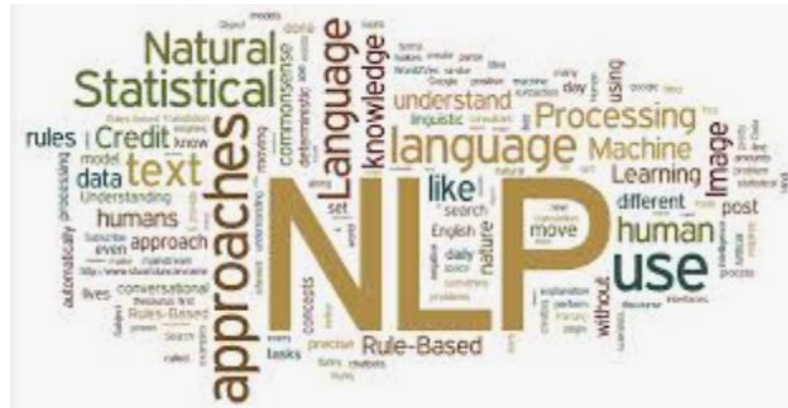
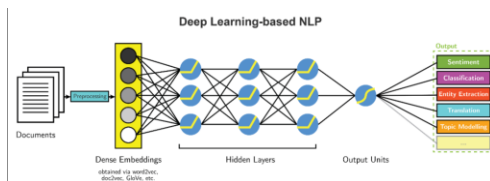


Natural Language Processing (NLP) in Medicine

Li Zhou, MD, PhD, FACMI, FIAHSI, FAMIA

Professor of Medicine

Division of General Internal Medicine and Primary Care
Brigham and Women's Hospital, Harvard Medical School



Email: lzhou@bwh.harvard.edu
Website: <http://mterms.bwh.harvard.edu/>



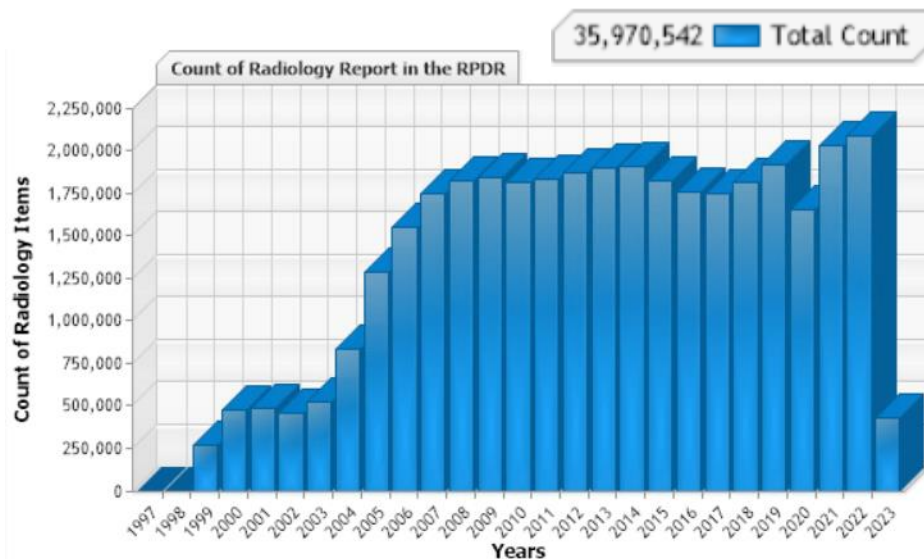
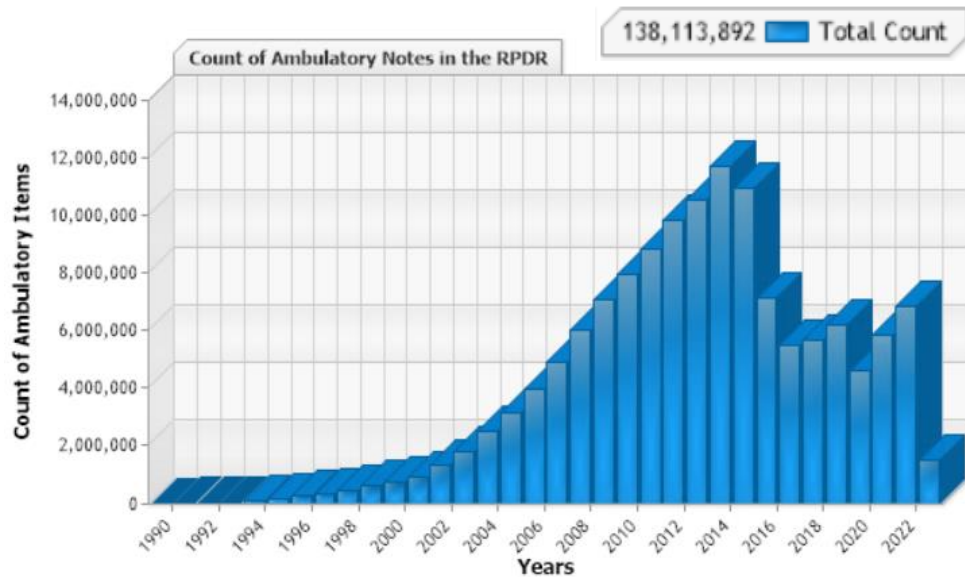
Free-text Clinical Data in EHR

NLP in Biomedicine

■ A significant portion of biomedical information is stored in textual form.

■ Electronic Health Records

- Ambulatory notes
- Admission notes
- Progress notes
- Discharge summaries
- Radiology reports
- Pathology reports
- Free-text entries and comments
-



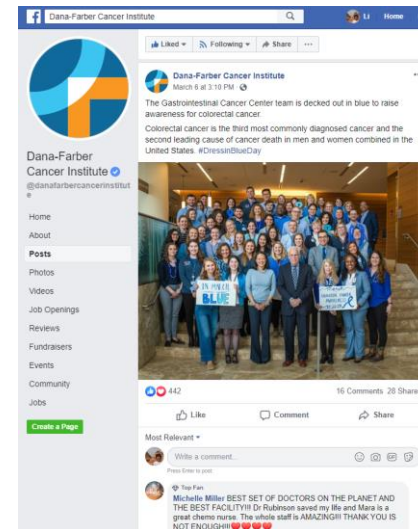


The Cloud, the Crowd, and Big Data

NLP in Biomedicine

- Biomedical literature
- Books, guidelines, surveys
- Wikipedia

- Social media (Twitter, Facebook, Reddit, blogs)
- News, reports
- Open innovation contests



- Topaz M, Lai K, Dhopeswarkar N, Seger DL, Sa'adon R, Goss F, Rozenblum R, Zhou L. Clinicians' reports in electronic health records versus patients' concerns in social media: a pilot study of adverse drug reactions of aspirin and atorvastatin. Drug Saf. 2016
- Tang C, Zhou L, et al;. Comment Topic Evolution on a Cancer Institution's Facebook Page. Applied clinical informatics 2017
- Blumenthal K, Topaz M, Zhou L, et al. Mining Social Media Data to Assess the Risk of and Soft Tissue Infections from Allergen Immunotherapy. J Allergy Clin Immunol. 2019
- Hua Y, Jiang H, Lin S, Yang J, Plasek JM, Bates DW, Zhou L. Using Twitter Data to Understand Public Perceptions of Approved versus Off-label Use for COVID-19-related Medications. J Am Med Inform Assoc. 2022. PMID: 35775946;



Speech Recognition

NLP in Biomedicine

Dictation



Voice-enabled care (virtual medical assistants; chatbots)



Speech and Diseases

- Zhou L, et al. Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. JAMA Network Open. 2018
- Blackley SV, et al. Speech Recognition for Clinical Documentation from 1990 to 2018: A Systematic Review. JAMIA 2019
- Goss FR, et al. A Clinician Survey of Using Speech Recognition for Clinical Documentation in the Electronic Health Record. Int J Med Inform (IJMI). 2019.
- Blackley SV, et al. Physician Use of Speech Recognition versus Typing in Clinical Documentation: A Controlled Observational Study. IJMI, 2020.



NLP Tasks

Natural Language Understanding

- Part-of-Speech Tagging (POS)
- Named entity recognition (NER)
- Information extraction
- Text/document classification
- Information retrieval
- Grammar and spelling checking and correction
- Relationship extraction
- Coreference resolution
- Sentiment analysis
- Speech recognition

Natural Language Generation

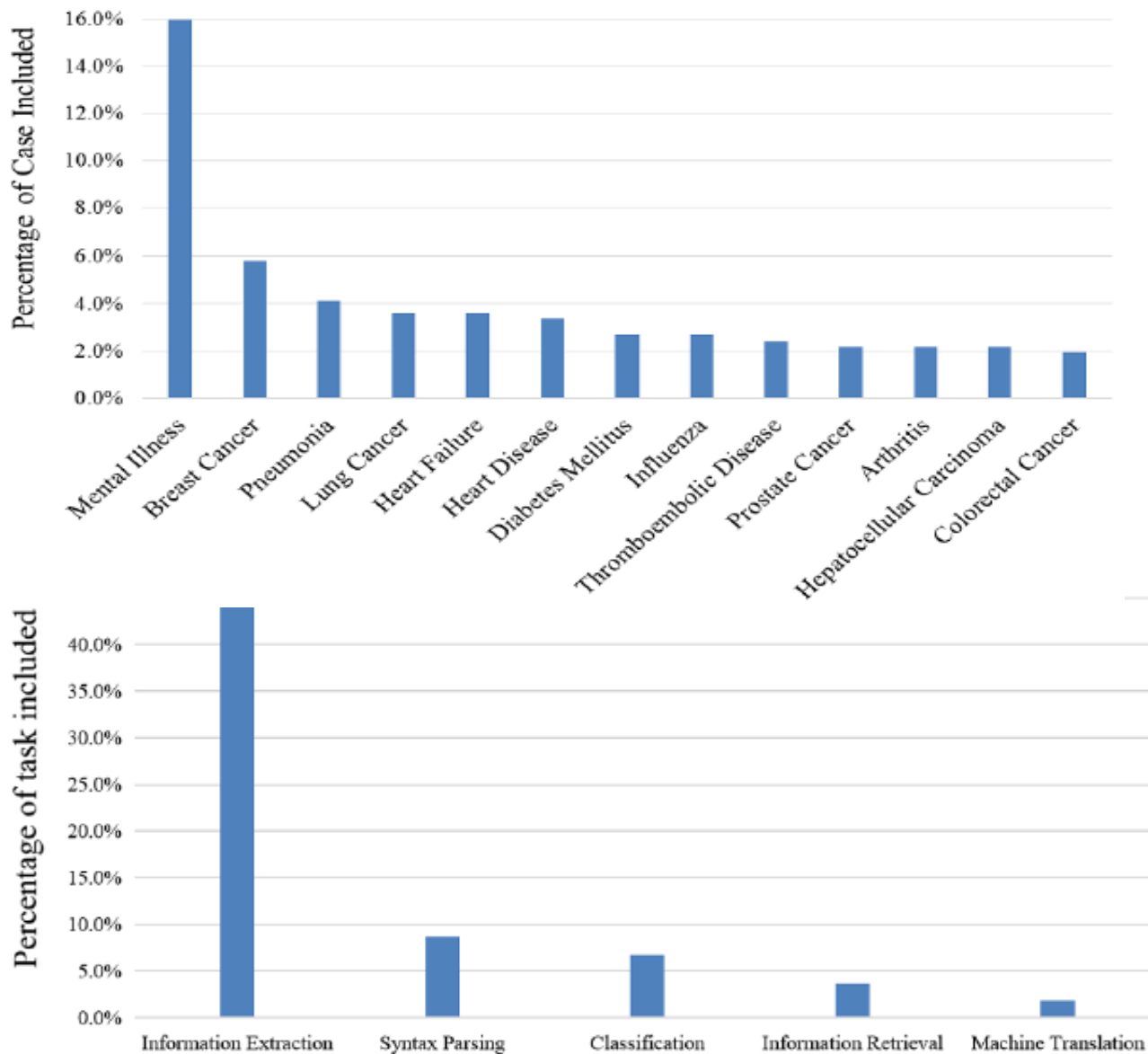
- Dialogue generation (chatbot)
- Text generation
- Text summarization
- Language translation
- Speech synthesis
- Image captioning
- Data-to-Text generation



Clinical Domains and Tasks

NLP in Biomedicine

Among 2336 NLP articles between 1999-2018 (Wang J, JMIR, 2020)





Deep Learning Methods in Clinical NLP

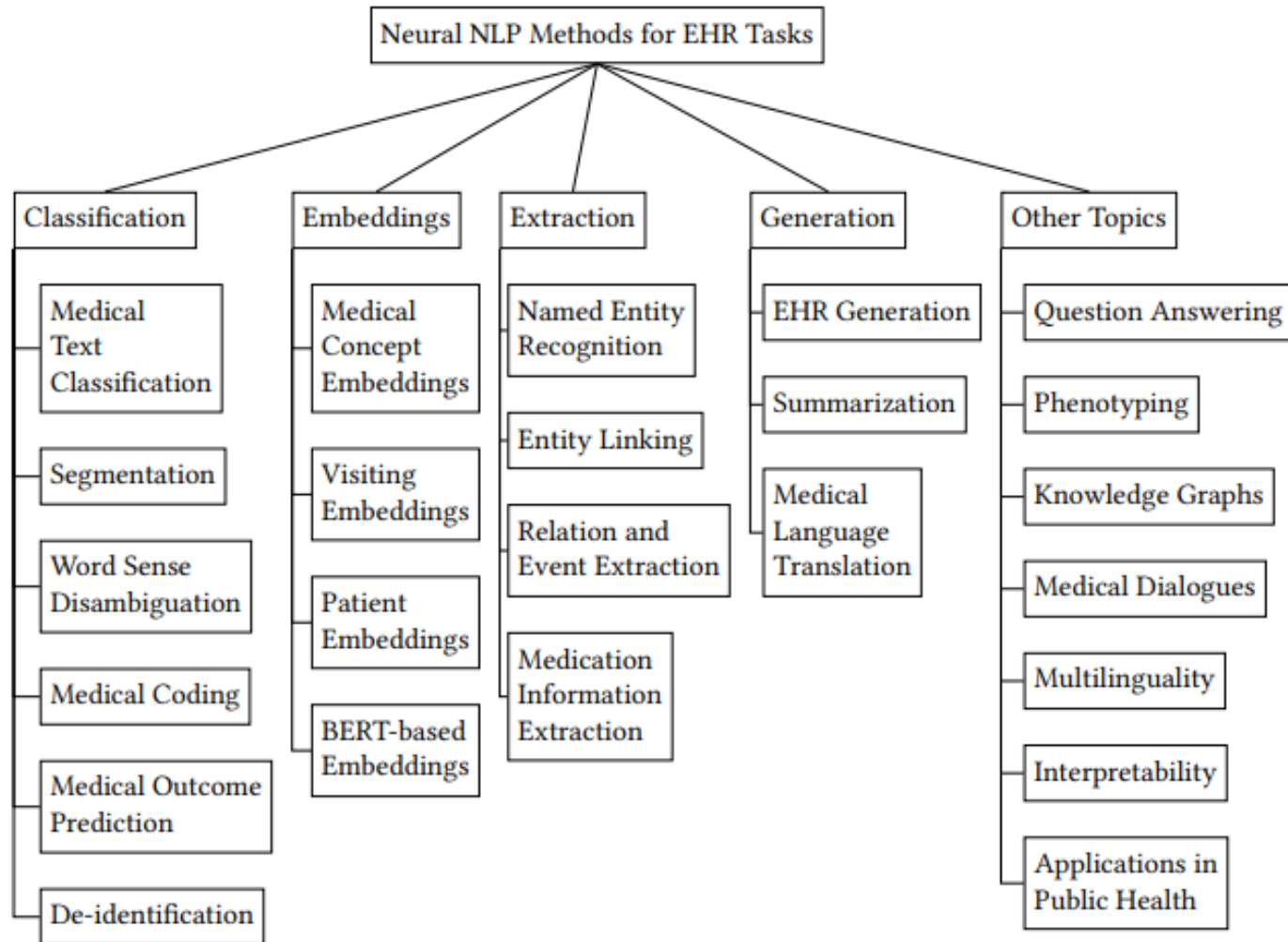
NLP in Biomedicine

Deep Learning Methods in
Clinical NLP
(n=212 articles)
(Wu S, JAMIA 2019)

