Received: 28 February 2024 | Revised: 15 April 2024 | Accepted: 17 April 2024

DOI: 10.1111/acem.14937

AEM Academic Emergency Medicine

SYSTEMATIC REVIEW

# Using natural language processing in emergency medicine health service research: A systematic review and meta-analysis

Hao Wang MD, PhD[1] | Naomi Alanis MS[1] | Laura Haygood MLIS[2] | Thomas K. Swoboda MD[3] | Nathan Hoot MD[1] | Daniel Phillips MD[1] | Heidi Knowles MD[1] | Sara Ann Stinson MLIS[4] | Prachi Mehta MD[1] | Usha Sambamoorthi PhD[5]

[1]Department of Emergency Medicine, JPS Health Network, Fort Worth, Texas, USA

[2]Health Sciences Librarian for Public Health, Brown University, Providence, Rhode Island, USA

[3]Department of Emergency Medicine, The Valley Health System, Touro University Nevada School of Osteopathic Medicine, Las Vegas, Nevada, USA

[4]Mary Couts Burnett Library, Burnett School of Medicine at Texas Christian University, Fort Worth, Texas, USA

[5]College of Pharmacy, University of North Texas Health Science Center, Fort Worth, Texas, USA

**Correspondence**

Hao Wang, Department of Emergency Medicine, John Peter Smith Health Network, 1500 S. Main St., Fort Worth, TX 76104, USA.
Email: hwang@ies.healthcare

**Funding information**

National Institute of Health Common Fund, Grant/Award Number: 1OT2OD032581-01; National Institute on Minority Health and Health Disparities, Grant/Award Number: 5S21MD012472-05

## Abstract

**Objectives:** Natural language processing (NLP) represents one of the adjunct technologies within artificial intelligence and machine learning, creating structure out of unstructured data. This study aims to assess the performance of employing NLP to identify and categorize unstructured data within the emergency medicine (EM) setting.

**Methods:** We systematically searched publications related to EM research and NLP across databases including MEDLINE, Embase, Scopus, CENTRAL, and ProQuest Dissertations & Theses Global. Independent reviewers screened, reviewed, and evaluated article quality and bias. NLP usage was categorized into syndromic surveillance, radiologic interpretation, and identification of specific diseases/events/syndromes, with respective sensitivity analysis reported. Performance metrics for NLP usage were calculated and the overall area under the summary of receiver operating characteristic curve (SROC) was determined.

**Results:** A total of 27 studies underwent meta-analysis. Findings indicated an overall mean sensitivity (recall) of 82%–87%, specificity of 95%, with the area under the SROC at 0.96 (95% CI 0.94–0.98). Optimal performance using NLP was observed in radiologic interpretation, demonstrating an overall mean sensitivity of 93% and specificity of 96%.

**Conclusions:** Our analysis revealed a generally favorable performance accuracy in using NLP within EM research, particularly in the realm of radiologic interpretation. Consequently, we advocate for the adoption of NLP-based research to augment EM health care management.

KEYWORDS

electronic health records, emergency medicine, natural language processing, performance accuracy, precision, recall

**Supervising Editor:** Richard Sinert

# INTRODUCTION

The utilization of artificial intelligence and machine learning (AI/ML) techniques in the medical field has become increasingly prevalent.[1,2] At present, AI/ML is employed for various purposes, including disease diagnosis, prognosis prediction, and hospitalization determination.[3–5] During the COVID-19 pandemic, studies utilized diverse clinical parameters to forecast infection severity among COVID-19–positive patients.[6,7] Various ML algorithms have been applied to predict hospitalization, critical care conditions, and mortality rates.[6,7] AI/ML has also found extensive use in forecasting outcomes for specific conditions such as sepsis, coronary artery diseases, and fever.[5,8,9]

Electronic health records (EHRs) contain data that can be categorized into structured forms (e.g., age, gender, race) and unstructured forms. The latter encompasses narrative text, including histories, current medical conditions, discharge summaries, and image reports from health care providers. Structured data in EHR are formatted by health care providers through the selection of itemized boxes or options. This often requires more effort from humans to enter information but is easier for computers to process and interpret. Unstructured data include text from dictation or typing and can convey a narrative more clearly for humans reading the chart. However, it poses a challenge for computers to analyze, due to the context-sensitive nature of human language and variability across providers.[10,11] The goal of natural language processing (NLP) is to enable computers to investigate and reason using human languages as input.[10,11]

NLP operates by breaking sentences into words to make them comprehensible for computers, using the following methods: Tokenization breaks down a sentence into individual words, and stop words, such as “the” and “is,” are often removed. Stemming and lemmatization simplify words to their base or root forms, while part-of-speech tagging labels words as nouns, verbs, adjectives, etc. Several open-source NLP programs exist that implement these steps, though some researchers develop custom NLP software. A complete NLP program, combining these various technologies, aims to understand human language accurately.

