Natural Language Processing in FDA Post-Marketing Activities: Case studies and Lessons Learned

Yong Ma, PhD
Lead Mathematical Statistician
OB/OTS/CDER/FDA

ASA Safety Working Group Quarterly Scientific Webinar – Q3 2025
Topic: Reimagining Drug Safety with AI: A Quantitative Leap Forward (Part 2)
October 7, 2025

Disclaimer

This presentation reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mentions of organizations, companies, stakeholders or commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

This material was presented at the ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop on September 25, 2025

Overview

Five NLP Projects (2018-Present)

- **Project 1-2:** Pharmacovigilance (FAERS enhancement)
- **Project 3-4:** Pharmacoepidemiology (EHR processing)
- **Project 5:** Public health emergency response (COVID-19 surveillance)

The FAERS Challenges

Missing Demographic Data Issue

- FAERS: FDA's adverse event reporting system
- Age data missing in 22% of reports (2002) → 44% (2018)
- Critical impact on pediatric safety monitoring
- Similar issues with gender, weight, race, and ethnicity data

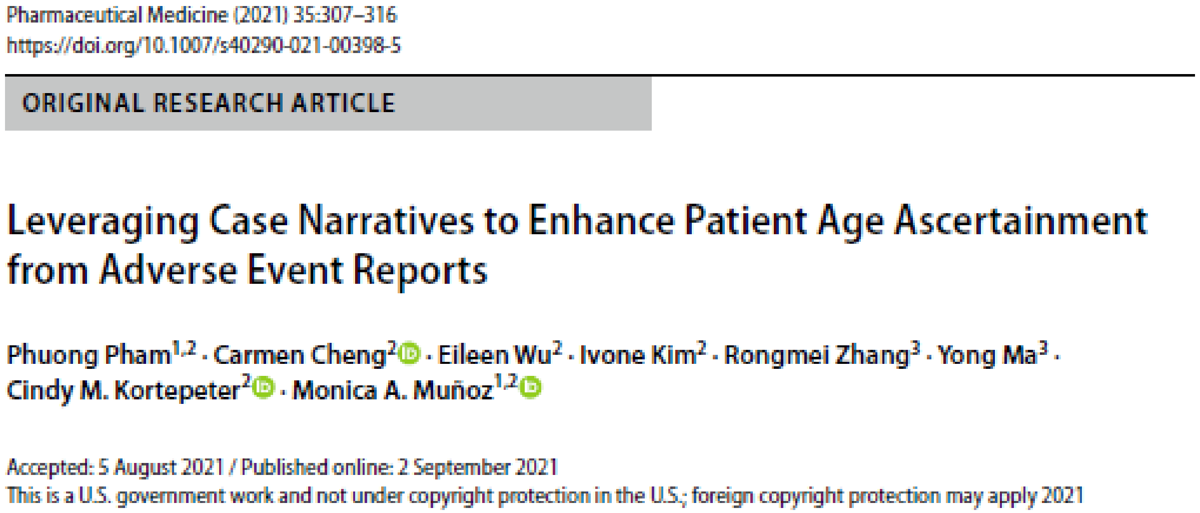
Duplicate Reporting Issue

- Multiple reporting of the same AE
- Inflated case counts → overestimated reporting rates and incorrect risk assessments
- Signal prioritization problems

Project 1 - Age Extraction Solution

Rule-Based NLP for Missing Age Data

- Simple algorithm searching for numbers + "years" or "years old"
- Fallback to months conversion when years unavailable
- Extracted first age found in narrative text
- No training data required, works across different FAERS runs



Pharmaceutical Medicine (2021) 35:307–316
https://doi.org/10.1007/s40290-021-00398-5

ORIGINAL RESEARCH ARTICLE

Leveraging Case Narratives to Enhance Patient Age Ascertainment from Adverse Event Reports

Phuong Pham[1,2] · Carmen Cheng[2] · Eileen Wu[2] · Ivone Kim[2] · Rongmei Zhang[3] · Yong Ma[3] · Cindy M. Kortepeter[2] · Monica A. Muñoz[1,2]

