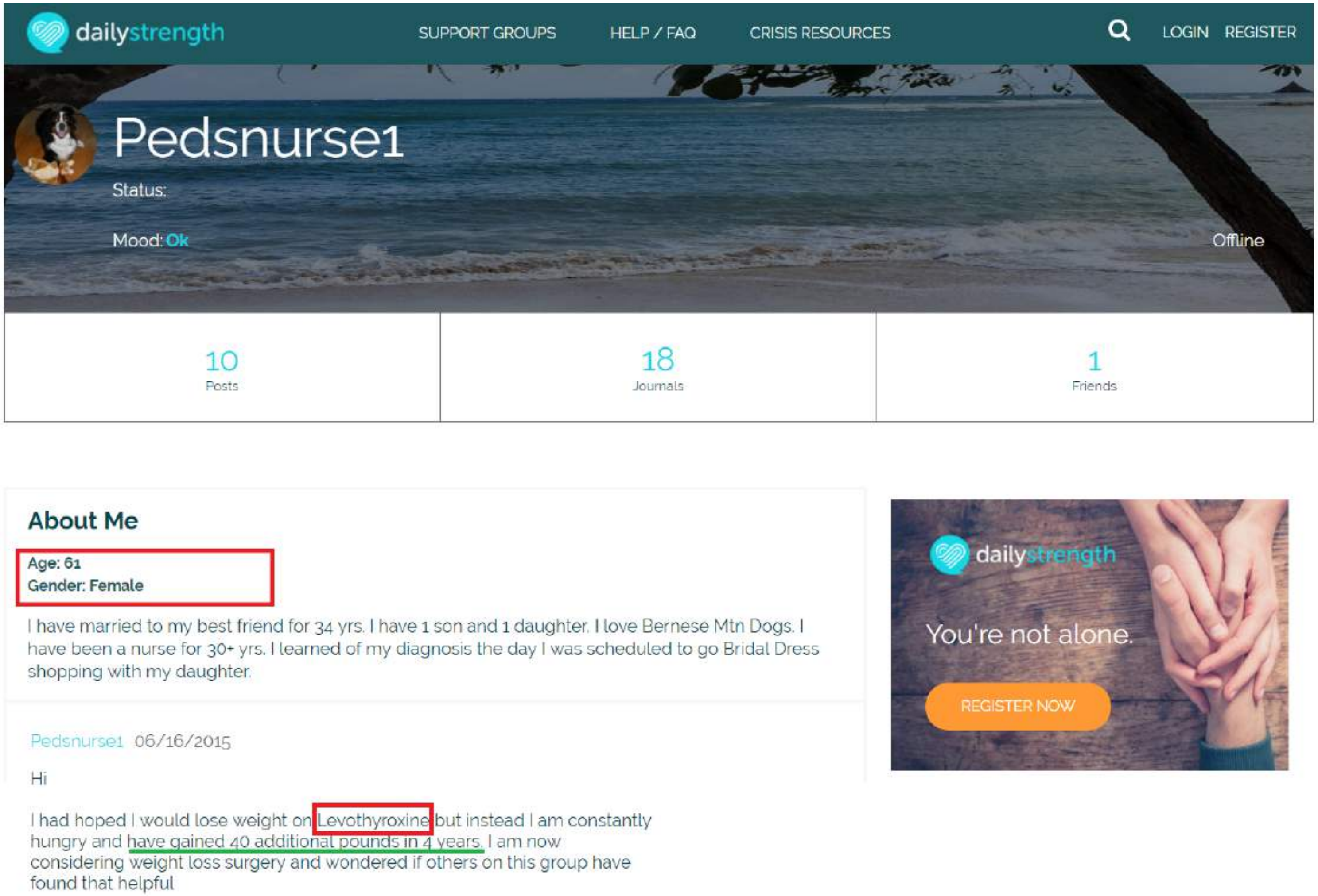
Advices and Guidelines on How to Use This Wealth Of Data

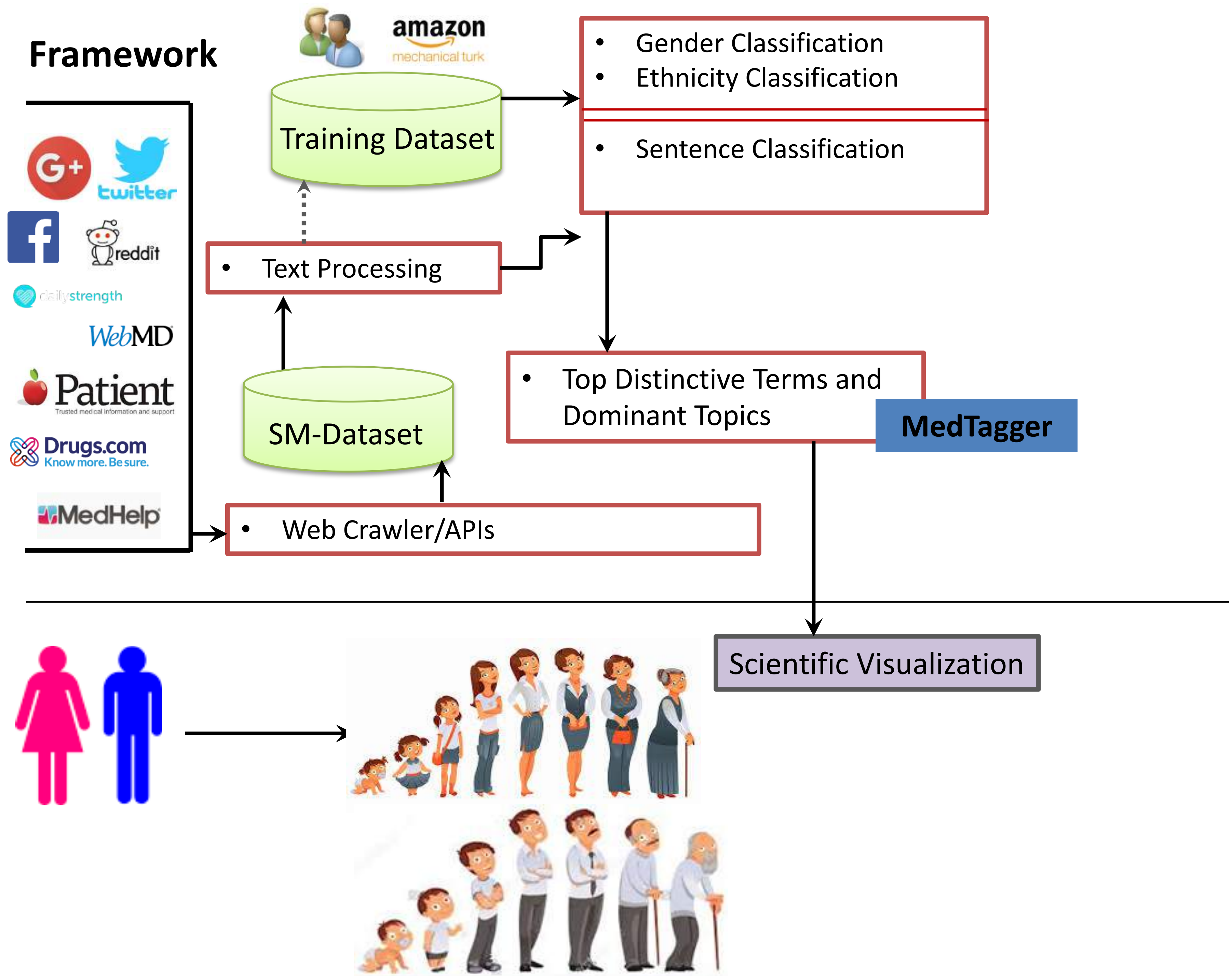
- Social Media Guidelines Issued by Health Care Institutions
- IRBs Protocols
- License agreements of the social media

Qualitative Analysis of Social Media

Source	Gender Availability	Age Availability	Ethnicity Availability	Location Availability
Twitter	No	No	No	Yes
Google+	Yes	Yes	No	Yes
Drugs.com	No	No	No	No
DailyStrength	Yes	Yes	No	Yes
WebMD	Yes	Yes	No	No
MedHelp	No	No	No	No
Patient.info	Yes	No	No	No

Qualitative Analysis of Social Media




Framework
amazon
mechanical turk
Training Dataset
Gender Classification
Ethnicity Classification
Sentence Classification
twitter
reddit
dailystrength
WebMD
Patient
Trusted medical information and support
Drugs.com
Know more. Be sure.
MedHelp
Text Processing
SM-Dataset
Top Distinctive Terms and Dominant Topics
MedTagger
Web Crawler/APIs
Scientific Visualization

Gender Classification

Gender API: https://gender-api.com
gender-guesser: https://pypi.org/project/gender-guesser
genderize.io: https://genderize.io
NameAPI: https://www.nameapi.org
NamSor: 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**

An Example: Good Reviews

"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 its completely worth it because it provides amazing pain relief, and its 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."

1.0

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!!!"

1.0

Col. Davis (taken for 2 to 5 years) **December 20, 2017**

Sentence Classification: ADEs vs Non-ADEs

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

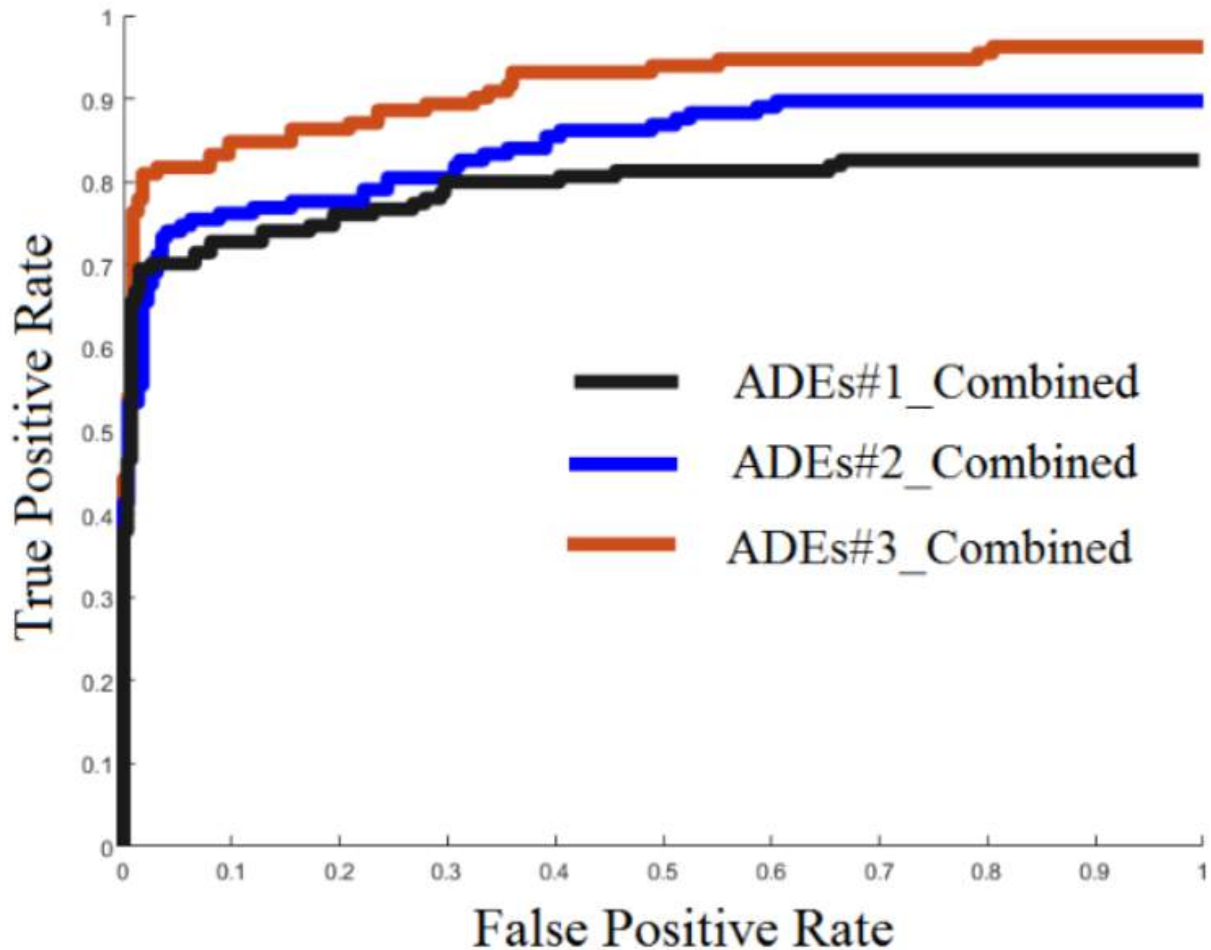
[a]ADEs: adverse drug events.

[b]bigNN: big data neutral network.

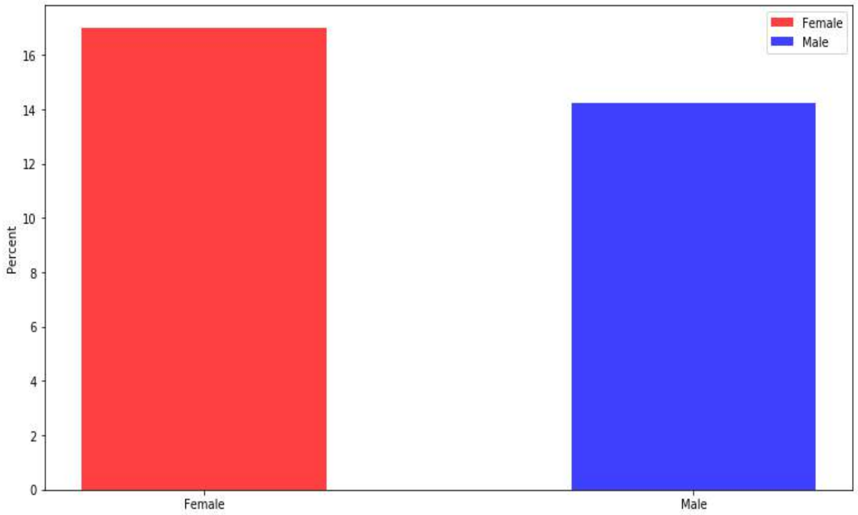
[c]BoW: bag-of-words.

[d]SVM: support vector machine.

Sentence Classification: ADEs vs Non-ADEs



Visualization Results: An Example



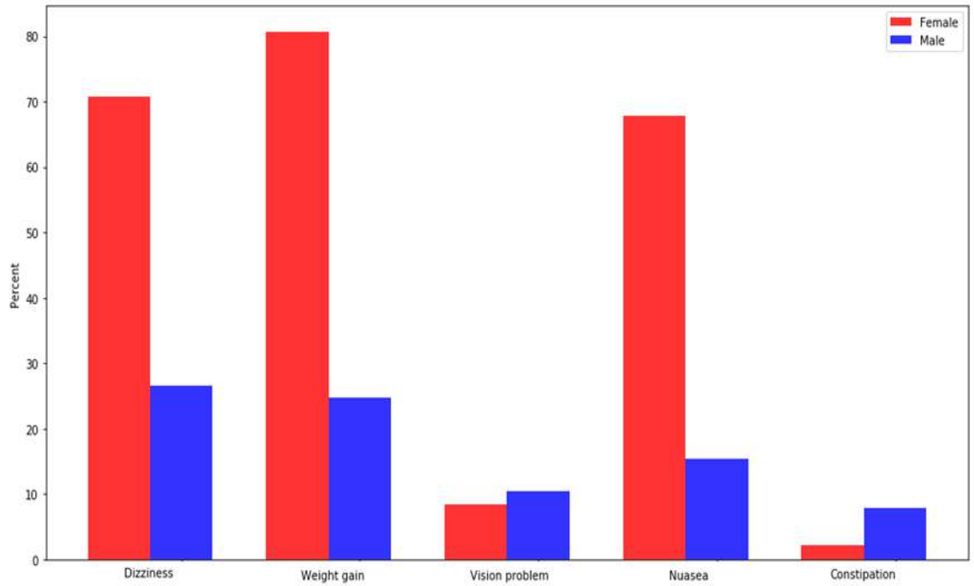


Figure 1. (Left) Women were more likely than men to report side effects from gabapentin. Women: 239 posts out of 1407 associated with gabapentin (including side effects, indication, and/or other topics)Men: it was 281 out of 1,973 posts.

(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.

Visualization Results: An Example

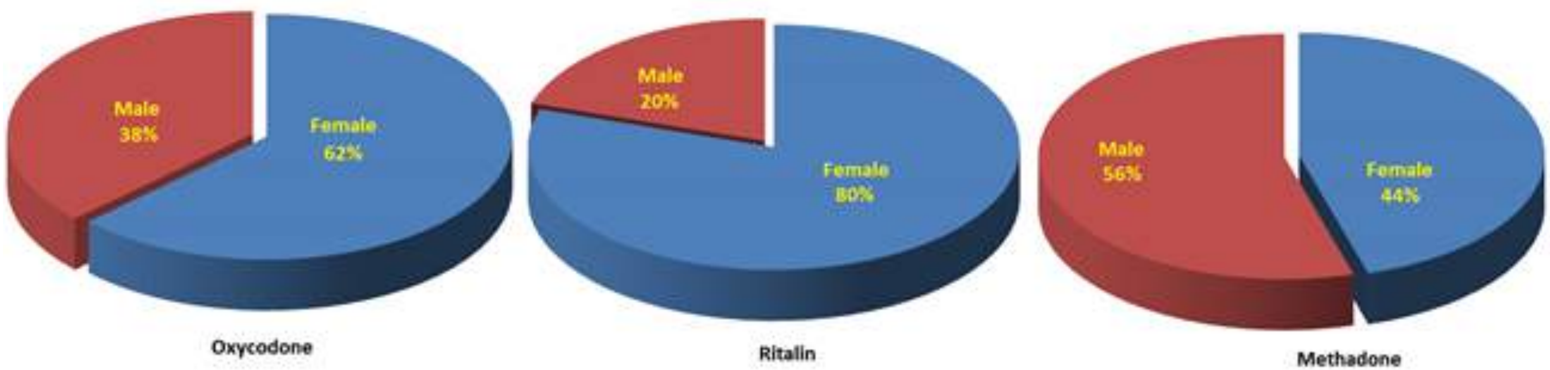


Figure 2. Percentwise proportion of women and men who did shared their pain-related experiences within three opioid class medications in Twitter, within last 30 days. One can see the number of Oxycodone and Ritalin tweets generated by women is greater than those generated by men. For Methadone, it shows 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

University of Minnesota
College of Pharmacy

University of Minnesota
Institute for Health Informatics

University of Minnesota
Clinical and Translational Science Institute

Outline

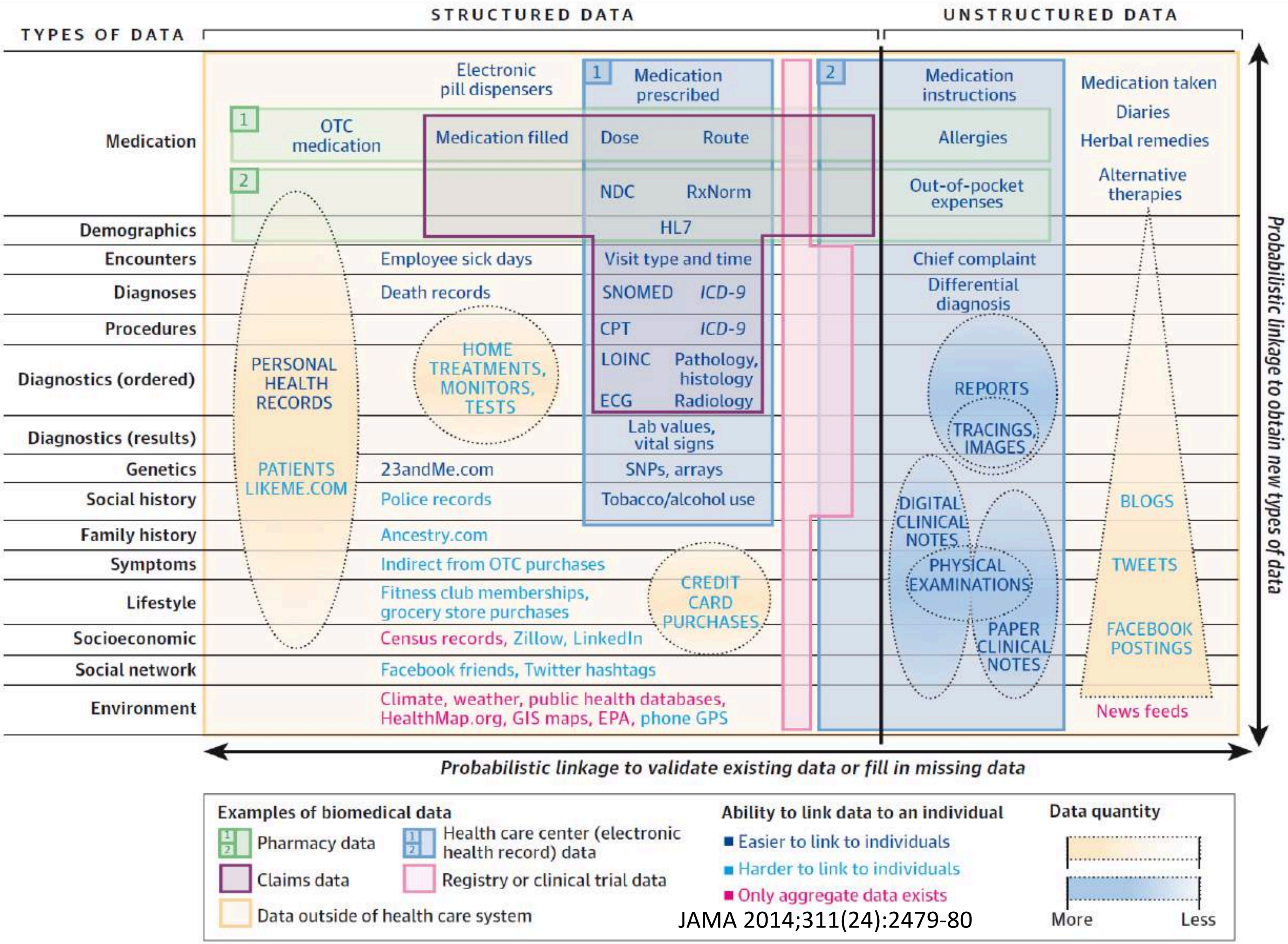
- Background of NLP to Support Clinical Research
- NLP Systems and Tools for Clinical Research
- Use Case 1: NLP to Support Dietary Supplement Safety Research
- Use Case 2: NLP to Support Mental Health Research

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



JAMA 2014;311(24):2479-80

Structured vs. Unstructured Data

Relational databases

NLP Tools Required

Diagnosis codes

Fake ID	ENTRY_DAT	CODE
34068	5/13/2001	41.85
37660	8/6/2002	79.99
140680	8/31/2003	79.99
23315	5/14/2003	112
75936	7/9/2004	117.9

Lab tests

Fake ID	TEST	ENTRY_DAT	VALU
3536	pO2	1/23/1996	314
72921	LDL	2/5/1996	34
102460	pCO2	1/26/1996	45
135043	HDL	1/25/1996	35
135432	MonAb	1/24/1999	0.16

Structured

Problem lists:

---- Medications known to be prescribed:
Keppra 750 mg 1/2 tab q am and pm
Dexilant 60 mg by mouth daily
aspirin 325 mg 1 tablet by mouth daily
clopidogrel 75 mg tablet 1 tablet by mouth daily

---- Known adverse and allergic drug reactions:
Sulfa Drugs

---- known significant medical diagnoses:
Seizure disorder
Aneurysm
Heartburn

---- Known significant operative and invasive procedures:
2003 Appendectomy
2005 Stents put in **DATE [Aug 29 05]

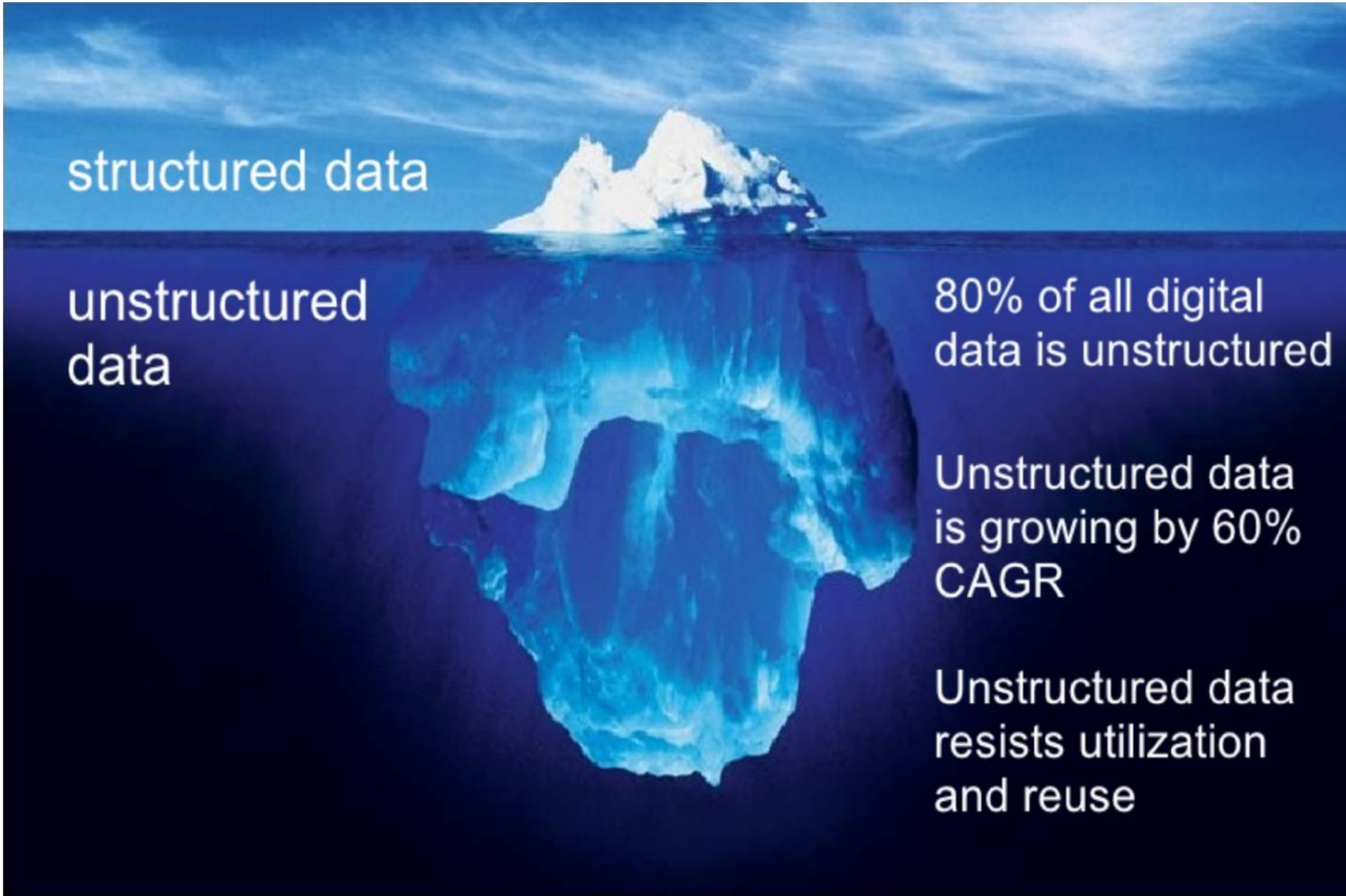
Semi-structured

Clinical notes

EXAM: BILATERAL DIGITAL SCREENING MAMMOGRAM WITH CAD, **DATE[Mar 16 01]: COMPARISON: **DATE[Jul 01 01] TECHNIQUE: Standard CC and MLO views of both breasts were obtained. FINDINGS: The breast parenchyma is heterogeneously dense. The pattern is extremely complex with postsurgical change seen in the right upper outer quadrant and scattered benign-appearing calcification seen bilaterally. A possible asymmetry is seen in the superior aspect of the left breast. The parenchymal pattern otherwise remains stable bilaterally, with no new distortion or suspicious calcifications. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy.. LEFT: Possible developing asymmetry superior aspect left breast for which further evaluation by true lateral and spot compression views recommended. Ultrasound may also be needed.. RECOMMENDATION: Left diagnostic mammogram with additional imaging as outlined above.. A left breast ultrasound may also be needed. BI-RADS Category 0: Incomplete Assessment - Need additional imaging evaluation. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy....