Architecture	Method	Freq.
RNN	LSTM	109
	GRU	16
	Vanilla RNN	5
	Tree-LSTM	1
	CNN-LSTM	3
CNN	CNN	80
	CNN-LSTM	3
FFNN	NN	22
Embeddings Only	Embeddings	21
Other	Autoencoder	3
	DBN	3
	Other DL	3
	Capsule	1
	Memory Network	1
	RecursiveNN	1
	Transformer	1



Neural NLP



Li I, et al. Neural Natural Language Processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46, 100511. <https://doi.org/10.1016/j.cosrev.2022.100511>



NLP Methods

NLP in Biomedicine

	Description	Methods	Tasks	Pros	Cons
Rule-Based	Uses predefined rules and patterns to process text	Regular expressions, Lexicon models	<ul style="list-style-type: none"> Syntax parsing NER POS tagging 	<ul style="list-style-type: none"> Simple to implement High interpretability Fast execution for specific tasks 	<ul style="list-style-type: none"> Extensive domain knowledge needed Poor scalability Limited to known rules
Statistical Methods	Leverages statistical theories to infer linguistic structures from data	Naive Bayes, TF-IDF	<ul style="list-style-type: none"> Text classification Sentiment analysis Topic modeling 	<ul style="list-style-type: none"> Good at handling ambiguity Can learn from data 	<ul style="list-style-type: none"> Requires large datasets Feature engineering needed Limited context understanding
Machine Learning	Involves algorithms that can learn from and make predictions on data	SVM, Decision Trees, Random Forests	<ul style="list-style-type: none"> Classification Regression Clustering 	<ul style="list-style-type: none"> Versatile Can handle various types of data 	<ul style="list-style-type: none"> Requires feature engineering Prone to overfitting Interpretability can be challenging
Deep Learning	Employs neural networks with multiple layers to model complex language patterns	RNNs, CNNs, Transformer models	<ul style="list-style-type: none"> Machine translation Speech recognition Text generation 	<ul style="list-style-type: none"> Excels in capturing complex patterns High accuracy Scalable with data volume 	<ul style="list-style-type: none"> Large datasets needed High computational cost Overfitting risk
Large Language Models	Utilizes very large neural networks trained on vast amounts of text data to understand and generate human-like text	GPT, BERT, Transformer-based architectures	<ul style="list-style-type: none"> Question answering Text summarization Advanced text generation and understanding 	<ul style="list-style-type: none"> State-of-the-art performance Deep understanding of context Flexible across many tasks 	<ul style="list-style-type: none"> Significant computational resources Potential for biased outputs Interpretability challenges



I2b2/n2c2 NLP Challenges

NLP in Biomedicine

- 📄 2006, 1) De-identification and 2) Smoking
- 📄 2008, Obesity
- 📄 2009, Medication extraction
- 📄 2010, Relations (of medical problems, tests, treatments)
- 📄 2011, 1) Co-reference (anaphora) resolution and 2) Sentiment classification (emotions in suicide notes)
- 📄 2012, Temporal relations
- 📄 2014, 1) De-identification and 2) Identifying risk factors for heart disease over time
- 📄 2016, 1) De-identification and 2) RDoc classification (determine symptom severity based on a patient's initial psychiatric evaluation)
- 📄 2018, 1) Cohort selection for clinical trials and 2) Adverse drug events and medication extraction in EHRs
- 📄 2019, 1) Clinical semantic textual similarity 2) Family history 3) Clinical concept normalization 4) Novel data use
- 📄 2022, 1) Contextualized medication event extraction, 2) Social determinants of health, 3) Progress note understanding: assessment and plan reasoning

<https://n2c2.dbmi.hms.harvard.edu/>

The challenges and data sets are now administered through the [DBMI Data Portal](#).

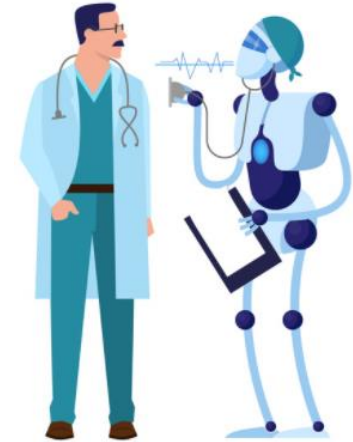
I2b2: Informatics for Integrating Biology and the Bedside; n2c2: National NLP Clinical Challenges



Could you help me?

NLP in Biomedicine

- Identify the long-term symptoms resulting from COVID-19?
- Generate a cohort of patients who had severe cutaneous adverse reactions caused by vancomycin?
- Find “need to know” clinical information from Epic EHR relevant to the patient’s chief complaints?
- Read/interpret a pathology report to find abnormal cancer screening results and tell me when to follow up?
- Transcribe and summarize my conversation with the patient?
- Detect and correct errors in my notes dictated by Dragon?
-

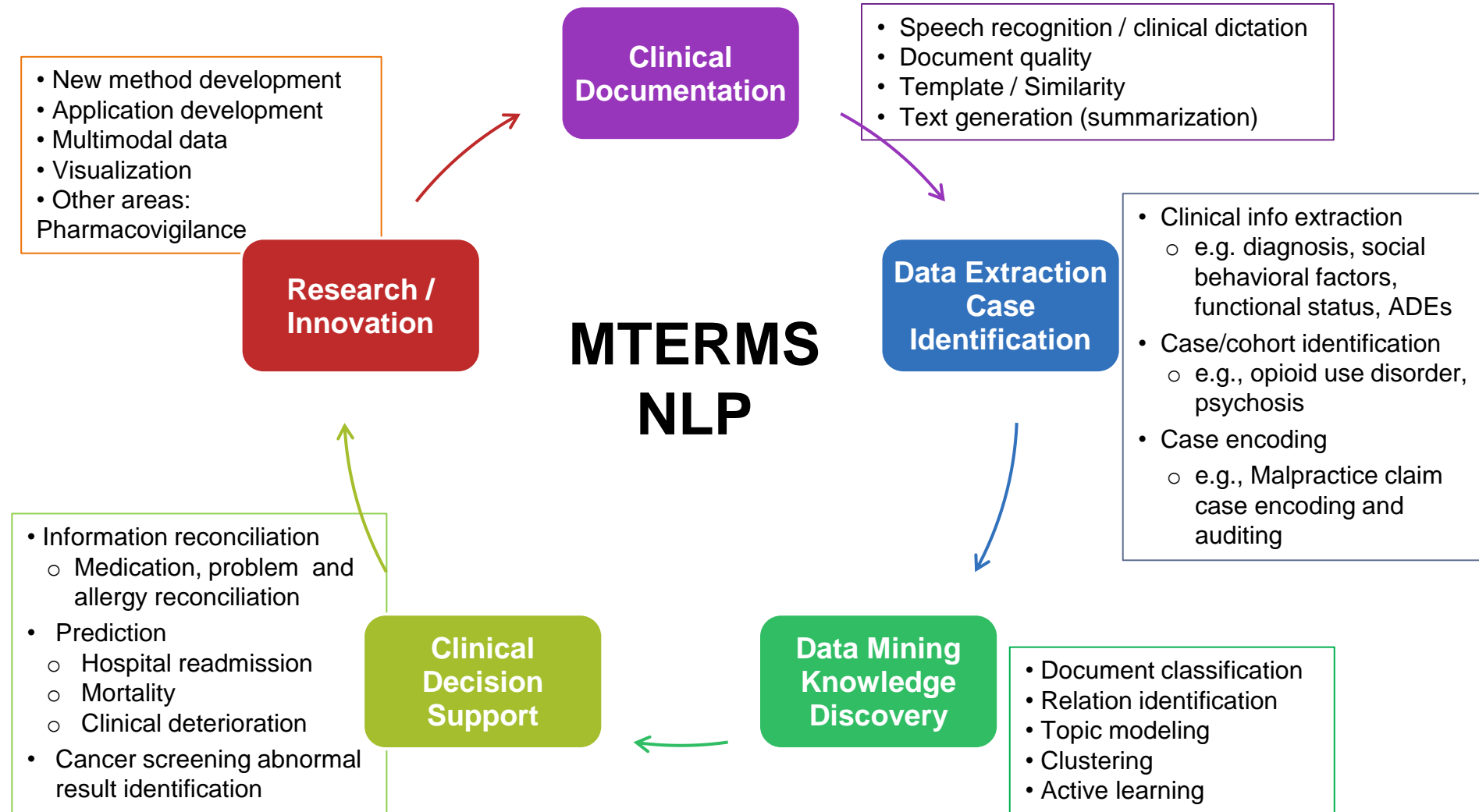


Natural Language Processing (NLP)



MTERMS Research Areas

NLP in Biomedicine





MTERMS Applications

NLP in Biomedicine



Real-time pilots (integrated with Epic)

- Allergy reconciliation module
 - Medication reconciliation module (in LMR)
- Cancer Screening Follow-up (primary care)
- Patient clinical deterioration based on nursing notes and EHR (inpatient)



Near real-time

- Patient mortality predication to improve palliative care intervention



Research projects

- Allergic and adverse reactions
- Opioid use disorder patient identification
- Gunshot intention classification
- Malpractice cases (coding + similar cases)
- Psychosis identification
- Confounding factors for pharmacoepidemiology studies
- Dementia/cognitive decline
- PASCLex: Post-Acute Sequelae of COVID-19 (PASC) Symptom
- Using Twitter data to understand public perceptions of approved vs. off-label user for COVID-19-related medications
- Examination of stigmatizing language in the electronic health record
- Early Detection of Cognitive Decline Using Large Language Models (LLMs)



NLP service to support real-time (or near real-time) applications: system architecture

NLP in Biomedicine

Objectives

- Improving data interoperability
- Integrating NLP with EHRs
- Providing clinical decision support
- Improving patient safety

MTERMS NLP Services

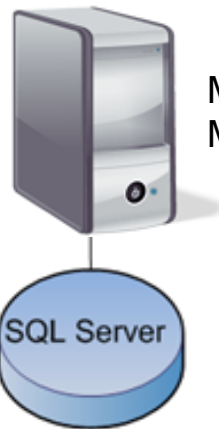
- MTERMS Natural Language Processing
- Batch Processing & Summarization
- Knowledge Base
- Web Application



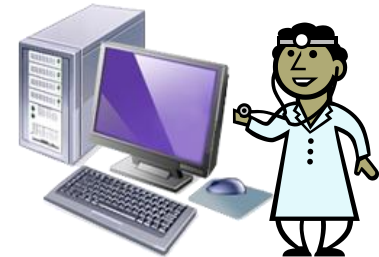
MGB Web Services

- Get Schedules
- Get Medications
- Get Allergy
- Get Notes
- Get Other Data

MGB Data Repository
& Web Service



MGB Session Service
MTERMS Web Service



EHR Web Application



Natural Language Processing to Identify Abnormal Breast, Lung, and Cervical Cancer Screening Test Results from Unstructured Reports to Support Timely Follow-up

Courtney J. Diamond, BS^a, John Laurentiev, MS^b, Jie Yang, PhD^b, Amy Wint, MSc^a, Kimberly A. Harris, MM^a, Tin H. Dang, BA^a, Amrita Mecker, BS^a, Emily B. Carpenter, BS^a, Anna N. Tosteson, ScD^c, Adam Wright, PhD^d, Jennifer S. Haas, MD, MSc^a, Steven J. Atlas, MD, MPH^a, and Li Zhou, MD, PhD^b

^a Department of General Internal Medicine, Massachusetts General Hospital, Boston, MA, United States

^b Department of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, United States

^c The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, Lebanon, NH, United States

^d Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

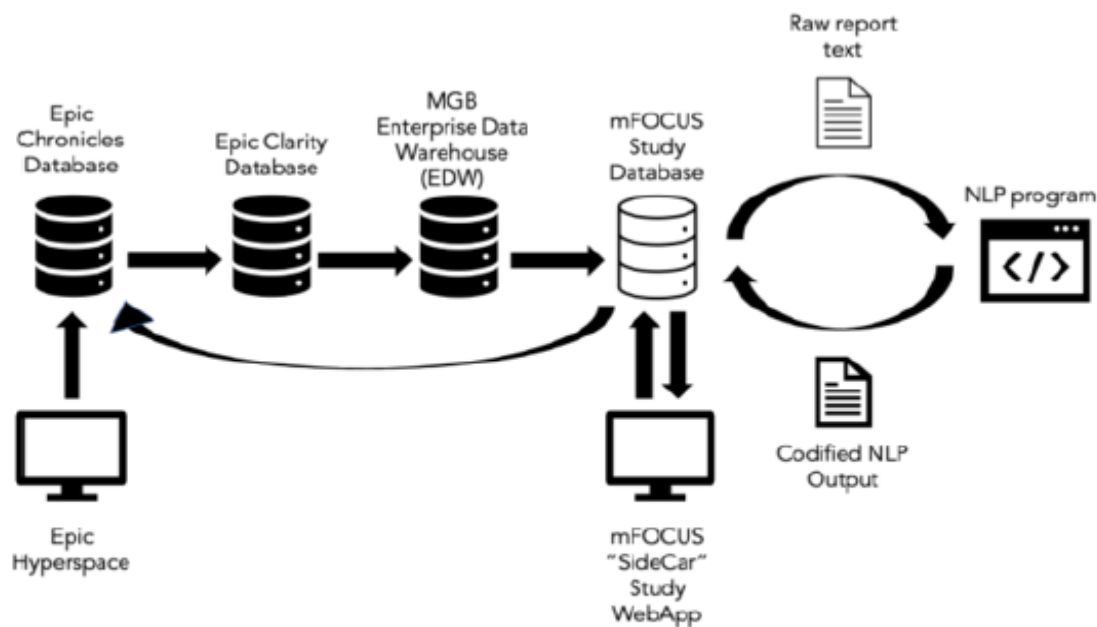


Figure 1—mFOCUS system architecture



NLP Performance

NLP in Biomedicine

Table 3— Performance Metrics

Text source	Precision	Recall	F-Measure
Mammogram- BIRADS score (n = 1000)	1.000	0.986	0.993
LDCT scan- LungRADS score (n = 1000)	1.000	0.999	0.999
Pap smear- primary cytology (n = 1000)	1.000	0.983	0.991
Pap smear- other cytology (n = 500)	1.000	0.970	0.985
Pap smear- HPV test result (n = 500)	1.000	1.000	1.000
Pap smear- HPV genotype (n = 500)	1.000	0.978	0.989
Anatomic pathology- colposcopy (n = 500)	0.975	0.845	0.905

JAMA | Original Investigation

A Multilevel Primary Care Intervention to Improve Follow-Up of Overdue Abnormal Cancer Screening Test Results A Cluster Randomized Clinical Trial

Steven J. Atlas, MD, MPH; Anna N. A. Tosteson, ScD; Adam Wright, PhD; E. John Orav, PhD; Timothy E. Burdick, MD, MSc, MBA; Wenyan Zhao, PhD; Shoshana J. Hort, MD; Amy J. Wint, MSc; Rebecca E. Smith, MS; Frank Y. Chang, MS; David G. Aman, BA; Mathan Thillaiyapillai, MS; Courtney J. Diamond, MA; Li Zhou, MD, PhD; Jennifer S. Haas, MD, MSc

Key Points

Question Can a primary care intervention comprising electronic health record (EHR) reminders and patient outreach with or without patient navigation improve timely follow-up of overdue abnormal cancer screening test results?

Findings Among 11 980 patients in 44 primary care practices, completion of follow-up for an abnormal breast, cervical, colorectal, or lung cancer screening test result within 120 days of study enrollment was higher among patients exposed to EHR reminders, outreach, and navigation (31.4%) or EHR reminders and outreach (31.0%) than those exposed to EHR reminders only (22.7%) or usual care (22.9%).