Accepted: 5 August 2021 / Published online: 2 September 2021

Age Extraction Results

In the testing sample

- **98.5%** accuracy, **92.9%** specificity, **94.9%** precision

When applied to the 2002-2018 reports

- 1 million additional reports with age data extracted
- Overall missing age data: 37% → 27%
- Pediatric cases under 6 years: more than doubled reports with known ages
- Remaining 27% limitation: information simply not in reports

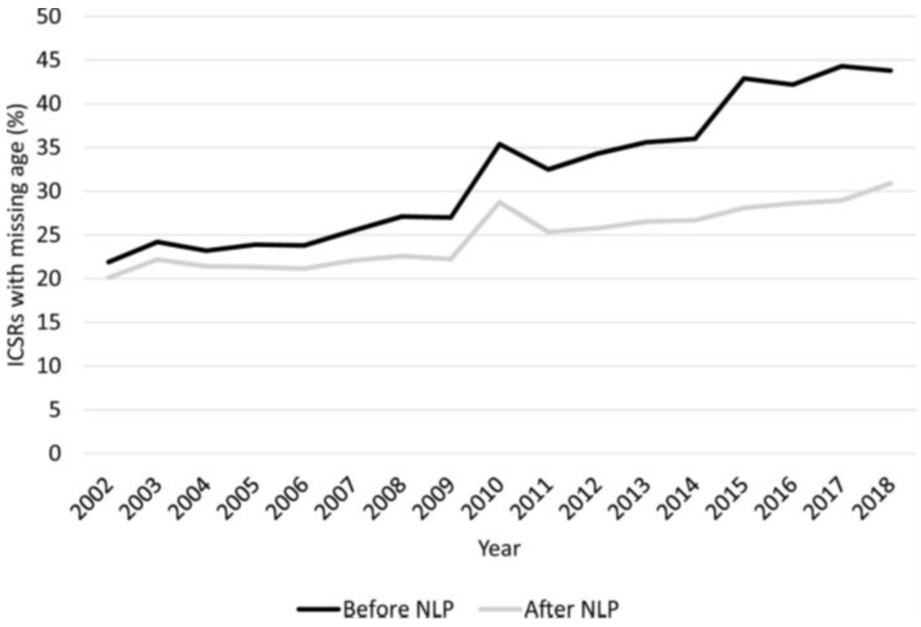


Figure 1. Percentage of FAERS ICSRs with missing age before and after NLP implementation.

Expanding Demographics Extraction

Beyond Age: Gender, Weight, Race, Ethnicity

- Tailored mini-algorithms for each demographic variable
- Gender algorithm: 98.6% sensitivity, 33% reduction in missing data
- Weight, race, ethnicity: high specificity but low sensitivity
- Key insight: "You can't extract what doesn't exist"

frontiers | Frontiers in Drug Safety and Regulation

TYPE Brief Research Report
PUBLISHED 14 November 2022
DOI 10.3389/fdsfr.2022.1020943

Check for updates

OPEN ACCESS

EDITED BY
G. Niklas Norén,
Uppsala Monitoring Centre, Sweden

REVIEWED BY
Luis Pinheiro,
European Medicines Agency,
Netherlands
Xiaofei Ye,
Second Military Medical University,
China

*CORRESPONDENCE
Vivian Dang,
Vivian.Dang@fda.hhs.gov

SPECIALTY SECTION
This article was submitted to Advanced Methods in Pharmacovigilance and Pharmacoepidemiology, a section of the journal Frontiers in Drug Safety and Regulation

RECEIVED 16 August 2022
ACCEPTED 26 October 2022
PUBLISHED 14 November 2022

Evaluation of a natural language processing tool for extracting gender, weight, ethnicity, and race in the US food and drug administration adverse event reporting system