Unstructured

Structured vs. Unstructured Data



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 <problem> previous stroke </problem> ; <problem> hypertension </problem> ; <problem> COPD </problem> , stable ; <problem> renal carcinoma </problem> ; presenting after <problem> a fall </problem> and possible <problem> syncope </problem> .
While walking , she accidentally fell to her knees and did hit <problem> her head on the ground </problem> , near <problem> her left eye </problem> .
<problem> Her fall </problem> was not observed , but the patient does not profess <problem> any loss of consciousness </problem> , recalling the entire event.
The patient does have a history of <problem> previous falls </problem> , one of which resulted in <problem> a hip fracture </problem> .
She has had <treatment> physical therapy </treatment> and recovered completely from that .
<test> Initial examination </test> showed <problem> bruising </problem> around the left eye , normal lung examination , normal heart examination , normal neurologic function with a baseline decreased mobility of <problem> her left arm </problem> .
The patient was admitted for <test> evaluation </test> of <problem> her fall </problem> and to rule out <problem> syncope </problem> and possible <problem> stroke </problem> with <problem> her positive histories </problem> .
<test> DIAGNOSTIC STUDIES: All x-rays </test> including <problem> left foot , right knee , left shoulder and cervical spine </problem> showed no <problem> acute fractures </problem> .
<problem> The left shoulder did show old healed left humeral head and neck fracture </problem> with <problem> baseline anterior dislocation </problem> .
<test> CT of the brain </test> showed no <problem> acute changes </problem> , <problem> left periorbital soft tissue swelling </problem> .
<test> CT of the maxillofacial area </test> showed no <problem> facial bone fracture </problem> .
<test> Echocardiogram </test> showed normal left ventricular function , <test> ejection fraction </test> estimated greater than 65% .

http://text-machine.cs.uml.edu/cliner/

Relationship Extraction (RE)

- Determine relationships between entities or events

"We used <u>hemofiltration</u> to **treat** a <u>patient</u> with digoxin overdose that was complicated by refractory hyperkalemia." [PMID: 3718110]

Relationship: Hemofiltration-**TREATS**-Patients

https://semrep.nlm.nih.gov

Negation Identification (NegIde)

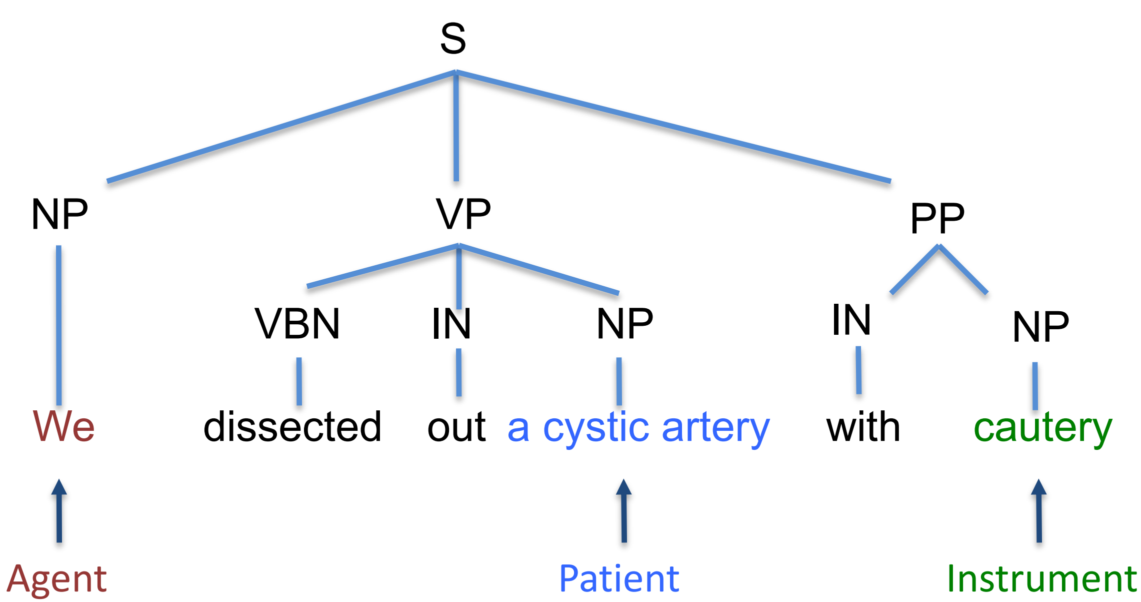
- Identify pertinent Negatives from narrative clinical reports

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

Journal of Biomedical Informatics **34,** 301–310 (2001)

Semantic Role Labeling

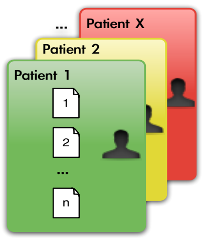
- Detect the semantic role played by each noun phrase associated with the verb of a sentence
  - Agent: Noun Phrase (NP) before the verb
  - Patient: NP after the verb
  - Instrument: NP in a Prepositional Phrase (PP)



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)

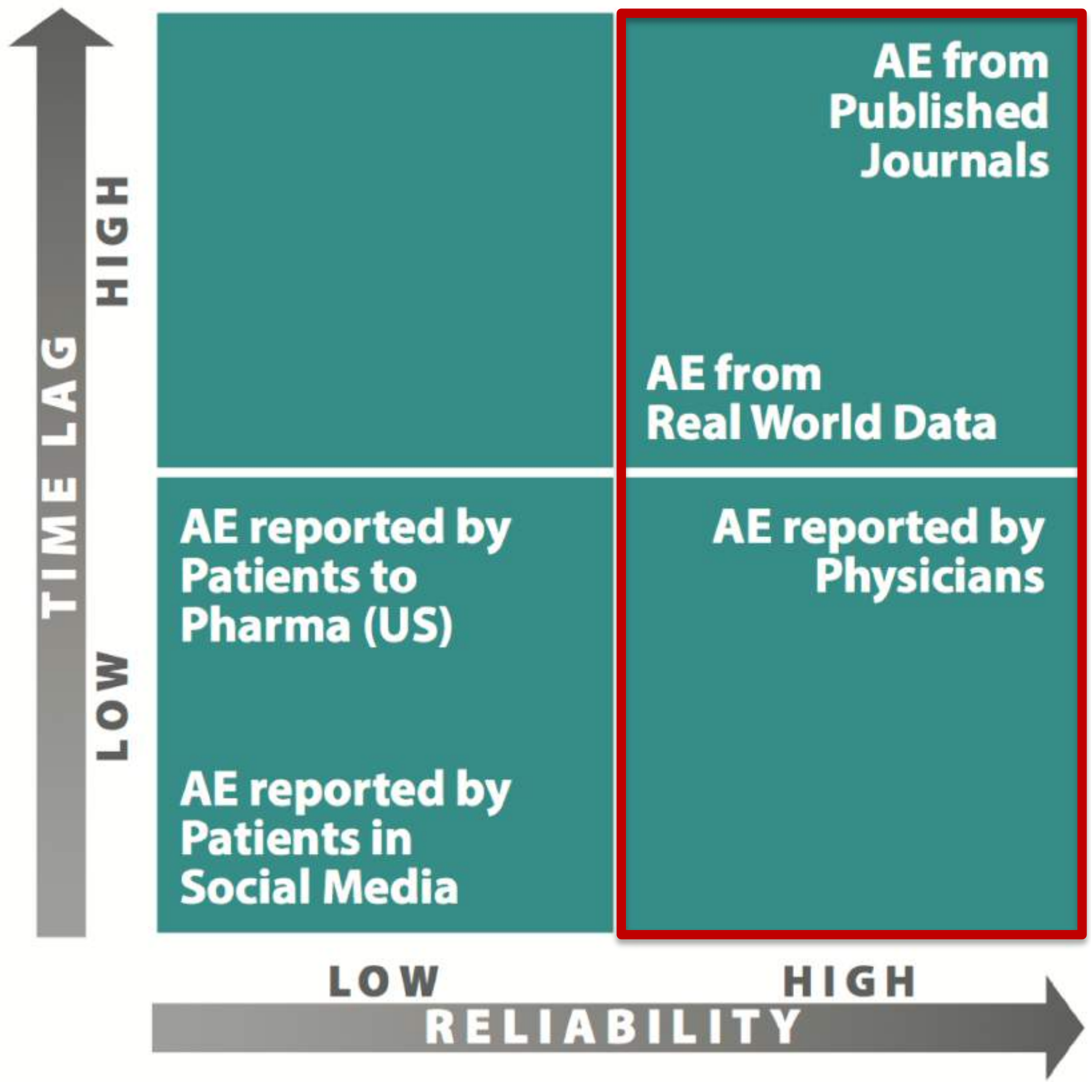
- Social history
- Function score
- Medical history
- Smoking status

Biomedical knowledge
(Subject – 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

J. Am. Med. Inform. Assoc., 14 (5) (2007), p. 550
Springer International Publishing (2016), 10.1007/978-3-319-44564-9_24 pp. 255–266
Computational Linguistics, San Diego, California (2016), pp. 1052-1062

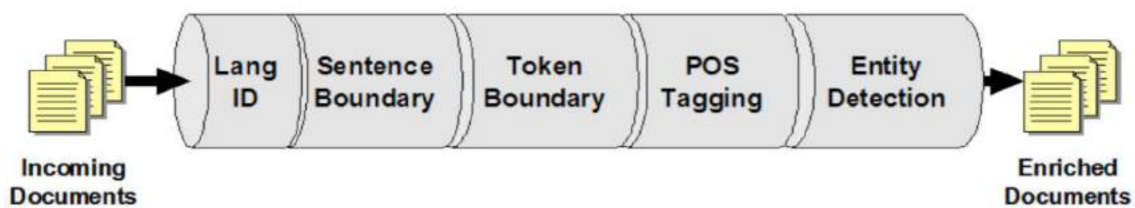
Open source NLP Systems

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

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)

Unstructured Information Management Architecture
An Apache Project



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.

J. Biomedical Inform. 2018, 88: 11-19

NLP-PIER (Patient Information Extraction for Research)

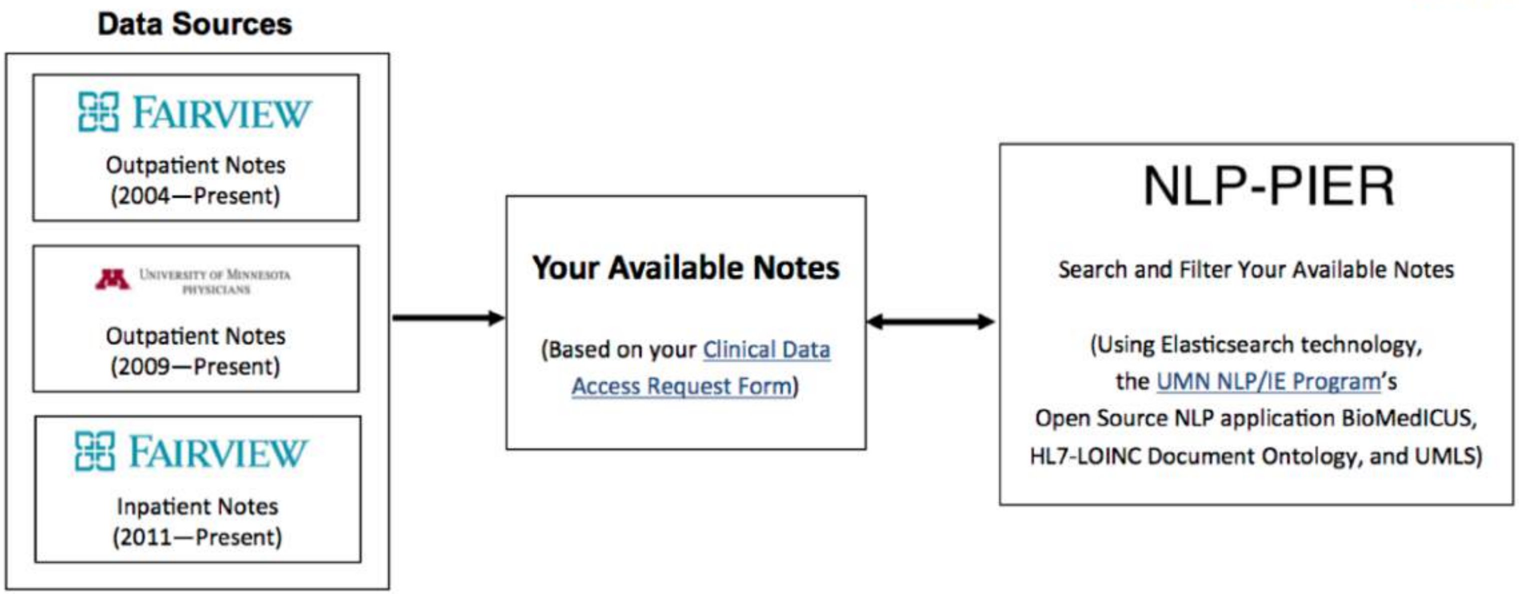
NLP-PIER

Login

Searching Clinical Notes for Research

A collaboration between the UMN NLP/IE Program, Clinical Translational Science Institute (CTSI), MN Supercomputing Institute (MSI), Academic Health Center-Information Services (AHC-IS), and Fairview Health Services Information Technology (FHS-IT)

Search →



University of Minnesota
Driven to Discover℠

AMIA Joint Summit. 2016: 150–159.

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 elasticsearch 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 — **Select Query Template**
 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 equivalant operators

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

mrn:xxxxxxxxxxx
 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

cuis:C0033213
 Find notes tagged with UMLS CUIs (Concept Unique Identifier). Can be combined using logical AND / OR operators, e.g., cuis:C0039796 OR cuis: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.

Epic Categories	Filter category	Enabled	Filter values displayed	Description
	Department Id		5	CDR department identifier
	Encounter Center		5	Encounter center name in Epic
	Encounter Center Type		5	Encounter center type in Epic
	Encounter Clinic Type		5	Encounter Clinic type in Epic
	Encounter Department		5	Encounter department in Epic
	Encounter Department Specialty		5	Specialty name in Epic
	Encounter Id		5	Epic visit number
	Filing Date		5	Date note was filed
	Filing Datetime		5	Date/time note was filed
	Mrn		5	Epic patient identifier
	Patient Id		5	CDR Patient ID
	Prov Id		5	Provider ID in Epic
	Prov Name		5	Provider name in Epic
	Prov Type		5	Provider type in Epic name
	Provider Id		5	CDR department identifier
	Service Date		5	Date of Service
	Service Id		5	CDR encounter identifier
	Text Source Format		5	plain text, rich text, format of analyzed note

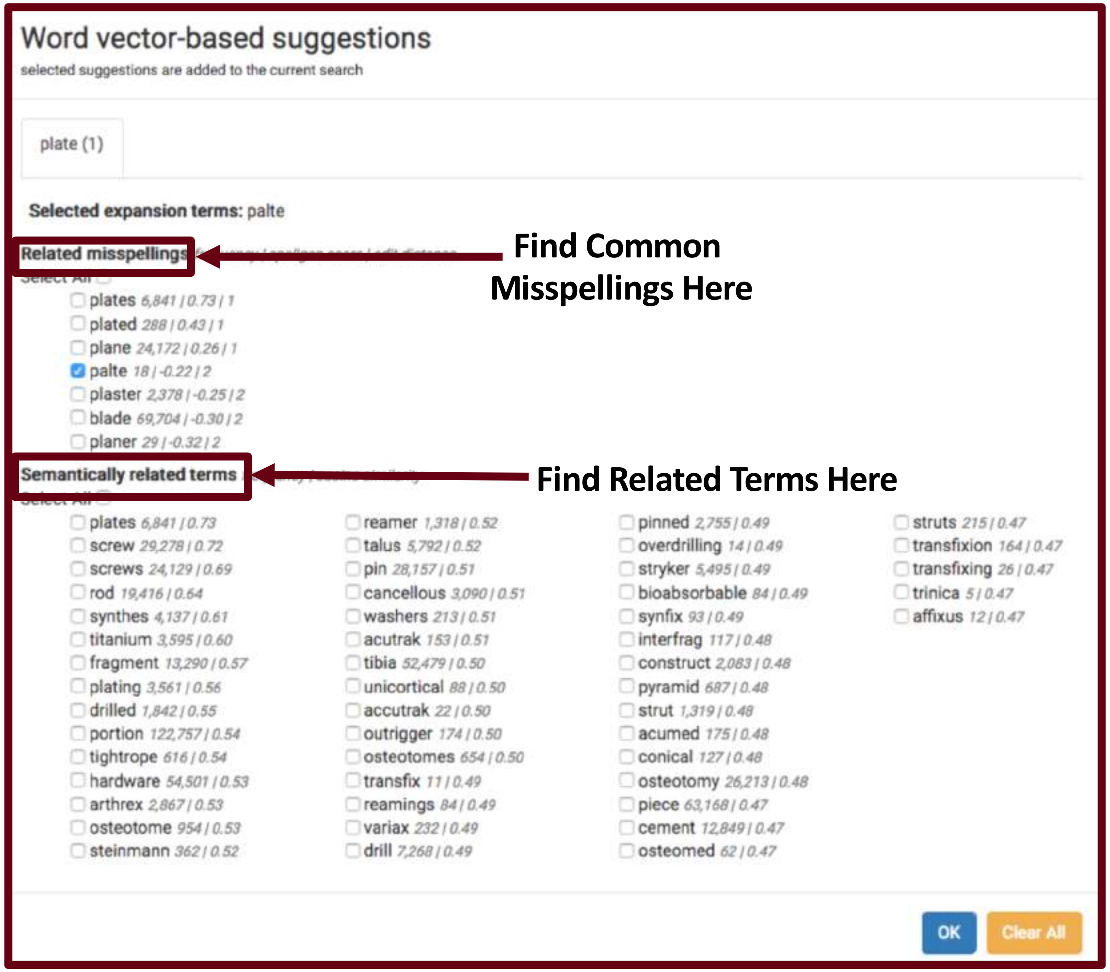
HL7 LOINC	Filter category	Enabled	Filter values displayed	Description
	Kod		5	Kind of Document axis in HL7-LOINC DO
	Role		5	Role axis in HL7-LOINC DO
	Setting		5	Setting axis in HL7-LOINC DO
	Smd		5	Subject Matter Domain in HL7-LOINC DO
	Tos		5	Subject Matter Domain in HL7-LOINC DO

NLP Annotations	Filter category	Enabled	Filter values displayed	Description
	Low Confidence Medical Concepts		10	UMLS CUIs identified by BioMedICUS NLP pipeline, lower confidence detection
	Medical Concepts		20	UMLS CUIs identified by BioMedICUS NLP pipeline

Change number of displayed filters here

Select type of filters that are enabled here

Query Expansion Word Vectors



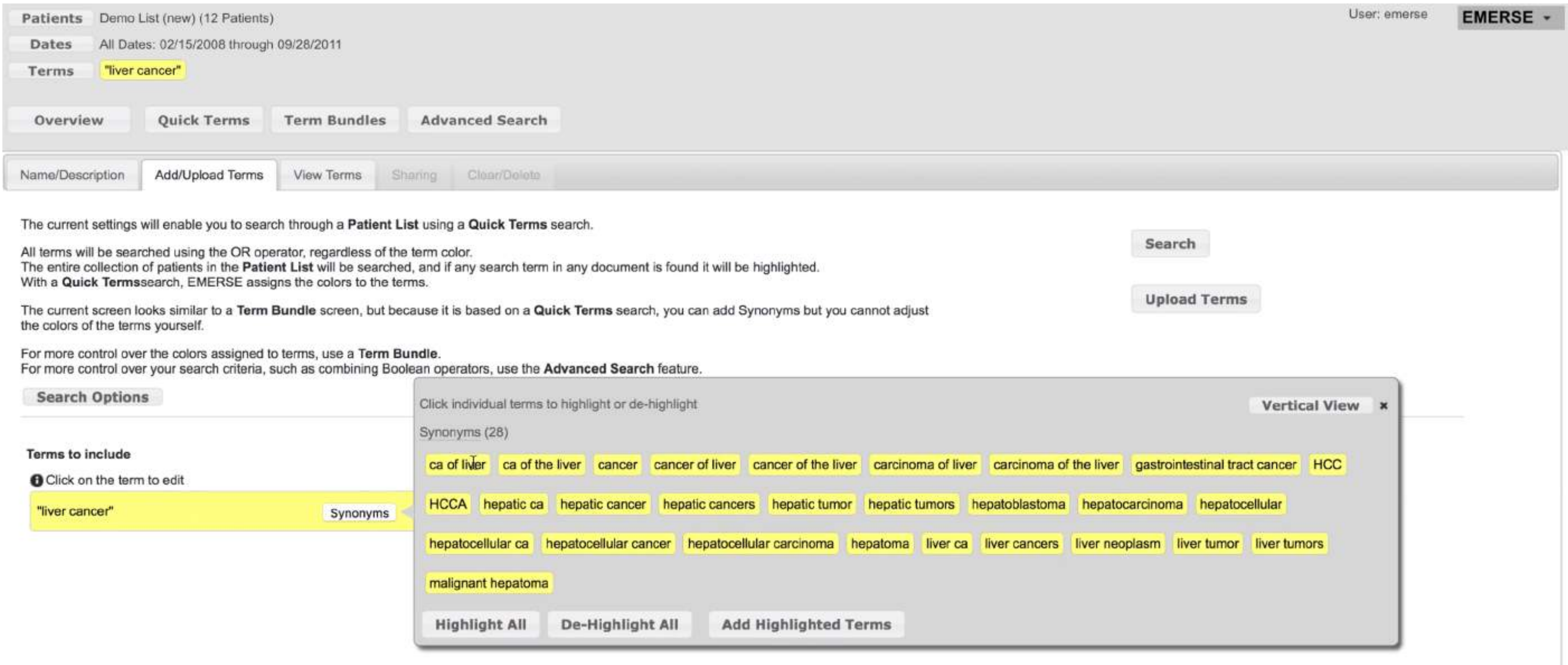
EMERSE (Electronic Medical Record Search Engine)

- 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

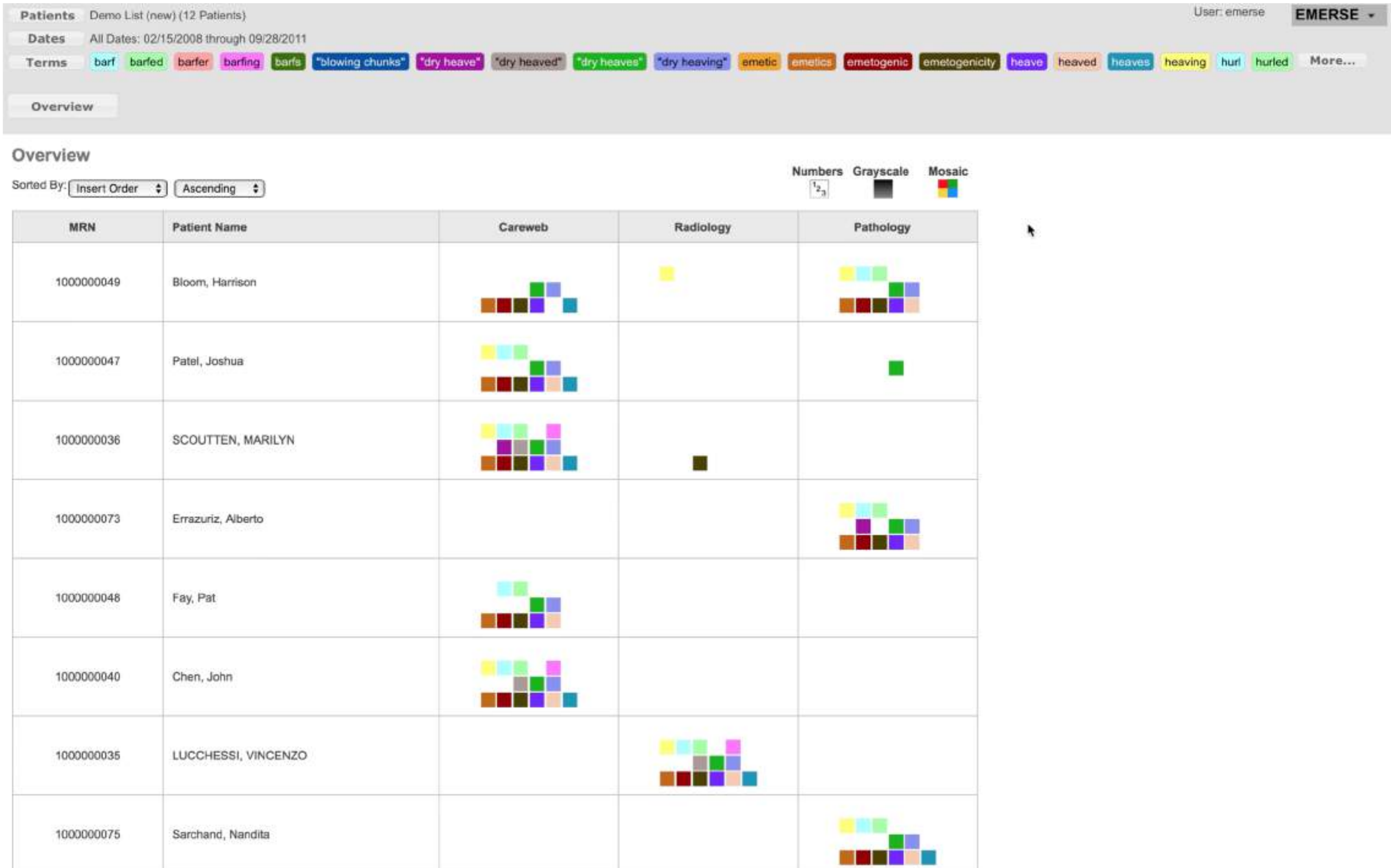
https://research.medicine.umich.edu/our-units/data-office-clinical-translational-research/self-serve-data-tools/emerse