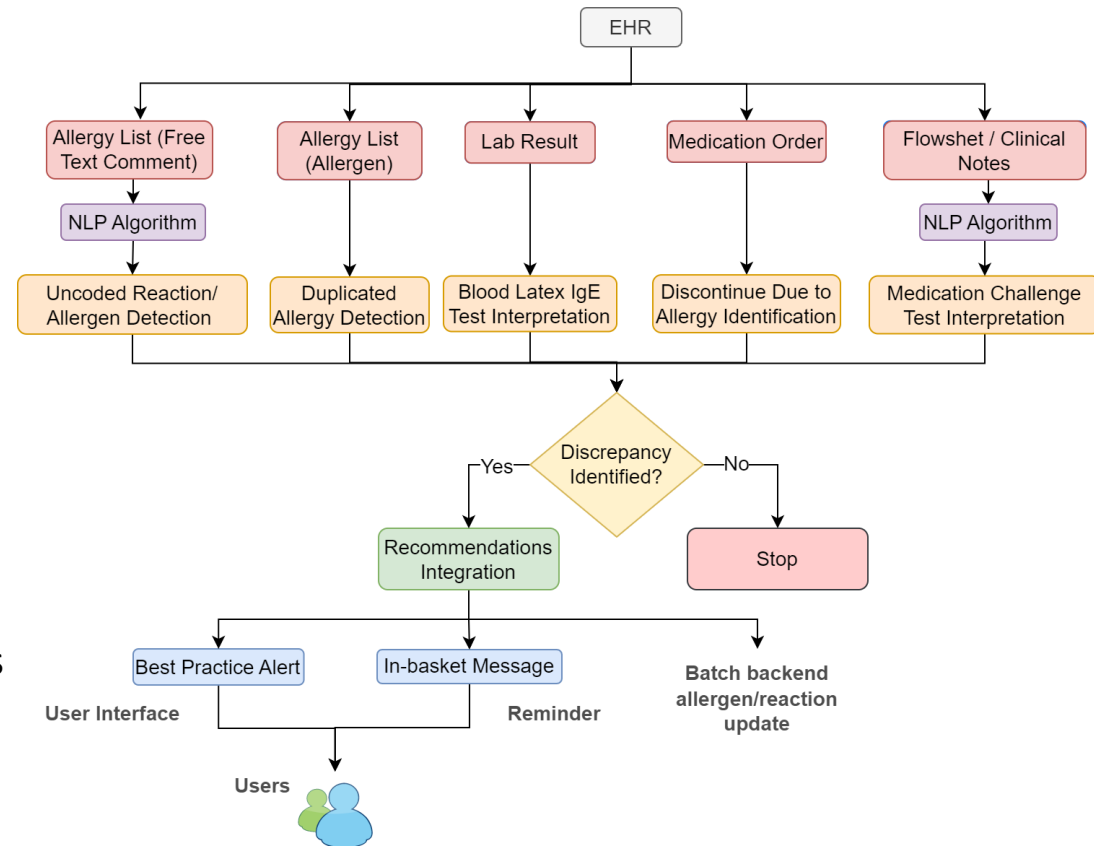
Meaning Systems-based outreach in primary care settings can improve the timely follow-up of abnormal cancer screening results, but gaps in follow-up care need to be addressed if the full benefits of preventive cancer screening are to be realized.



Allergy Reconciliation across the EHR

NLP in Biomedicine

- Accurate and complete allergy documentation in the EHR is essential to guide clinical decision-making.
- Patient allergy information often exists in several locations in the EHR, and patients' allergy lists are often inaccurate or incomplete
- Automatically identify discrepancies in allergy information from across the EHR
- User interface – providers can accept, reject, or modify suggested changes
 - Automatically add free-text reactions to allergy list
- Integrated with Epic in pilot study with 111 BWH providers



*Preliminary unpublished data



> [Front Allergy](#). 2022 May 10;3:904923. doi: 10.3389/falgy.2022.904923. eCollection 2022.

Reconciling Allergy Information in the Electronic Health Record After a Drug Challenge Using Natural Language Processing

Ying-Chih Lo ^{1 2}, Sheril Varghese ¹, Suzanne Blackley ³, Diane L Seger ^{1 3},
Kimberly G Blumenthal ^{2 4}, Foster R Goss ⁵, Li Zhou ^{1 2}

Affiliations + expand

PMID: 35769562 PMCID: PMC9234873 DOI: 10.3389/falgy.2022.904923

[Free PMC article](#)

Abstract

Background: Drug challenge tests serve to evaluate whether a patient is allergic to a medication. However, the allergy list in the electronic health record (EHR) is not consistently updated to reflect the results of the challenge, affecting clinicians' prescription decisions and contributing to inaccurate allergy labels, inappropriate drug-allergy alerts, and potentially ineffective, more toxic, and/or costly care. In this study, we used natural language processing (NLP) to automatically detect discrepancies between the EHR allergy list and drug challenge test results and to inform the clinical recommendations provided in a real-time allergy reconciliation module.

Methods: This study included patients who received drug challenge tests at the Mass General Brigham (MGB) Healthcare System between June 9, 2015 and January 5, 2022. At MGB, drug challenge tests are performed in allergy/immunology encounters with routine clinical documentation in notes and flowsheets. We developed a rule-based NLP tool to analyze and interpret the challenge test results. We compared these results against EHR allergy lists to detect potential discrepancies in allergy documentation and form a recommendation for reconciliation if a discrepancy was identified. To evaluate the capability of our tool in identifying discrepancies, we calculated the percentage of challenge test results that were not updated and the precision of the NLP algorithm for 200 randomly sampled encounters.

Results: Among 200 samples from 5,312 drug challenge tests, 59% challenged penicillin reactivity and 99% were negative. 42.0%, 61.5%, and 76.0% of the results were confirmed by flowsheets, NLP, or both, respectively. The precision of the NLP algorithm was 96.1%. Seven percent of patient allergy lists were not updated based on drug challenge test results. Flowsheets alone were used to identify 2.0% of these discrepancies, and NLP alone detected 5.0% of these discrepancies. Because challenge test results can be recorded in both flowsheets and clinical notes, the combined use of NLP and flowsheets can reliably detect 5.5% of discrepancies.

Conclusion: This NLP-based tool may be able to advance global delabeling efforts and the effectiveness of drug allergy assessments. In the real-time EHR environment, it can be used to examine patient allergy lists and identify drug allergy label discrepancies, mitigating patient risks.

- Allergy labels are common, often incorrect, and potentially harmful.

Up to 15% of hospitalized patients, 6% to 10% of the general population report a penicillin allergy. Of these individuals, 94% can tolerate penicillin after formal allergy testing



Allergy List

Allergen	Severity	Reactions	Comments
Amoxicillin Sodium	Medium	Rash	On 01/03/21 tolerated amoxicillin 250 mg without immediate reaction.
Sulfamethoprime DS	Low	Nausea	

Flowsheet

Med Challenge	01/03/2021
Premedication	NA
Challenge Medication	250 mg Amoxicillin x 1
Dose	
Time	2:05 PM
O ₂ Saturation	100
BP	110/70
RR	18
HR	70
Symptoms	No
Rash	0
Sneezing/itching	0
Nasal Congestion	0
Larynx	0
Wheezing	0
GI Subjective Complaints	0
Cardiovascular	0
Symptom Score	0
This Challenge Test Was:	0

Clinical Note

Reason for Consult: Amoxicillin Allergy
 HPI : xxx
 Past Medical History : xxx
 Family History : xxx
 Social / Environmental History : xxx
 Allergies :
 - Amoxicillin - Skin rash
 - Sulfamethoxazole - GI upset
 ROS : xxx
 Physical Exam : xxx
 Skin Test Result :

 Patient tolerated the testing well and did not have any symptoms.
 Oral Challenge Test :

 Patient was asymptomatic and well-appearing with a stable exam at the time of leaving the clinic.
 Assessment :

 Skin testing was negative and challenge today was tolerated.

FIGURE 1 | Scenario of allergy information discrepancy in EHR. Allergy list, flowsheets, and clinical notes are different locations in the EHR that store allergy information. A negative result of drug challenge test may not be updated to the allergy list accordingly. Sometimes, the physician would leave a comment instead of removing the allergen from the list, as pictured in this figure.

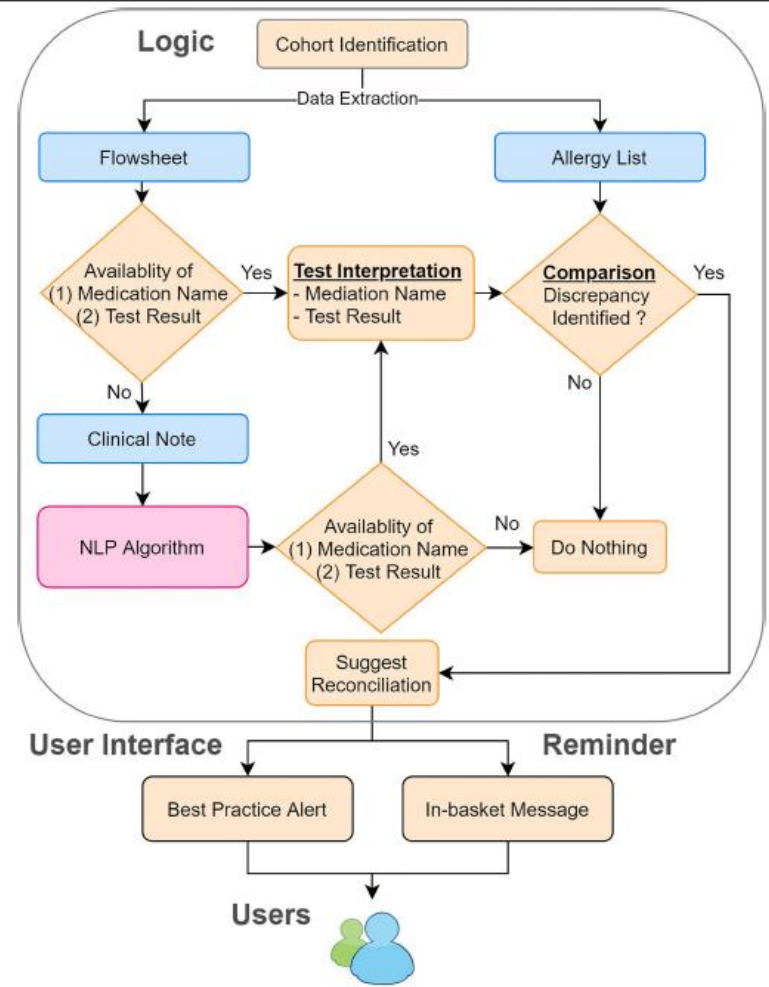


FIGURE 2 | System architecture of the reconciliation module. We combined the information derived from the flowsheets and clinical notes. We then compared this information to the allergy list to identify the discrepancies. If any discrepancies were found, we sent in-basket messages weekly to remind the physician to reconcile the allergy discrepancies by using our tool.



NLP in Biomedicine

Allergy Reconciliation

MGB Allergy Reconciliation Module [FAQ](#) Please provide feedback

To reconcile - approve ✓, reject ✖ or edit ✎ suggested changes.

Agent	Reason Type	Suggestion/action	View
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> CEPHALOSPORINS	Update	RASH was found in the comment for this allergen. Would you like to add Rash to the reactions preview for this allergen?	
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Pen-VK	Delete	The patient was given Pen VK under observation in the allergy clinic on 2018-12-08 and did not have symptoms consistent with allergy. Please review the patient record and consider removing PENICILLINS(Pen VK) from the allergy list if appropriate.	PreviewClinical Note

FIGURE 4 | User interface of the recommendation for challenge tests. We provide the reason in addition to the suggested action, such as “Add” and “Delete” for the user to make decision. We also include a hyperlink (right-hand side) to the clinical notes in case the user wants to know more about the reaction.

Results: Among 200 samples from 5,312 drug challenge tests, 59% challenged penicillin reactivity and 99% were negative. 42.0%, 61.5%, and 76.0% of the results were confirmed by flowsheets, NLP, or both, respectively. The precision of the NLP algorithm was 96.1%. Seven percent of patient allergy lists were not updated based on drug challenge test results. Flowsheets alone were used to identify 2.0% of these discrepancies, and NLP alone detected 5.0% of these discrepancies. Because challenge test results can be recorded in both flowsheets and clinical notes, the combined use of NLP and flowsheets can reliably detect 5.5% of discrepancies.



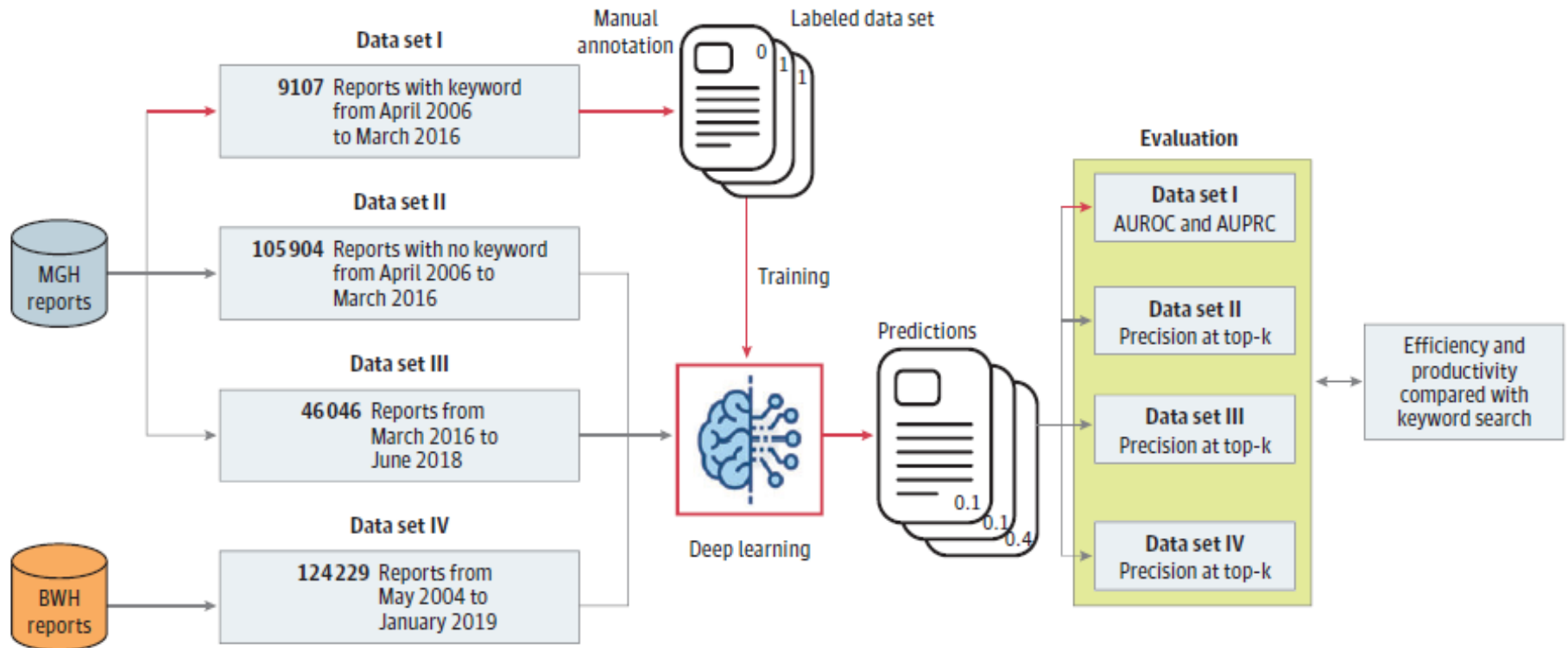
Deep Learning to Detect Allergy Events from Hospital Safety Reports

- 📄 Allergy safety knowledge is limited by case identification challenges.
- 📄 Hospital safety event reporting systems are integral to the detection of patient safety signals in health care, but still lacking are processes to analyze them in a manner that allows for timely feedback to health care professionals.
- 📄 We developed an AI method, a hierarchical attention-based deep neural network (DNN), that automatically reads the free-text description of safety reports and identifies cases describing allergic reactions.

Yang J, Wang L, Phadke NA, Wickner PG, Mancini CM, Blumenthal KG, Zhou L. Development and Validation of a Deep Learning Model for Detection of Allergic Reactions Using Safety Event Reports Across Hospitals. *JAMA Network Open*. 2020 Nov 2;3(11):e2022836



Study Design and Datasets



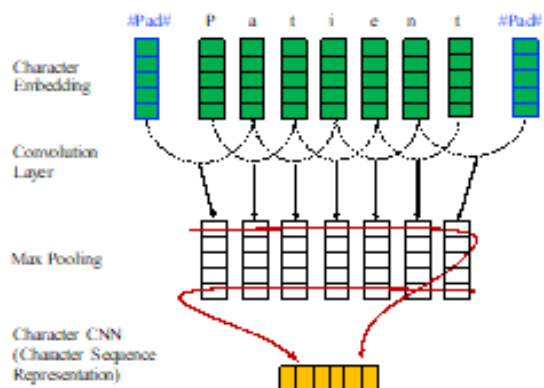
- Our model was trained on the free-text descriptions of 9,107 labeled reports extracted using expert-curated keywords from MGH's safety event reporting system.
- We then used the model to automatically identify allergy events from nearly 300,000 reports from MGH and BWH across 15 years.