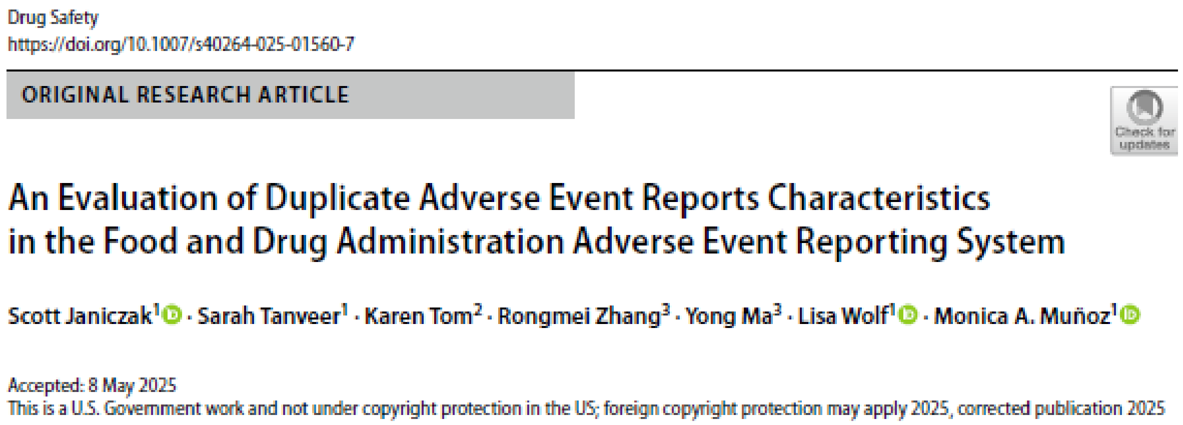
Vivian Dang [1]*, Eileen Wu[1], Cindy M. Kortepeter [1], Michael Phan[1], Rongmei Zhang[2], Yong Ma[2] and Monica A. Muñoz [1]

[1]Division of Pharmacovigilance, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States, [2]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

Project 2 - Duplicate Detection

Addressing FAERS Duplicate Reports Challenge

- Multiple reporters submitting similar narratives for same events
- Distorts safety signal detection and inflates counts
- Solution: Sentence-BERT (SBERT) for semantic similarity analysis
- Used all-MPNet-base-v2 variant for narrative embeddings



Drug Safety
https://doi.org/10.1007/s40264-025-01560-7

ORIGINAL RESEARCH ARTICLE

Check for updates

An Evaluation of Duplicate Adverse Event Reports Characteristics in the Food and Drug Administration Adverse Event Reporting System

Scott Janiczak[1] · Sarah Tanveer[1] · Karen Tom[2] · Rongmei Zhang[3] · Yong Ma[3] · Lisa Wolf[1] · Monica A. Muñoz[1]

Accepted: 8 May 2025
This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025, corrected publication 2025

Duplicate Detection Results

- Confirmed duplicates: Median cosine similarity **0.87**
- Non-duplicates: Median cosine similarity **0.48**
- At 0.73 threshold: **96% sensitivity** and **96% specificity**
- May serve as a decision support tool for expert review

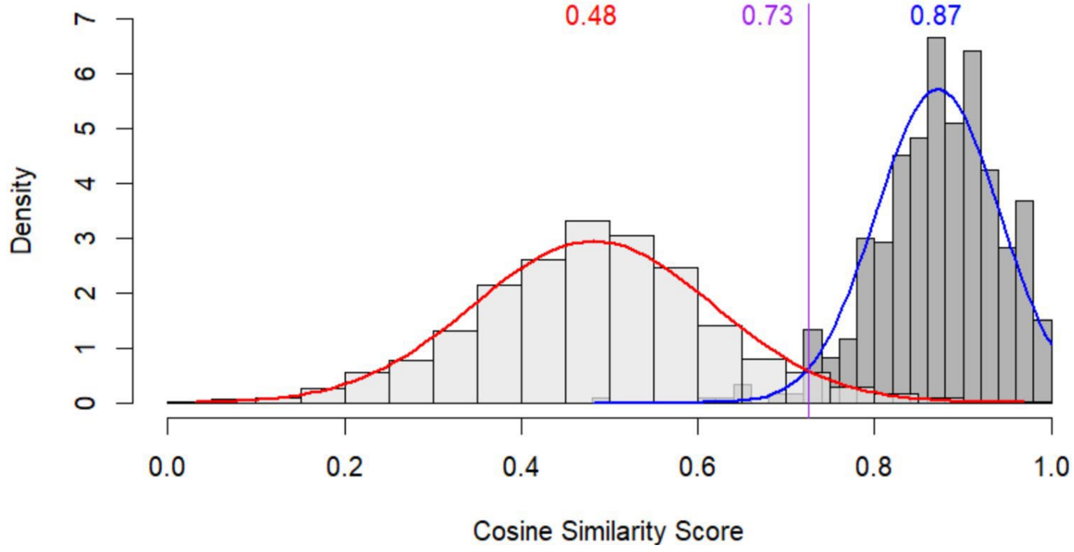


Figure: Distribution of cosine similarity analysis of narrative text. reprinted from:

Janiczak S, Tanveer S, Tom K, Zhang R, Ma Y, Wolf L, Muñoz MA. An Evaluation of Duplicate Adverse Event Reports Characteristics in the Food and Drug Administration Adverse Event Reporting System. Drug Saf. 2025 Jun 4; Licensed under CC BY 4.0.

The EHR challenge

Unstructured data

- Outcome data hidden in in clinical notes
- Covariate information also in the narratives

Large amount of data makes manual processing cumbersome and expensive

Project 3 - Anaphylaxis Identification

Advanced ML for Critical Safety Events

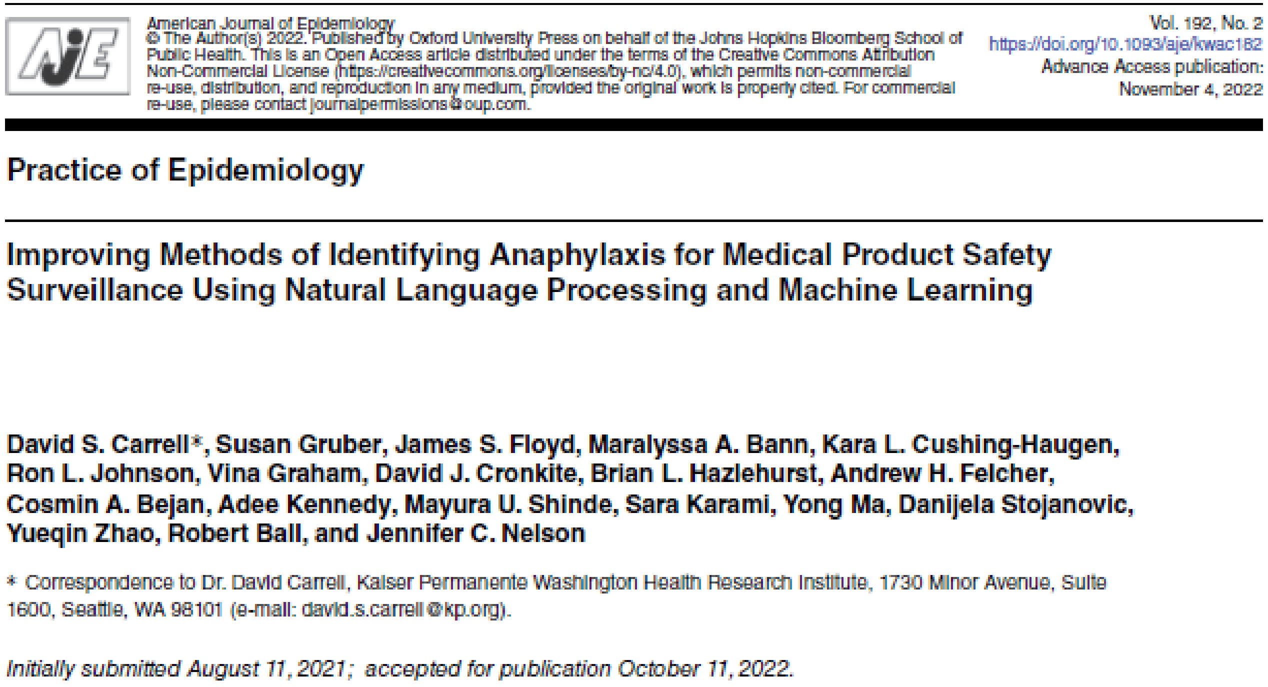
- Challenge: Identifying anaphylaxis in electronic health records
- Complex clinical presentations and "rule-out" coding practices
- Existing algorithms: Only 63% positive predictive value
- FDA ARIA requirement: ≥80% threshold needed