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

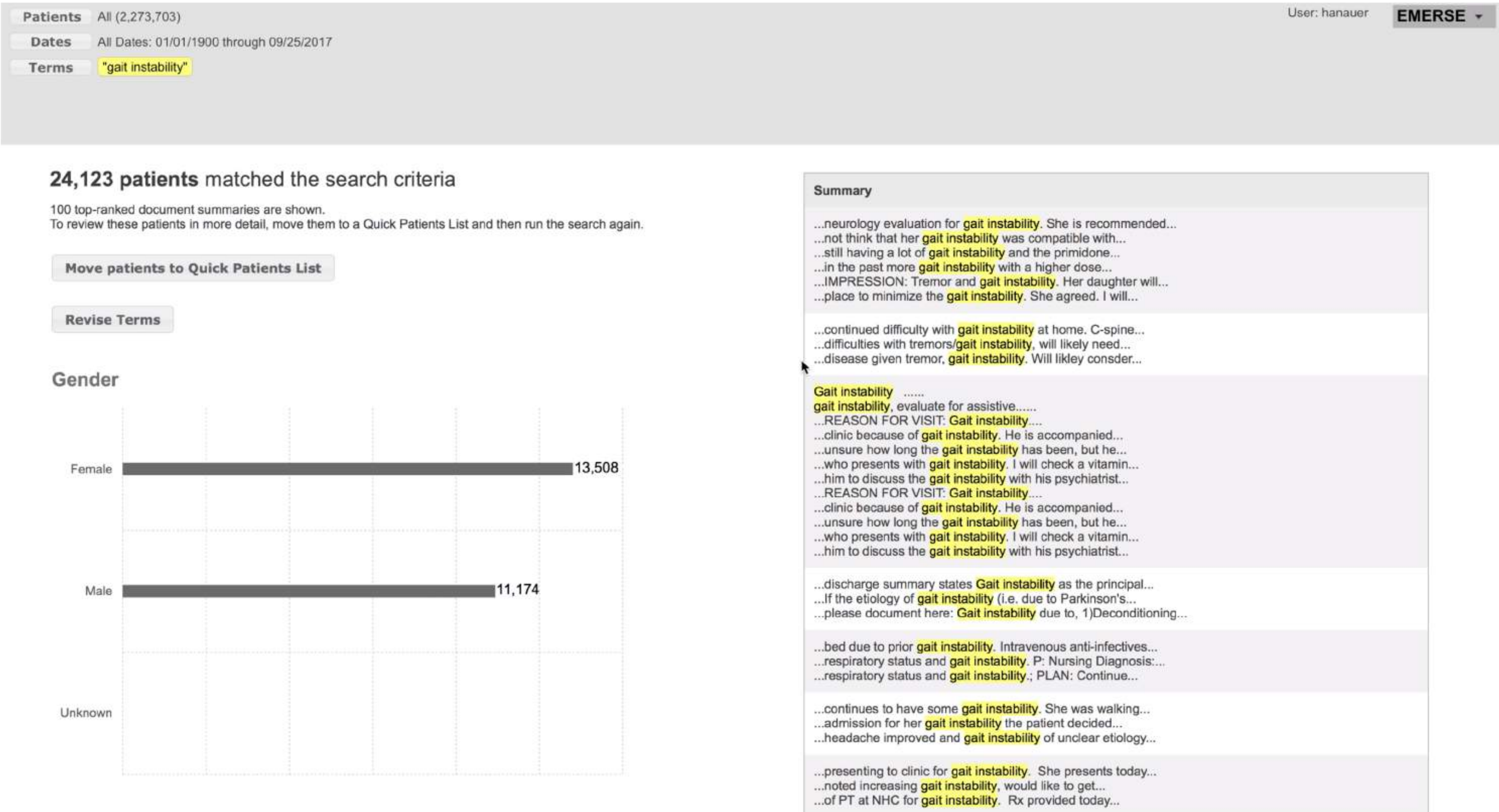
Search with Synonyms



Visualization with NER



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)

NIH National Center for Complementary and Integrative Health

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
Use of complementary and alternative medicine by children in Europe: Published data and expert perspectives. Complement Ther Med. 2013 4;21.
Kaufman, Kelly, JAMA. 2002;287(3):337-344.
Dietary Supplement Use Among U.S. Adults Has Increased Since NHANES III (1988–1994). 2014(Nov 4, 2014). CDC.

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)

Kaufman, Kelly, JAMA. 2002;287(3):337-344.
Geller et al. New England J Med. 2015; 373:1531-40.
Gurley BJ. Molecular nutrition & food research. 2008, 52(7):772-9.

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)

Department of Health and Human Services, Food and Drug Administration. New dietary ingredients in dietary supplements — background for industry. March 3, 2014
Dietary Supplement and Nonprescription Drug Consumer Protection Act. Public Law 109-462, 120 Stat 4500.

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.)

1. Institute of Medicine. Committee on the use of complementary and alternative medicine by the American . complementary and alternative medicine in the united states. National Academies P, editor. Washington, D.C: National Academies Press; 2005.
2. Bent S. Herbal medicine in the United States: review of efficacy, safety, and regulation: grand rounds at University of California, San Francisco Medical Center. J Gen Intern Med. 2008;23(6):854-9

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

Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. BioMed research international. 2014;2014.
Pakhomov SV, Finley G, McEwan R, Wang Y, Melton GB. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics. 2016 Aug 16;32(23):3635-44.

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

calcium	**chamomile**	**cranberry**	**dandelion**	**flaxseed**	**garlic**	**ginger**
ginkgo	**ginseng**	**glucosamine**	**lavender**	**melatonin**	**turmeric**	**valerian**

Method Overview

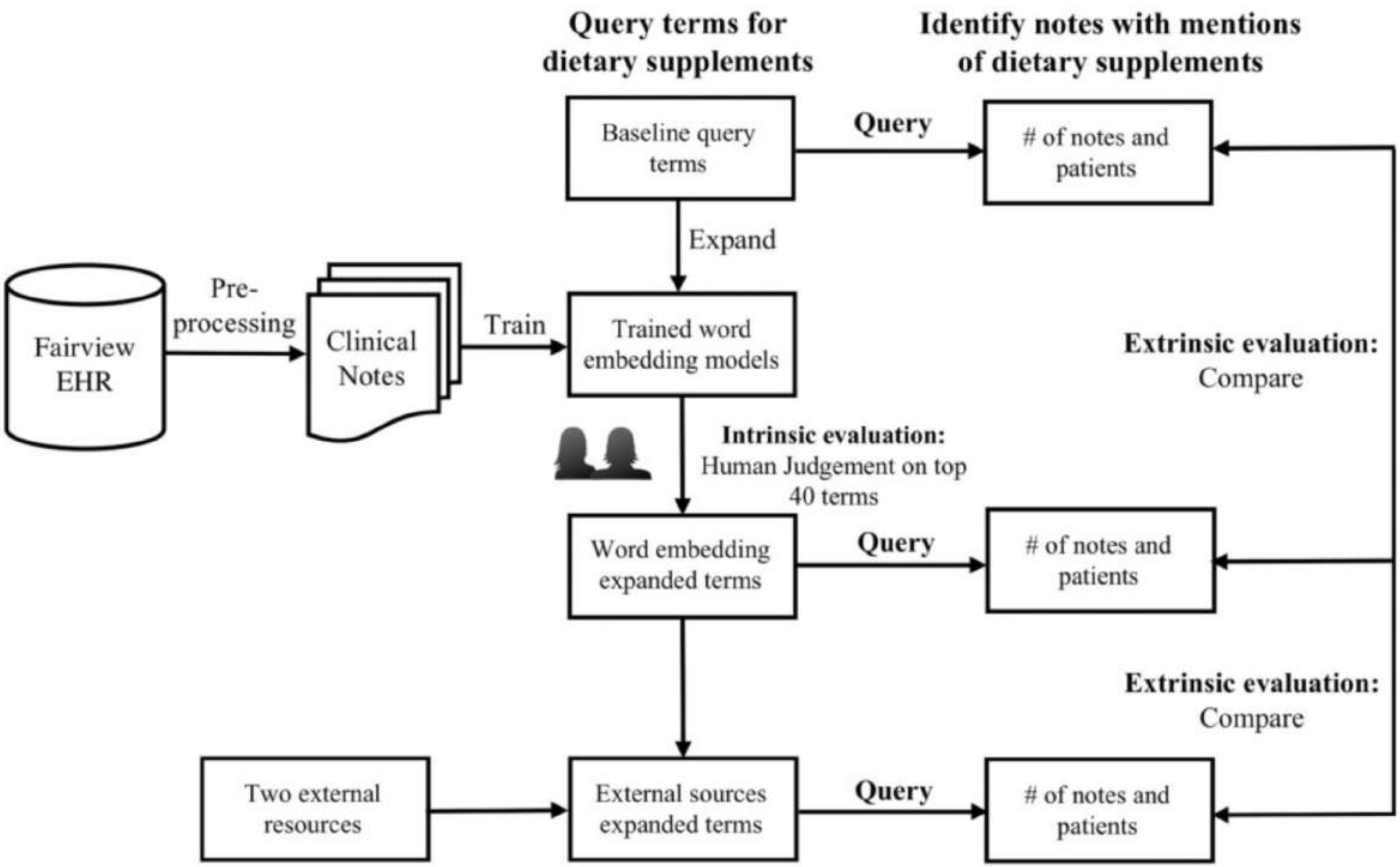


Figure 1. The overview and workflow of the method. EHR: electronic health record.

Model Training

- Corpus size

Table 1. The number of semantically similar terms identified by human experts based on 40 top-ranked terms by word2vec for each 14 DS from 7 corpora

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

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)

Results

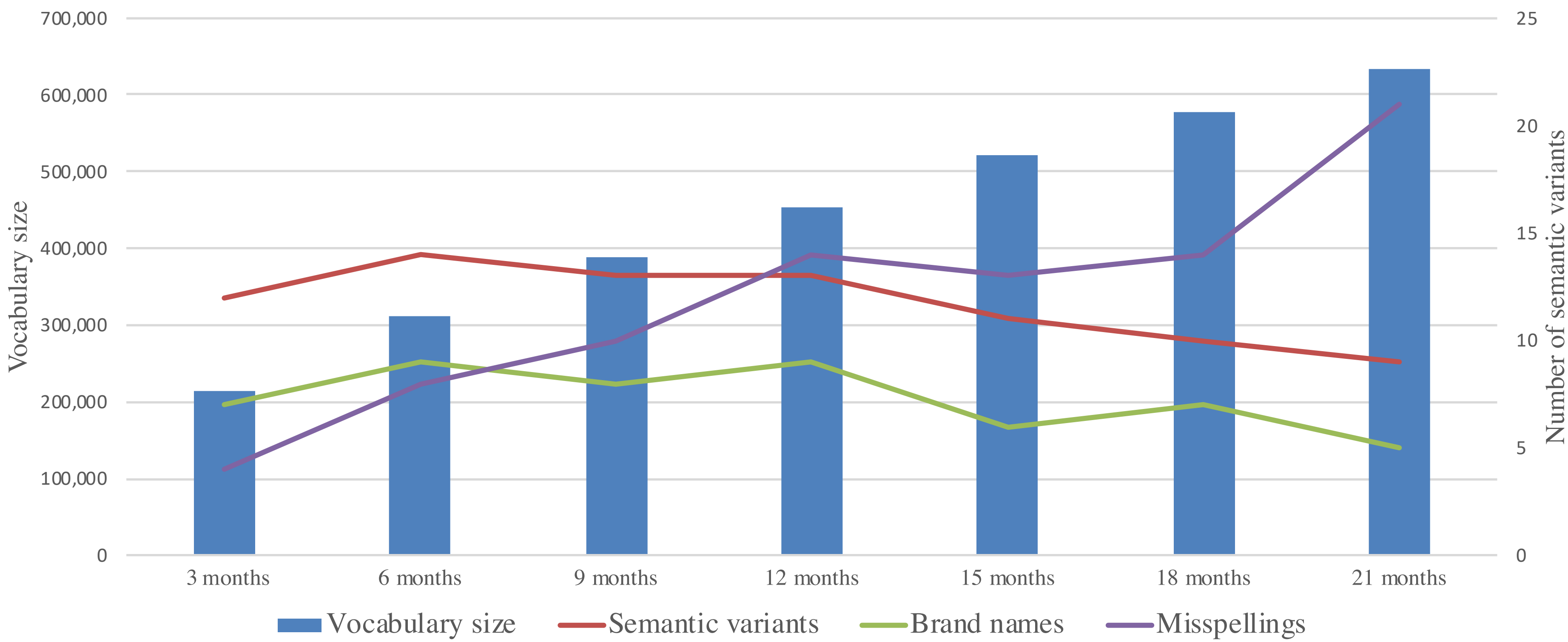


Figure 1. The number of semantic similar terms identified by human experts based on 40 top-ranked terms by word2vec for each DS from 7 corpora

Results: Query Expansion Examples

Initial Query	word2vec Expanded Query	Expanded Examples
Black cohosh	**Misspelling**: black kohosh, black kohash; **Brand name**: remifemin Estroven Estrovan estraven icool amberen amberin Estrovera EstroFactor	• Please try black cohash 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**: Melantonin, melotonin **Brand name**: alteril, neuro sleep	• Can try melantonin 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

Queries		Number of clinical notes				Number of patients			
Dietary supplements	Number of expanded terms	Base query	Expanded query	Additional records found	Percentage increase (%)	Base query	Expanded query	Additional patients found	Percentage increase (%)
Black cohosh	3	13,782	23,641	9,859	71.54	5,560	8,833	3,273	**58.87**
Calcium	10	7,024,626	7,053,856	29,230	0.42	950,282	950,992	710	0.07
Cranberry	3	187,586	189,239	1,653	0.88	71,860	72,499	639	0.89
Dandelion	2	4,316	4,375	59	1.37	2,155	2,191	36	1.67
Fish oil	1	1,305,996	1,311,777	5,781	0.44	192,326	195,015	2,689	1.40
Folic acid	3	839,710	1,058,627	218,917	26.07	107,897	159,768	**51,871**	**48.07**
Garlic	1	91,342	92,481	1,139	1.25	28,657	28,784	127	0.44
Ginger	0	88,870	88,870	0	0	53,867	53,867	0	0
Ginkgo	3	19,020	27,502	8,482	44.60	5,202	7,080	1,878	36.10
Ginseng	2	9,663	10,748	1,085	11.23	3,754	4,112	358	9.54
Glucosamine	5	468,774	469,925	1,151	0.25	68,013	68,106	93	0.14
Green tea	2	29,810	29,816	6	0.02	12,853	12,856	3	0.02
Melatonin	1	647,389	647,601	212	0.03	101,994	102,041	47	0.05
Milk thistle	1	18,930	19,298	368	1.94	3,245	3,279	34	1.05
Saw palmetto	1	38,934	38,947	13	0.03	6,708	6,709	1	0.01
Turmeric	3	25,172	37,959	12,787	50.80	6,583	10,758	4,175	63.42
Valerian	2	15,023	15,330	307	2.04	6,435	6,589	154	2.39
Vitamin E	0	384,072	384,072	0	0	68,284	68,284	0	0

Results: Comparison with External Source

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

Dietary supplements	Number of external source terms	Number of word embedding expanded terms	Number of clinical notes				Number of patients			
			External source query	Word embedding query	Additional records found	Percentage increase (%)	Base query	Word embedding query	Additional records found	Percentage increase (%)
Calcium	15	12	7453873	7543569	89696	1.20	1000906	1002211	1305	0.13
Cranberry	21	3	196944	198625	1681	0.85	76697	77327	630	0.82
Flaxseed	10	2	169349	169343	-6	0.00	45229	45222	-7	-0.02
Ginkgo	6	3	20275	28093	7818	38.56	5855	7791	1936	33.07
Turmeric	18	3	35719	48749	13030	36.48	8962	13486	4524	50.48

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***)
Fan Y, et al., *BMC Medical Informatics & Clinical Decision*. 2018, 18(Suppl 2): 51.

Results: Performance comparison

Type	Features	Decision tree			Random forest			Naïve Bayes			SVM			Maximum Entropy		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Type 1	raw uni[a]	0.819	0.817	0.816	0.858	0.853	0.853	0.770	0.757	0.755	0.818	0.816	0.815	0.850	0.849	0.849
Type 2	uni	0.846	0.845	0.844	0.878	0.876	0.876	0.793	0.784	0.783	0.837	0.835	0.834	0.874	0.873	0.873
Type 3	tf-idf	0.862	0.857	0.857	0.862	0.857	0.857	0.763	0.704	0.701	0.844	0.839	0.839	0.840	0.831	0.831
Type 4	bi[a]	0.760	0.720	0.716	0.760	0.720	0.716	0.715	0.707	0.702	0.735	0.719	0.720	0.749	0.739	0.739
Type 5	uni + bi	0.872	0.864	0.863	0.872	0.864	0.863	0.815	0.808	0.807	0.881	0.877	0.876	0.890	0.888	0.887
Type 6	uni + bi+tri[a]	0.863	0.852	0.850	0.863	0.852	0.850	0.815	0.808	0.808	0.880	0.876	0.875	0.887	0.883	0.882
Type 7	indi[a] only	0.848	0.847	0.846	0.861	0.860	0.860	0.860	0.849	0.848	0.851	0.849	0.849	0.862	0.859	0.859
Type 8	uni + bi+indi	0.860	0.860	0.860	0.875	0.865	0.864	0.813	0.803	0.801	0.899	0.897	0.897	0.895	0.903	**0.902**
Type 9	uni + bi+tri + indi	0.860	0.857	0.857	0.872	0.861	0.860	0.813	0.803	0.801	0.899	0.897	0.897	0.905	0.903	**0.902**

[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

Fan Y, et al., *BMC Medical Informatics & Clinical Decision*. 2018, 18(Suppl 2): 51.

Performance comparison

The Performance of Maximum Entropy with Type 8 in Test Set

Status	Number	Precision	Recall	F-measure
Continuing	233	0.86	0.95	0.90
Discontinued	166	0.94	0.89	0.92
Started	178	0.92	0.91	0.91
Unclassified	173	0.92	0.84	0.88
Total (weighted)	750	0.91	0.90	0.90

The Performance of Classifier on 25 dietary supplements

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

Fan Y, et al., *BMC Medical Informatics & Clinical Decision*. 2018, 18(Suppl 2): 51.

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

Models	Precision	Recall	F-measure (weighted)
Word-CNN	0.808	0.804	0.803
Bi-LSTM	0.882	0.880	0.879
Stacked Bi-LSTM	0.918	0.916	0.916

1.3. Mining Biomedical Literature to Discovery Drug-Supplement Interactions (DSIs)

THE WALL STREET JOURNAL.

Rui Zhang ▾ WSJ+

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate Search

How Your Supplements Interact With Rx Drugs | Sloan Kettering's Quest to Prove Exercise Can Inhibit Cancer | WHAT'S YOUR WORKOUT How a Diva Trains for Opera's Ironman | 20 ODD QUESTIONS Designer Mark McNairy on Bow Ties and Maseratis

LIFE | HEALTH | THE INFORMED PATIENT

How Your Supplements Interact With Prescription Drugs

St. John's Wort, lavender, garlic and others can alter drug potency, cause side effects

Millions of people consume supplements that may impact the way the prescription drugs they also take are metabolized b

PHOTO: GETTY IMAGES

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.

FOX NEWS Health

Home Video Politics U.S. Opinion Business Entertainment Tech Science Health Travel Lifestyle World On Air

MEDICATIONS

What you should know about how supplements interact with prescription drugs

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

Literature-based Discovery

PubMed

Big Data:
28 million abstracts

X1-Y1 X2-Y2 … Xm-Ym	&	Y1-Z1 Y3-Z2 … Yn-Zn	→	X1-Z1 Y5-Z8 … Xk-Zt

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**)

Named entity recognition (NER), Relationship extraction

Echinacea - INHIBITS - CYP450

&

CYP450 - INTERACTS_WITH - Toremifene

Echinacea - <Potentially Interacts With> - Toremifene

Results: Selected Interactions

Drug/Supplement	Predicate	Gene/Gene Class	Predicate	Supplement/Drug	Known
Echinacea	INH	CYP450	INT	Docetaxel	Y
Echinacea	INH	CYP450	INT	Toremifene	N
Echinacea	STI	CYP1A1	INT	Exemestane	N
Grape seed extract	INH	CYP3A4	INT	Docetaxel	N
Kava preparation	STI	CYP3A4	INT	Docetaxel	Y

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

Semantic predication	Citations
Echinacea INHIBITS CYP450	We have shown that ECHINACEA preparations and some common alkylamides weakly ***inhibit*** several cytochrome P450 (CYP) isoforms, with considerable variation in potency. (19790031)
Grape seed extract INHIBITS CYP3A4	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)
Melatonin INHIBITS Cyclooxygenase-2	Moreover, Western blot analysis showed that melatonin ***inhibited*** LPS/IFN-gamma-induced expression of COX-2 protein, but not that of constitutive cyclooxygenase. (18078452).
CYP450 INTERACTS_WITH Toremifene	Tamoxifen and toremifene are ***metabolised*** by the cytochrome p450 enzyme system, and raloxifene is metabolised by glucuronide conjugation. (12648026)
CYP3A INHIBITS Docetaxel	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)

1.4. Active Learning to Reduce Annotation Costs for NLP Tasks

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

Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. *J Biomed Inform* 2015; 58: 11–8.
Chen Y, Cao H, Mei Q, Zheng K, Xu H. Applying active learning to supervised word sense disambiguation in MEDLINE. *J Am Med Inform Assoc* 2013; 20 (5): 1001–6.

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

Vasilakes J, Rizvi R, Melton G, Pakhomov S, Zhang R. *J Am Med Info Assoc Open*. 2018

Method Overview

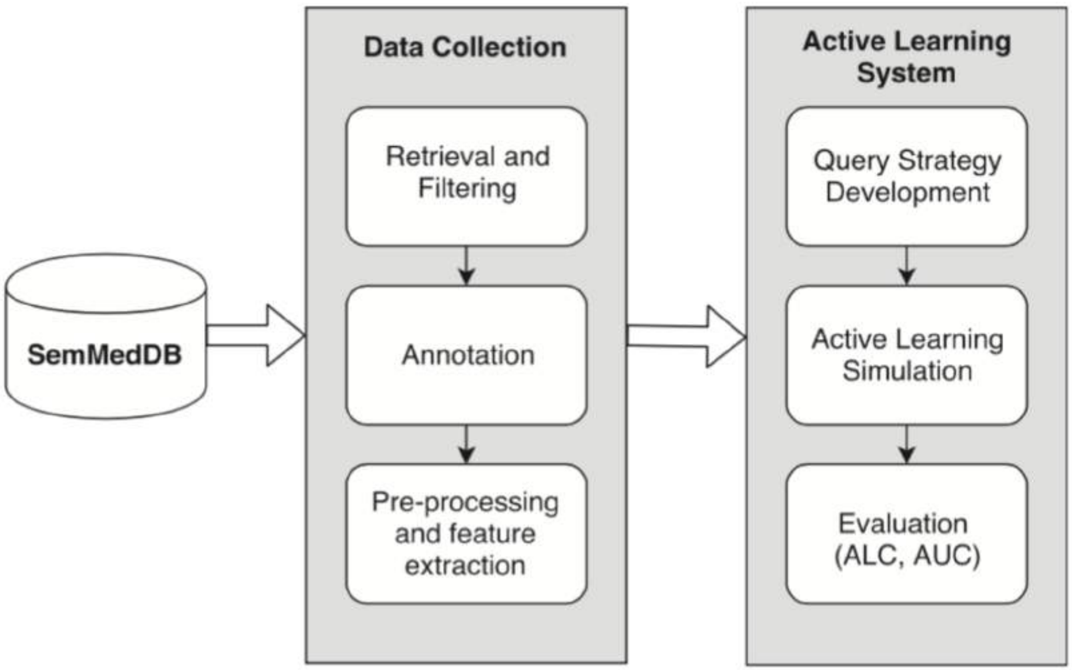


Figure 1. An overview of the active learning system development process.

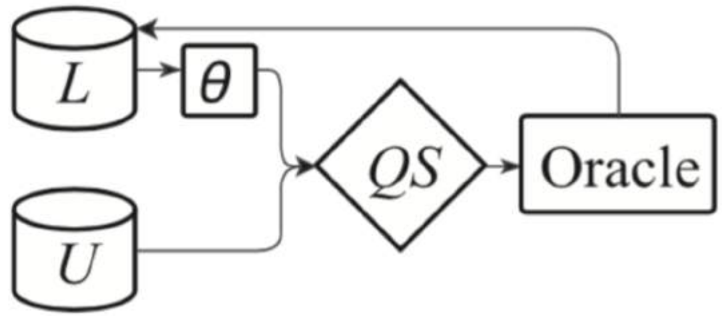


Figure 2. The active learning process. From an initial labeled set L, train the ML model θ, choose the most informative example from the unlabeled set U using the query strategy QS and the updated θ, query the oracle for its label, and update L.