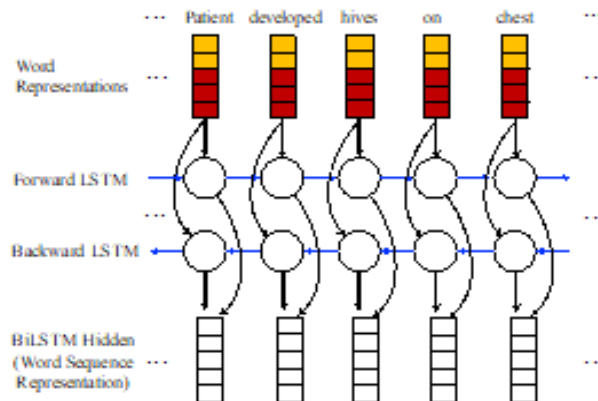
Deep Learning Model

NLP in Biomedicine

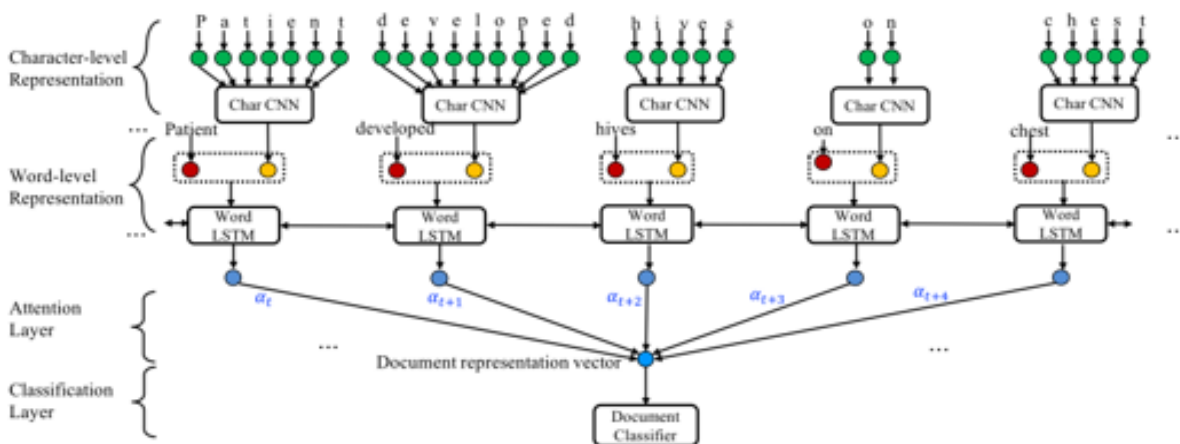
a.



b.



c.



The *first* layer is a character-level encoder, which aims to capture lexical variations (e.g., misspelling) of a word. It encoded the character sequence within each word using a single layer CNN. Each character of a word was represented using randomly initialized character embedding, which was fed as the input of the character-level CNN. The output of the CNN was then fed into a max-pooling function to create a fix-dimension vector for the word.

In the *second* layer, each word vector was concatenated with the word's embedding that was pretrained on all MGH reports using word2vec. On top of the concatenated word representation, a LSTM network was built to utilize the contextual information of the whole report and generate an output vector for each word.

Because different words within a report may have different levels of contribution in distinguishing the report, we added an attention model as a *third* layer to assign a unique weight for each word, which was calculated based on the LSTM output vector.

We computed a weighted sum of the LSTM output vectors of all the words in the report to generate a report representation vector, which was then fed into the *fourth* layer, the classifier. The classifier was trained using the cross-entropy loss function and the Stochastic Gradient Descent (SGD) optimizer.

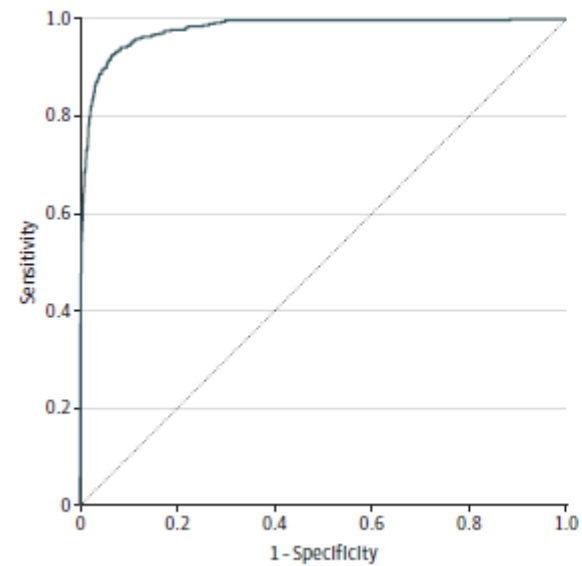
The output of the classifier was a vector representing the probability of whether or not a report described an allergy event.



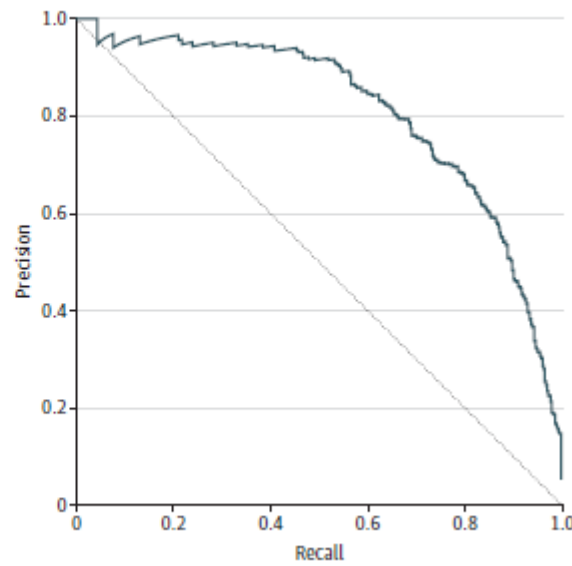
Model Performance

NLP in Biomedicine

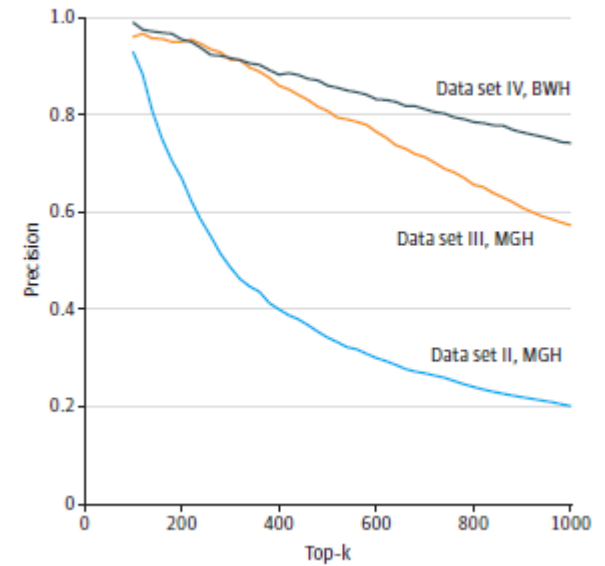
A Area under the receiver operating characteristic for data set I



B Area under the precision-recall curve for data set I



C Precision at top-k



The deep learning model achieved an AUROC of 0.979 (95%CI, 0.973-0.985) and an area under the precision-recall curve of 0.809 (95%CI, 0.773- 0.845).



Efficiency and Productivity

Data set	Measures	Keyword-search approach	Attention-based DNN model
II	Cases to review	0	1627
	True cases	0	184
	Precision, %	NA	11.3
III	Cases to review	10 131	1984
	True cases	570	625
	Precision, %	5.6	31.5
IV	Cases to review	15 896	5800
	True cases	1344	1569
	Precision, %	8.5	27.1
Total	Cases to review	26 027	9411
	True cases	1914	2378
	Precision, %	7.4	25.3

Compared with the keyword-search approach, the deep learning model **reduced the number of cases for manual review by 64%** and **identified 24% more cases** of confirmed allergic reactions.



Interpretation

NLP in Biomedicine

A Attention to words contributing to the prediction of positive cases

PT immediately started sneezing x2 after injection of **isovue 300**. He developed a stuffy nose 4 minutes after injection. Took him to RN station to be examined by RN and Radiologist. He developed a **hive** on left arm and heavy 15 min after injection.

PT received 100cc's of **Isovue 300**. Immediately following the injection the PT **experienced itchy eyes, face and throat**. He also had a racing heart and difficulty breathing. Dr's Smith and Brown responded immediately. Vitals BP 130/80, HR 95. 200ml normal saline was hung and 60mg po Benedryl was administered by Dr. Smith. He was monitored here for approximately 20 minutes. As symptoms resolved we notified his nurse.

Gd injected at 17:20. Pt started **coughing** and **c/o throat**. **Radiologist** evaluated pt. and asked that patient be given oxygen 3 L, BP 120/80, 60mg benadryl p.o given.

B Attention to words contributing to the prediction of negative cases

An order to rule out a **groin hematoma** was written for both Radiology as well as the vascular lab. It was performed in the vascular lab and we **did not** realize that we were about to repeat the exam until the patient told us that **she had already** had the exam earlier in the day.

Medication came up in the **tube** system, there was one **10ml syringe** with the appropriate dose and **amount** (was labeled correctly), the second Medication (same as the first) had the same lable but it was **on** a 60ml bag of D5W. Medication **not given**, pharmacy called and safety report filed.

Pt **ambulated** to BR with RN. Once in BR, pt was instructed to pull string when ready. Pt pulled string, RN was in BR with pt, PT stated she felt **SOB** but not light headed or dizzy. Pt stood up with RN and was walking out of the BR and **started** to fall forward, RN caught pt and directed to the floor. **BP 87/51** SPO2 100% HR 76. MD at bedside. After some IVF, pt **SBP 116** and was able to walk assisted to bed with staff.

Regarding interpretability of the model, these attention heatmaps demonstrate how much attention the model gives and to which words when making positive and negative allergy event predictions. Darker color represents a higher attention weight.



Deep Learning for Mortality Prediction in Selecting Patients for Earlier Palliative Care Interventions

NLP in Biomedicine

One of the largest challenges in expanding palliative care is identifying those patients who can benefit the most

- Which patients will benefit from which interventions and when?
 - ✓ Predict patients' clinical trajectories
 - ✓ Identify those who need palliative care
 - ✓ Determine the right time to start the interventions
- Existing population management algorithms generally target patients with high healthcare utilization
- Most clinicians hesitate to provide prognosis information to patients



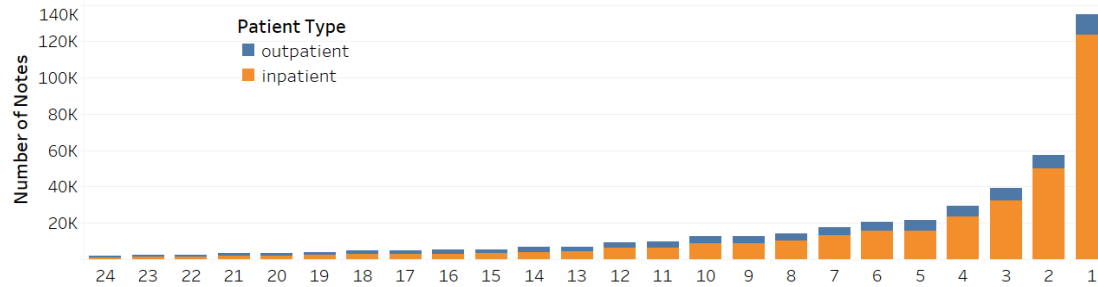
There is an urgent need to leverage information technology and the EHR to provide decision support for healthcare providers



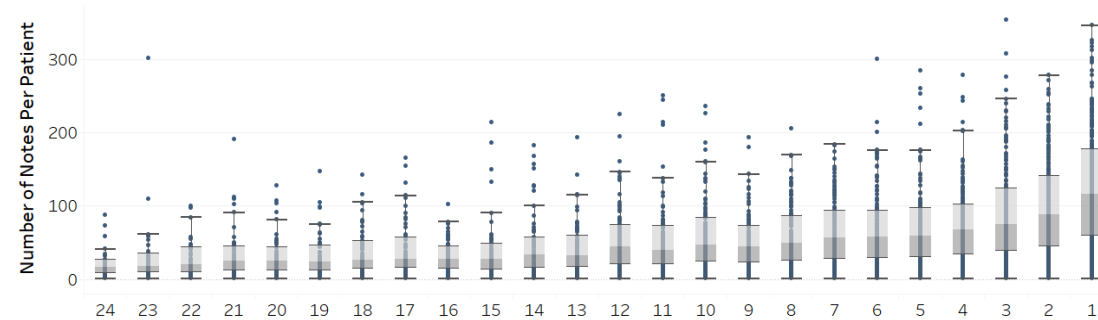
Large amount of free-text EHR data

NLP in Biomedicine

A. Total number of notes by month to death



B. Distribution of the number of notes per patient by month to death



Healthy Brain Severe Dementia



Dementia

- > 5.5 million Americans in 2017
- Sixth leading cause of death
- One of the costliest disease
 - \$259 billion for elderly per year

Number of clinical notes per patient with dementia by month over the last two years of life (a total of 432,007 notes of 7,875 patients) (Wang L, Zhou L, et al. AMIA 2019)



Latent Topic Modeling

NLP in Biomedicine

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,⁸ two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments

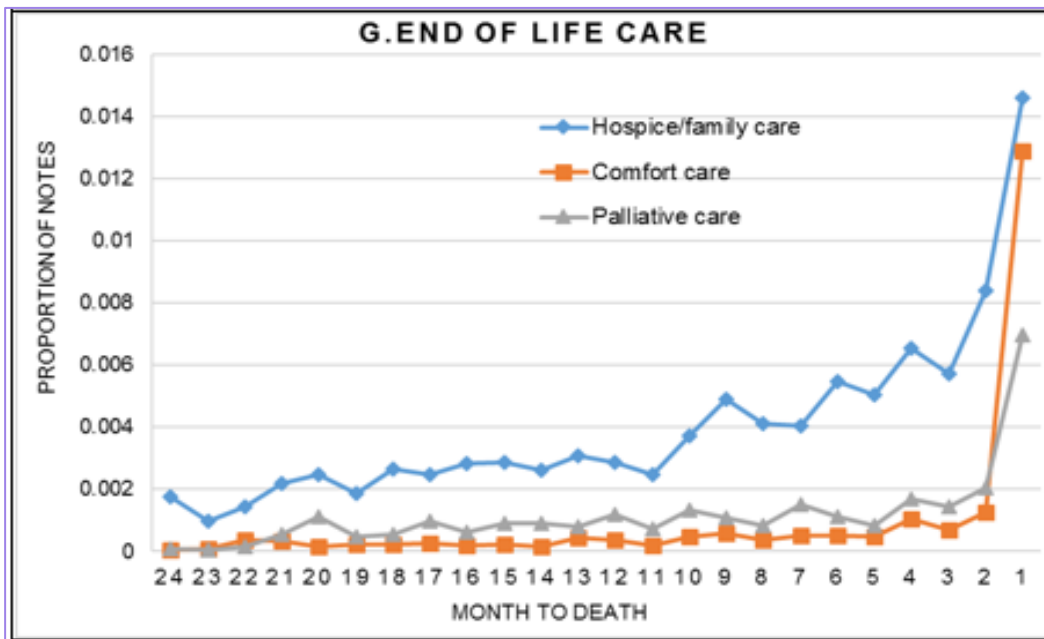
Topic modeling applies statistical-based unsupervised machine learning approaches to discover abstract topics that occur in a collection of documents. The topics are clusters of similar words. Each document may have multiple topics with different proportions.

Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), pp.77-84.