Anaphylaxis NLP Methodology

Sophisticated Feature Engineering Approach

- Custom dictionary with clinical expert review
- UMLS concepts from published literature
- Enriched with synonyms from UMLS Metathesaurus and manual review
- ConText-like algorithm for affirmative mentions
- 468 candidate NLP-derived covariates engineered
- 100 covariates selected through expert judgment

—
UMLS: Unified Medical Language System



AJE
American Journal of Epidemiology

Vol. 192, No. 2
https://doi.org/10.1093/aje/kwac182
Advance Access publication:
November 4, 2022

Practice of Epidemiology

Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning

David S. Carrell*, Susan Gruber, James S. Floyd, Maralyssa A. Bann, Kara L. Cushing-Haugen, Ron L. Johnson, Vina Graham, David J. Cronkite, Brian L. Hazlehurst, Andrew H. Felcher, Cosmin A. Bejan, Adee Kennedy, Mayura U. Shinde, Sara Karami, Yong Ma, Danijela Stojanovic, Yueqin Zhao, Robert Ball, and Jennifer C. Nelson

* Correspondence to Dr. David Carrell, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: david.s.carrell@kp.org).

Initially submitted August 11, 2021; accepted for publication October 11, 2022.

Anaphylaxis Results

- NLP-enhanced models: **AUC 0.71** vs. structured data only: **AUC 0.62**
- At 66% sensitivity threshold: **79% positive predictive value**
- Substantial improvement over existing 63% PPV methods
- Demonstrates value of unstructured data integration

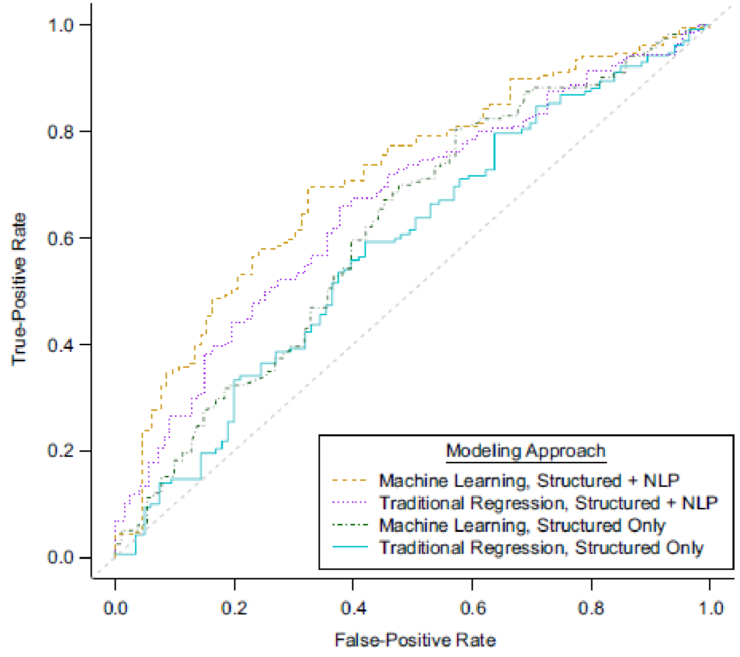


Figure Weighted cross-validated area under the receiver operating characteristic curve for Kaiser Permanente Washington algorithms identifying actual anaphylaxis events in Kaiser Permanente Washington data (2015–2019) using the best machine-learning approach applied to structured and all natural language processing (NLP) data, traditional logistic regression approach applied to structured and all NLP data, machine-learning approach applied to structured data only, and traditional logistic regression approach applied to structured data only

Project 4 - MOSAIC-NLP

Multi-source **O**bservational **S**afety study for **A**dvanced **I**nformation **C**lassification using **NLP**

Large-Scale study to demonstrate value, scalability and transportability in pharmacoepidemiology study

- ***Study Design****:* Retrospective cohort study
- ***Study Data****:* EHR-claims linked structured and unstructured data (2015-2022)
- ***Study Cohort****:* Patients with asthma newly initiating montelukast (monotherapy) or inhaled corticosteroids (comparator)
- ***Study Outcomes****:* Neuropsychiatric events