In emergency medicine (EM), timely disease recognition, appropriate treatment, and determining patient dispositions are critical. Accurate disease recognition facilitates treatment and syndromic surveillance for contagious diseases posing public threats.[10] Interpreting a patient's clinical condition from narrative text can guide treatment and disposition decisions. Common NLP applications in EM focus on monitoring syndromic surveillance; interpreting imaging findings; and identifying specific diseases, events, or syndromes.[10–12] As the health care industry increasingly relies on technology, it becomes imperative to understand nuances of NLP in the context of EM.

Before applying NLP to unstructured EHR data, it is essential to evaluate its performance. Suboptimal performance can yield erroneous predictions and diminish credibility of the technology among users. This systematic review and meta-analysis primarily focused on evaluating the performance of NLP to identify and categorize unstructured ED data, emphasizing three key areas: monitoring syndromic surveillance (e.g., respiratory illness, influenza, or gastrointestinal illness), determining image interpretations, and recognizing specific diseases/events/syndromes.

# METHODS

## Database search and eligibility criteria

We built a search strategy around the concepts of NLP, health services research, and ED. The strategy was drafted in Ovid Medline and translated to the following databases: Embase, CENTRAL via Cochrane, Scopus, and ProQuest Dissertations & Theses Global. We filtered the literature to encompass studies published between 1990 and August 2023. Additional manual searches were conducted to avoid missing any references cited in previously published NLP review papers. For the full strategy, see the supplemental files.

Eligible studies consisted of original prospective, retrospective, and randomized controlled studies from peer-reviewed journals, preprint, print, online ahead-of-print publications, and open-access journals. Of the included studies, the ED notes encompassed diverse document types such as triage, history of present illness, medical decision, radiology, and discharge summaries. To be included in the analysis, these notes were required to be limited to ED care. Moreover, studies were eligible for inclusion if they satisfied all the following criteria: (1) applied NLP in EM; (2) used raw clinical text as input for analysis; (3) employed a criterion standard definition to permit calculation of performance; and (4) if NLP was not open-resourced, segregated data into training and testing sets.

Exclusion criteria comprised: (1) duplicated studies; (2) studies exclusively using clinical notes unrelated to the ED (e.g., notes from solely prehospital EMS or home health care); (3) review papers, editorial comments or perspectives, notes extracted from social media platforms like Twitter or YouTube; (4) methodology studies solely addressing derivation or validation of open-resource NLP software; (5) studies not addressing NLP performance on testing data; (6) studies lacking validation or partially validating the final NLP results; (7) studies solely comparing performance between training and testing data without a defined criterion standard from NLP; and (8) studies reporting mean performance accuracies without case-specific accuracies.

## Data extraction and quality assessment

### Study selection and extraction

Six independent reviewers (N.A., T.S., N.H, D.P., H.K., and P.M.) conducted a comprehensive screening of search results in two distinct phases, an initial screening based on titles and abstracts (Phase 1) and then a thorough examination of full-text articles

(Phase 2). At least two independent reviewers assessed each article. Advancement from Phase 1 to Phase 2 relied on mutual agreement between both reviewers to include the individual study. In case of disagreement, a third reviewer (HW) was consulted to make a final determination regarding the study's eligibility. Instances where data were missing, incomplete, or deemed inadequate for calculating sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and positive and negative likelihood ratios (LR+ and LR−) prompted sending at least two separate emails to the study's corresponding author, requesting access to the raw study data. Studies were excluded when no response was received.

### Outcome measurements

We measured three key functions of utilizing NLP including: (1) identification of cases for syndromic surveillance (e.g., respiratory illness, gastrointestinal illness); (2) extraction of specific findings from radiology interpretations (e.g., pulmonary embolism, pneumonia); and (3) recognition of diseases, events, or syndromes (e.g., sepsis, seizures, allergies).

### Risk of bias and quality assessment of individual studies

We utilized the Critical Appraisal Skills Program (CASP) diagnostic checklist tool for studies' quality assessment (see detail in Table S2).[13,14] To address potential meta-biases, including publication bias and small-study effects, several analyses were performed. Using funnel plots for assessing publication bias in diagnostic test studies may produce misleading results.[15,16] Publication bias was evaluated in our study using a regression of diagnostic log odds ratio against 1/square root (effective sample size), weighted by effective sample size. Significant asymmetry ($p<0.10$) indicated the presence of publication bias. Small-study effects were assessed using Harbord's and Peter's tests.[17,18]

## Data synthesis and analysis

To assess study quality, we employed kappa statistics to measure inter-rater variability. The study populations' data were aggregated to evaluate the overall performance (e.g., sensitivity, specificity, PPV, NPV, LR+, and LR−) to identify and categorize unstructured ED data. Subsequently, we determined the performance of three main NLP applications. Due to anticipated study heterogeneity, we employed a random-effects model. To assess heterogeneity, summary receiver operating characteristic analysis (SROC) was performed for a visual assessment of the threshold effect, a form of diagnostic meta-analysis heterogeneity analysis. Furthermore, our study utilized LR to calculate posttest probability using a Bayes nomogram, depicting this information via a Fagan plot.[19] This plot illustrated the pretest probability (set at 50%) through LR+ and LR− to determine posttest probability.[19] We chose a pretest probability of 50% to represent a neutral uncertainty.