Topics

NLP in Biomedicine



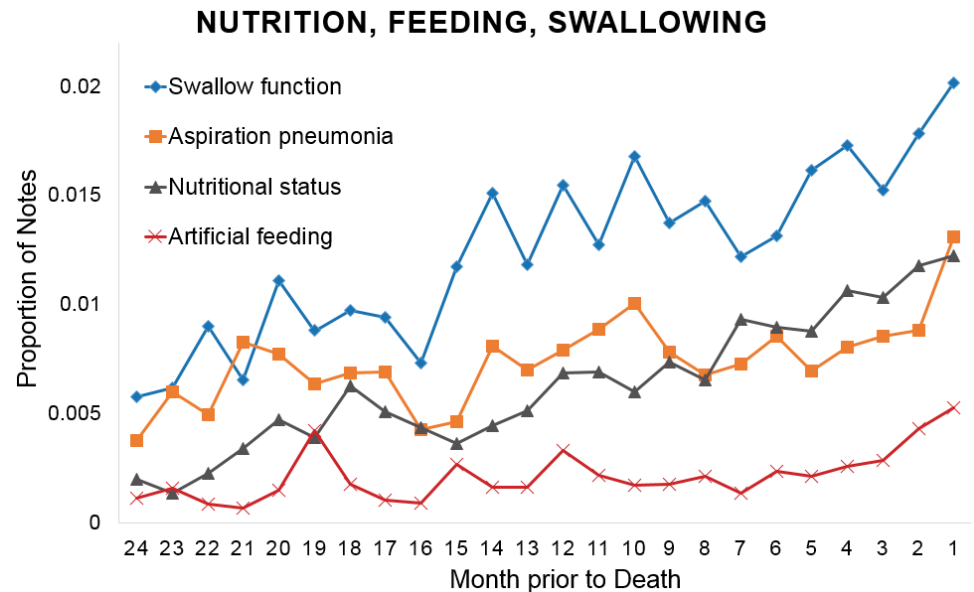
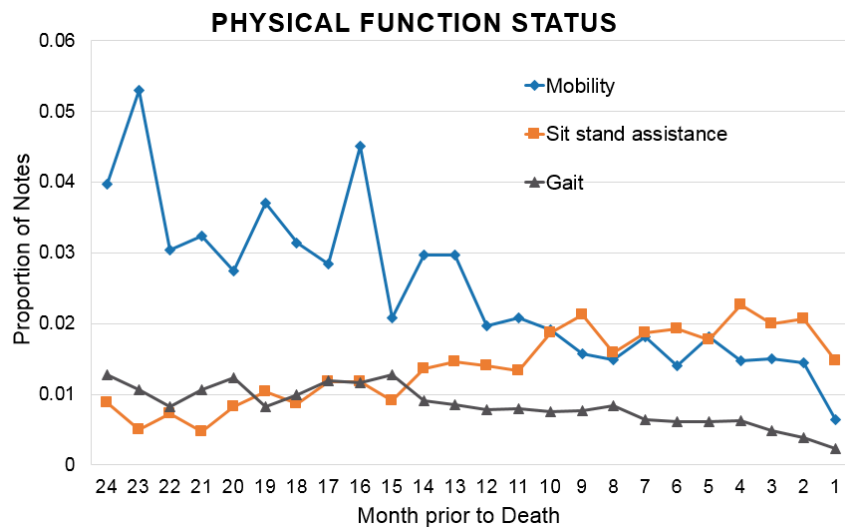
End of life care	Family/hospice care	care family hospice home dnr dementia palliative dni discussion goal intake daughter failure admission thrive
	Comfort care	comfort prn care cmo morphine family hospice measure transition pain comfortable palliative palliative dni dilaudid dnr
	Palliative care	care palliative pain prn continue family delirium time comfort review symptom well management agitation follow

Wang L, Lakin J, Riley C, Korach Z, Frain L, Zhou L. Disease Trajectories and End-of-Life Care for Dementias: Latent Topic Modeling and Trend Analysis Using Clinical Notes. AMIA Annu Symp Proc. 2018 (Distinguished Paper Award)



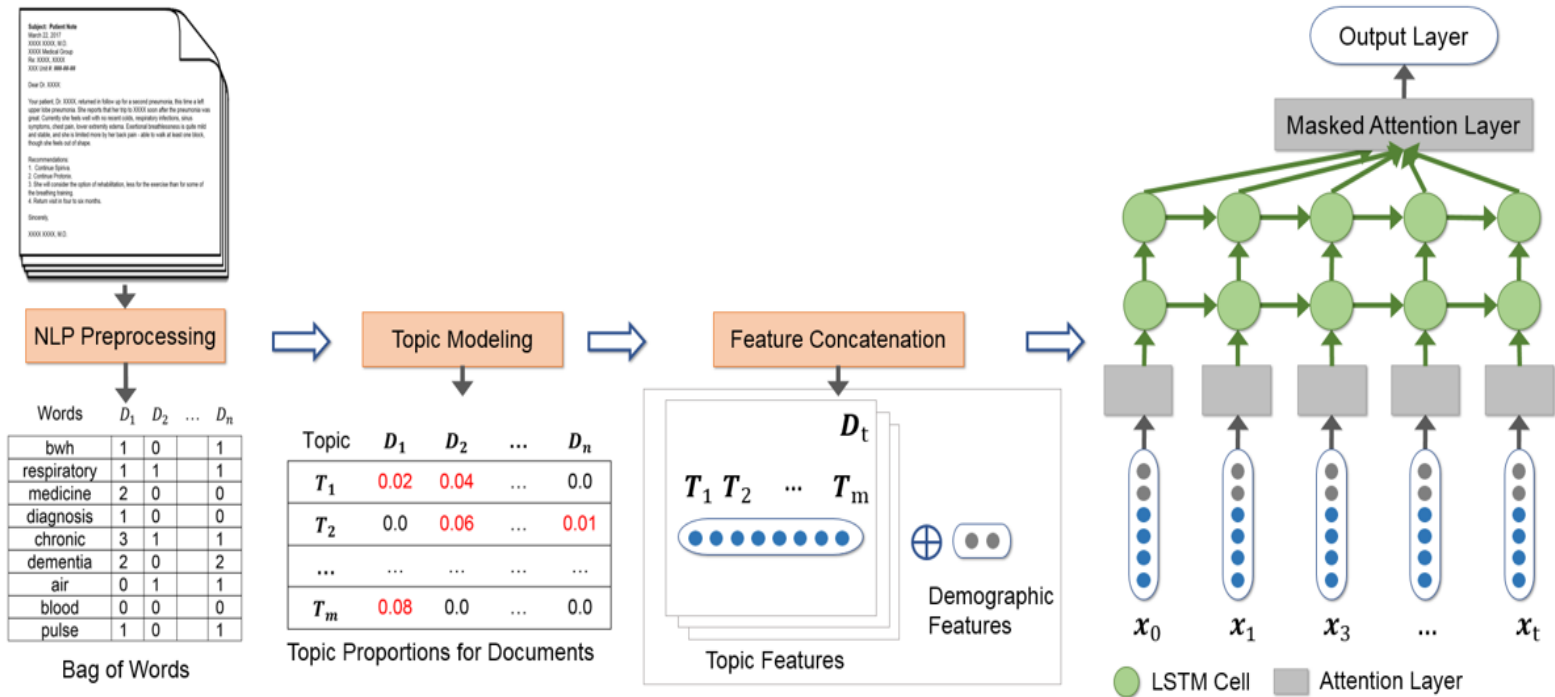
Topics in clinical notes during patients' last two years of life

NLP in Biomedicine



Wang L, Lakin J, Riley C, Korach Z, Frain L, Zhou L. Disease Trajectories and End-of-Life Care for Dementias: Latent Topic Modeling and Trend Analysis Using Clinical Notes. *AMIA Annu Symp Proc.* 2018 Dec 5;2018:1056-1065 (Distinguished paper award).

Deep learning for mortality prediction

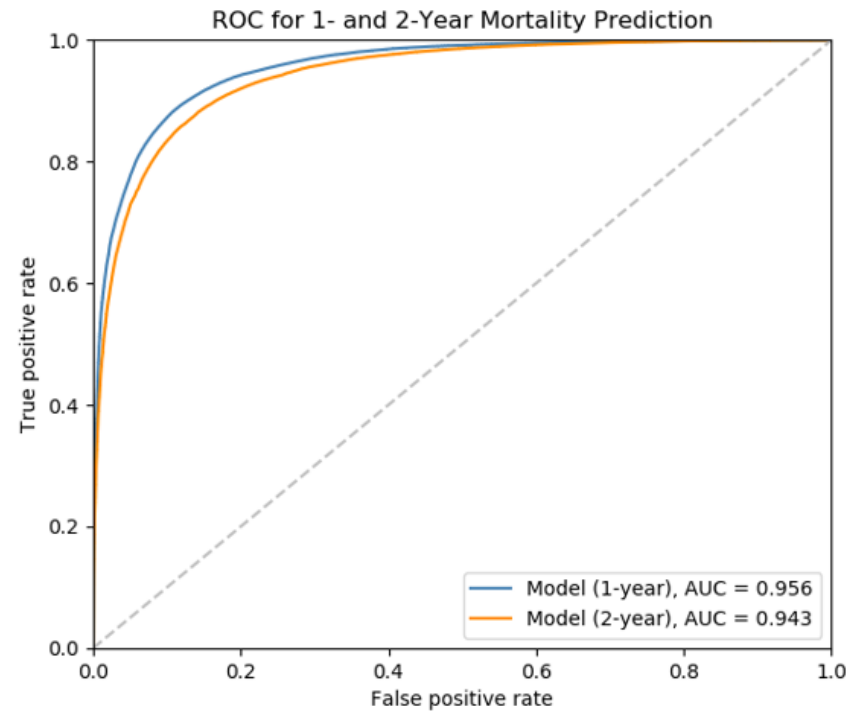


Wang L, Sha L, Lakin JR, Bynum J, Bates DW, Hong P, Zhou L. Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients With Dementia for Earlier Palliative Care Interventions. *JAMA Netw Open*. 2019.2(7):e196972.



Results

Our study shows promising results in patient stratification for clinical practice

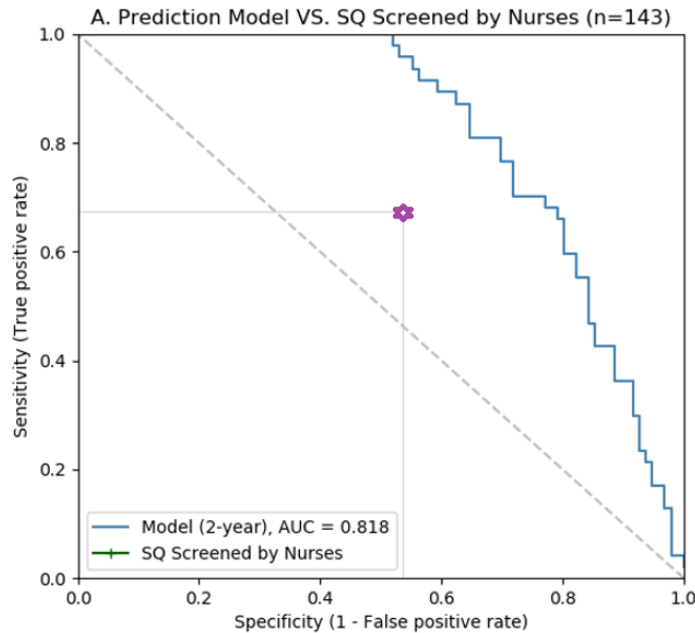


Wang L, Sha L, Lakin J, Bynum J, Bates DW, Hong P, Zhou L. JAMA Network Open, 2019.

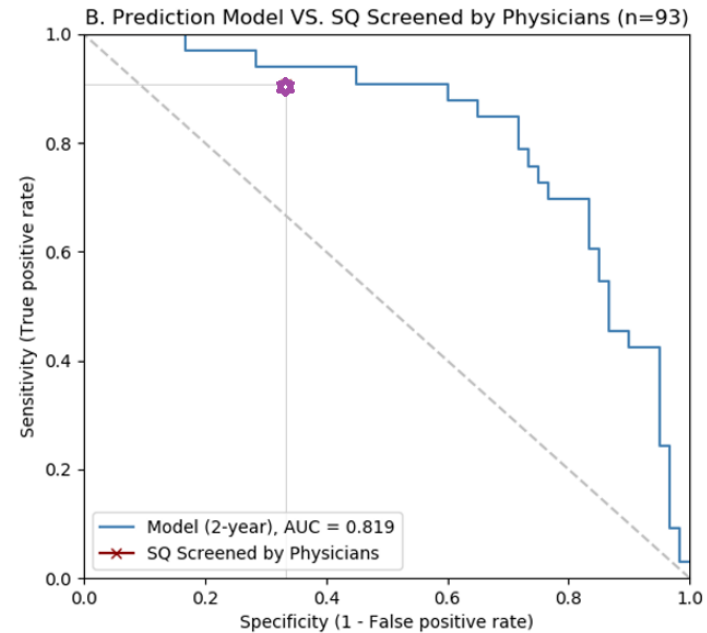


Results

Our mortality prediction model performs better than clinician screening



Sensitivity: 0.674
Specificity: 0.792 (model) vs 0.536 (SQ)



Sensitivity: 0.909
Specificity: 0.525 (model) vs 0.330 (SQ)



Early Detection of Cognitive Decline Using Large Language Models (LLMs)

- 📄 Early detection of cognitive decline in elderly individuals can facilitate clinical trial enrollment and timely medical interventions.
- 📄 Clinical notes within EHRs contain critical information on cognitive decline, detailing symptoms like memory loss, language difficulties, and impaired daily activities.
- 📄 NLP offers a powerful tool to identify these early signs of decline, which may not be coded in diagnoses.
- 📄 We applied, evaluated and compared advanced NLP techniques for identifying evidence of cognitive decline in clinical notes.



Data Sources and Setting

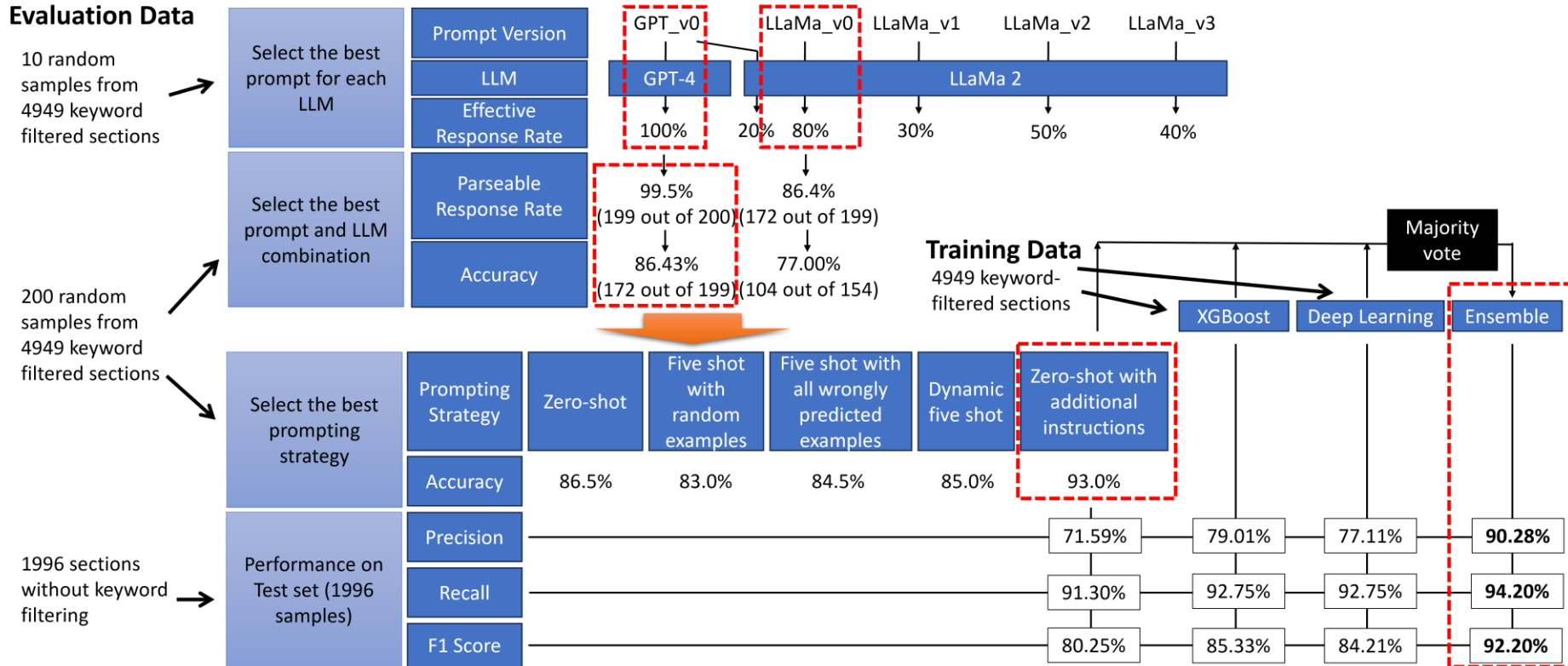
- Data:
 - Four years prior to initial mild cognitive impairment (MCI) diagnosis in 2019 for patients aged 50 years and older.
 - Model development: 4,949 sections filtered by keywords.
 - Model testing: a random sample of 1,996 sections not subjected to keyword filtering.
- Models:
 - XGBoost.
 - CNN+LSTM+Attention.
 - LLMs: GPT-4 and Llama 2
 - Ensemble with majority vote.