Mosaic-NLP Methodology

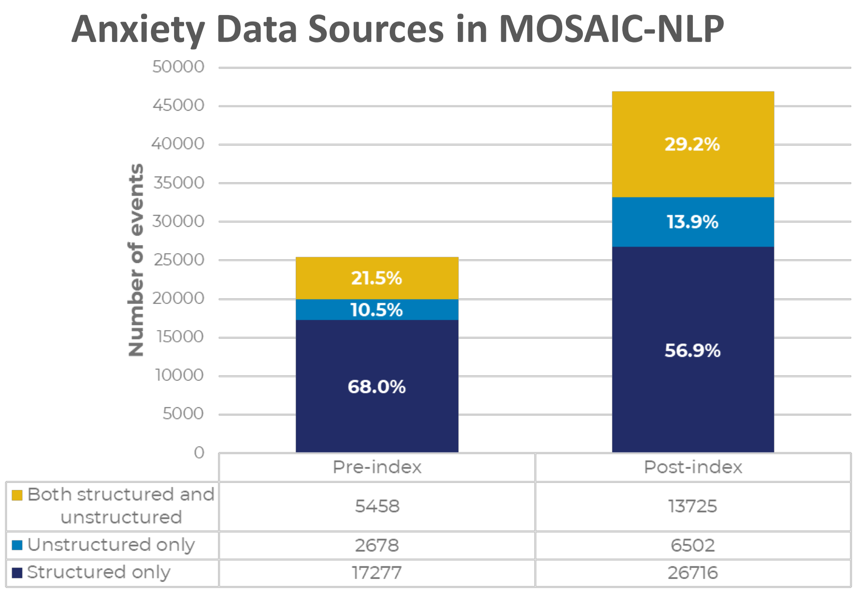
Named entity recognition (NER) framework

- Entity Extraction
- Entity Model Training
- Entity Model Tuning

MOSAIC-NLP Results

- Enhanced detection of suicidality, attempt, and self-harm
- Enhanced detection of covariates

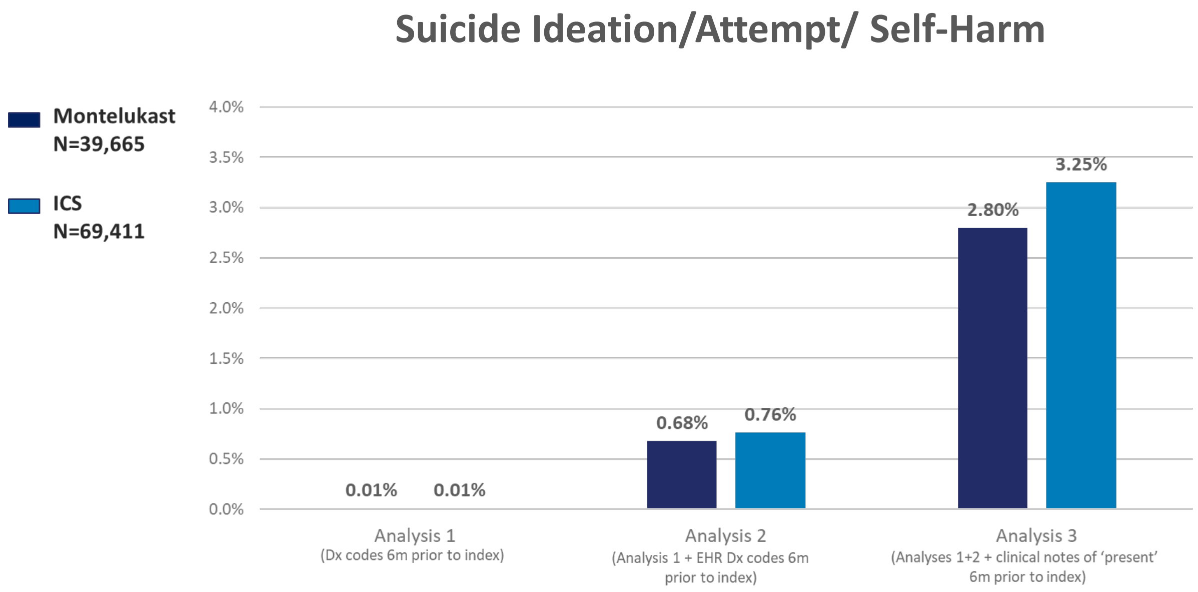
Anxiety Data Sources in MOSAIC-NLP



	Pre-index	Post-index
🟨 Both structured and unstructured	5458	13725
🟦 Unstructured only	2678	6502
■ Structured only	17277	26716