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

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

Type	Query strategy	Substance interactions		Clinical medicine	
		ALC	$\|L\|$ @ 0.80 AUC	ALC	$\|L\|$ @ 0.80 AUC
Baseline	Passive	0.590	1295	0.491	2473
	SM	0.597	1218	0.541	2093
Uncertainty	LC	0.606	1051	0.543	2043
	LCB2	0.607	1060	0.542	2089
	D2C	0.623	891	0.548	2166
Representative	Density	0.622	905	0.547	2136
	Min-Max	0.634	657	**0.550**	**2127**
Combined	ID ($\beta = 0.01$)	0.626	771	0.534	2157
	ID ($\beta = 1$)	**0.642**	**546**	0.542	2146
	ID ($\beta = 100$)	0.635	653	**0.550**	2174
	ID (dynamic β)[a]	0.641	587	0.549	2180

When L is small and U is large:

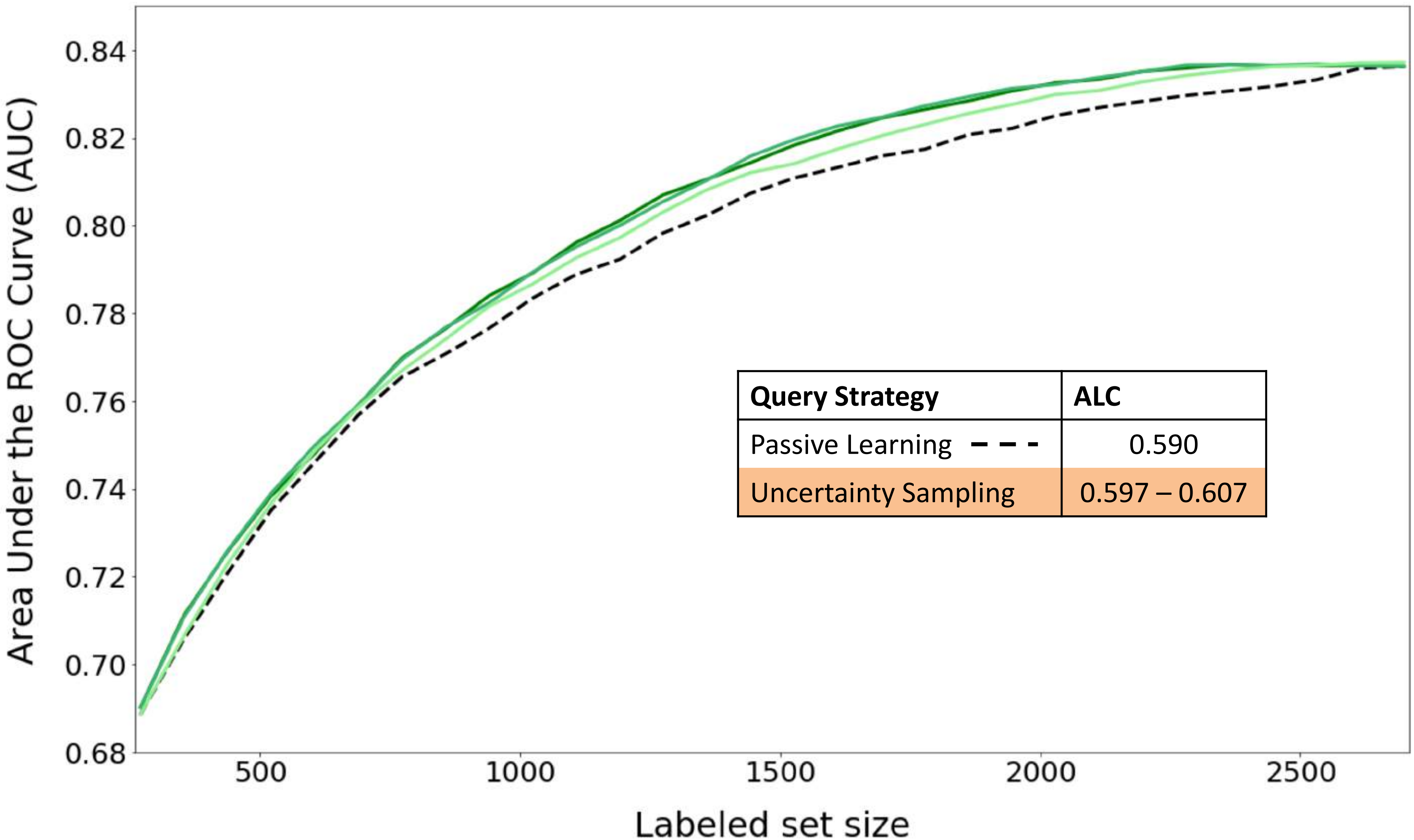
- it is unlikely that L is representative of U
- given that L is small and unrepresentative, the prediction model trained on L is likely to be poor.

$$\beta = \frac{2|U|}{|L|}$$

|U| is the size of the current unlabeled set
|L| is the size of the current labeled set

Results

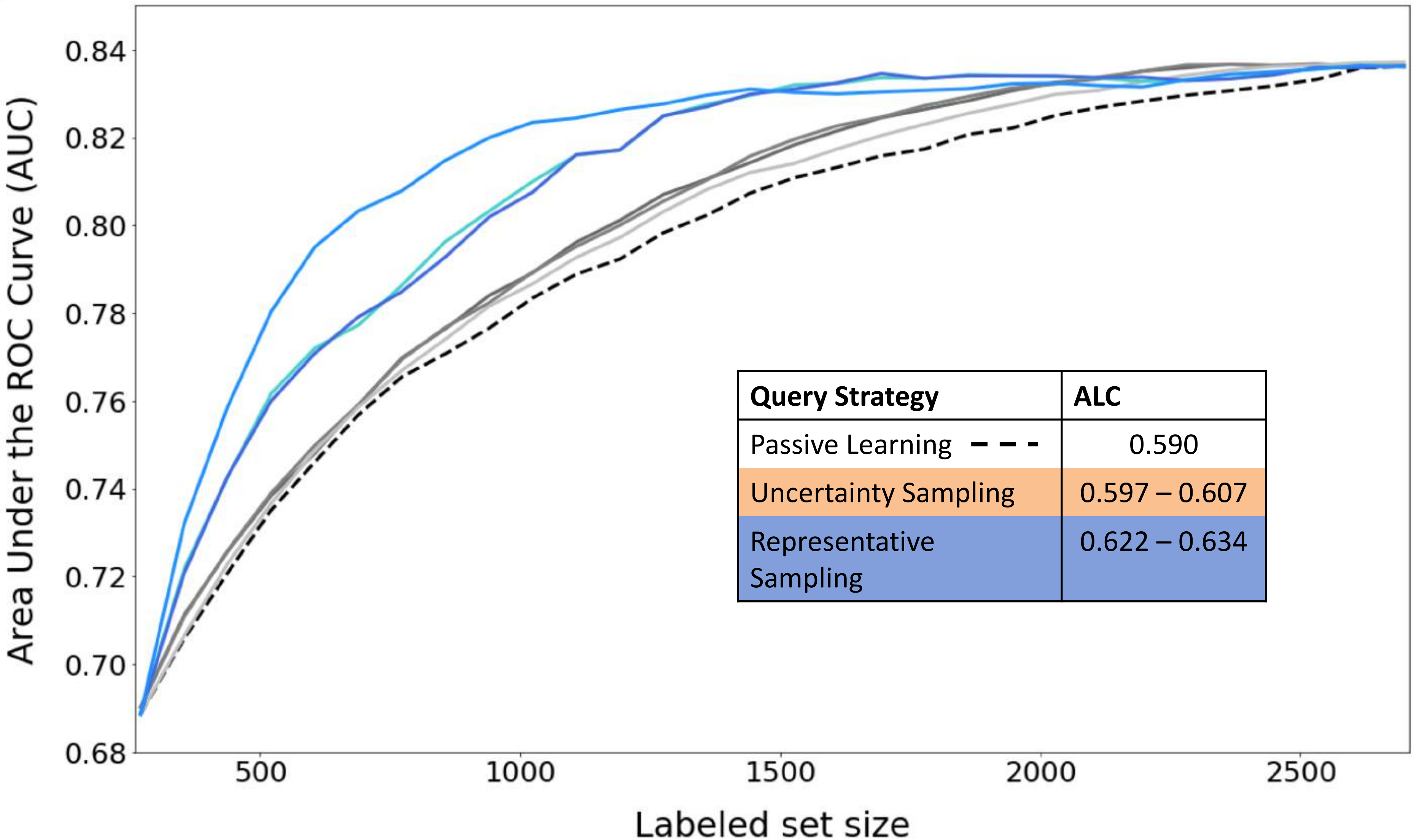
Uncertainty Sampling



Query Strategy	**ALC**
Passive Learning - - -	0.590
Uncertainty Sampling	0.597 – 0.607

Results

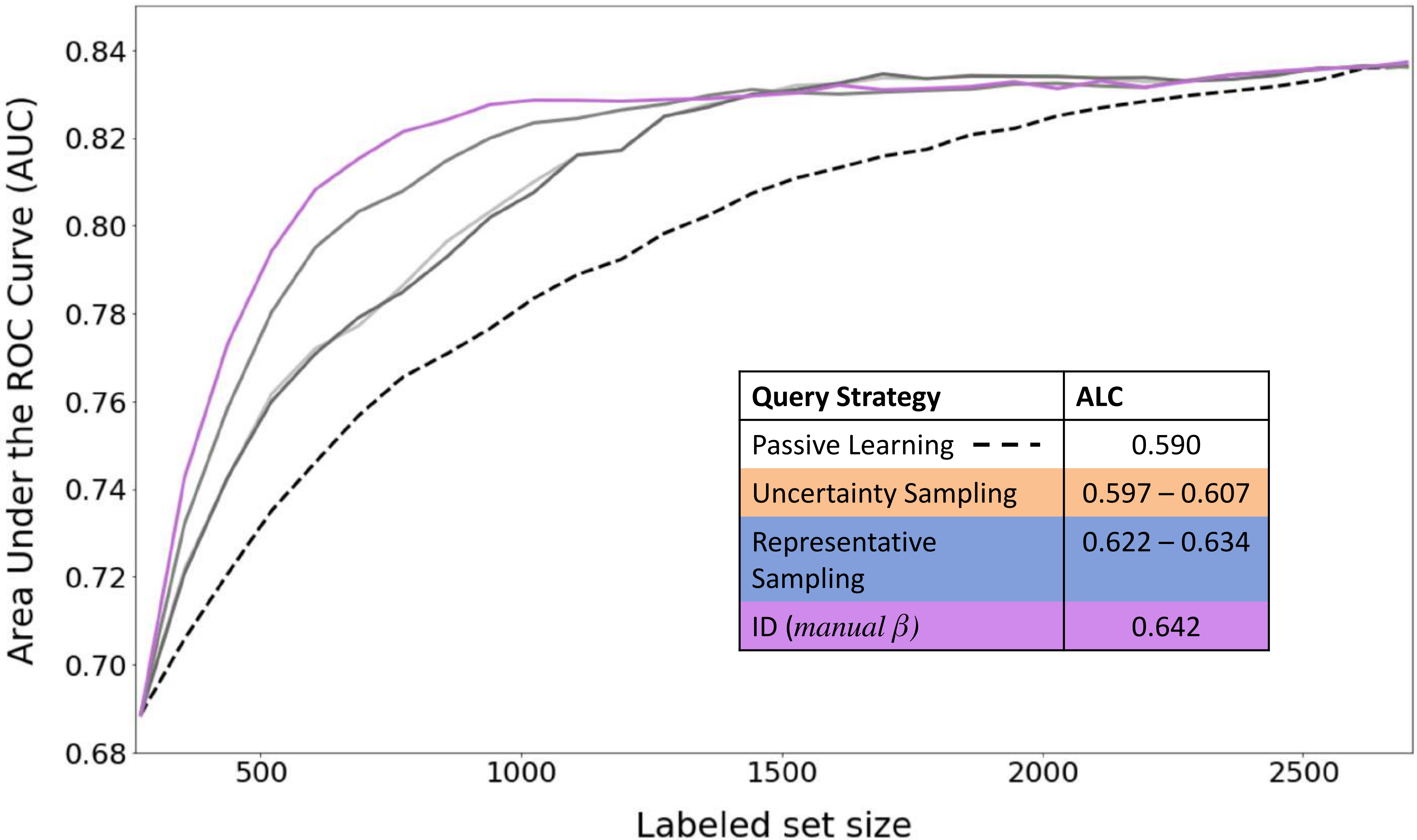
Representative Sampling



Query Strategy	ALC
Passive Learning – – –	0.590
Uncertainty Sampling	0.597 – 0.607
Representative Sampling	0.622 – 0.634

Results

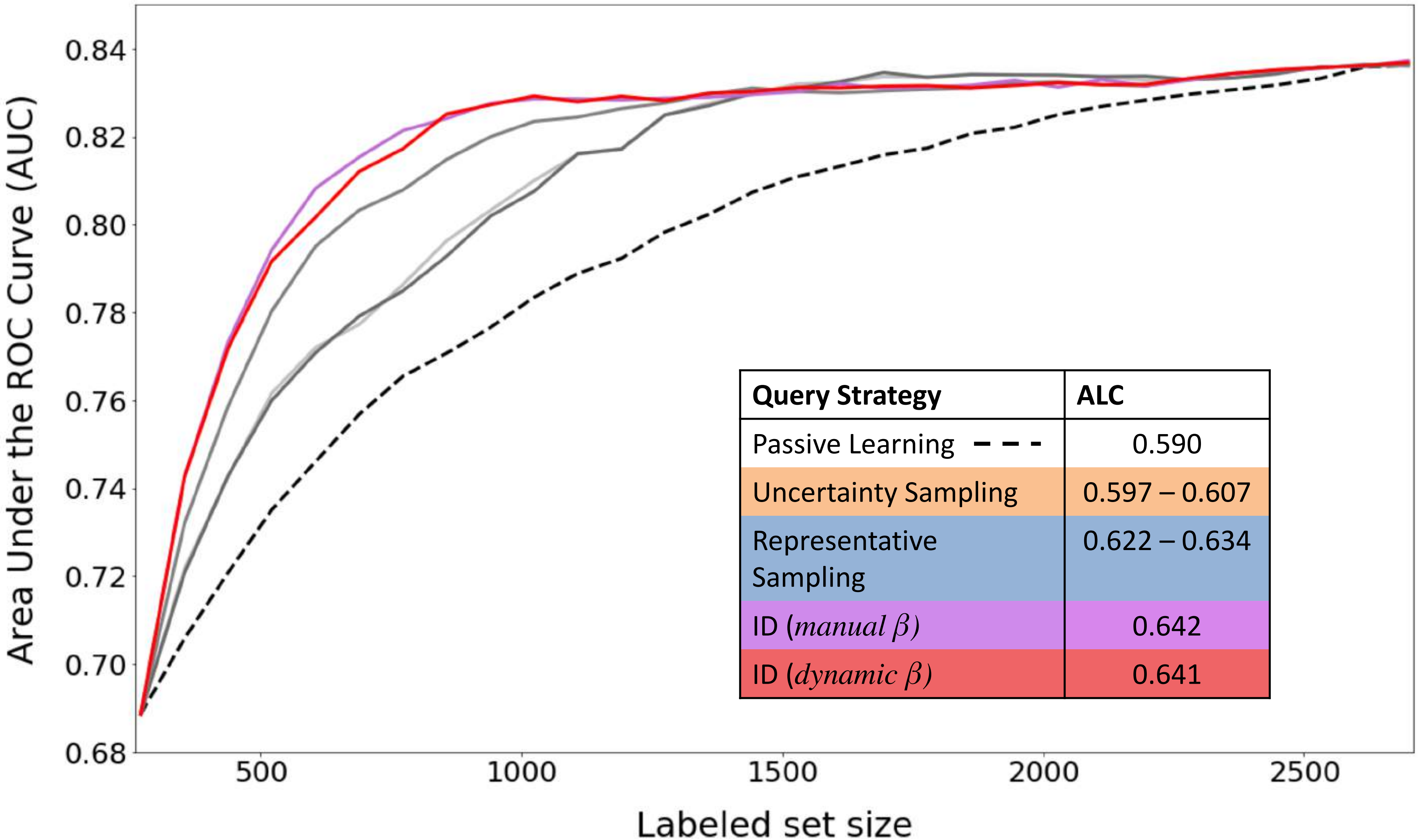
Combined Sampling



Query Strategy	ALC
Passive Learning - - -	0.590
Uncertainty Sampling	0.597 – 0.607
Representative Sampling	0.622 – 0.634
ID (*manual* β)	0.642

Results

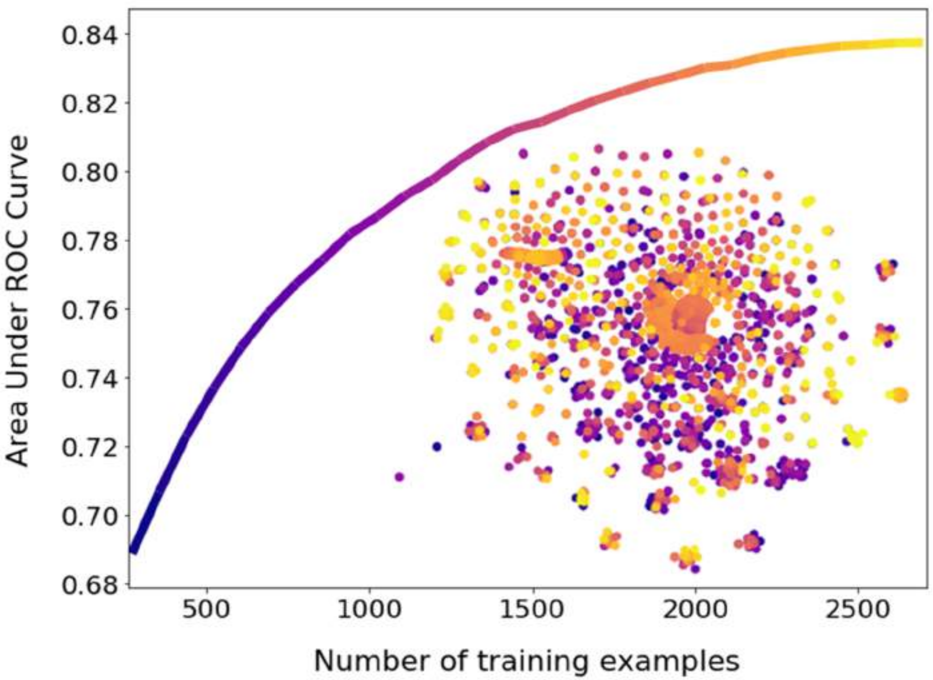
Dynamic β



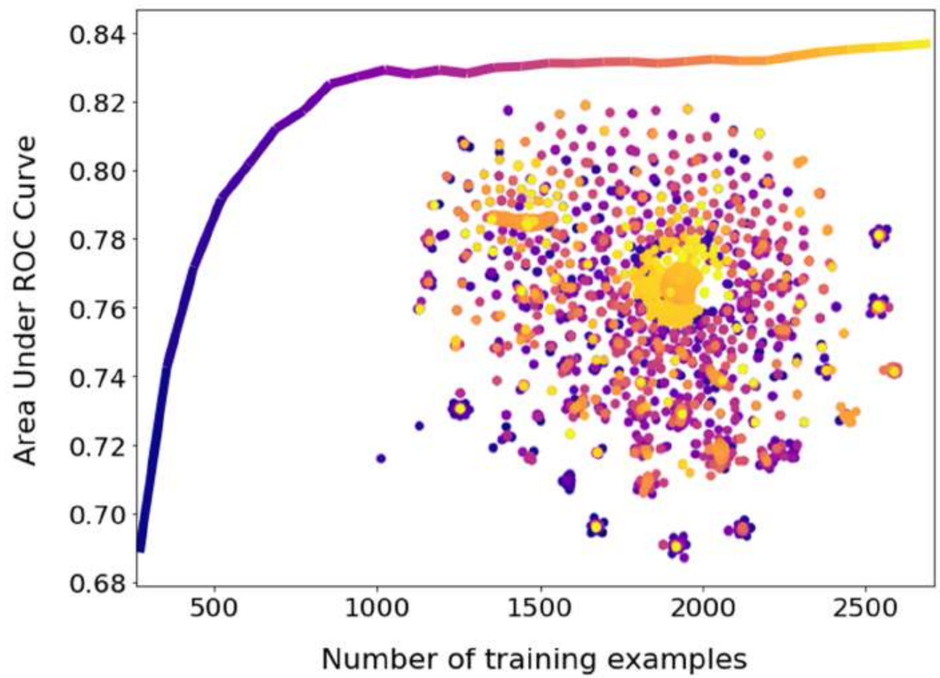
Query Strategy	ALC
Passive Learning — — —	0.590
Uncertainty Sampling	0.597 – 0.607
Representative Sampling	0.622 – 0.634
ID (*manual* β)	0.642
ID (*dynamic* β)	0.641

Performance Analysis

Uncertainty Sampling
(worst performing)



Representative Sampling
(best performing)



Vasilakes J, Rizvi R, Melton G, Pakhomov S, Zhang R. *J Am Med Info Assoc Open*. 2018

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

Jackson RG, et al. BMJ Open 2017;6:e012012

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

SMI concept	Keyword strings	Modifier strings	Lax or strict modifiers	SNOMED-CT (SCTID)†
Aggression	aggress*			61372001
Agitation	agitat*			106126000
Anhedonia	anhedon*			28669007
Apathy	apath*			20602000
Arousal	arous*			(none)
Blunted or flat affect	Affect	blunt*, flat*, restrict*	Optional	6140007/932006/39370001
Catalepsy	catalep*			247917007
Catatonic syndrome	catatoni*			247917007
Circumstantial speech	circumstan*			18343006
Deficient abstract thinking	Concrete			71573006
Delusions	delusion*			2073000
Derailment of speech	derail*			65135009
Diminished eye contact	eye contact			412786000
Disturbed sleep	Sleep	not, poor*, interrupt*, nightmare*, disturb*, inadequat*, disorder*, prevent*, stop*, problem*, difficult*, reduced*, less*, impair*, erratic*, unable*, worse*, depriv*	Optional	26677001

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 κ of 0.83)

Table 4 Comparison of the hybrid approach and context alone across all symptoms (excluding catalepsy, echopraxia and mutism in SMI cohort)

Statistic	Model	P%	R%	F1
Mean	ConText + ML	83	78	0.80
	ConText	71	97	0.79
Median	ConText + ML	90	85	0.88
	ConText	84	98	0.91

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

Data

- 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)

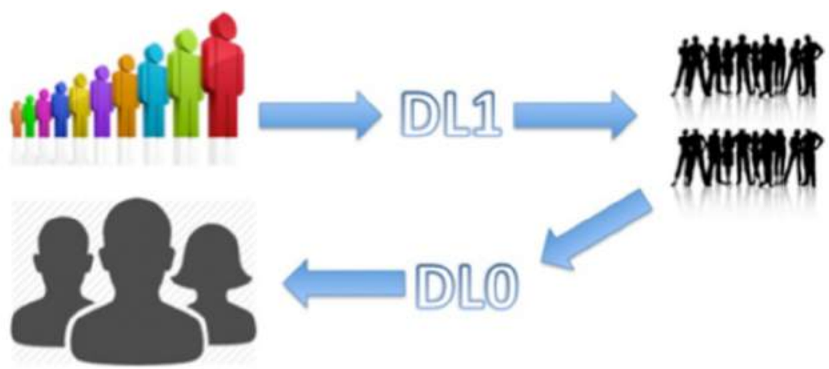


Figure 1 The more sensitive DL1 method was initially applied. Following DL1, the more specific DL0 model was then used on the documents selected with DL1. DL, deep learning paradigm.

Performance Comparison

Table 2 Performance of DL0 considering a fivefold cross-validation

	Predicted 0s	Predicted 1s
True 0s	639	18
True 1s	56	45

Sensitivity 44.5%; specificity 97%.

Table 3 Performance of DL1 considering a fivefold cross-validation

	Predicted 0s	Predicted 1s
True 0s	47	53
True 1s	11	90

Sensitivity 89%; specificity 53%.

Table 5 Performance of DL1+0 considering a fivefold cross-validation

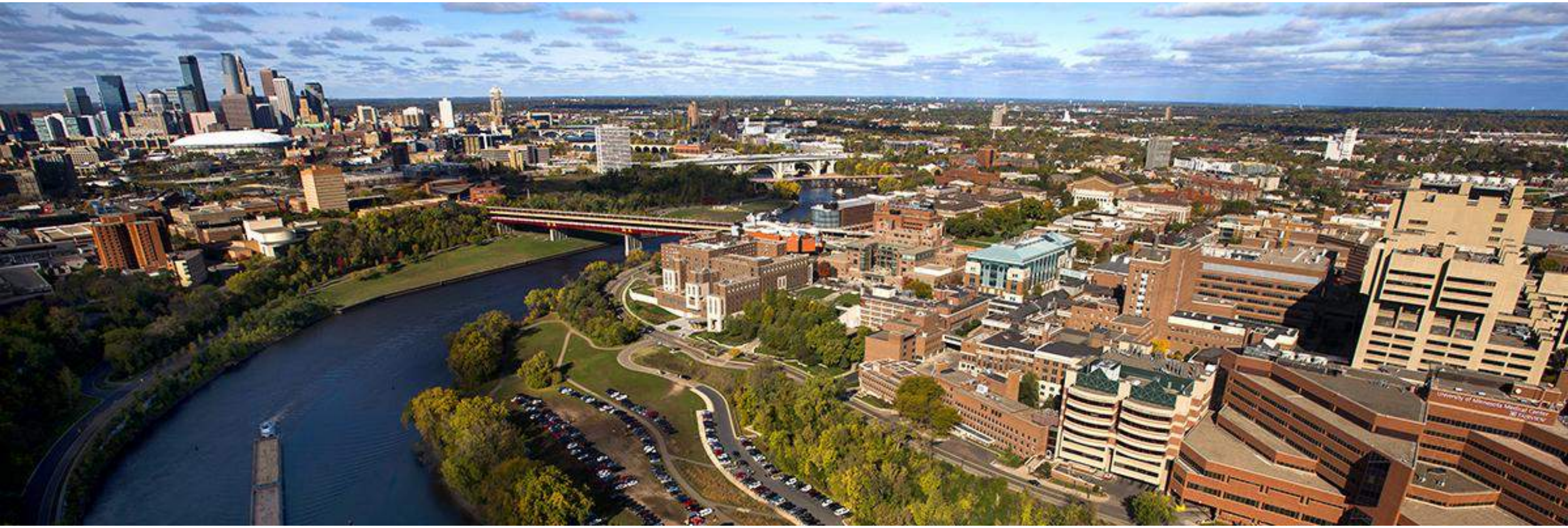
	Predicted 0s	Predicted 1s
True 0s	73	5
True 1s	8	17

Sensitivity 93.5%; specificity 68%; positive predictive value (precision) 77%.

- 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



Rui Zhang, Ph.D.