Preliminary data, please don't distribute



Results

Evaluation Data

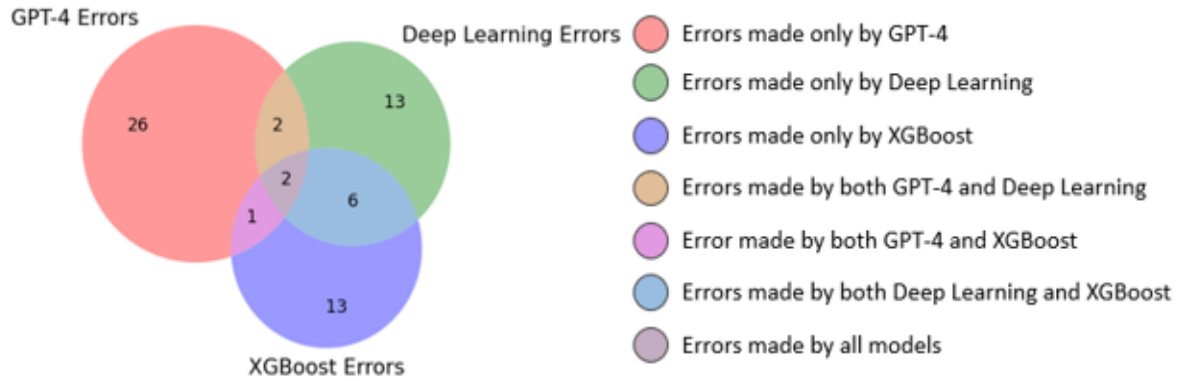


Preliminary data, please don't distribute



Error Analysis

NLP in Biomedicine



Possible reasons	Examples	GPT-4	LSTM	XGBoost	Areas in Venn Diagram
Signs/symptoms caused by other factors or clinical conditions	Pt is presenting with several days of malaise, poor appetite, slow cognition , and found to have pancreatitis c/b hypotension, now hemodynamically stable.	X	X	X	
Errors caused by negation or other contextual information	1 reports significant deficits that affect daily life ?2 3 = Referral does not indicate any cognitive deficits.	✓	X	X	
Errors caused by ambiguity (e.g., mCi stands for millicurie instead of mild cognitive decline)	The patient was injected with 10 mCi and 32.6. mCi of Tc-99m Sestamibi at rest and during peak stress, respectively, and SPECT imaging was performed.	✓	X	X	
Infer /amplify the nuanced information	Pt remains at heightened risk for falls and demos poor safety judgment . I/E that pt hx of falls and strength and balance impairments place pt at heightened risk for another fall. Pt remains reluctant but son Carl states he is working on putting a plan in place.	X	✓	✓	



PASCLex: Post-Acute Sequelae of COVID-19 (PASC) Symptom

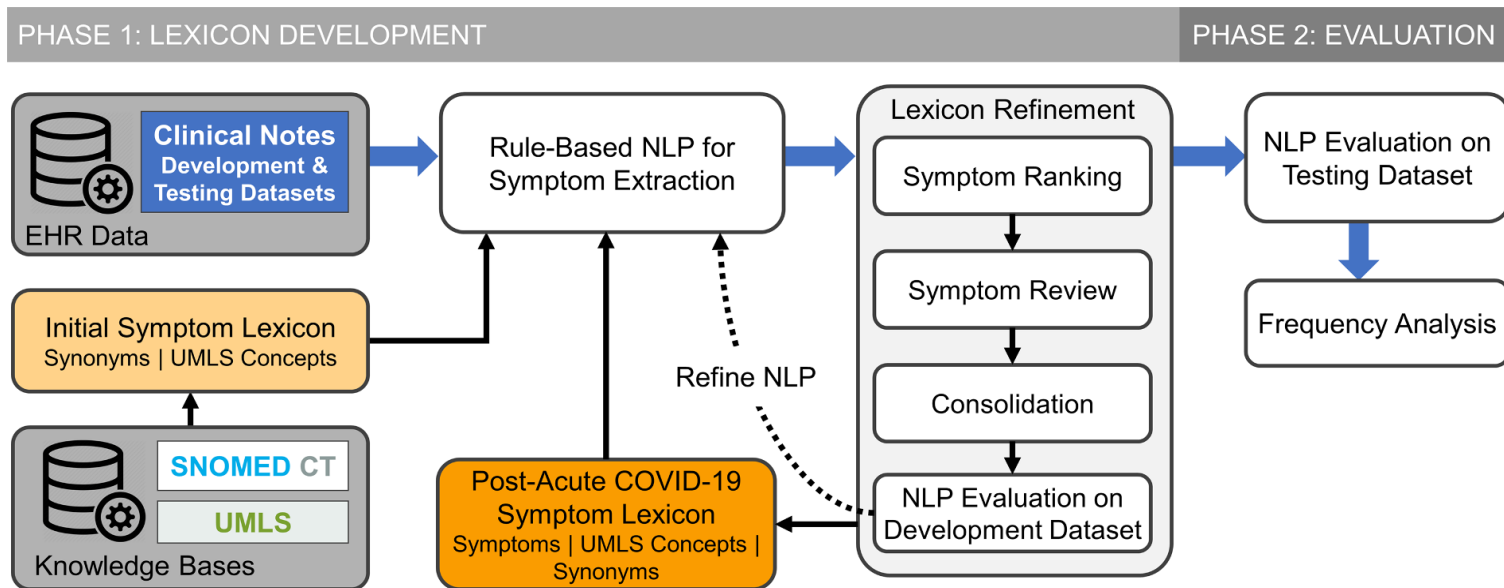
- 📄 **PSAC syndrome or long COVID:** some patients have persistent symptoms and/or develop delayed or long-term complications after their recovery from acute COVID-19.
- 📄 Most early studies on PASC symptoms relied on patient survey data, manual chart, and in person follow-up.
 - Simple size, reporting bias
- 📄 Longitudinal EHR data serve as a rich data source for studying PASC symptoms
 - Structured data (lab results or diagnosis codes)
- 📄 NLP can automatically identify relevant symptoms and complications at different clinical stages from large volumes of longitudinal notes of a large patient cohort
- 📄 To capture wide variation in potential symptoms, a comprehensive lexicon encoded with a standard terminology is crucial for NLP tool development and utility and future EHR-based PASC analytics and research.



PASCLex: Methods

NLP in Biomedicine

- Ontology-driven, EHR-guided and NLP-assisted approach
- PASC symptom lexicon was derived from 328,879 clinical notes of 26,177 COVID-19 patients documented between day 51-100 after their first positive COVID-19 test.



Wang L, Foer D, MacPhaul E, Lo YC, Bates DW, Zhou L. PASCLex: A Comprehensive Post-Acute Sequelae of COVID-19 (PASC) Symptom Lexicon Derived from Electronic Health Record Clinical Notes. Journal of Biomedical Informatics. 2021 Nov 13:103951. <https://pubmed.ncbi.nlm.nih.gov/34785382/>



PASCLex

NLP in Biomedicine

PASCLex includes 355 symptoms (and 16,466 synonyms) consolidated from 1,520 Unified Medical Language System® (UMLS) concepts.

Selected examples of post-acute COVID-19 symptoms, consolidated Unified Medical Language System (UMLS) concepts, and synonyms from electronic health record clinical notes.

Symptoms	Consolidated UMLS concepts	Examples of Synonyms in Clinical Notes
Fatigue	C0015672:Fatigue, C0231218: Malaise, C0015674:Chronic Fatigue Syndrome, C0023380:Lethargy, C0392674: Exhaustion, C0024528:Malaise And Fatigue, C0424585:Tires Quickly, C0849970:Tired, C0439055:Tired All The Time, C2732413:Postexertional Fatigue, C3875100:Fatigue Due To Treatment, C4075947:Occasionally Tired	Fatigue, tiredness, malaise, tired, fatigued, lethargy, ill feeling, feeling unwell, feel tired, lethargic
Loss of appetite	C0003125:Anorexia Nervosa, C0232462:Decrease In Appetite, C0426587:Altered Appetite, C1971624:Loss Of Appetite, C0566582:Appetite Problem	Loss of appetite, decreased appetite poor appetite, appetite changes, change in appetite, decrease in appetite, appetite loss
Sleep apnea	C0037315:Sleep Apnea, C0003578:Apnea, C0020530:Hypersomnia With Sleep Apnea, C0751762:Primary Central Sleep Apnea, C1561861:Organic Sleep Apnea, C2732337:Sleep Hypoventilation	Sleep apnea, apneas, apnea, sleep disturbance, sleep disturbances, sleep problems, sleep disorder

The post-acute COVID-19 symptom lexicon can be accessed at: https://github.com/bylinn/Post_Acute_COVID19_Symptom_Lexicon.



Common PACS in clinical notes

NLP in Biomedicine

50 most common post-acute COVID-19 patient symptoms in electronic health record clinical notes by symptom frequency, and corresponding precision of natural language processing (NLP) performance for unique symptom extraction.

Top 1–25 Symptoms	% frequency of symptoms	Precision	Top 26–50 Symptoms	% frequency of symptoms	Precision
Pain	43.1	0.94	Insomnia	11.2	0.94
Anxiety	25.8	0.98	Pain in extremities	10.7	1.0
Depression	24.0	0.90	Paresthesia	10.7	0.92
Fatigue	23.4	1.0	Peripheral edema	10.5	0.98
Joint pain	21.0	0.98	Palpitations	10.3	0.94
Shortness of breath	20.8	0.94	Diarrhea	10.3	0.92
Headache	20.0	0.92	Itching	9.4	0.92
Nausea and/or vomiting	19.9	1.0	Erythema	9.2	0.98
Myalgia	19.0	0.96	Lower urinary tract symptoms	8.7	0.98
Gastroesophageal reflux	18.6	0.94	Lymphadenopathy	8.3	0.96
Cough	17.5	0.92	Edema	7.9	0.88
Back pain	16.9	0.98	Weight gain	7.3	0.98
Stress	15.1	0.86	Sinonasal congestion	7.1	0.96
Fever	14.7	0.94	Pain in throat	6.4	0.98
Swelling	14.7	0.90	Abnormal gait	5.9	1.0
Bleeding	14.7	0.90	Respiratory distress	5.8	0.82
Weight loss*	14.2	0.98	Visual changes	5.8	0.92
Abdominal pain	14.1	0.98	Chills	5.6	0.86
Dizziness or vertigo	14.0	0.94	Urinary incontinence	5.6	0.96
Chest pain	12.5	0.90	Sleep apnea	5.4	0.94
Weakness	12.3	0.94	Confusion	5.4	0.98
Constipation	11.9	0.96	Hearing loss	5.2	1.0
Skin lesion	11.9	0.94	Problem with smell or taste	5.0	0.94
Wheezing	11.9	0.98	Difficulty swallowing	4.9	0.98
Rash	11.4	0.82	Loss of appetite	4.8	0.96

Research and Applications

Using Twitter data to understand public perceptions of approved versus off-label use for COVID-19-related medications

Yining Hua^{1,2}, Hang Jiang³, Shixu Lin⁴, Jie Yang⁴, Joseph M. Plasek^{1,2}, David W. Bates^{1,2}, and Li Zhou^{1,2}

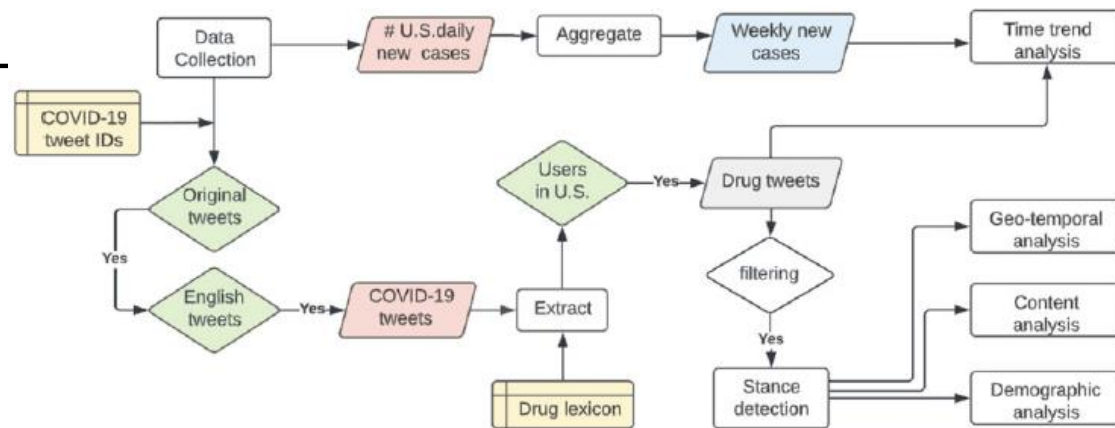


Figure 1. A comprehensive multimodal pipeline to study the public perception of drugs during the COVID-19 period.

ABSTRACT

Objective: Understanding public discourse on emergency use of unproven therapeutics is essential to monitor safe use and combat misinformation. We developed a natural language processing-based pipeline to understand public perceptions of and stances on coronavirus disease 2019 (COVID-19)-related drugs on Twitter across time.

Methods: This retrospective study included 609 189 US-based tweets between January 29, 2020 and November 30, 2021 on 4 drugs that gained wide public attention during the COVID-19 pandemic: (1) Hydroxychloroquine and Ivermectin, drug therapies with anecdotal evidence; and (2) Molnupiravir and Remdesivir, FDA-approved treatment options for eligible patients. Time-trend analysis was used to understand the popularity and related events. Content and demographic analyses were conducted to explore potential rationales of people's stances on each drug.

Results: Time-trend analysis revealed that Hydroxychloroquine and Ivermectin received much more discussion than Molnupiravir and Remdesivir, particularly during COVID-19 surges. Hydroxychloroquine and Ivermectin were highly politicized, related to conspiracy theories, hearsay, celebrity effects, etc. The distribution of stance between the 2 major US political parties was significantly different ($P < .001$); Republicans were much more likely to support Hydroxychloroquine (+55%) and Ivermectin (+30%) than Democrats. People with healthcare backgrounds tended to oppose Hydroxychloroquine (+7%) more than the general population; in contrast, the general population was more likely to support Ivermectin (+14%).

Conclusion: Our study found that social media users with have different perceptions and stances on off-label versus FDA-authorized drug use across different stages of COVID-19, indicating that health systems, regulatory agencies, and policymakers should design tailored strategies to monitor and reduce misinformation for promoting safe drug use. Our analysis pipeline and stance detection models are made public at <https://github.com/ningkko/COVID-drug>.