Pre-index (unmatched covariates) – Had Anxiety (n = 25,413); Post-index (unmatched outcomes)- Had Anxiety event (n = 46,943)

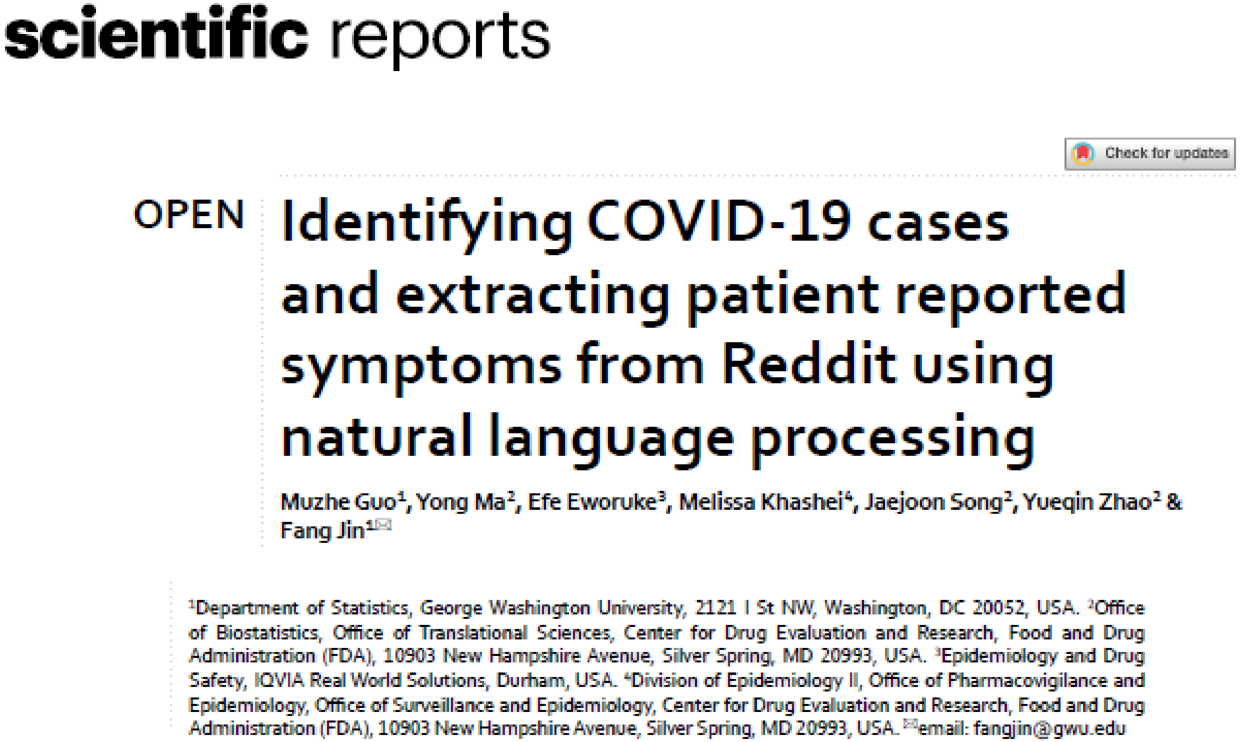
Suicide Ideation/Attempt/ Self-Harm



Analysis 1 = claims only data; Analysis 2 = claims + EHR structured data; Analysis 3 = claims + EHR structured data + EHR unstructured data

Project 5 - COVID-19 Social Media Surveillance

- Traditional surveillance struggling with real-time reporting
- Reddit as source for patient-reported symptoms
- Two-stage pipeline: case identification + symptom extraction
- BERT-Large model achieving **91.2%** accuracy for case identification



scientific reports

Check for updates

OPEN

Identifying COVID-19 cases and extracting patient reported symptoms from Reddit using natural language processing

Muzhe Guo[1], Yong Ma[2], Efe Eworuke[3], Melissa Khashei[4], Jaejoon Song[2], Yueqin Zhao[2] & Fang Jin[1]✉