When studies analyzed the same data using different AI/ML algorithms, only the data exhibiting the best performance were chosen for the meta-analysis, to avoid including duplicate study patients. To mitigate potential selection biases, we conducted a sensitivity analysis using the worst-reported performance from the studies. Variation in performance might stem from different NLP algorithms, varied outcomes using the same NLP algorithms, or both.

This systematic review and meta-analysis were managed using Covidence software, a tool for systematic review screening and data extraction. STATA statistical software version 14.2 was utilized for meta-analysis. This study was registered with PROSPERO (International Prospective Register of Systematic Reviews, CRD42023477884).

## Reporting guidelines

This study report followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[20]

# RESULTS

Through a comprehensive literature search (Table S1), 596 studies were initially screened. After titles and abstracts were reviewed, 207 studies were excluded. An additional 190 studies were excluded with the review of full-text articles, leaving 33 studies for quality assessment. Among these, four studies were excluded by the quality assessment due to lacking reference standards,[21–24] and an additional two studies were omitted due to not being able to calculate NLP performance.[25,26] Consequently, a total of 27 studies were included in the final analysis (Table 1; Figure 1).[27–53]

## Risk of bias

High agreement was observed in the quality assessment among raters (inter-rater variability test, $\kappa=0.8630$, $p>0.05$), indicating no statistically significant differences in inter-rater variability. The quality assessment for all 33 included studies was summarized (Table S2).

To mitigate potential biases in assessed studies, such as publication and small-study effects, publication bias analysis was utilized. The bias coefficient was −14.06 with a [95% CI of −51.00 to 22.87], and a $p$-value of 0.440 was obtained, indicating no evidence of publication bias (Table S3). Harbord and Peters tests were employed for small-study bias analysis. The Harbord test yielded an estimated intercept of 2.72 with a standard error of 3.29 and a $p$-value of 0.261, indicating no statistically significant evidence of small-study effects. The Peters test produced a bias coefficient of −263.24 with a $p$-value of 0.311, aligning with the findings from the Harbord test (Table S3).

**TABLE 1** General information of included articles in this study.

| Year of publication | First author | Type of study | Total number of patients or images | Purpose of NLP | NLP Applications | Unstructured data used for NLP | Criterion standard comparisons |
|---|---|---|---|---|---|---|---|
| 2007 | Travers[27] | Retrospective | 3699 | Surveillance study | Acute respiratory illness | ED chief complaint | Manual review |
| 2012 | Raja[28] | Retrospective | 179 | Radiology interpretation | Pulmonary embolism | ED radiology report | Manual review |
| 2013 | Travers[29] | Retrospective | 500 | Surveillance study | Gastrointestinal syndromes | ED notes | Manual review |
| 2013 | Yadav[30] | Retrospective | 1855 | Radiology interpretation | Orbital fractures | ED radiology report | Manual review |
| 2013 | Dutta[31] | Retrospective | 1635 | Radiology interpretation | Incidental image findings | ED radiology report | Manual review |
| 2013 | Wagholikar[32] | Retrospective | 99 | Radiology interpretation | Limb fractures | ED radiology report | Manual review |
| 2014 | Haas[33] | Retrospective | 485 | Surveillance study | Respiratory illness | ED chief complaint, triage notes | Manual review |
| 2015 | Koopman[34] | Retrospective | 2378 | Radiology interpretation | Fracture, dislocation, foreign body abnormal findings | ED radiology report | Manual review |
| 2016 | Yadav[35] | Retrospective[a] | 1042 | Radiology interpretation | Pediatric traumatic brain injury | ED radiological report | Manual review |
| 2016 | Pestian[36] | Prospective | 60 | Identify diseases/events/syndromes | Suicidal adolescents | Response to the Questionnaires | CSSRS screening (Columbia Suicidal Severity Rating Scale) |
| 2018 | Jones[37] | Retrospective | 100 | Identify diseases/events/syndromes | Pneumonia | ED notes | Manual review |
| 2019 | Patterson[38] | Retrospective | 500 | Identify diseases/events/syndromes | Falls in older adults | ED notes | Manual review |
| 2020 | Fernandes[39] | Retrospective | 70,750 | Identify diseases/events/syndromes | Hospital mortality and cardiopulmonary arrest | ED triage notes | Hospital mortality and cardiopulmonary arrest |
| 2020 | Osborne[40] | Retrospective | 264 | Identify diseases/events/syndromes | Gout flare | ED chief complaint | Manual review |
| 2021 | Bouchouar[41] | Retrospective | 19,023 | Surveillance study | Influenza like illness | ED chief complaint, clinical triage notes, and discharge diagnosis text | ICD-10 code with manual review |
| 2021 | Chartash[42] | Retrospective | 600 | Identify diseases/events/syndromes | Shared decision making and general practice decision | ED notes | Manual review |
| 2021 | Shung[43] | Retrospective | 2988 | Identify diseases/events/syndromes | Gastrointestinal bleeding | ED triage notes | Manual review |