Email: zhan1386@umn.edu

Research Lab: http://ruizhang.umn.edu/

University of Minnesota

Driven to Discover[SM]

MAYO CLINIC

Clinical Information Extraction

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

Learning Objects

1) Understand challenges of EHRs

challenges

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

technology

3) Explore clinical documentation variations and IE-based NLP tool portability
   - Case study of NLP tool portability (a... ascertainment)

application

Electronic Health Records

Structured Data

Retrieve using query

Unstructured Data

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

Extract using NLP

Challenges of EHR

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

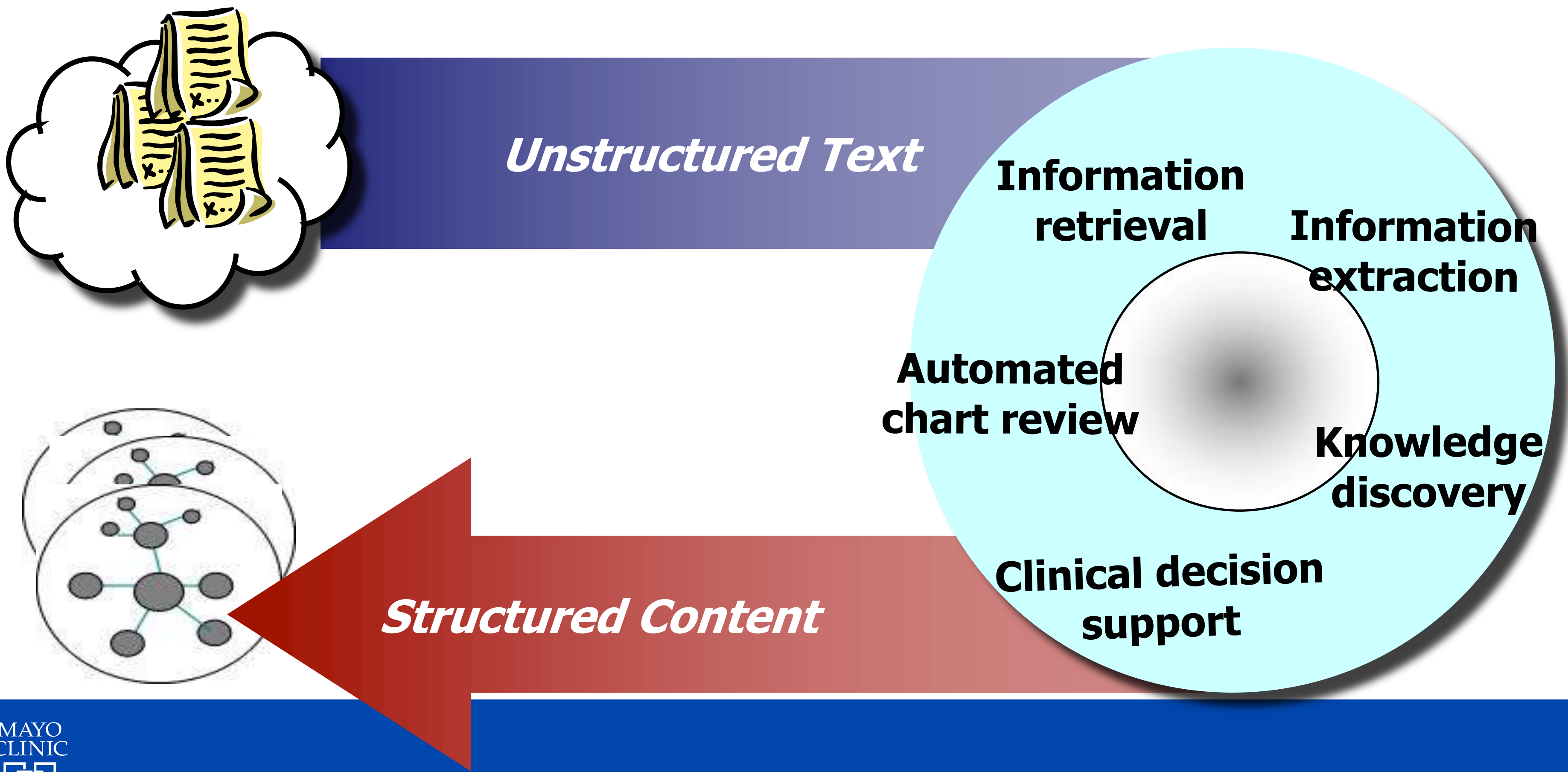
  Clinical IE
- Variability
  - Clinical practice and workflow vary across institutions
  - Clinical language is not homogenous

  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

Image courtesy of National Institutes of Health



Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Clinical information extraction applications: A literature review

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng[1], Saeed Mehrabi[2], Sunghwan Sohn, Hongfang Liu*

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

ARTICLE INFO

Keywords:
Information extraction
Natural language processing
Application
Clinical notes
Electronic health records

ABSTRACT

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.

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

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.

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.

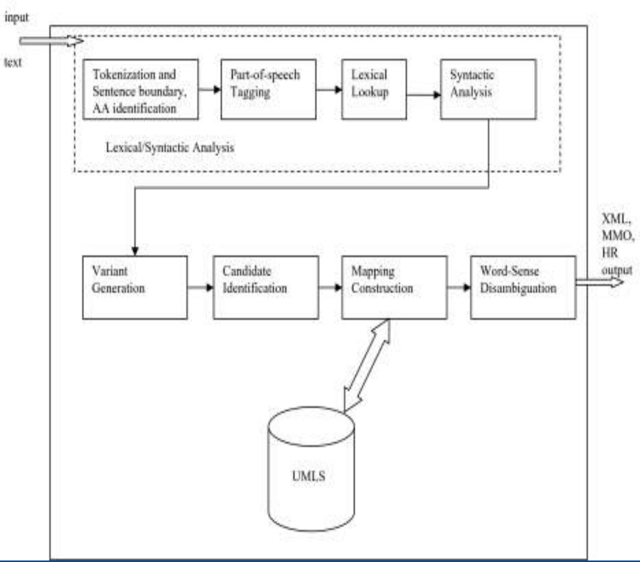
Conclusions: Clinical IE has been used for a wide range of applications, however, there is a considerable gap between clinical studies using EHR data and studies using clinical IE. This study enabled us to gain a more concrete understanding of the gap and to provide potential solutions to bridge this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

Clinical IE
- methodologies

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

Dictionary Lookup

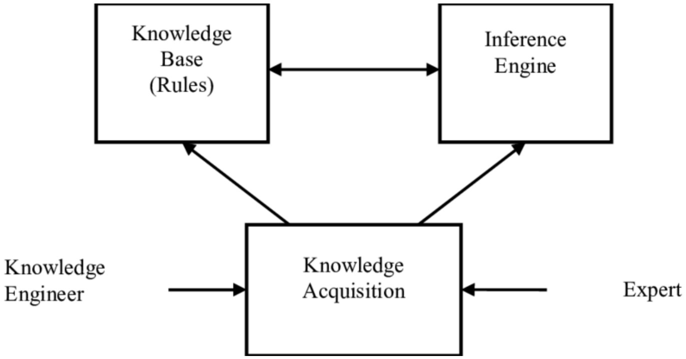
- Map medical text to the concepts in dictionary
- Dictionary resources
  - Existing: UMLS, SNOMED-CT, RxNorm, etc.
  - Custom-built
- Predefined concepts (by medical experts)
- Can follow the standard when using (inter)national resources
  - Enable interoperability between computer systems
- Tools: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



An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun

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

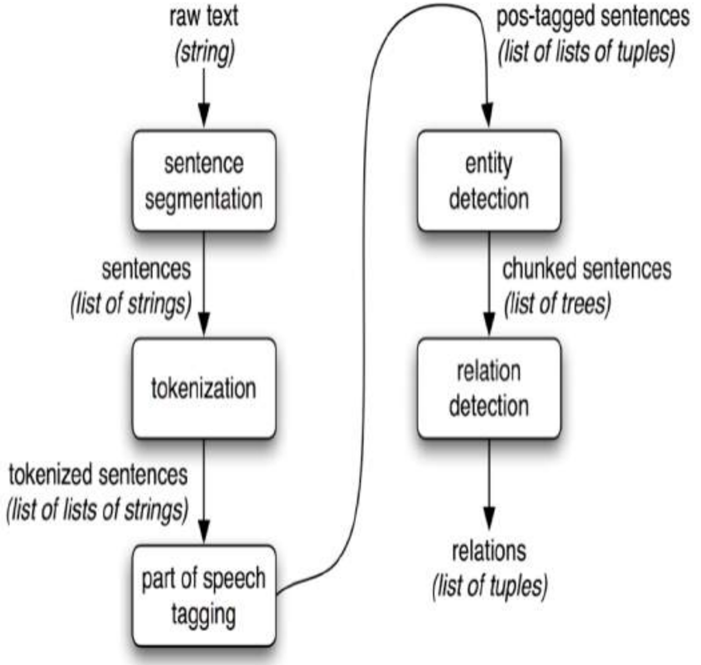


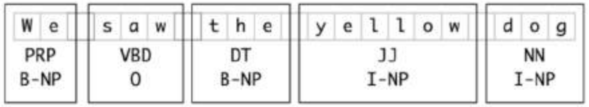
Chemical threats diagnosis expert system (CTDES) *in* Advanced Materials Research · November 2012

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

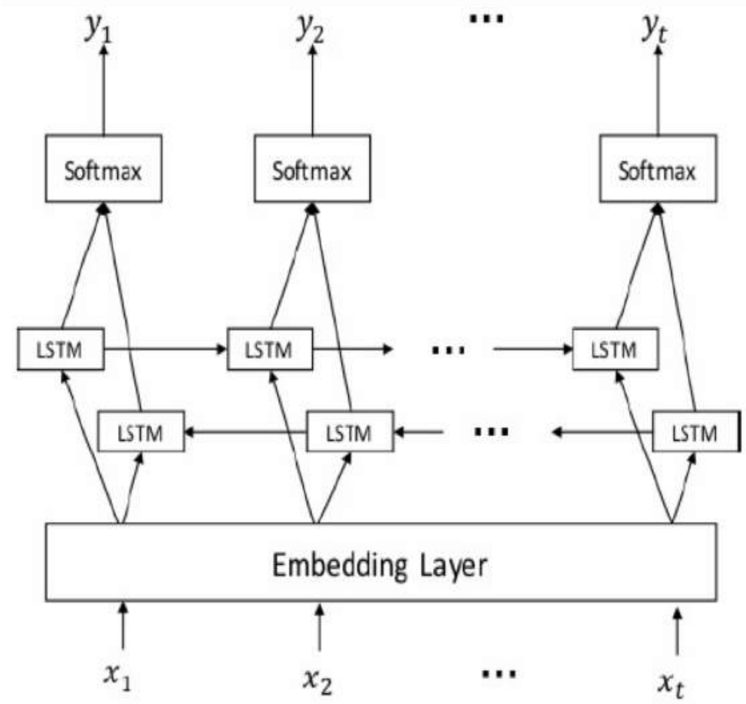




https://www.nltk.org

Deep Learning

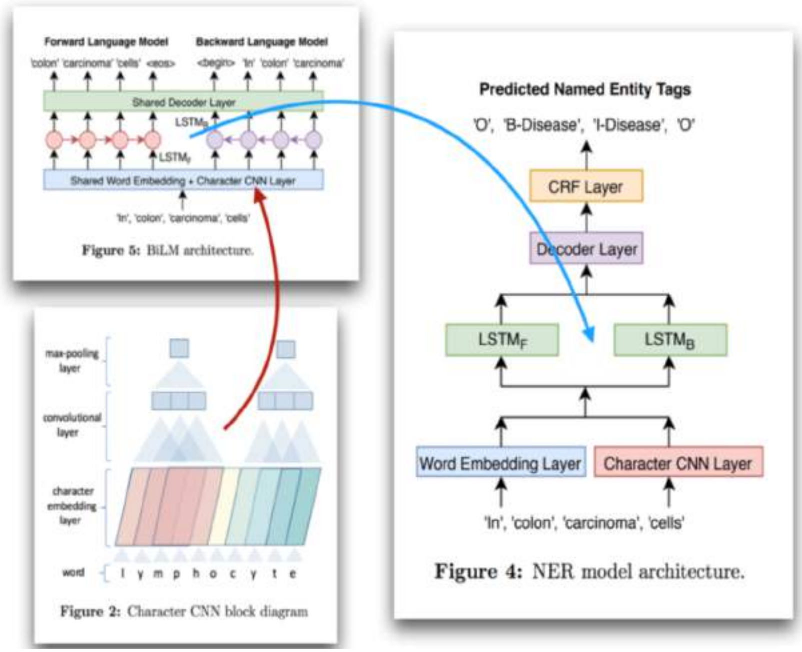
- No feature engineering
- Popular techniques
  - RNN (based on LSTM), CNN
- Applications
  - named entity recognition, relation extraction
- DL Information extraction: Words a sequence of tokens - embedding layer in RNN - softmax to classify token's entity



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?

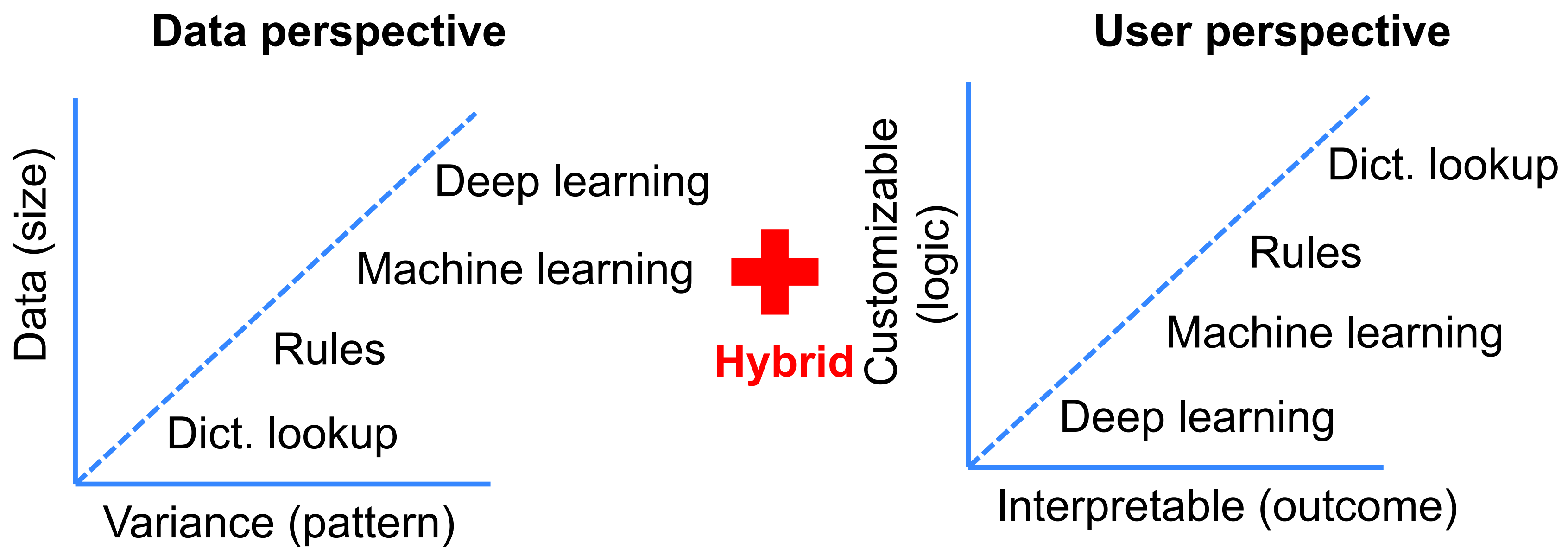
Not about selecting fancy technology, but about understanding the strengths / weaknesses and the nature of your project.

Rule based Input + Rule = Output

ML based Input + Output = Rule

MAYO CLINIC

Right approach? Cont.



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

word-level similarity

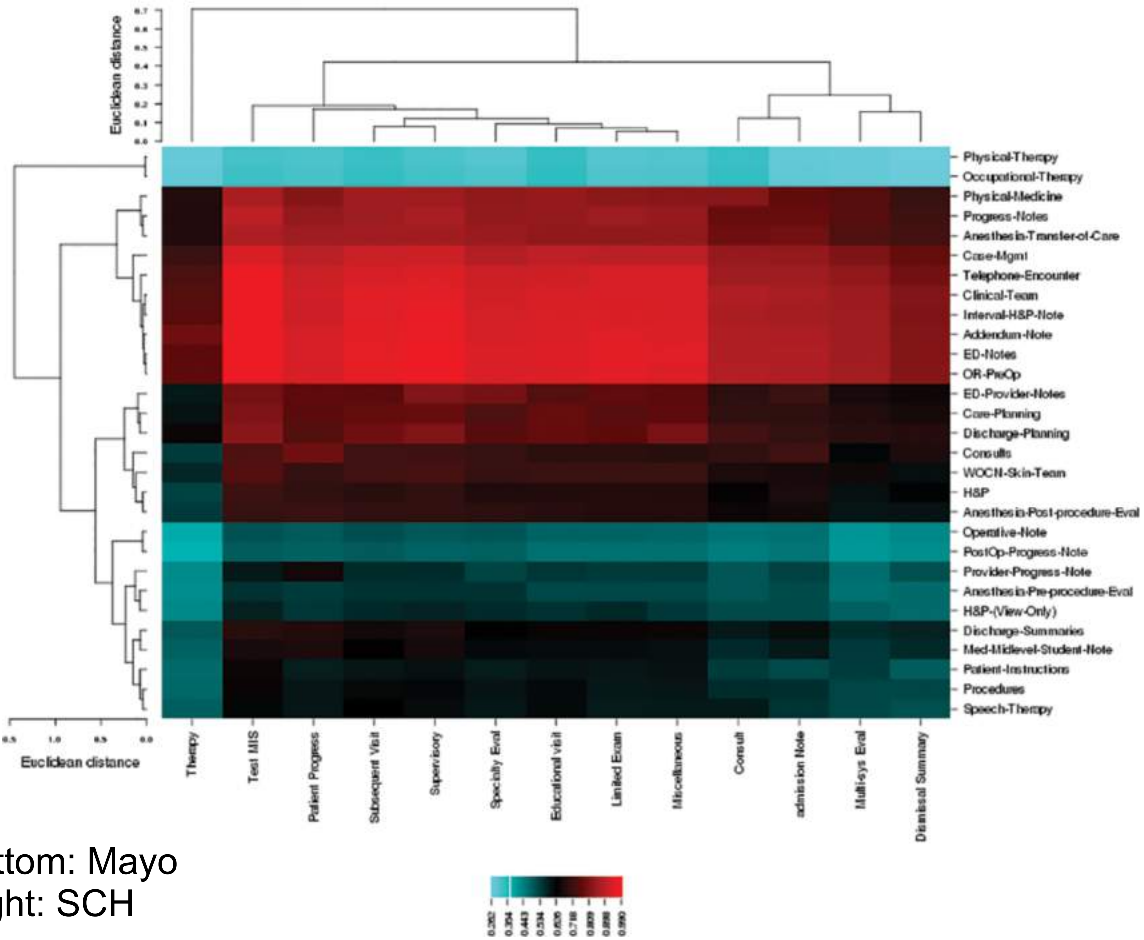
Data source	tf-idf	tf-ipf	topic
Whole corpus	0.669	0.581	0.944
Asthma-related concepts	0.971	0.855	NA

semantic similarity

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”

Bottom: Mayo
Right: SCH

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

Patients were considered to have definite asthma if a physician had made a diagnosis of asthma and/or if each of the following three conditions were present, and they were considered to have probable asthma if only the first two conditions were present:

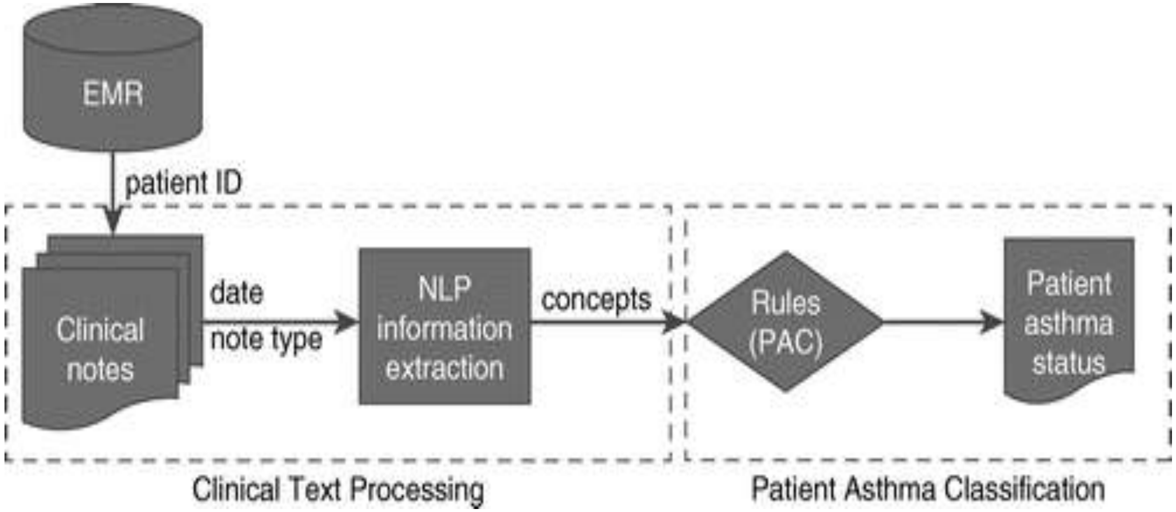
1. History of cough with wheezing, and/or dyspnea, OR history of cough and/or dyspnea plus wheezing on examination,
2. Substantial variability in symptoms from time to time or periods of weeks or more when symptoms were absent, and
3. Two or more of the following:
   - Sleep disturbance by nocturnal cough and wheeze
   - Nonsmoker (14 years or older)
   - Nasal polyps
   - Blood eosinophilia higher than 300/uL
   - Positive wheal and flare skin tests OR elevated serum IgE
   - History of hay fever or infantile eczema OR cough, dyspnea, and wheezing regularly on exposure to an antigen
   - Pulmonary function tests showing one FEV1 or FVC less than 70% predicted and another with at least 20% improvement to an FEV1 of higher than70% predicted OR methacholine challenge test showing 20% or greater decrease in FEV1
   - Favorable clinical response to bronchodilator

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

MAYO CLINIC

NLP-PAC

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

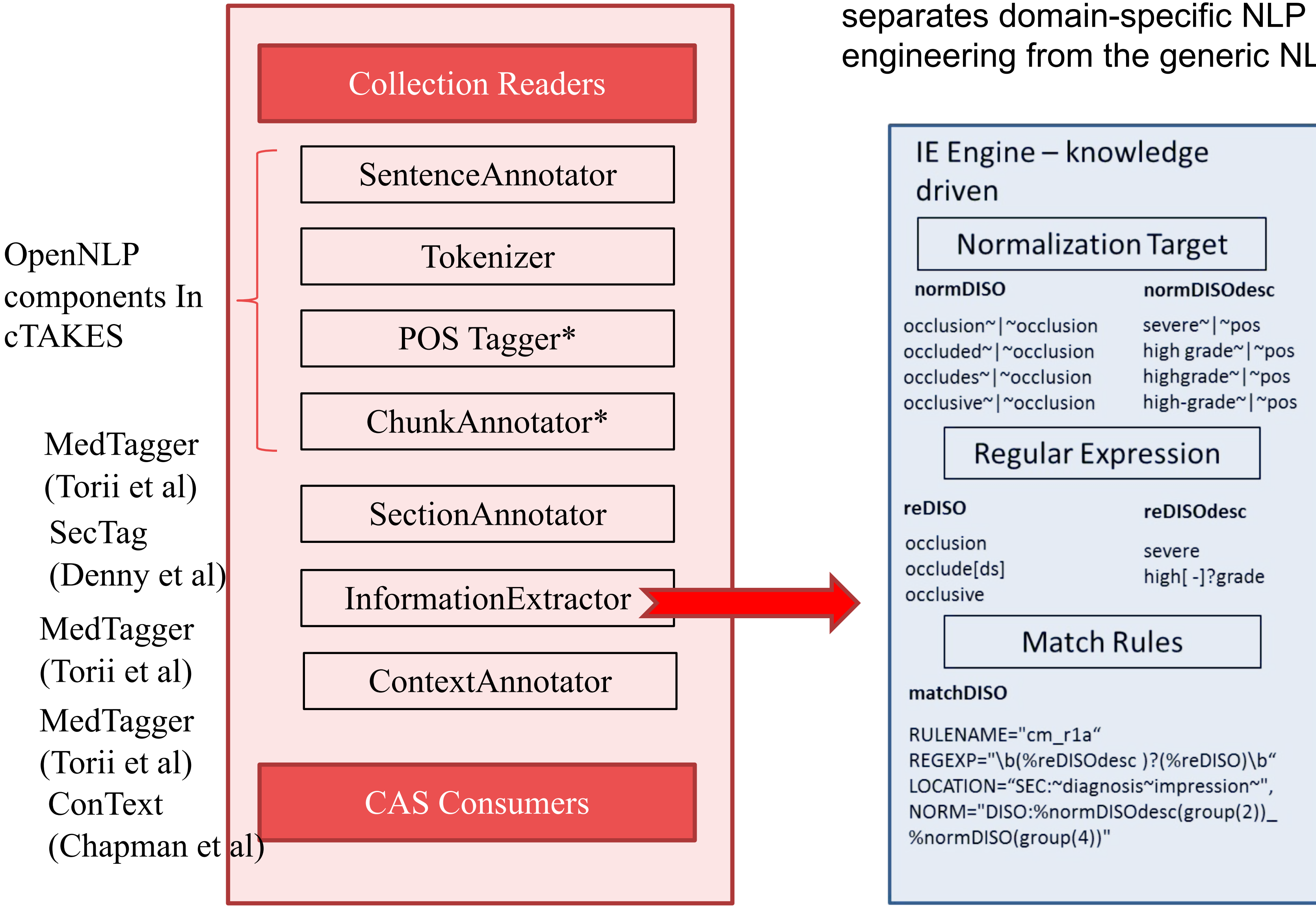


mcn|ASTHMA|2004****|PhD:
C1:C2|<PhD>docname::2011
3::Asthma::#1
Asthma<C1>docname::2011
2::wheezing::Brief summary
of clinical history and reason
for ED eval: ….wheezing,
respiratory distress and
respiratory rate in the 90's.

Application of a Natural Language Processing Algorithm to Asthma Ascertainment. An Automated Chart Review. Am J Respir Crit Care Med. 2017 Aug 15; 196 (4):430-437.