[1]Department of Statistics, George Washington University, 2121 I St NW, Washington, DC 20052, USA. [2]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration (FDA), 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA. [3]Epidemiology and Drug Safety, IQVIA Real World Solutions, Durham, USA. [4]Division of Epidemiology II, Office of Pharmacovigilance and Epidemiology, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, Food and Drug Administration (FDA), 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA. ✉email: fangjin@gwu.edu

QuadArm Framework

Four-Step Symptom Extraction Process

1. **BERT/BioBERT** question-answering for symptom identification
2. **Word embeddings** expansion using GoogleNews word2vec
3. **Adaptive Rotation Clustering (ARC)** for semantic grouping
4. **UMLS mapping** for standardized medical terminology

COVID-19 Results

Tracking Pandemic Evolution

- Symptom trends across early, Delta, and Omicron waves
- Decreased loss of smell reports over time
- Increased sore throat mentions
- Consistent with CDC epidemiological reports

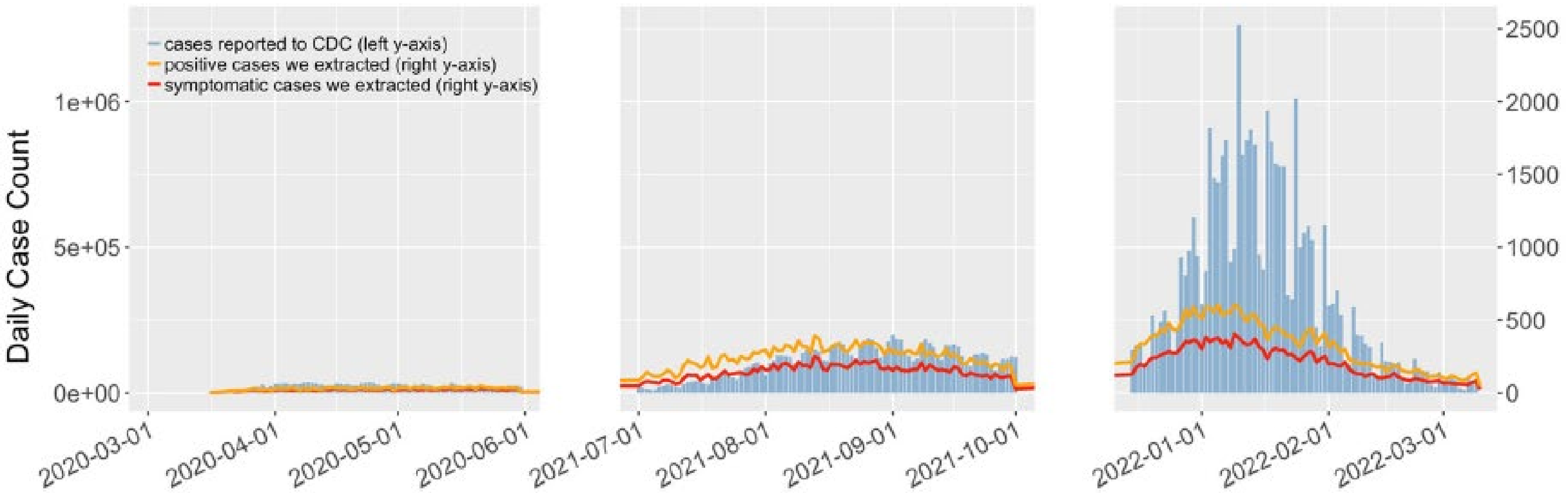


Figure 4. Daily trends in number of COVID-19 cases reported to the CDC and we extracted, for the corresponding three periods.

Take Home Message

- NLP is a powerful tool across pharmacovigilance, pharmacoepidemiology, and public health surveillance
- Text sources vary: spontaneous reports, clinical notes, social media
- Methodology continues to evolve over time
- Automated processing essential for scalibility and accuracy

Future Directions
Continued Innovation in Healthcare NLP

- Evolution toward more sophisticated language models
- Integration of structured and unstructured data sources
- Real-time surveillance capabilities expansion
- Enhanced patient safety through improved data extraction, processing and modeling

FDA U.S. FOOD & DRUG
ADMINISTRATION