(Continues)

TABLE 1 (Continued)

| Year of publication | First author | Type of study | Total number of patients or images | Purpose of NLP | NLP Applications | Unstructured data used for NLP | Criterion standard comparisons |
|---|---|---|---|---|---|---|---|
| 2021 | Sax[44] | Retrospective | 150 | Identify diseases/events/syndromes | Acute heart failure | ED triage notes, chief complaints | Manual review |
| 2022 | Cohen[45] | Retrospective | 70 | Identify diseases/events/syndromes | Suicidal ideation | ED interviews' note | CSSRS screening |
| 2022 | Irvin[46] | Retrospective | 461 | Radiology interpretation | Unilobar vs. multilobar pneumonia, pleural effusion, pneumonia | ED radiology report | Manual review |
| 2022 | Yaeger[47] | Retrospective | 1190 | Identify diseases/events/syndromes | Infants with fever | ED notes including triage, nursing, physician, resident notes | Manual review |
| 2022 | Rozova[48] | Retrospective | 78,045 | Identify diseases/events/syndromes | Self-harm | ED triage notes | Manual review |
| 2022 | Kooragayala[49] | Retrospective | 398 | Radiology interpretation | Worrisome pancreatic lesions | ED radiology report | Manual review |
| 2022 | Gordon[50] | Retrospective | 355 | Radiology interpretation | Intracranial mass effect | ED radiology report | Manual review |
| 2023 | Evans[51] | Retrospective | 282 | Radiology interpretation | Incidental findings from trauma CT | ED radiology report | Manual review |
| 2023 | Dotson[52] | Retrospective | 341 | Radiology interpretation | Incidental pulmonary nodules | ED radiology report | Manual review |
| 2023 | MacPhaul[53] | Retrospective | 685 | Identify diseases/events/syndromes | Assault | ED notes | Manual review |

Abbreviation: NLP, natural language processing.

[a]Secondary data analysis from prospectively collected data.

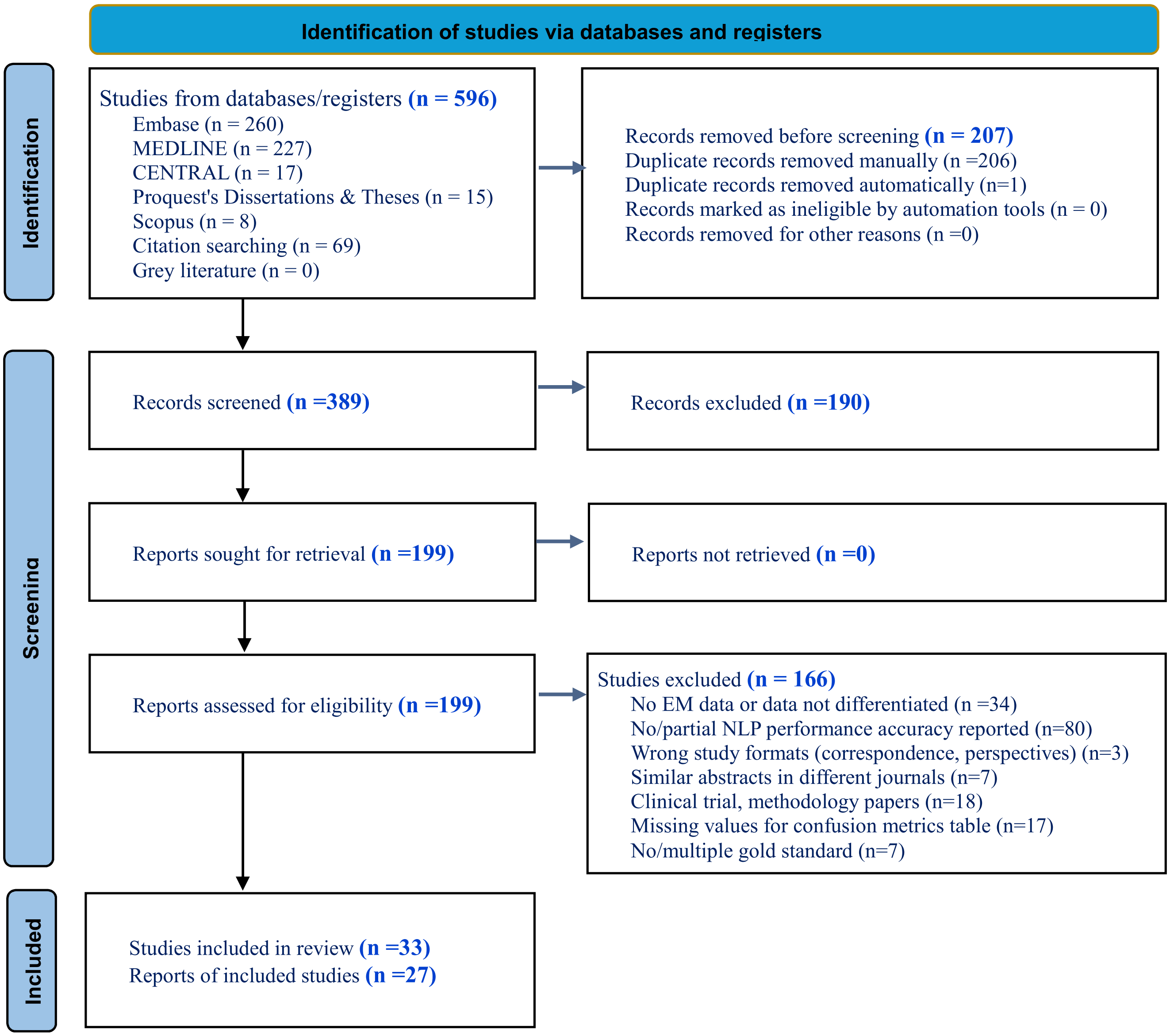
Identification of studies via databases and registers
Identification
Studies from databases/registers (n = 596)
Embase (n = 260)
MEDLINE (n = 227)
CENTRAL (n = 17)
Proquest's Dissertations & Theses (n = 15)
Scopus (n = 8)
Citation searching (n = 69)
Grey literature (n = 0)
Records removed before screening (n = 207)
Duplicate records removed manually (n =206)
Duplicate records removed automatically (n=1)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n =0)
Screening
Records screened (n =389)
Records excluded (n =190)
Reports sought for retrieval (n =199)
Reports not retrieved (n =0)
Reports assessed for eligibility (n =199)
Studies excluded (n = 166)
No EM data or data not differentiated (n =34)
No/partial NLP performance accuracy reported (n=80)
Wrong study formats (correspondence, perspectives) (n=3)
Similar abstracts in different journals (n=7)
Clinical trial, methodology papers (n=18)
Missing values for confusion metrics table (n=17)
No/multiple gold standard (n=7)
Included
Studies included in review (n =33)
Reports of included studies (n =27)