MedTagger

separates domain-specific NLP knowledge engineering from the generic NLP process

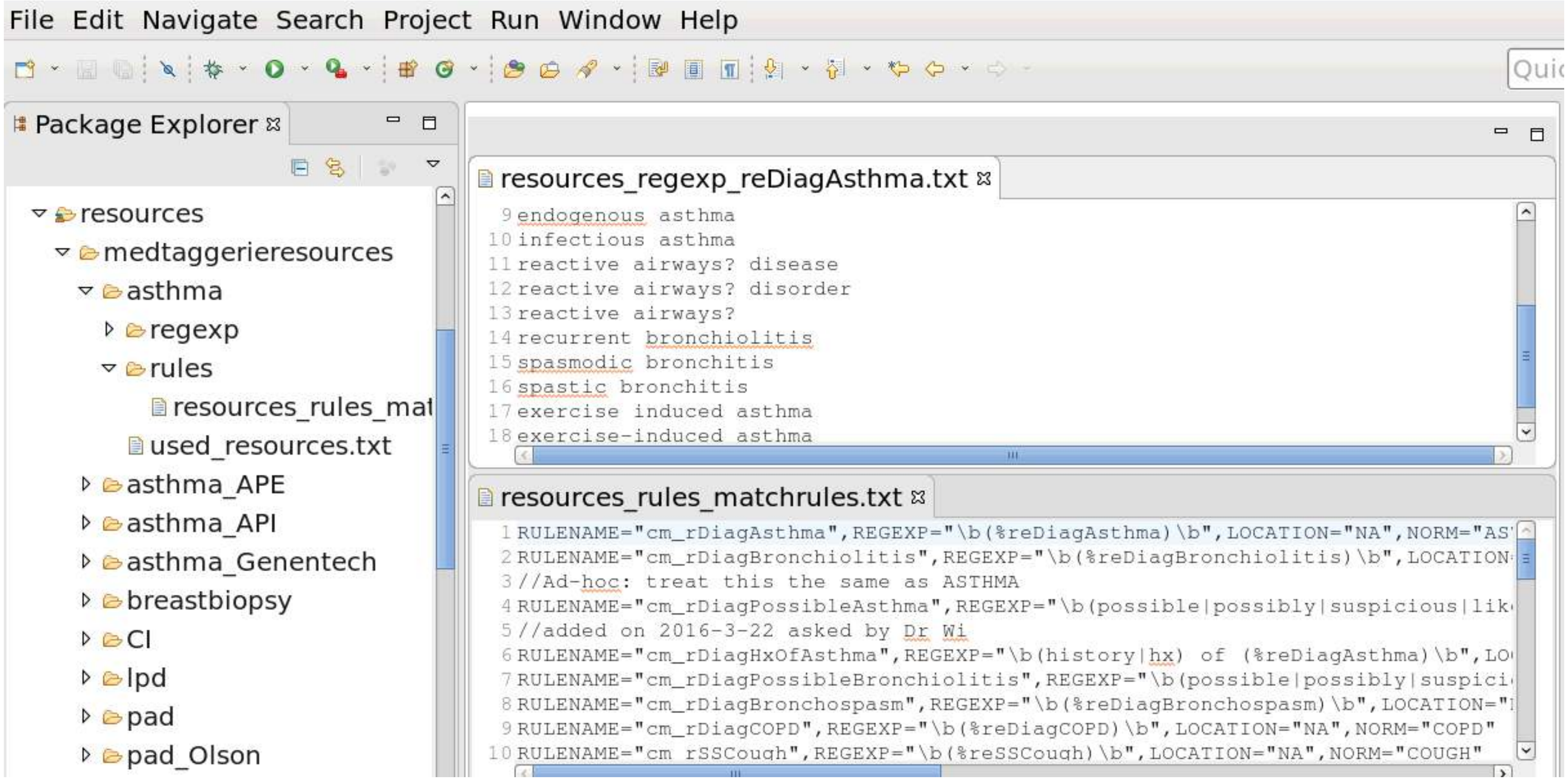


MAYO CLINIC

* Indicates optional components

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

Out-of-box

Refinement

Metrics	Mayo (N=497)	SCH stage 1 (prototype, N=298)	SCH stage 2 (refinement, N=298)
sensitivity	0.972	0.840	0.920
specificity	0.957	0.924	0.964
PPV	0.905	0.788	0.896
NPV	0.988	0.945	0.973
F-score	**0.937**	**0.813**	**0.908**

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

Volume 25, Issue 3
March 2018

Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions FREE

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

Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 353–359, https://doi.org/10.1093/jamia/ocx138

MAYO CLINIC

Patient Cohort Retrieval using Electronic Health Records

Yanshan Wang

Research Associate

Department of Health Sciences Research

Mayo Clinic

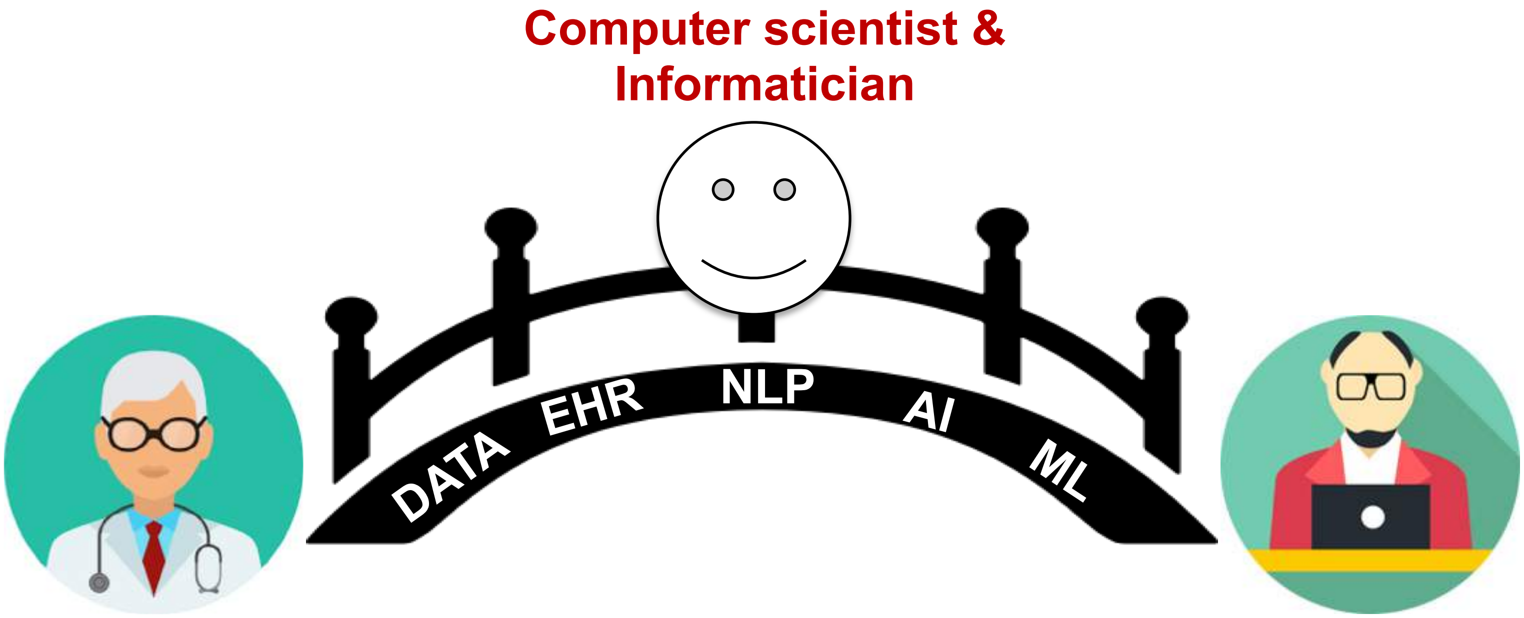
Agenda

- **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)

MAYO CLINIC

Introduction

- **What do we do?**



- **How can you contact me?**
  - Email: wang.yanshan@mayo.edu
  - LinkedIn, Twitter (yanshan_wang)

MAYO CLINIC

Why take this tutorial?

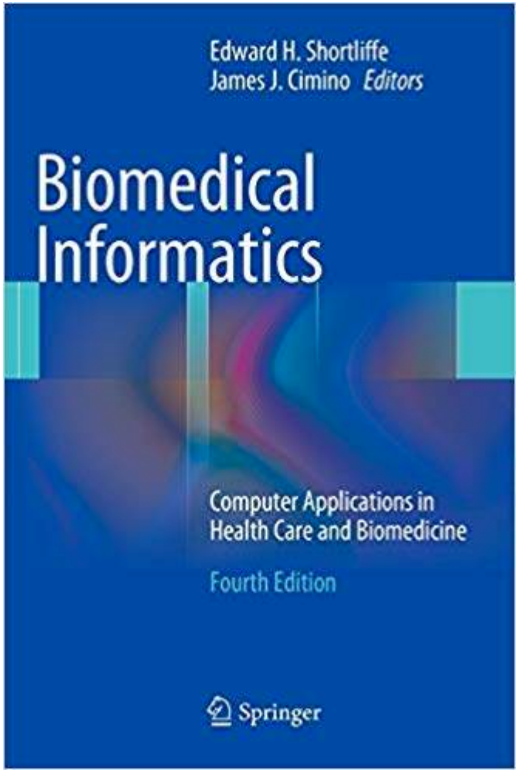
- **Patient cohort retrieval is still labor expensive today.**
- **Most information is embedded in unstructured EHRs.**
- **Natural language processing is under-utilized for cohort retrieval.**

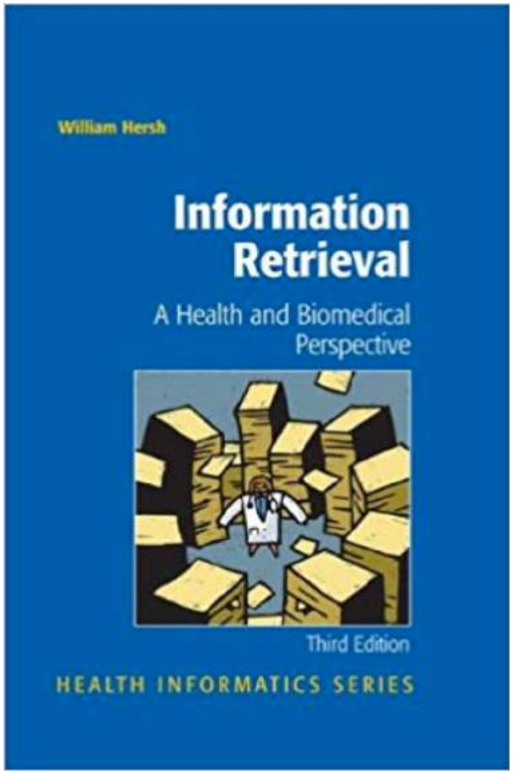
Goal of this tutorial

- 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**





Suggested reading

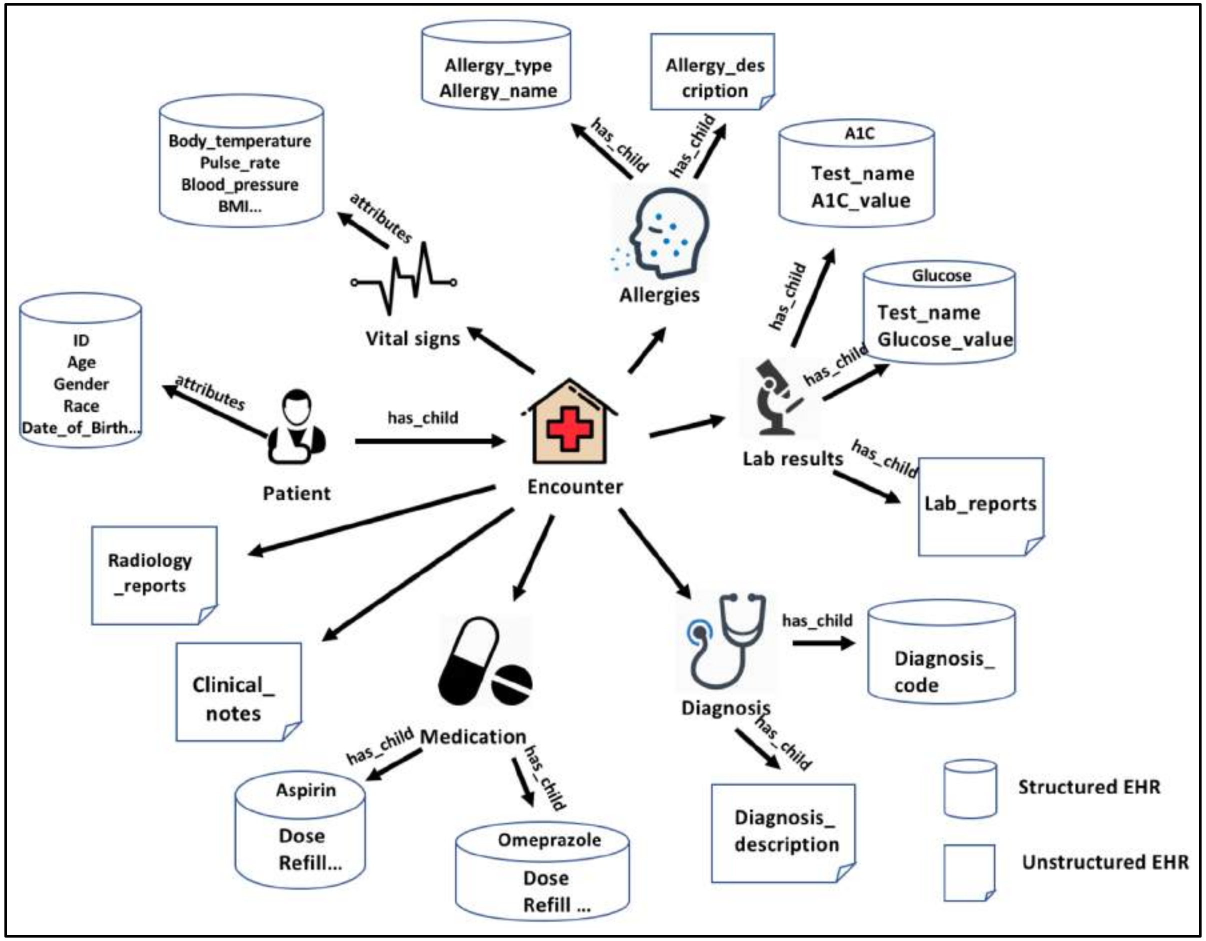
- **Papers**
  - A review of approaches to identifying patient phenotype cohorts using electronic health records. Shivade et al. 2013.
  - Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. Miotto et al. 2015.
  - A survey of practices for 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**



MAYO CLINIC

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].

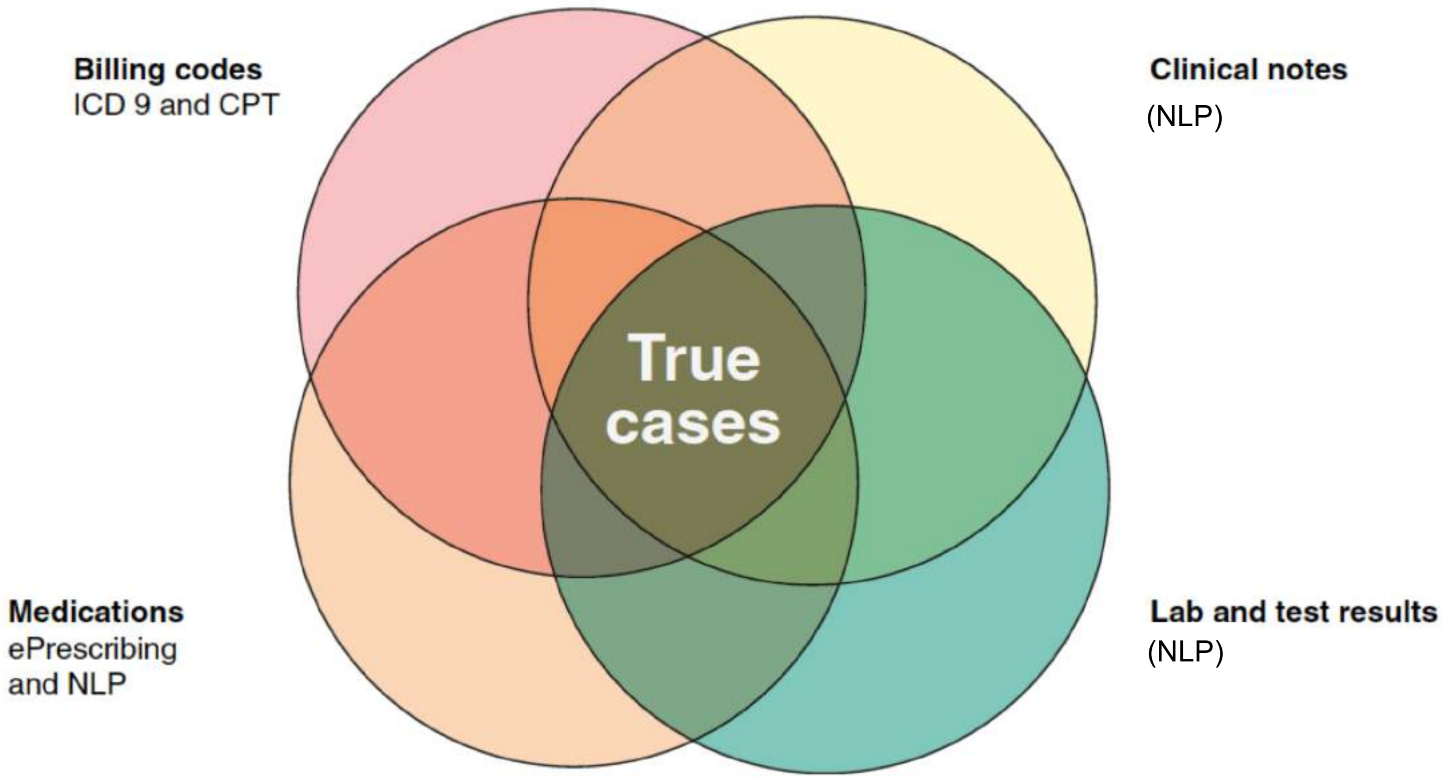
1. eMERGE network.
2. Wei, W. Q., & Denny, J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine, 7(1), 41.

MAYO CLINIC

Basic Concepts

-

-



Source: Wei, W. Q., & Denny, J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine, 7(1), 41.

MAYO CLINIC

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.

MAYO CLINIC

Basic Concepts

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

Criteria ***clinicaltrials.gov***

Inclusion Criteria:

- Alzheimer's disease (CDR 0.5, 1, & 2)
- Active study partner
- BMI > 21
- English speaking

Exclusion Criteria:

- BMI < 21
- Consume greater than 14 drinks of alcohol per week
- Insulin Dependent Diabetes Mellitus
- Diagnosis of active cancer
- Myocardial infarction or symptoms of coronary artery disease (e.g. angina) in last year

https://clinicaltrials.gov/ct2/show/NCT03690193?cond=alzheimer%27s+disease&rank=5

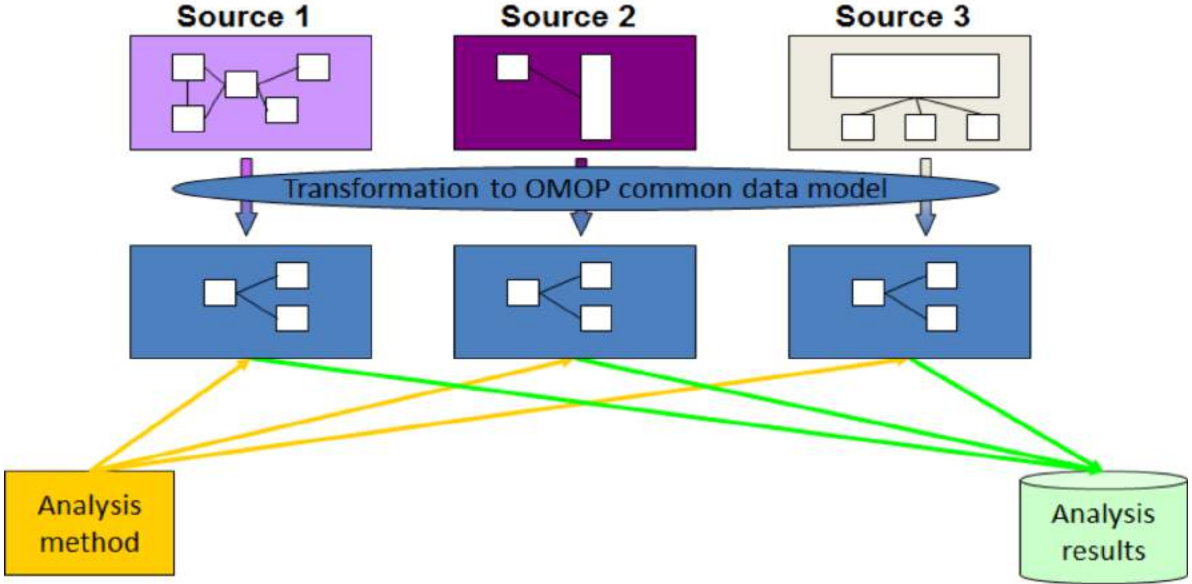
Basic Concepts

- **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

MAYO CLINIC

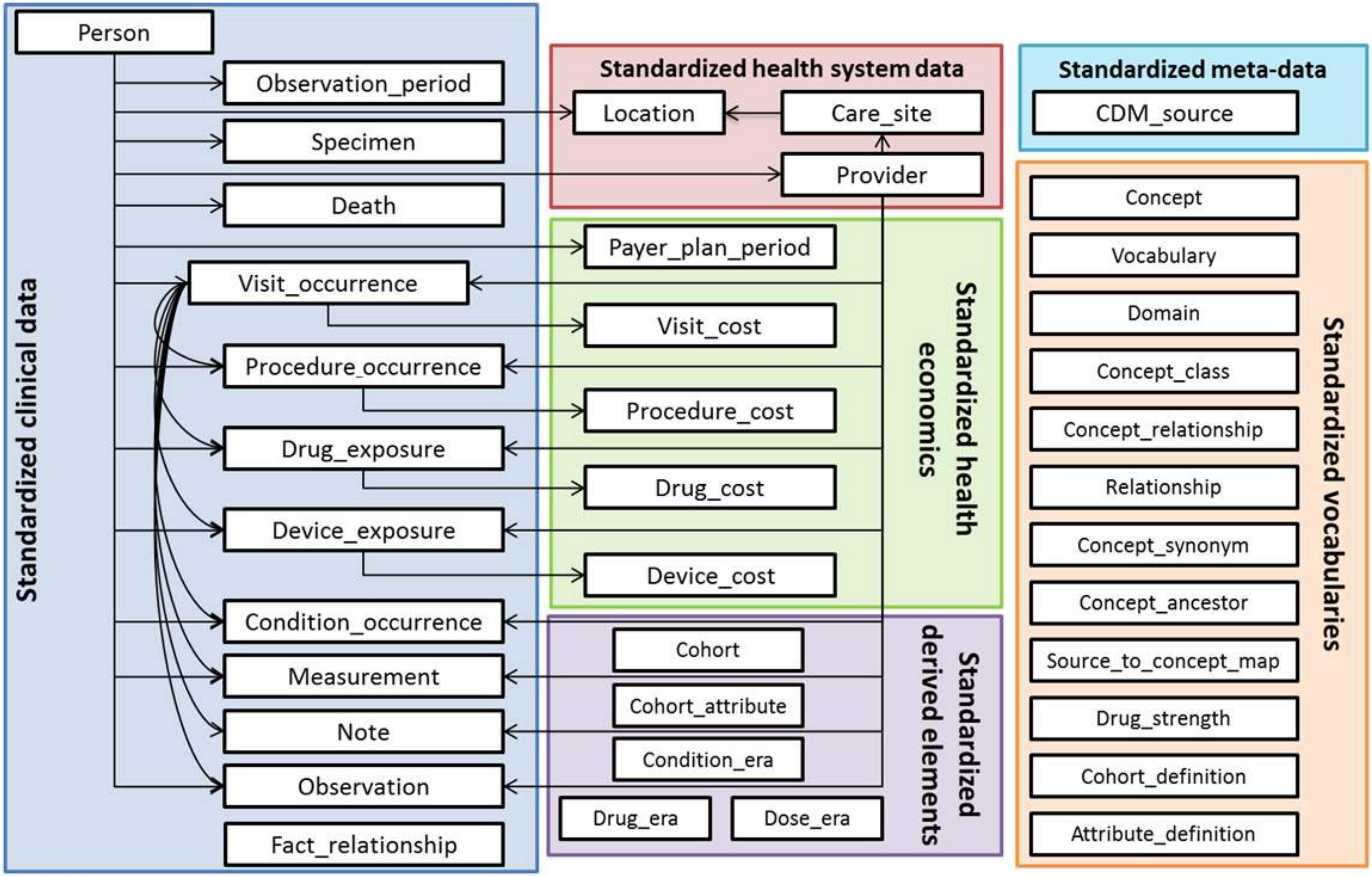
Basic Concepts

- **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)



MAYO CLINIC

OMOP CDM v. 5.0



Source: https://www.ohdsi.org/data-standardization/the-common-data-model/

MAYO CLINIC

Patient Cohort Retrieval for Clinical Trials using NLP

MAYO CLINIC

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 eligible[1]

<5%

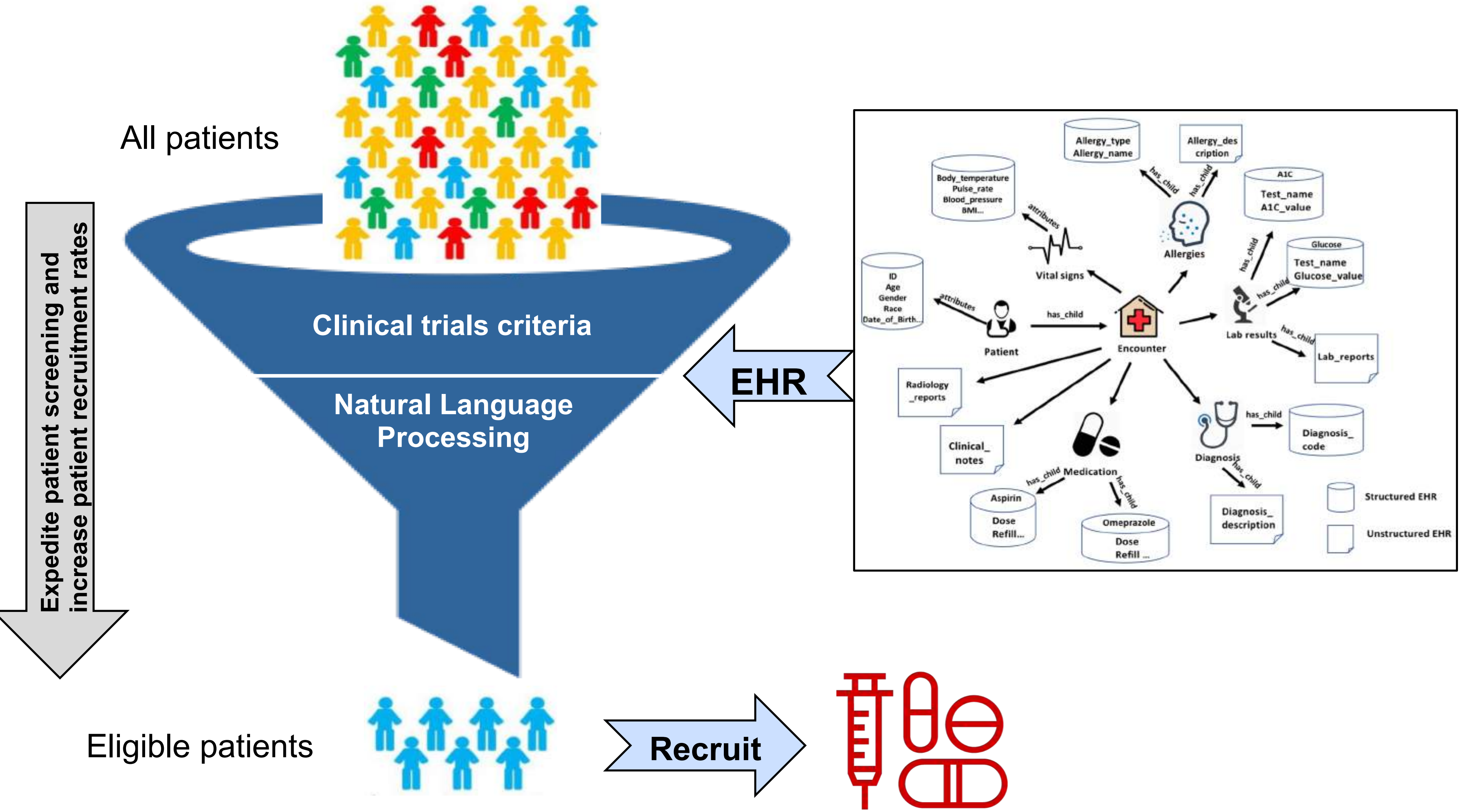
Cancer patients are participate[1]

85%

Clinical trials fail to retain enough patients[2]

1. Haddad TC, Helgeson J, Pomerleau K, Makey M, Lombardo P, Coverdill S, Urman A, Rammage M, Goetz MP, LaRusso N. Impact of a cognitive computing clinical trial matching system in an ambulatory oncology practice. American Society of Clinical Oncology; 2018.
2. Cote DN. Minimizing Trial Costs by Accelerating and Improving Enrollment and Retention. Global Clinical Trials for Alzheimer's Disease: Elsevier; 2014. p. 197-215.