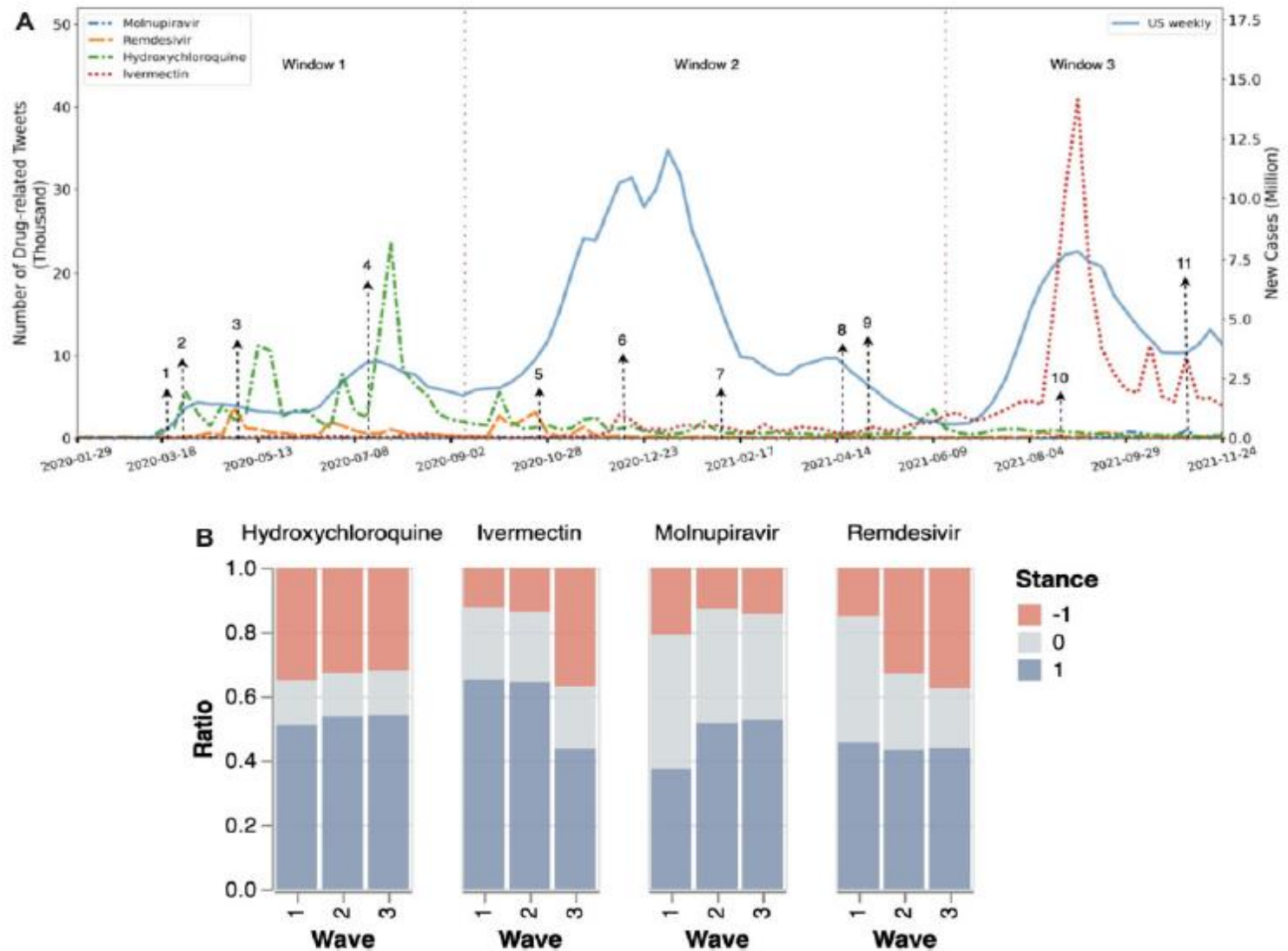


Figure 3. (A) The trends of (1) the number of tweets that mentioned COVID-19-related drugs: Hydroxychloroquine, Ivermectin, Molnupiravir, Remdesivir, and (2) weekly COVID-19 case counts (stepped line) in the United States. Wave boundaries are noted by dashed vertical lines. Major drug events are noted by numbers: (1) March 19, 2020: Trump declared Hydroxychloroquine a game-changer; (2) March 28, 2020: FDA approved a EUA to use Hydroxychloroquine for certain hospitalized patients; (3) May 1, 2020: FDA approved a EUA to use Remdesivir for severe patients; (4) July 15, 2020: FDA cautioned against the use of Hydroxychloroquine; (5) October 22, 2020: FDA approved Remdesivir for conditional use; (6) December 10, 2020: FDA cautioned against Ivermectin; (7) February 4, 2021: Merck cautioned against Ivermectin; (8) April 17, 2021: FDA clarified that Remdesivir was not approved; (9) May 1, 2021: FDA recalled a batch of Remdesivir vials, (10) August 21, 2021: FDA denounced Ivermectin as a COVID-19 treatment following an increase in overdoses; (11) November 4, 2021 Britain authorized Molnupiravir for COVID-19 treatment. (B) Distribution of percentage of tweets with positive (1, blue), neutral (0, gray), and negative (-1, red) stances for each drug. COVID-19: coronavirus disease 2019; EUA: Emergency Use Authorization FDA: US Food and Drug Administration.



NLP in Biomedicine

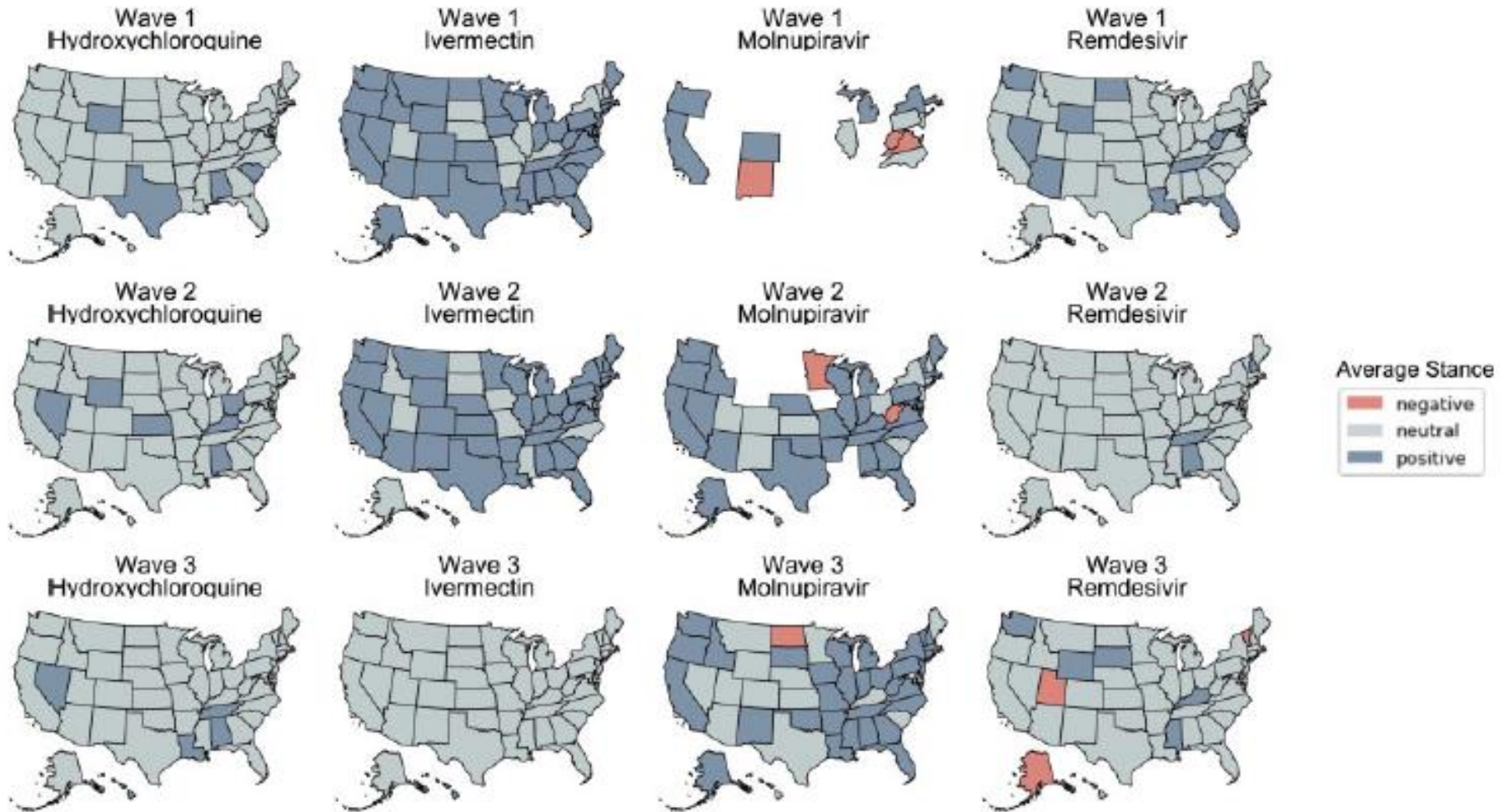
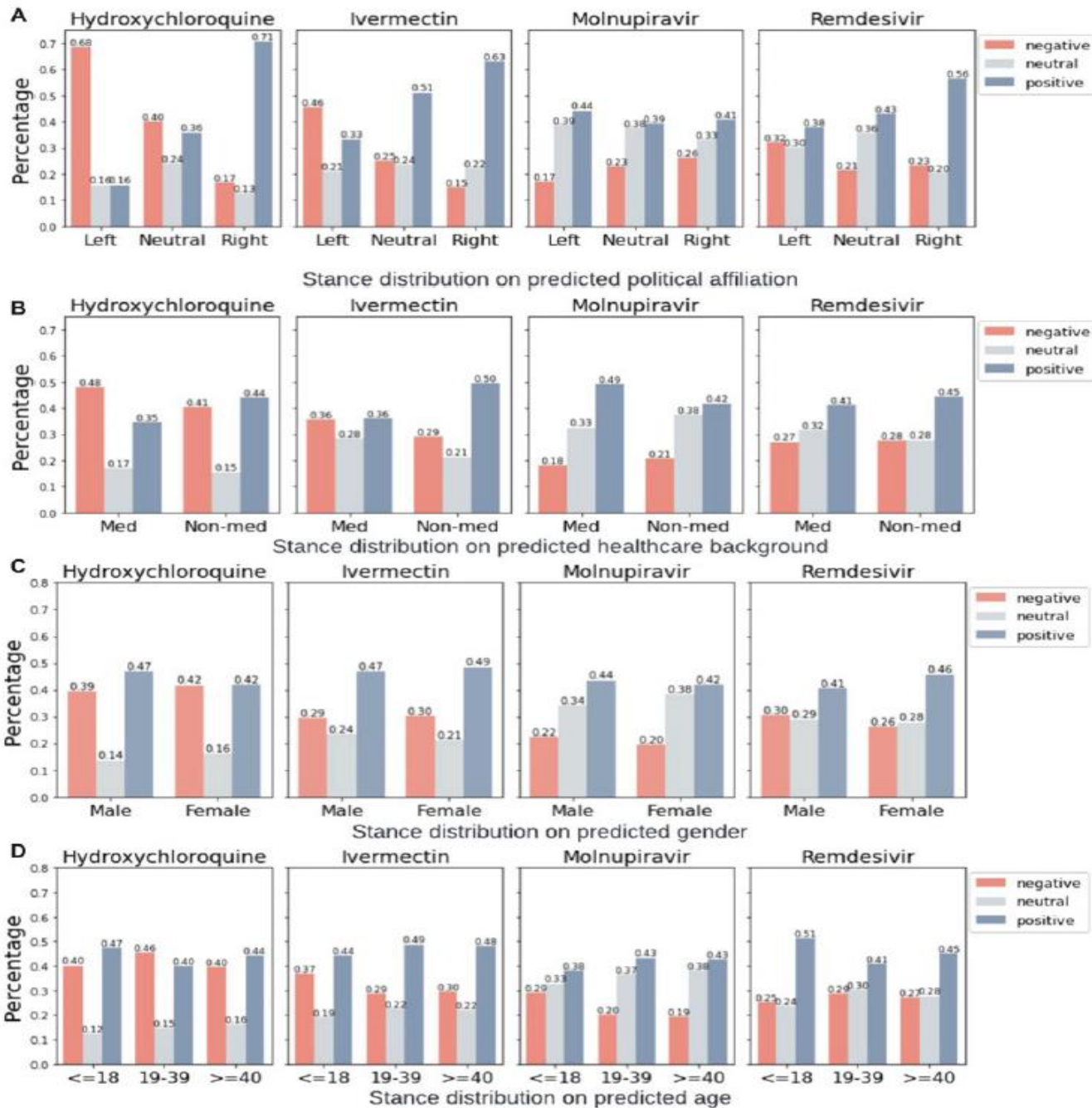


Figure 5. Longitudinal geo-temporal analysis of Tweeted sentiment of the 4 drugs by COVID-19 pandemic wave. The average sentiment of each state was classified into positive, neutral, and negative. COVID-19: coronavirus disease 2019.



- Hydroxychloroquine and Ivermectin were highly politicized, related to conspiracy theories, hearsay, celebrity effects, etc.
- The distribution of stance between the 2 major US political parties was significantly different ($P < .001$); Republicans were much more likely to support Hydroxychloroquine (+55%) and Ivermectin (+30%) than Democrats.
- People with healthcare backgrounds tended to oppose Hydroxychloroquine (+7%) more than the general population; in contrast, the general population was more likely to support Ivermectin (+14%).

Figure 6. Stance distribution on predicted partisanship, age, and medical background for each drug. The exact numbers of tweets can be found in [Supplementary Appendix SF](#).



Multi-modal AI and Generative AI

NLP in Biomedicine

nature communications

Article | [Open access](#) | [Published: 28 July 2023](#)

Knowledge-enhanced visual-language pre-training on chest radiology images

[Xiaoman Zhang](#), [Chaoyi Wu](#), [Ya Zhang](#), [Weidi Xie](#)  & [Yanfeng Wang](#) 

[Nature Communications](#) **14**, Article number: 4542 (2023) |

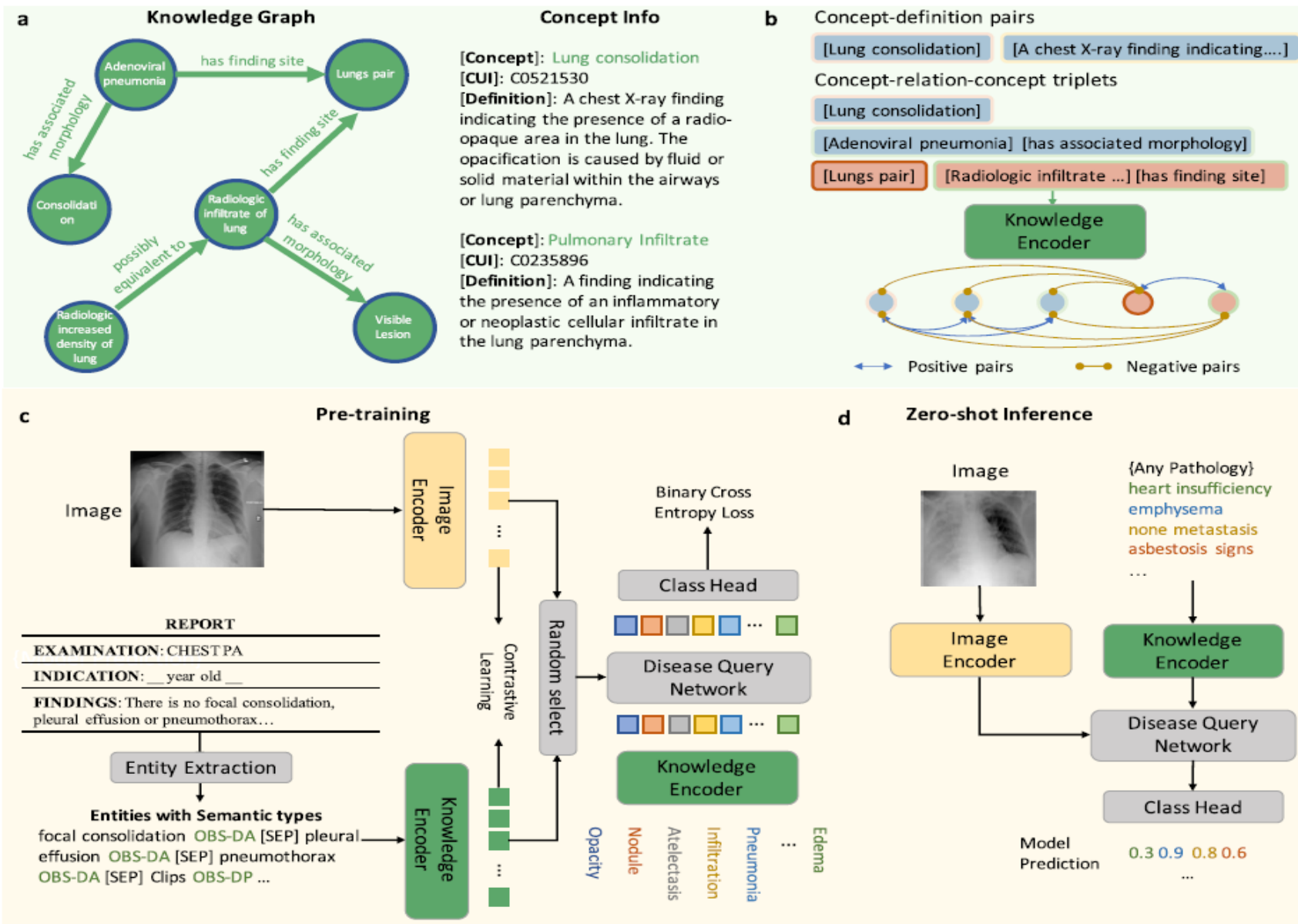


Fig. 1 | Overview of the KAD workflow. **a** Knowledge base used for training the knowledge encoder. It contains two parts, a knowledge graph consisting of concept-relation-concept triplets and a concept info list consisting of concept-definition pairs. **b** The knowledge encoder is trained to learn textual representations by maximizing similarities between positive pairs. **c** We first extract the clinical entities and relations from the radiology reports; this can be achieved by

heuristic rules, using an off-the-shelf reports information extraction toolbox (Entity Extraction), or ChatGPT, then we employ the pre-trained knowledge encoder to perform image-text contrastive learning with paired chest X-rays and extracted entities and optimize a Disease Query Network (DQN) for classification. **d** During the inference stage, we simply encode the disease name as a query input, and DQN will output the probability that the pathology is present in the input image.