FIGURE 1 PRISMA flow diagram: using NLP in EM research. NLP, natural language processing.

## Main results

Overall, this meta-analysis included 27 studies with 179,109 patients and 9025 images. It consisted of four studies involving 23,707 patients focused on syndromic surveillance,[27,29,33,41] 12 studies covering 155,402 patients targeting diseases/events/syndromes recognition,[36–40,42–45,47,48,53] and 11 studies involving 9025 images concentrating on radiology interpretations.[28,30–32,34,35,46,49–52] Detailed descriptions of each study are provided in Table 2. The mean sensitivity across studies was 0.87 (95% CI 0.82–0.91), specificity was 0.95 (95% CI 0.92–0.97), PPV was 0.18 (95% CI 0.18–0.18), NPV was 1.00 (95% CI 1.00–1.00), LR+ was 17.4 (95% CI 10.3–29.5), and LR− was 0.13 (95% CI 0.09–0.19). Figure 2 illustrates a forest plot displaying mean recall (sensitivity) and specificity. Upon subgroup analysis, it was noted that NLP's optimal performance lay in interpreting image reports (mean sensitivity of 0.93 and specificity of 0.96, Table 2).

## Sensitivity analysis

Among the 27 studies, variations arose due to different NLP algorithms used for outcome prediction and the measurement of diverse outcomes. To address this variability, a sensitivity analysis was conducted to ascertain the worst performance among these studies. Confusion matrices were unsuccessfully derived from two studies,[39,46] and the other 10 reported varying performance.[31,33,36,39–41,43,45–48,53]

The sensitivity analysis revealed that the mean sensitivity was 0.82 (95% CI 0.73–0.88), specificity was 0.95 (95% CI 0.92–0.97), PPV was 0.14 (95% CI 0.14–0.14), NPV was 1.00 (95% CI 1.00–1.00), LR+ was 15.8 (95% CI 10.1–24.6), and LR− was 0.19 (95% CI 0.13–0.29). These metrics represented a less optimal performance, when compared to the best-reported performance observed across the studies (Table 3).

## DISCUSSION

The manual conversion of unstructured data into structured data has historically posed significant challenges within research, including extensive time investment in data collection and training, different inter-rater variabilities, and a propensity for errors during manual processing.[54–56] Leveraging NLP to identify and classify unstructured data has notable advantages in the realm of AI/ML, by mitigating these issues.[57,58] In our investigation, we observed a generally favorable performance of NLP usage in EM research. Additionally, upon categorizing NLP applications into three key areas, we discovered that the highest performance was achieved in interpreting radiological reports. These findings encourage further exploration and research into the application of NLP for managing unstructured data.