NLP for Clinical Trials Eligibility Screening

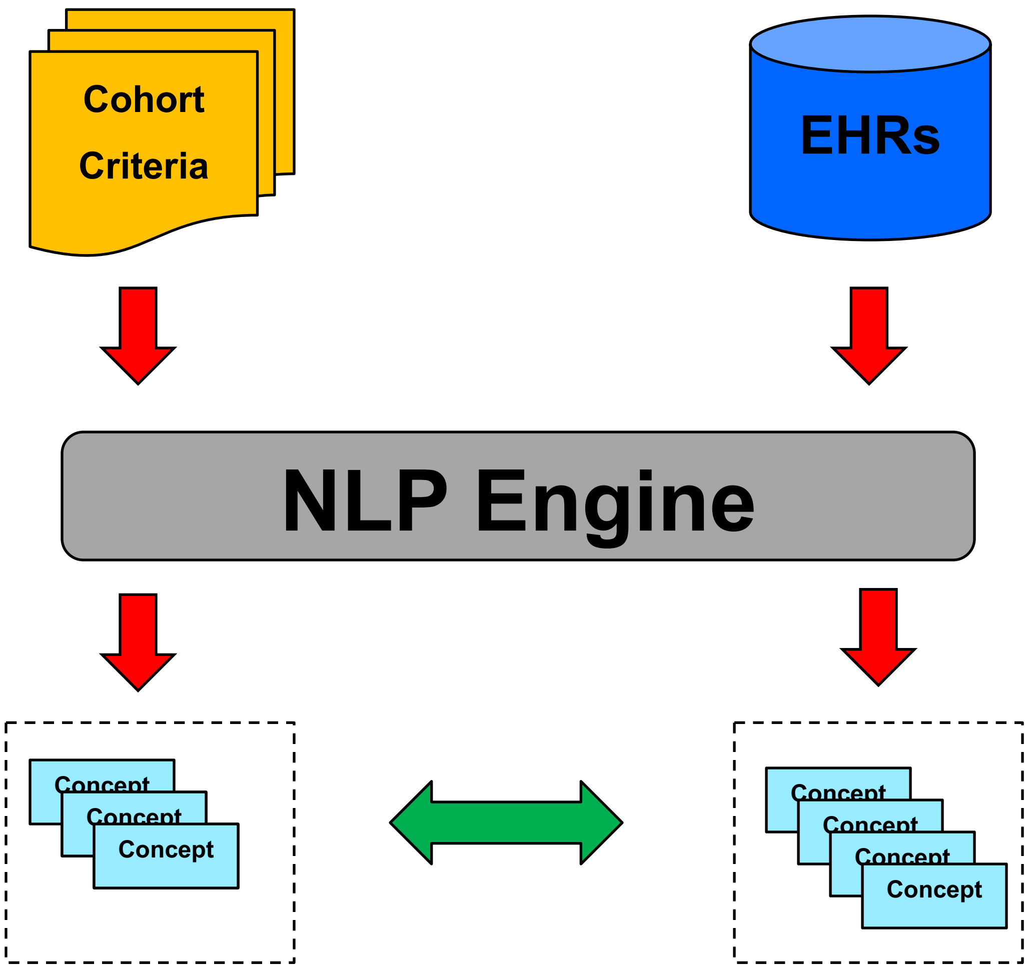


NLP Approaches for Cohort Retrieval

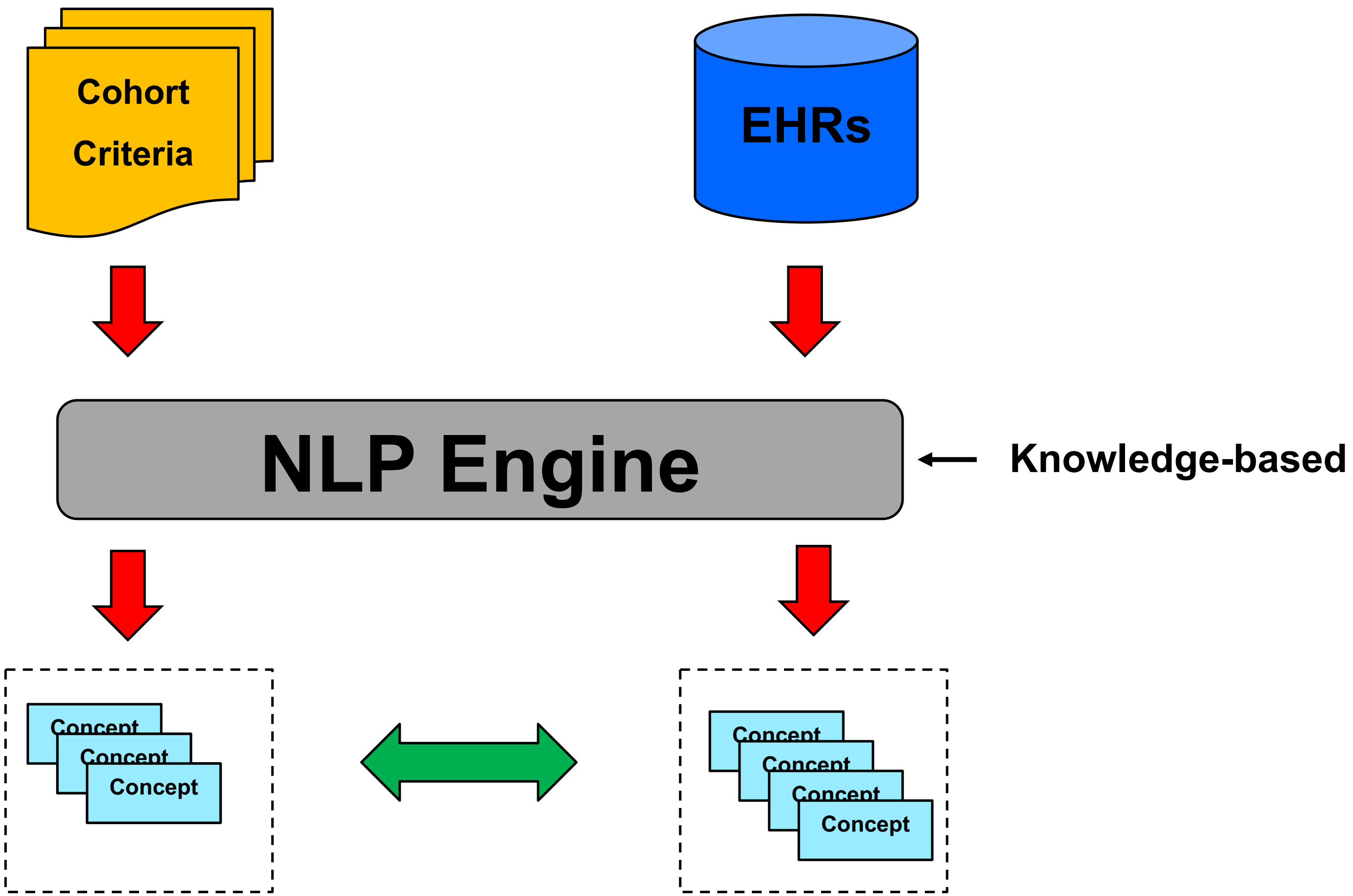
- Medical Concept Embedding
- Information Retrieval
- Deep Patient Representation

MAYO CLINIC

Medical Concept Embedding



Medical Concept Embedding



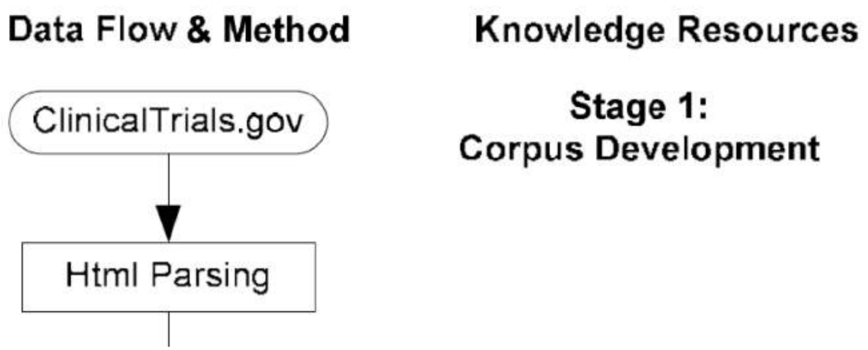
MAYO CLINIC

Medical Concept Extraction and Representation

- Luo, Zhihui, et al. "Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS." Summit on Translational Bioinformatics 2010 (2010): 26.
- Weng, Chunhua, et al. "EliXR: an approach to eligibility criteria extraction and representation." Journal of the American Medical Informatics Association 18.Supplement_1 (2011): i116-i124.

Semantic Lexicon Extraction

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

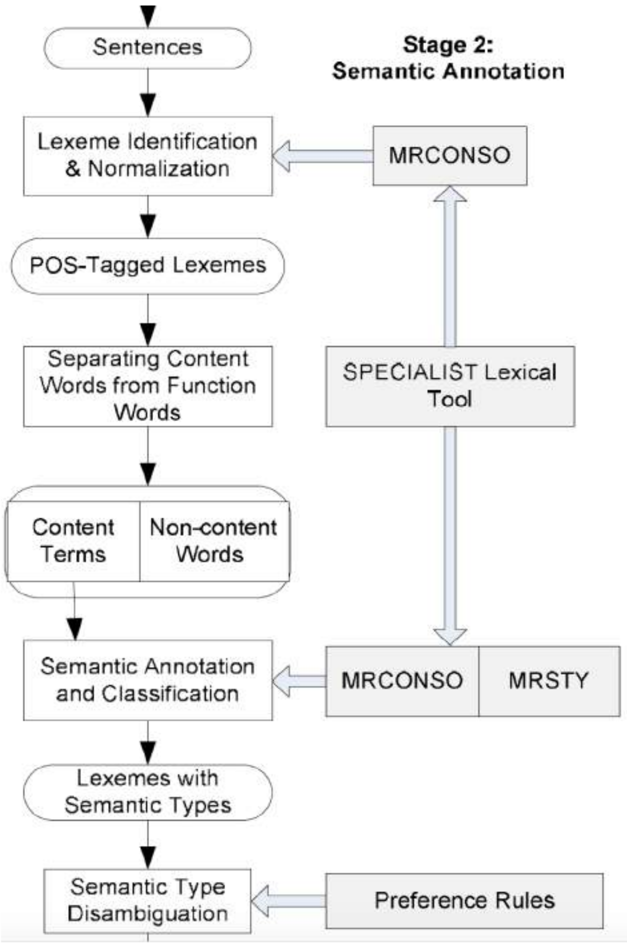


Luo, Zhihui, et al. "Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS." Summit on Translational Bioinformatics 2010 (2010): 26.

MAYO CLINIC

Semantic Lexicon Extraction

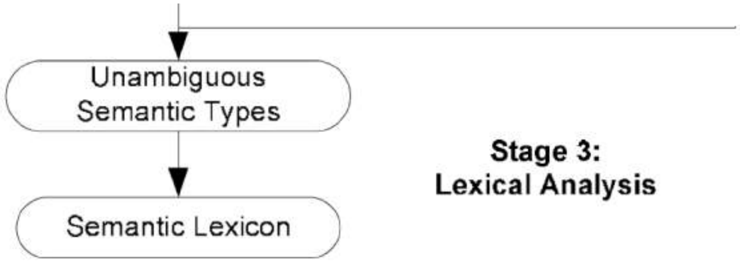
- **UMLS-based lexicon discovery from text**
  - Processed the corpus to identify UMLS-recognizable semantic units that matched the medical concepts in the MRCONSO table of the Metathesaurus of UMLS.
  - Used the Stanford tagger and the Penn Treebank tag set for part-of-speech (POS) tagging. All words 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.



Luo, Zhihui, et al. "Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS." Summit on Translational Bioinformatics 2010 (2010): 26.

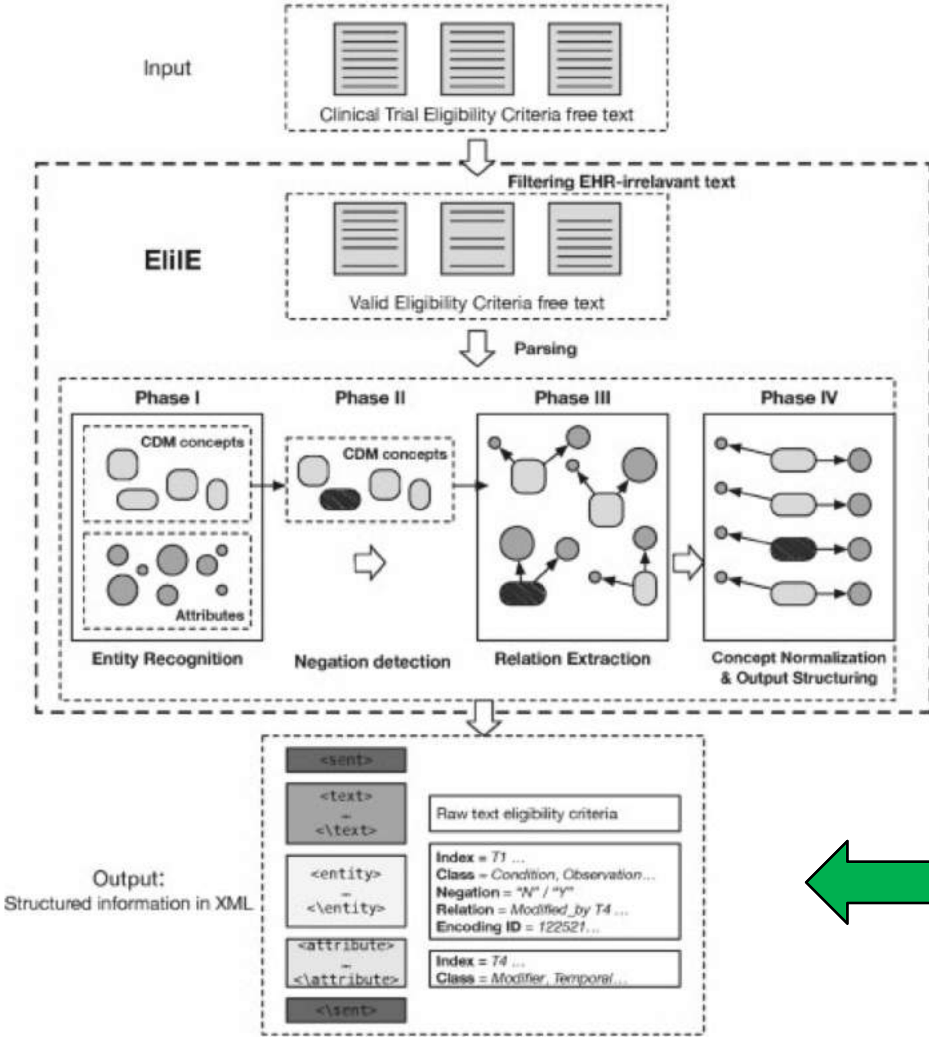
Semantic Lexicon Extraction

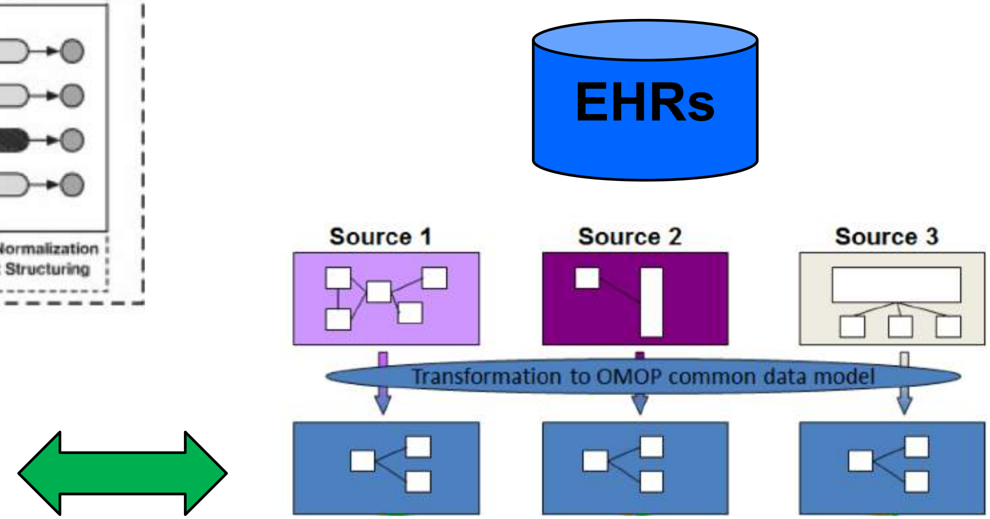
- **UMLS-based lexicon discovery from text**
  - Investigated the coverage of the sample corpus provided by annotation procedure, using the Metathesaurus, Semantic Network, and preference rules.



Luo, Zhihui, et al. "Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS." Summit on Translational Bioinformatics 2010 (2010): 26.

Eligibility criteria extraction and representation using CDM





Kang, Tian, et al. "EliIE: An open-source information extraction system for clinical trial eligibility criteria." Journal of the American Medical Informatics Association 24.6 (2017): 1062-1071.

Limitations of Concept Embedding

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

MAYO CLINIC

Deep Patient Representation

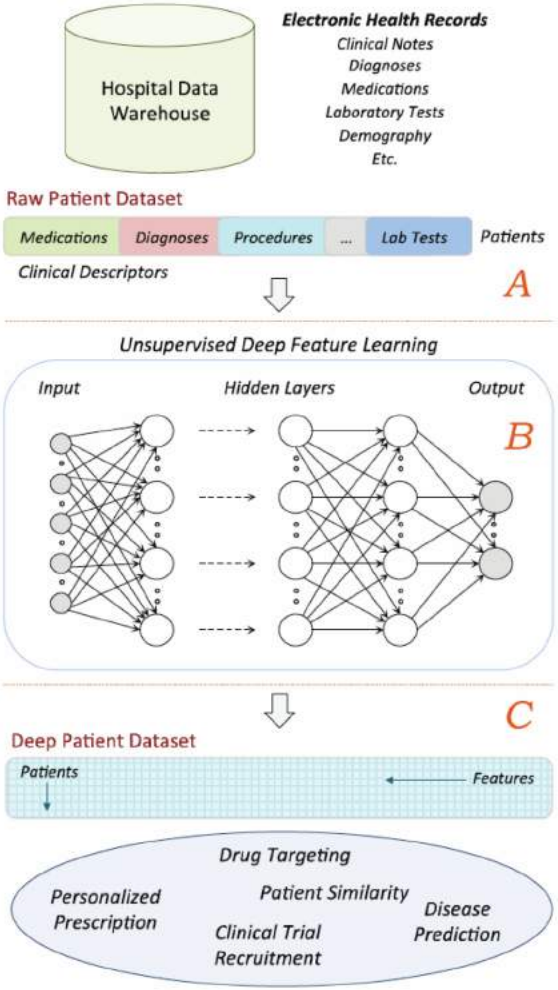
- Miotto, Riccardo, et al. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records." Scientific reports 6 (2016): 26094.
- Rajkomar, Alvin, et al. "Scalable and accurate deep learning with electronic health records." NPJ Digital Medicine 1.1 (2018): 18.

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



Slide courtesy of Dr. Miotto

Deep Patient: Learning

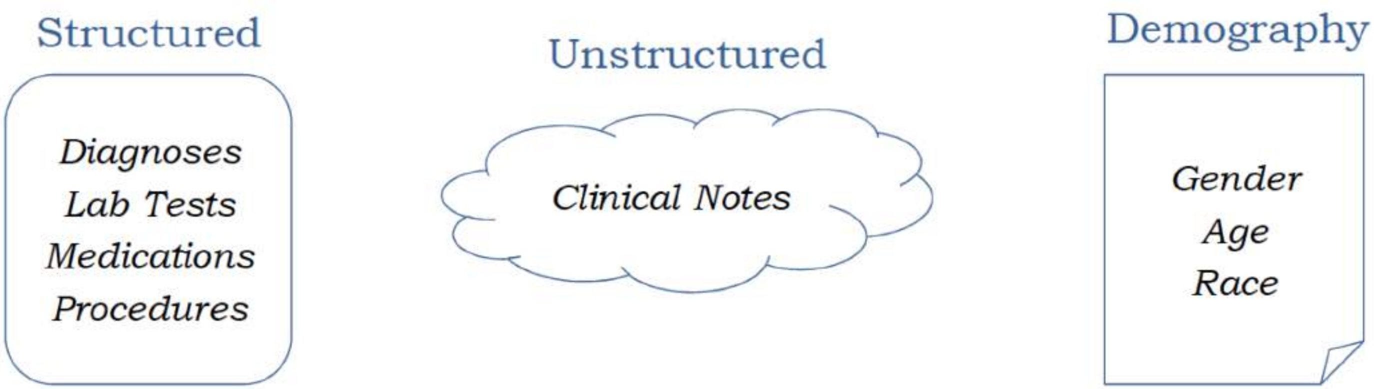
- Multi-layer neural network
  - ✓ each layer of the network produces a higher-level representation of the observed patterns, based on the data it receives as input from the layer below, by optimizing a local unsupervised criterion