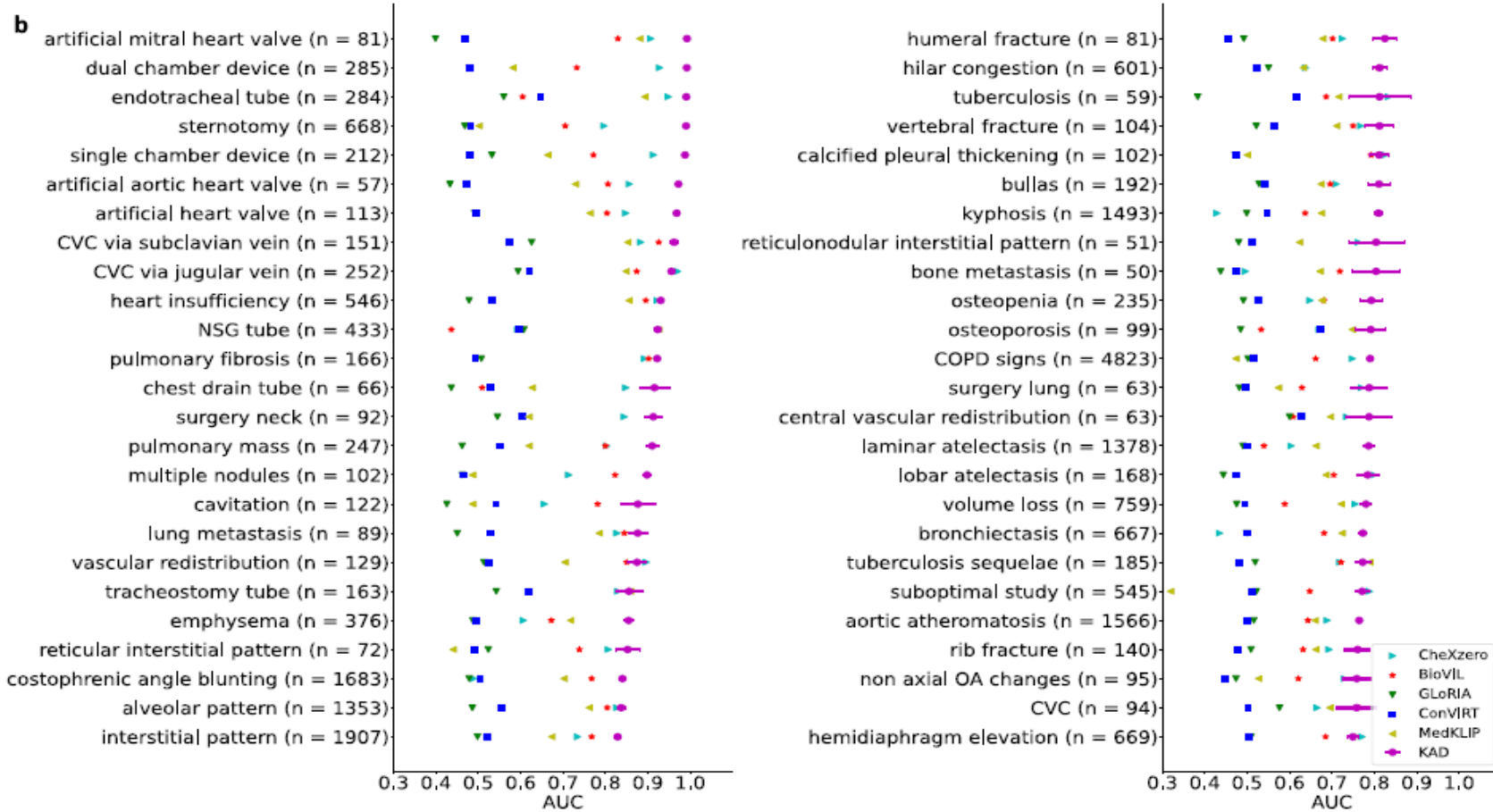
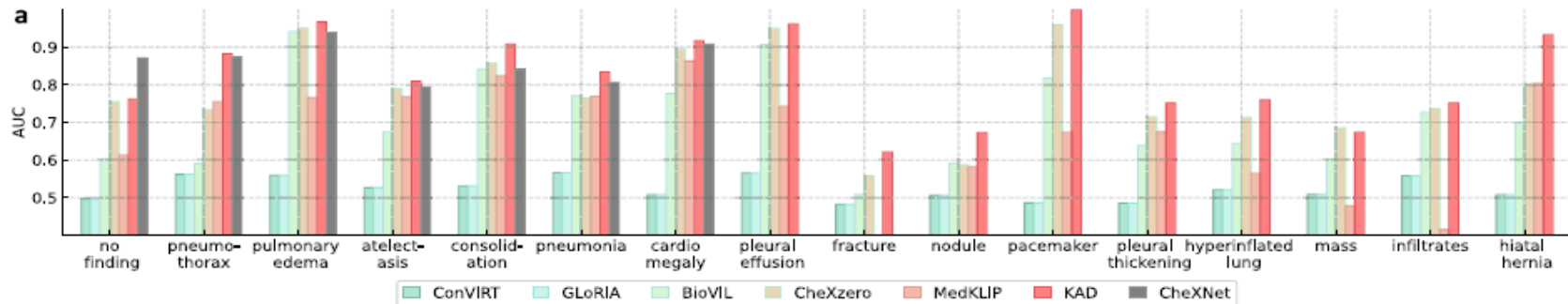


Fig. 2 | Comparison of KAD with SOTA medical image-text pre-training models under zero-shot setting on radiographic findings or diagnoses in the PadChest dataset. We evaluate model on the human-annotated subset of the PadChest dataset ($n = 39,053$ chest X-rays), and mean AUC and 95% CI of KAD are shown for each radiographic finding or diagnosis ($n > 50$). **a** Results of seen classes. Note that

CheXNet is a supervised model trained on the PadChest dataset. **b** Results of unseen classes. KAD achieves an AUC of at least 0.900 on 31 classes and at least 0.700 on 111 classes out of 177 unseen classes in the PadChest test dataset. Top 50 classes where ($n > 50$) in the test dataset ($n = 39,053$) are shown in the figure.



What's Next?

NLP in Biomedicine

- 📄 **Clinical Decision Support**
 - 📄 Retrieve relevant information, similar cases
 - 📄 Answer clinical/research questions
 - 📄 Identify high-risk, high-cost patients prospectively
- 📄 **Enhancing EHR functions**
 - 📄 Advanced search, spelling error correction, auto-fill, etc.
 - 📄 Improve clinical documentation and identify incomplete or inconsistent information
- 📄 **Text Summarization and Generation**
 - 📄 Summarize/generate a note, a specific condition, or the whole record
 - 📄 Text generation for QA, CDS and education
- 📄 **Speech Recognition**
 - 📄 Further improve usability and integration with clinical workflow and the EHR
- 📄 **Language and Diseases**
- 📄 **Multimodal data**
- 📄 **Others, e.g. computer-assisted coding**



NLP in Biomedicine

July 17, 2023

Comparison of History of Present Illness Summaries Generated by a Chatbot and Senior Internal Medicine Residents

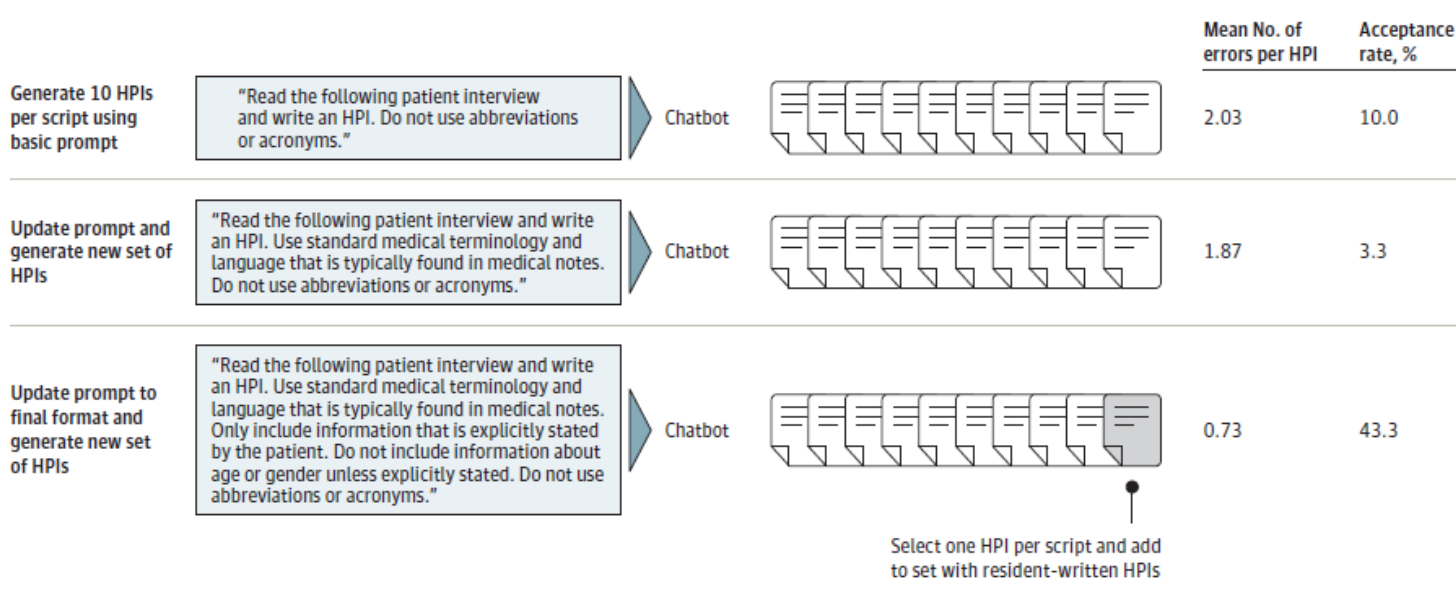
Ashwin Nayak, MD, MS¹; Matthew S. Alkaitis, MD, PhD¹; Kristen Nayak, MD¹; [et al](#)

[Author Affiliations](#) | [Article Information](#)

JAMA Intern Med. Published online July 17, 2023. doi:10.1001/jamainternmed.2023.2561

HPI generated by ChatGPT vs written by 4 Residents based on interview scripts of 3 patients with different types of chest pain

Figure. Prompt Engineering With the Chatbot



A total of 10 HPIs per script (30 total) were generated in each round of prompt engineering and assessed for major errors such as reported information in the HPIs that was not present in the source dialogue (hallucinations), the presence of atypical terminology, and structural issues. The HPIs without any major errors were considered acceptable. Prompts were refined in each round to minimize error and maximize acceptance. HPI indicates history of present illness.

Table. Composite and Domain-Specific HPI Grades^a





Grade type	Mean (SD) resident HPI grade (n = 120)	Mean (SD) chatbot HPI grade (n = 30)	Wilcoxon rank sum P value
Composite	12.18 (2.40)	11.23 (2.84)	.09
Detail-oriented	4.13 (0.86)	3.57 (1.04)	.006 ^b
Succinctness	3.93 (1.09)	3.70 (1.15)	.29
Organization	4.12 (0.91)	3.97 (0.96)	.43

Abbreviation: HPI, history of present illness.

^a HPIs were graded across 3 domains addressing level of detail, succinctness, and organization using a 5-point Likert scale with 5

representing the best grade on the domain. The composite grade was calculated as the total grade across the 3 domains.

^b Denotes statistical significance, defined as 2-sided $P < .05$.

-  HPIs generated by a chatbot or written by residents were graded similarly by attending physicians.
-  Chatbot’s performance was heavily dependent on prompt quality. Without robust prompt engineering, the chatbot frequently reported information in the HPIs that was not present in the source dialogue (“Hallucinations”).
 -  e.g., addition of patient age and gender, which none of the scripts specified.
-  The framework of model output generation needs to be evaluated to improve accuracy.

2. The patient presents with a complaint of discomfort in the chest that has been ongoing for the past 3 weeks. The discomfort is described as a 7 out of 10 in severity, usually starting in the middle of the chest and sometimes radiating to the left arm. The discomfort can come on at any time, is not related to exertion or eating, and may be worse with stress. The patient does not experience sharp pain, nausea, or vomiting. The discomfort lasts for a few minutes and is accompanied by shortness of breath. The patient reports no history of similar symptoms in the past.

3. Patient presents with 3 weeks of intermittent chest pain, which is 7 out of 10 in intensity, located in the center of the chest with radiation to the left arm, and has associated shortness of breath. The pain is usually sudden onset and lasts for several minutes. The pain is triggered by stress and sometimes occurs at random while at rest. Patient has not noted exertion to be a trigger but does not regularly exert himself. The pain resolves with rest and relaxation techniques. There is no relation to eating and the pain doesn't improve with antacids.

Capabilities of GPT-4 on Medical Challenge Problems

Harsha Nori¹, Nicholas King¹, Scott Mayer McKinney²,
Dean Carignan¹, and Eric Horvitz¹

¹Microsoft
²OpenAI

<https://arxiv.org/abs/2303.13375>



GPT-4, without any specialized prompt crafting, exceeds the passing score on USMLE by over 20 points and outperforms earlier general-purpose models (GPT-3.5) as well as models specifically fine-tuned on medical knowledge (Med-PaLM)

Research Letter

July 17, 2023

Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations

Eric Strong, MD¹; Alicia DiGiammarino, MS²; Yingjie Weng, MHS³; et al

[» Author Affiliations](#)

JAMA Intern Med. Published online July 17, 2023. doi:10.1001/jamainternmed.2023.2909

ONLINE



Compared chatbot vs medical student performance on clinical reasoning final examinations given to 1st and 2nd year students at Stanford School of Medicine.



GPT 4 outperformed first- and second-year students on clinical reasoning examinations (Scored a mean 4.2 points more than student; respective passing rates 93% vs 85%) and had significant improvement vs GPT 3.5.





Prompt engineering is important as chatbot's responses can be sensitive to rewording of prompts







As the medical community had to learn online resources and electronic medical records, the next challenge is learning judicious use of generative AI to improve patient care.

Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare

 Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E. Abdounour, Atul J. Butte,  Emily Alsentzer

doi: <https://doi.org/10.1101/2023.07.13.23292577>

-  GPT-4 does not appropriately model the demographic diversity of medical conditions, consistently producing clinical vignettes that stereotype demographic presentations.
-  The differential diagnoses created by GPT-4 for standardized clinical vignettes were more likely to include diagnoses that stereotype certain races, ethnicities, and gender identities.
-  Assessment and plans created by the model showed significant association between demographic attributes and recommendations for more expensive procedures as well as differences in patient perception.
-  These findings highlight the urgent need for comprehensive and transparent bias assessments of LLM tools like GPT-4 for every intended use case before they are integrated into clinical care.



NLP Timeline

NLP in Biomedicine