NLP can extract various medical concepts from clinical documentation, facilitating analysis or reasoning based on human language.[31,59] Substantial potential exists for leveraging EHR data for health care advancements through NLP-based decision support systems.[60,61] These systems, derived from NLP-based research, may reduce health care costs while enhancing clinical decision making.

Upon review of the literature, we noted a dearth of performance reports employing meta-analysis to gauge the performance of NLP in identifying and categorizing unstructured data. Currently, there exists only one such meta-analysis, which evaluated postoperative complications within the surgical domain.[62] The authors reported an overall sensitivity of 92% and specificity of 99%. Our study, although focused on EM research, yielded similar results. Such consistent findings support the potential of NLP to bolster health care efficiency and precision across various medical domains.

Our analysis indicated an overall lower PPV in the range of 0.13–0.14. This outcome was predominantly influenced by a single study with a substantial sample size, which contributed to over 80% of the weighting in the meta-analysis.[39] This study utilized triage notes to predict instances of mortality and cardiopulmonary arrest, employing NLP techniques for managing and analyzing these notes to predict said events.[39] A low PPV can be anticipated when the prevalence of the event, such as mortality and cardiopulmonary arrest, is exceedingly low (i.e., 0.48% in this study).[39] Upon the removal of this influential study from the meta-analysis, the overall mean PPV became 0.82. This revised finding aligns more closely with the PPV we observed when utilizing NLP for syndromic surveillance or interpreting radiology reports.

TABLE 2 Performance of NLP in EM.

| | Sensitivity (recall) | Specificity | PPV (precision) | NPV | LR+ | LR− |
|---|---|---|---|---|---|---|
| Overall | | | | | | |
| Pooled | 0.87 | 0.95 | 0.18 | 1.00 | 17.4 | 0.13 |
| 95% CI | 0.82–0.91 | 0.92–0.97 | 0.18–0.18 | 1.00–1.00 | 10.3–29.5 | 0.09–0.19 |
| Range | 0.33–1.00 | 0.55–1.00 | 0.02–0.95 | 0.72–1.00 | 1.78–286.46 | 0.01–0.67 |
| Subgroup, pooled (95% CI) | | | | | | |
| Syndromic surveillance | 0.73 (0.51–0.87) | 0.96 (0.82–0.99) | 0.85 (0.84–0.85) | 0.99 (0.99–0.99) | 16.3 (3.7–72.6) | 0.29 (0.14–0.57) |
| Radiology interpretation | 0.93 (0.90–0.94) | 0.96 (0.92–0.98) | 0.82 (0.81–0.83) | 0.99 (0.99–1.00) | 20.8 (11.3–38.2) | 0.08 (0.06–0.10) |
| Identifying diseases/ events/syndromes | 0.86 (0.79–0.91) | 0.94 (0.86–0.98) | 0.13 (0.13–0.13) | 1.00 (1.00–1.00) | 14.5 (5.9–35.6) | 0.15 (0.09–0.23) |

Abbreviations: LR, likelihood ratio; NLP, natural language processing; NPV, negative predictive value; PPV, positive predictive value.

TABLE 3 Sensitivity analysis of performance using NLP in EM.

| | Sensitivity (recall) | Specificity | PPV (precision) | NPV | LR+ | LR− |
|---|---|---|---|---|---|---|
| Overall | 0.82 (0.73–0.88) | 0.95 (0.92–0.97) | 0.14 (0.14–0.14) | 1.00 (1.00–1.00) | 15.8 (10.1–24.6) | 0.19 (0.13–0.29) |
| Subgroup | | | | | | |
| Syndromic surveillance | 0.58 (0.38–0.75) | 0.92 (0.86–0.95) | 0.46 (0.46–0.47) | 0.94 (0.93–0.94) | 7.0 (5.7–8.5) | 0.46 (0.30–0.70) |
| Radiology interpretation | 0.88 (0.75–0.95) | 0.96 (0.92–0.98) | 0.87 (0.86–0.87) | 0.98 (0.98–0.99) | 22.1 (11.5–42.3) | 0.12 (0.06–0.27) |
| Identifying diseases/ events/syndromes | 0.81 (0.73–0.87) | 0.94 (0.88–0.97) | 0.12 (0.11–0.12) | 1.00 (1.00–1.00) | 14.4 (6.8–30.8) | 0.20 (0.14–0.30) |

*Note*: Data are reported as pooled performance metrics with 95% CI.

Abbreviations: LR, likelihood ratio; NLP, natural language processing; NPV, negative predictive value; PPV, positive predictive value.