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

Slide courtesy of Dr. Miotto

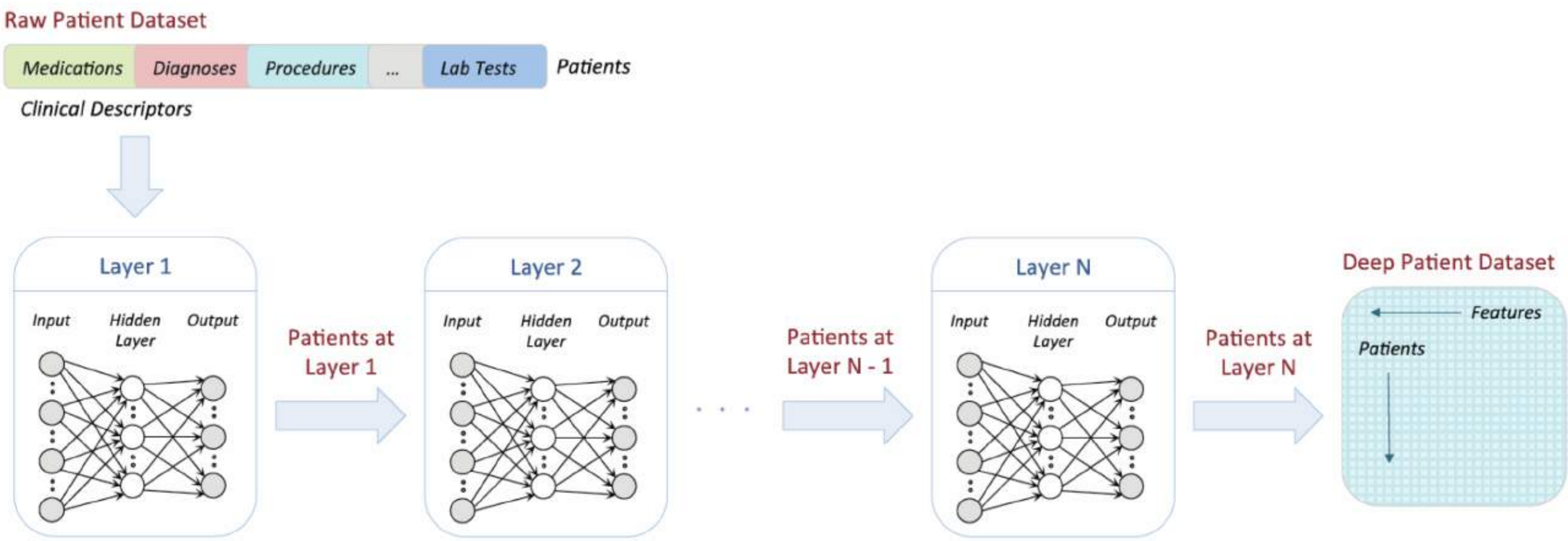
Deep Patient: Data Processing

- Patients data available in the data warehouse



- 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*

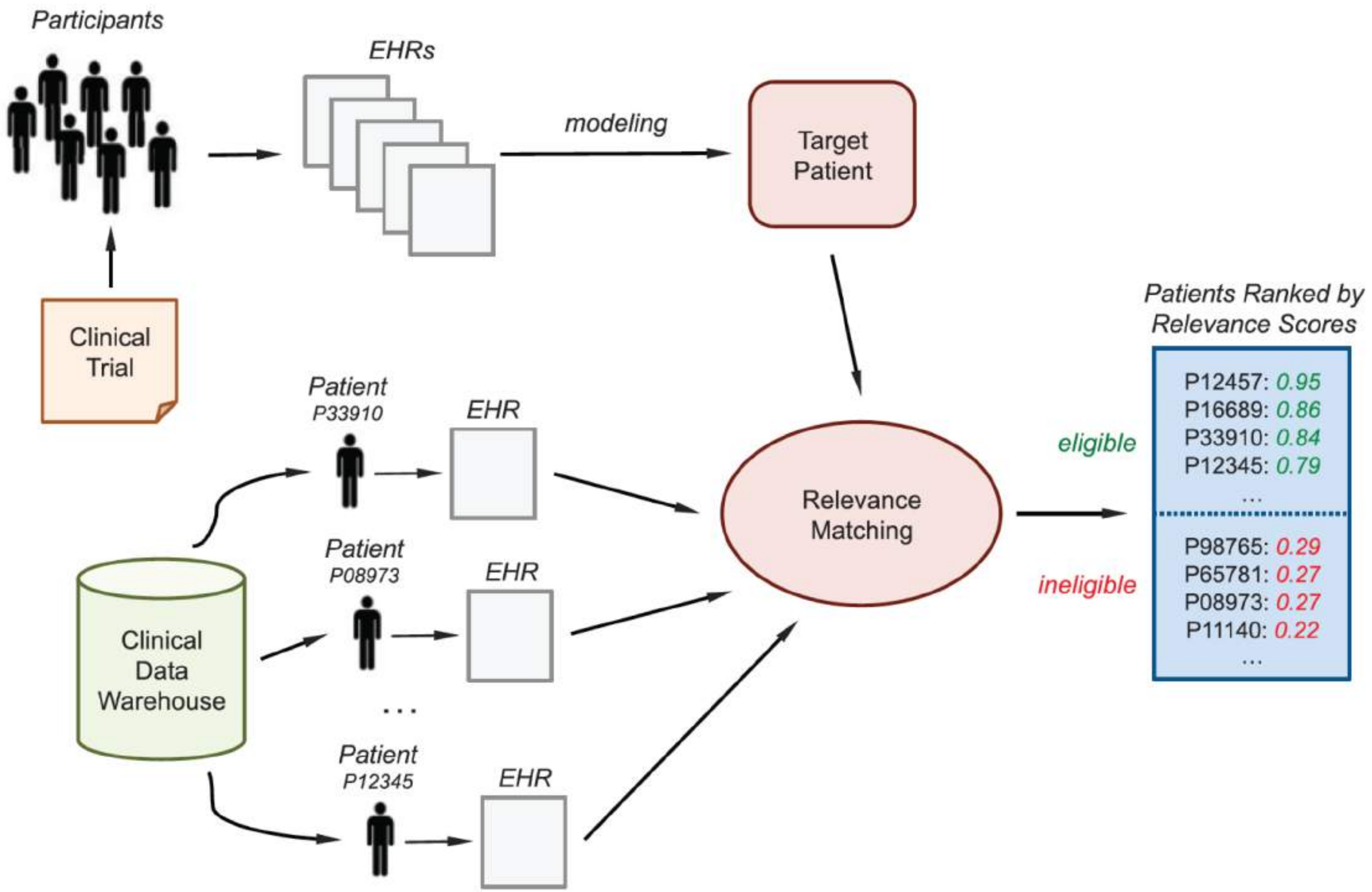
Deep Patient: Architecture



- 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

Slide courtesy of Dr. Miotto

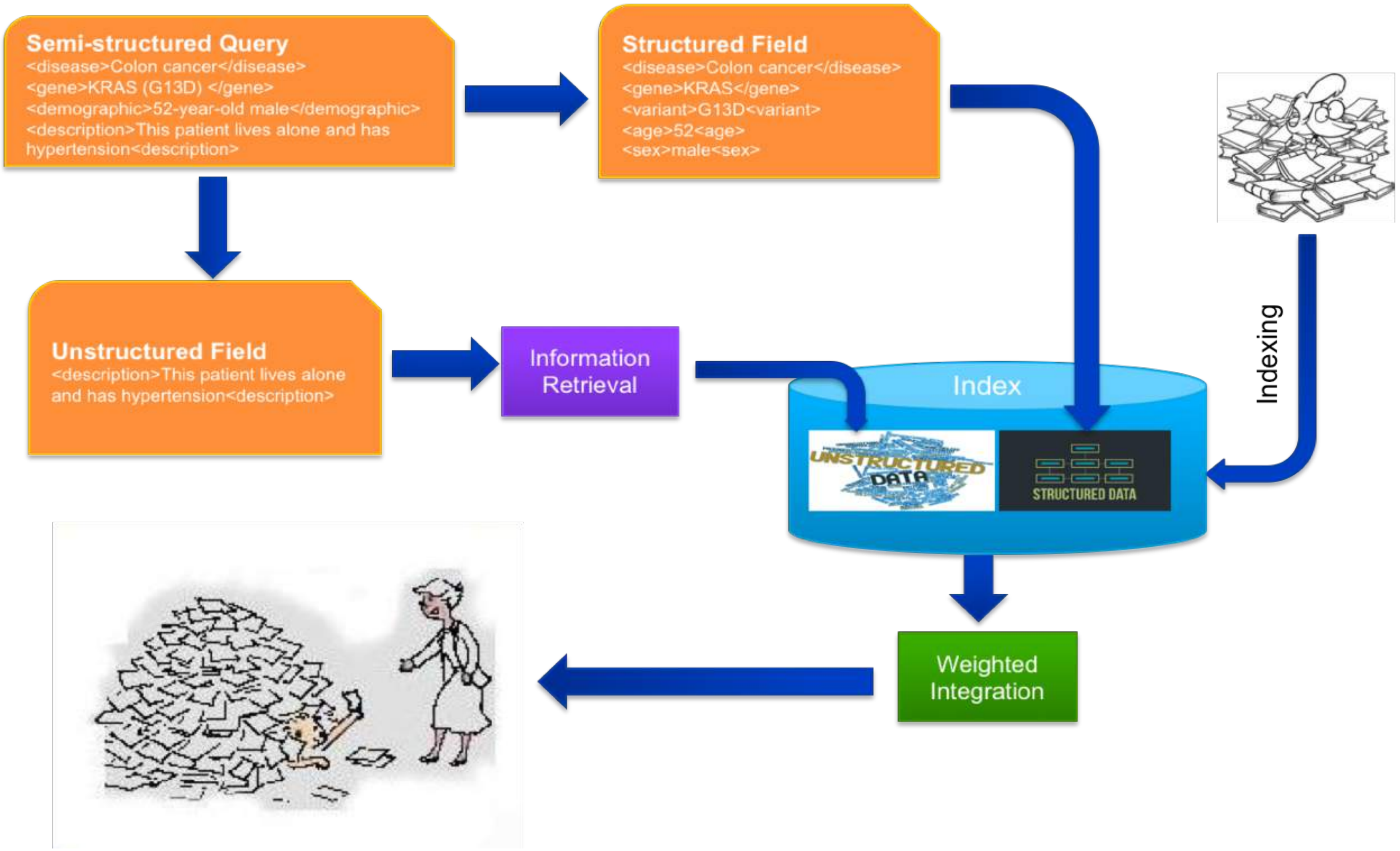
Deep Patient Representation for Cohort Retrieval



Miotto, Riccardo, and Chunhua Weng. "Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials." Journal of the American Medical Informatics Association 22.e1 (2015): e141-e150.

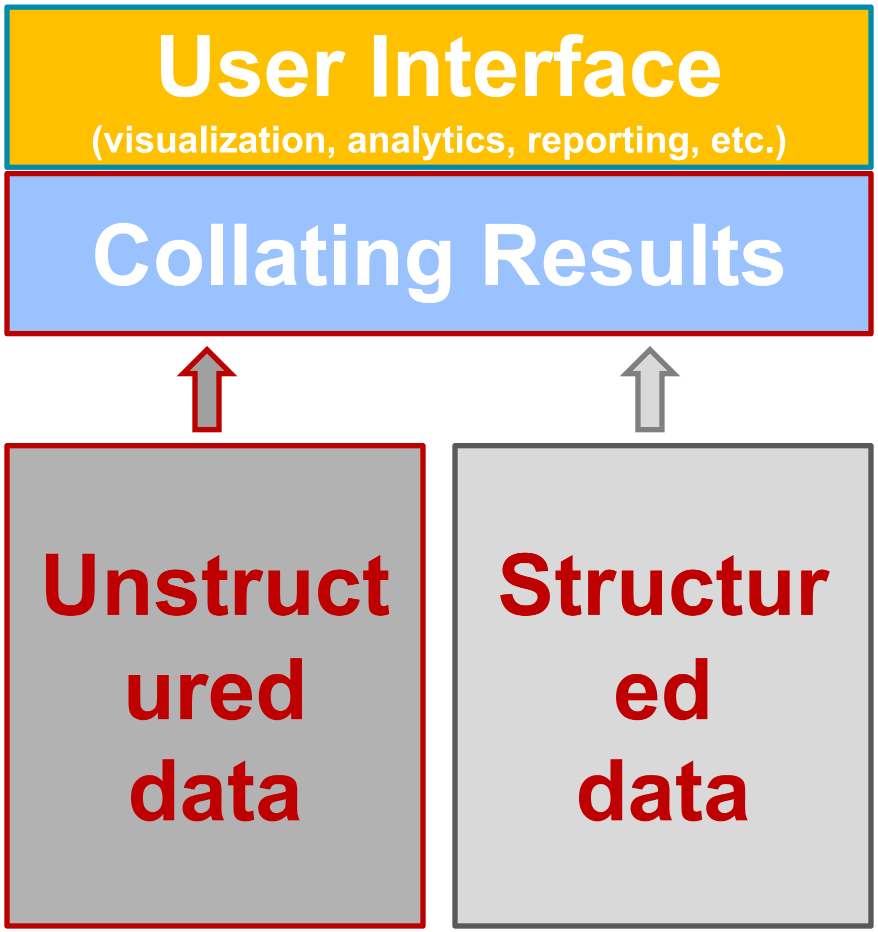
MAYO CLINIC

Information Retrieval for Cohort Retrieval

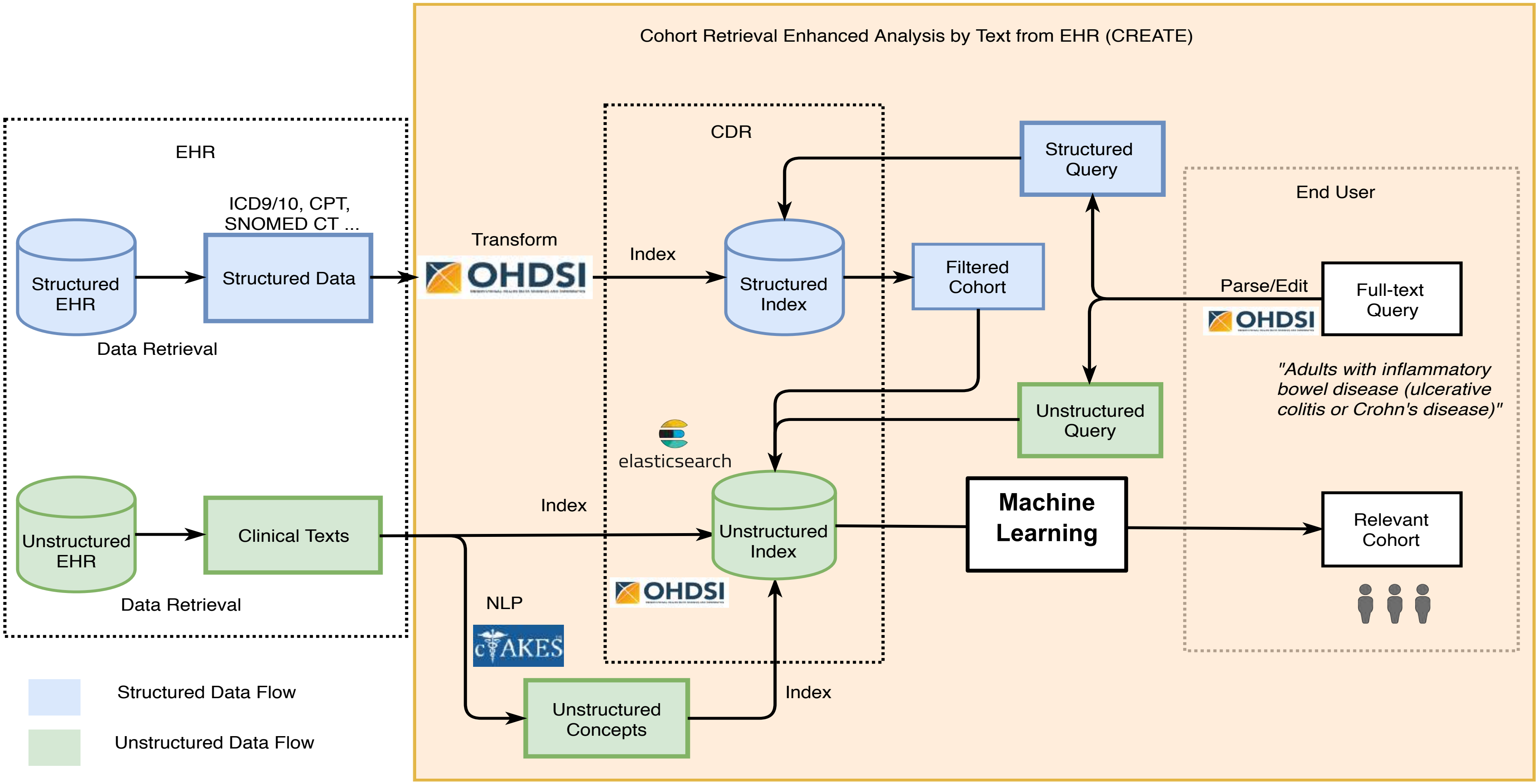


Wang et al. "Test Collections for EHR-based Clinical Information Retrieval." JAMIA Open. In press.

IR Tool for Cohort Retrieval

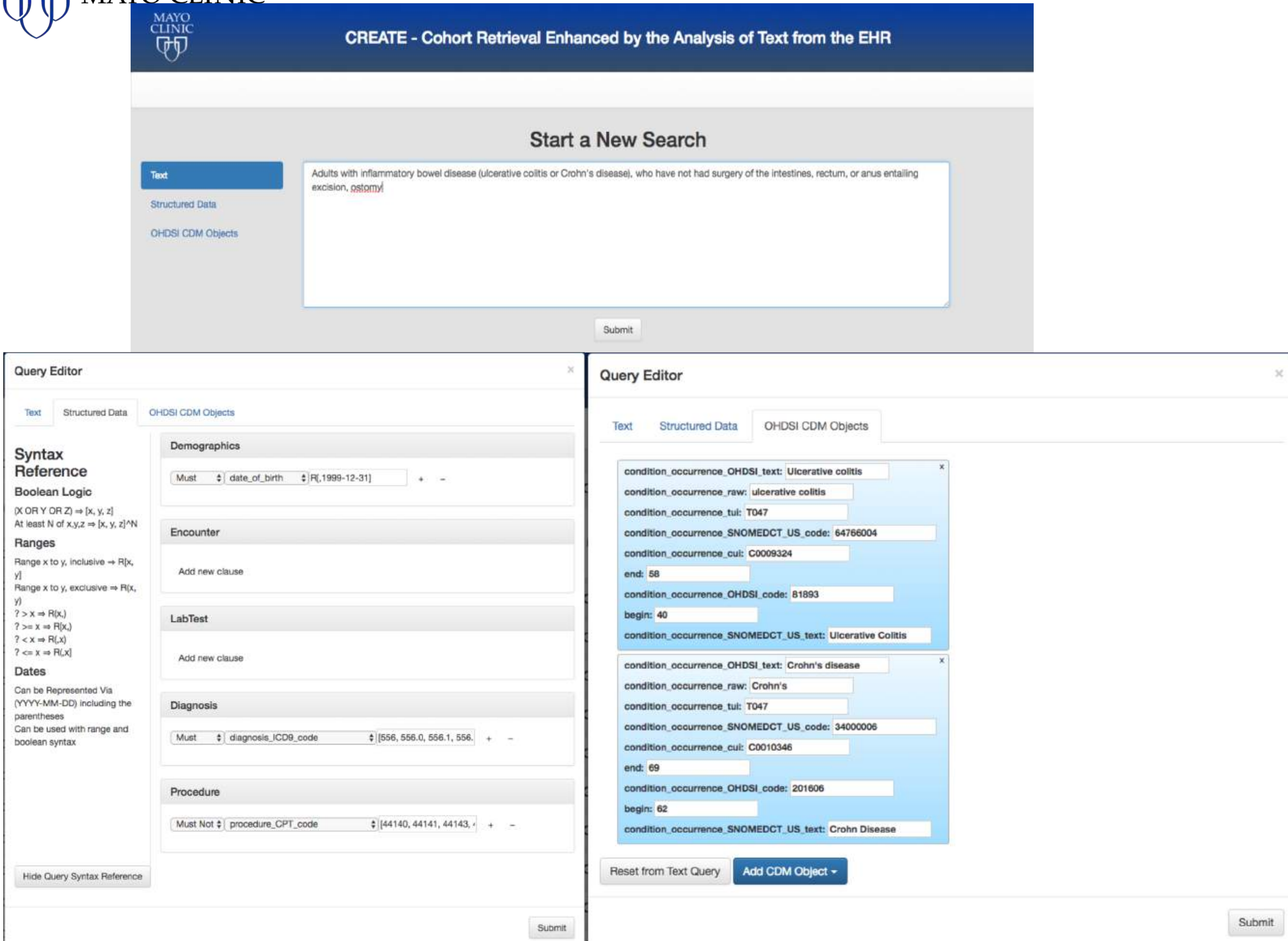


IR for Cohort Retrieval Prototype: CREATE



Liu, Sijia, et al. "CREATE: Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records using OMOP Common Data Model." arXiv preprint arXiv:1901.07601 (2019).

MAYO CLINIC


MAYO CLINIC
CREATE - Cohort Retrieval Enhanced by the Analysis of Text from the EHR
Start a New Search
Text
Structured Data
OHDSI CDM Objects
Adults with inflammatory bowel disease (ulcerative colitis or Crohn's disease), who have not had surgery of the intestines, rectum, or anus entailing excision, ostomy
Submit
Query Editor
Text
Structured Data
OHDSI CDM Objects
Syntax Reference
Boolean Logic
(X OR Y OR Z) ⇒ [x, y, z]
At least N of x,y,z ⇒ [x, y, z]^N
Ranges
Range x to y, inclusive ⇒ R[x, y]
Range x to y, exclusive ⇒ R(x, y)
? > x ⇒ R(x,)
? >= x ⇒ R[x,)
? < x ⇒ R(,x)
? <= x ⇒ R(,x]
Dates
Can be Represented Via (YYYY-MM-DD) including the parentheses
Can be used with range and boolean syntax
Hide Query Syntax Reference
Demographics
Must
date_of_birth
R[,1999-12-31]
Encounter
Add new clause
LabTest
Add new clause
Diagnosis
Must
diagnosis_ICD9_code
[556, 556.0, 556.1, 556.
Procedure
Must Not
procedure_CPT_code
[44140, 44141, 44143,
Submit
Query Editor
Text
Structured Data
OHDSI CDM Objects
condition_occurrence_OHDSI_text: Ulcerative colitis
condition_occurrence_raw: ulcerative colitis
condition_occurrence_tui: T047
condition_occurrence_SNOMEDCT_US_code: 64766004
condition_occurrence_cui: C0009324
end: 58
condition_occurrence_OHDSI_code: 81893
begin: 40
condition_occurrence_SNOMEDCT_US_text: Ulcerative Colitis
condition_occurrence_OHDSI_text: Crohn's disease
condition_occurrence_raw: Crohn's
condition_occurrence_tui: T047
condition_occurrence_SNOMEDCT_US_code: 34000006
condition_occurrence_cui: C0010346
end: 69
condition_occurrence_OHDSI_code: 201606
begin: 62
condition_occurrence_SNOMEDCT_US_text: Crohn Disease
Reset from Text Query
Add CDM Object
Submit

MAYO CLINIC

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 cords. 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

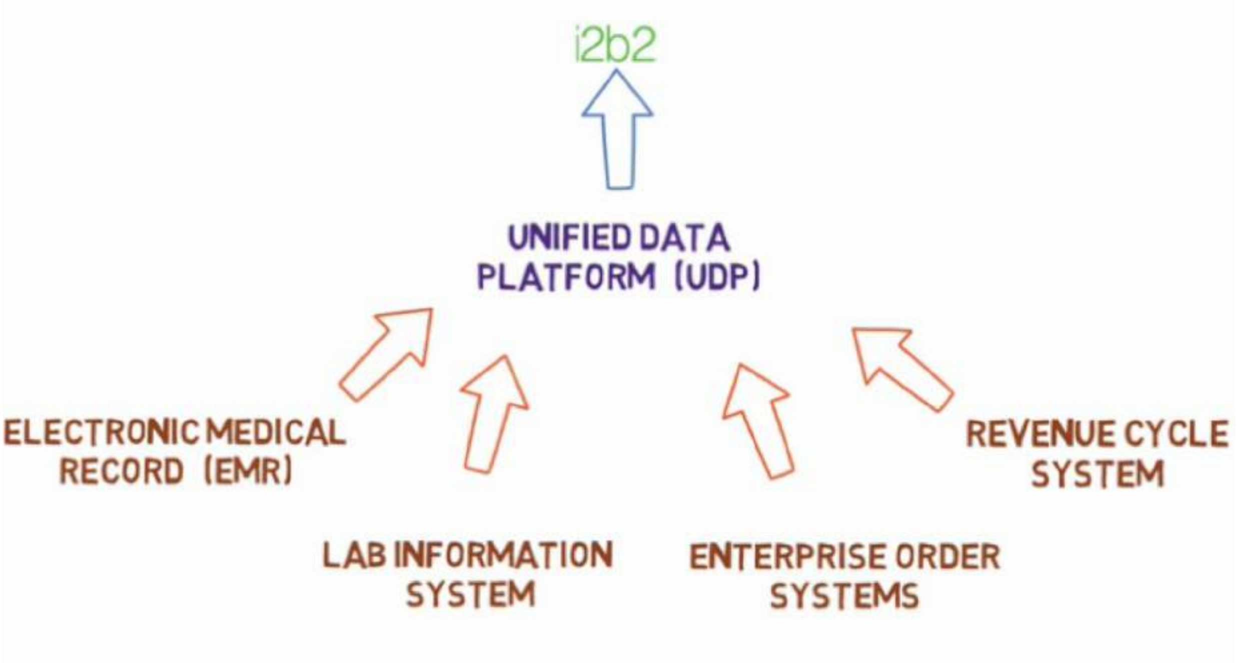
Exclusion criteria

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.

MAYO CLINIC

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.**



Criteria		ICD 9	ICD 10	CPT 4	Medication	Addressed by I2B2	Addressed by ACE
Inclusion							
1. Age greater than 50 years.						Yes	
2. Gastroesophageal reflux disease (any of 2.1, 2.2, 2.3)	**2.1 Gastroesophageal reflux disease defined by Dx**	530.81	K21.9			Yes	
	2.2 Gastroesophageal reflux disease defined by drug, duration of use >= 3 months over the last 5 years				omeprazole, esomeprazole, pantoprazole, rabeprazole, dexlansoprazole, lansoprazole, ranitidine, famotidine, cimetidine	(duration of use >= 3 months?)	
	2.3 Gastroesophageal reflux disease defined by prior endoscopic diagnosis of erosive esophagitis	530.19	K21.0	Not able to find specific code for esophagitis		No	No
3. Male gender						Yes	
4. Obesity defined as body mass index greater than equal to 30.						Yes	
5. Current or previous history of smoking						No	Partially
6. Family history of esophageal adenocarcinoma/cancer or Barrett's esophagus						No	Partially
7. Caucasian						Yes	
Exclusion							
1. Previous history of esophageal adenocarcinoma/cancer		150.9	C15.9			Yes	
2. previous history of endoscopic ablation for Barrett's esophagus.				43229, 43270 43228 43258		Yes	
3. Previous history of esophageal squamous carcinoma (included in 1)		150.9	C15.9			Yes	
4. Previous history of esophageal squamous dysplasia		622.10	N87.9			Yes	
5. Current Treatment (drug) with oral anticoagulation - warfarin					warfarin	Yes	
6. Current Treatment (drug) with oral anticoagulation - Coumadin. (included in 5)					Coumadin	Yes	
7. History of cirrhosis		571.5	K74.60			Yes	
8. History of esophageal varices		456.20	I85.00			Yes	
9. History of Barrett's esophagus		530.85	K22.7 K22.710 K22.711 K22.719			Yes	
10. History of endoscopy in the last 5 years				43235-43270		Yes	

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)

MAYO CLINIC

Thank you!

MAYO CLINIC

Clinical NLP: Challenges and Opportunities

MAYO CLINIC

Data!!

Data!!

MAYO CLINIC

Data!!

- HIPPA: 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

Data!!

- **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

Wang, Yanshan, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.

MAYO CLINIC

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

Available Under a Data Use Agreement

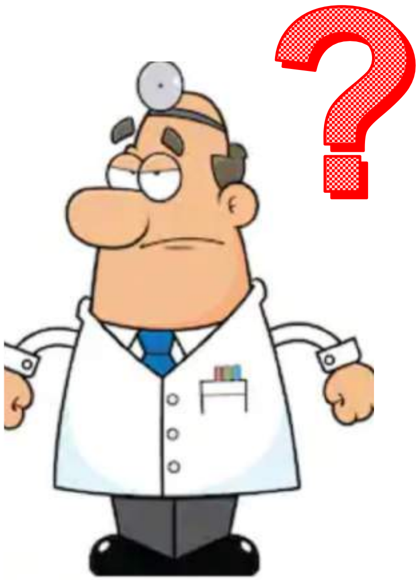
Wang, Yanshan, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.

MAYO CLINIC

Interpretation of Machine/Deep Learning

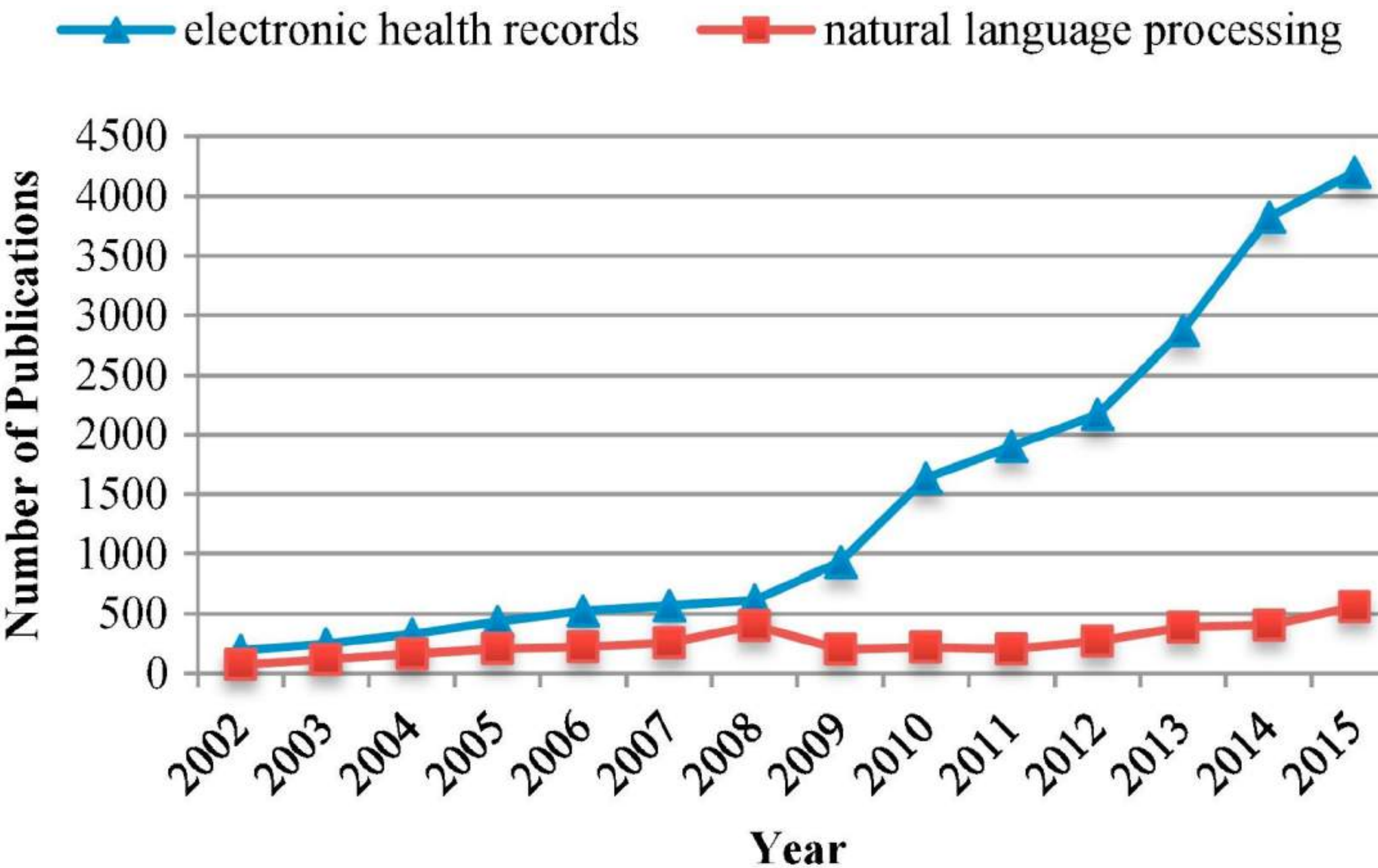
Input → **BLACK BOX** → Output

Why false positive??
Why false negative??



MAYO CLINIC

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.

Source: Wang, Yanshan, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.

MAYO CLINIC

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