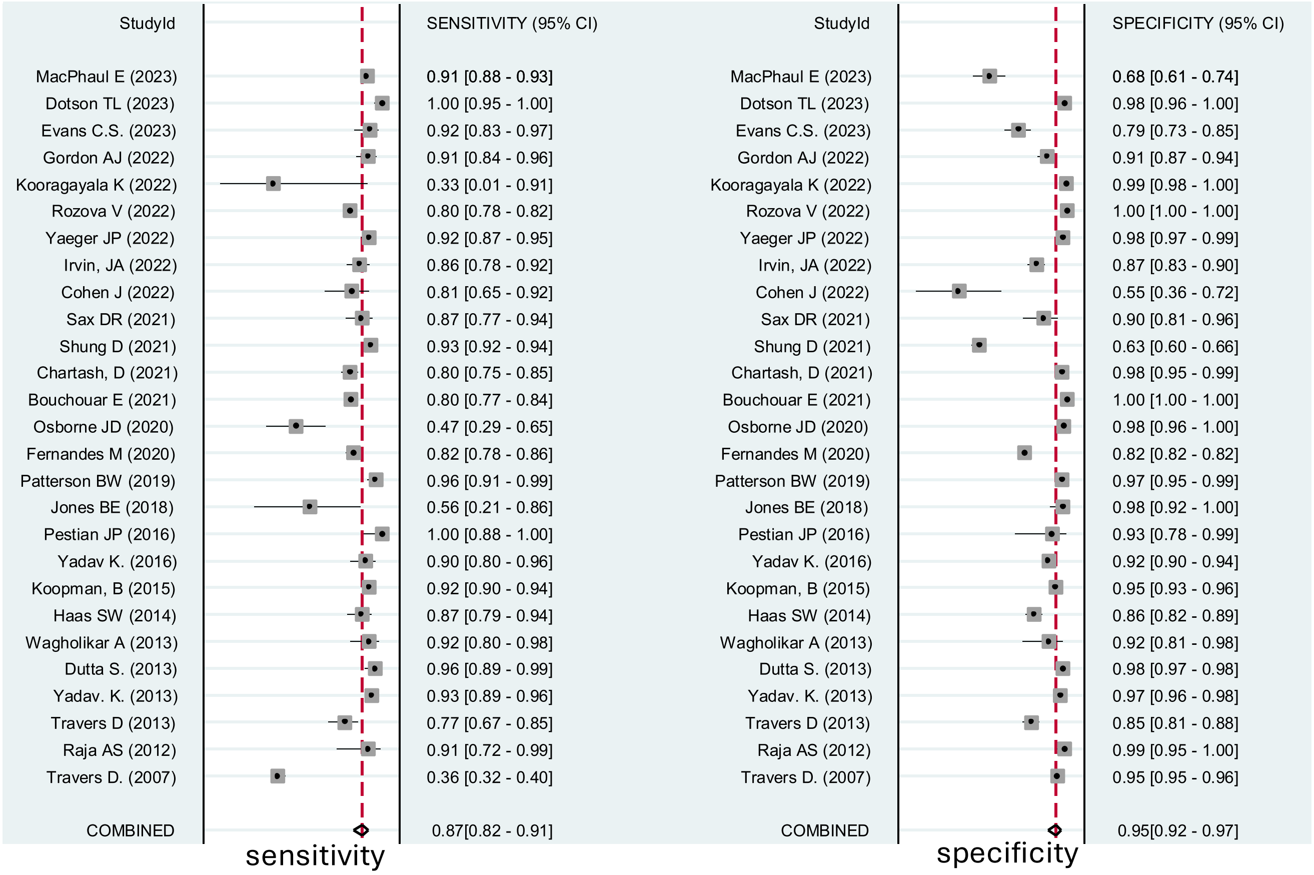

FIGURE 2 Performance (sensitivity and specificity) of using NLP in EM research. The SROC curve depicts mean operating sensitivity and specificity points (Figure 3). The area under the SROC was calculated as 0.96 (95% CI 0.94–0.98). Setting the pretest probability at 50%, following the use of NLP to identify unstructured ED data, the probability of correct prediction reached 95% (Figure 4). Utilizing NLP to exclude cases/disease/events at the same pretest probability (50%) resulted in a posttest probability of 12%, indicating fairly accurate classifications of unstructured data through NLP. NLP, natural language processing; SROC, summary receiver operating characteristic.

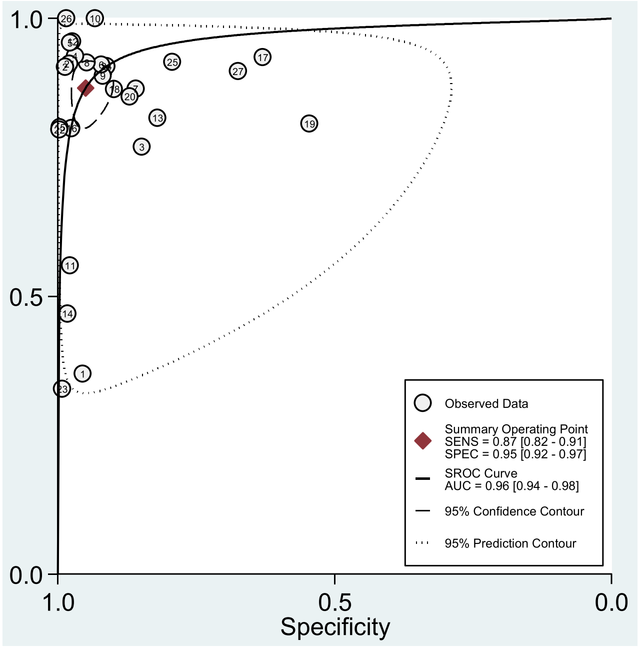

FIGURE 3 An SROC curve. SROC, summary receiver operating characteristic.

Considerable variations in performance among studies underscored the heterogeneity observed in our analysis, plausible mechanisms behind these variations may stem from the inherent challenge of interpreting human language and the utilization of diverse ML models.[45,63] Our sensitivity analysis highlighted the potential influence of language variance across different times, geographic regions, or events, contributing to this observed heterogeneity. It is imperative to exercise caution when applying open-source NLP programs to individual studies, emphasizing the necessity of validation before formal application.

Our study focused solely on assessing the performance of using NLP independently, while other studies in the literature adopt a fusion of NLP and additional AI/ML algorithms to handle both structured and unstructured data.[64,65] To offer a broader understanding of NLP usage, future studies could compare performance between both approaches.

## LIMITATIONS

Our study, despite a robust search strategy utilizing various databases and manually searching relevant NLP studies, may have missed

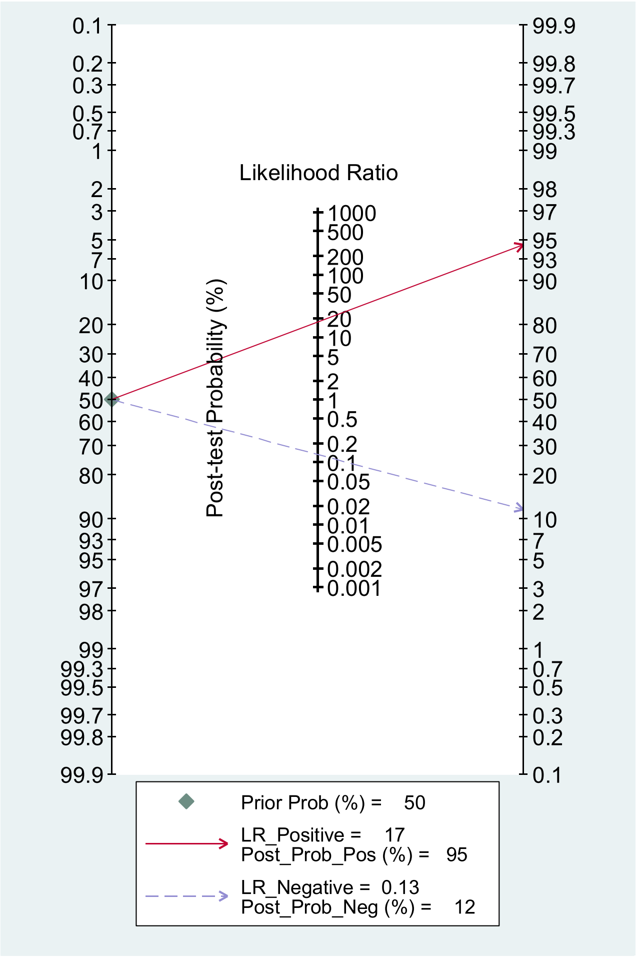

**FIGURE 4** Fagan plot.

other NLP studies related to EM, which might affect the overall findings of performance. Another major limitation of this meta-analysis was the potential high heterogeneity found between studies due to threshold effect, which may compromise the accuracy of the meta-analysis results in this study.[66,67] Moreover, several studies employed a blend of NLP and other AI/ML algorithms to handle both structured and unstructured data.[68,69] As our inclusion criteria were focused primarily on NLP accuracy, studies employing such combinations were excluded, potentially limiting the scope of NLP investigation. Many studies used sample sizes for NLP validation compared to their designated "criterion standard."[28,38,44] Typically, the criterion standard in most studies involved the consensus of manual review, due to the unstructured nature of data. Following validation, these studies applied NLP to larger data sets without additional validation against the criterion standard. Our meta-analysis only incorporated validated data, which might lead to an inaccurate performance assessment of NLP. Some studies have indicated degraded performance when prospectively applying models that were trained retrospectively, underscoring the potential bias in our analysis.[31,45,63] Lastly, although NLP is commonly used to handle unstructured data, our investigation focused solely on NLP studies within the field of EM. Because our meta-analysis was limited to EM, our results on NLP performance may not generalize to other specialties. Future studies concentrating on different outcomes, purposes, or diverse medical domains are necessary to validate and expand upon our findings.

## CONCLUSIONS

Our study identified generally favorable performance when employing natural language processing in emergency medicine research, with higher performance for radiologic interpretation when compared with syndromic surveillance and disease/event/syndrome recognition. We advocate for further natural language processing research to augment health care efficiency and precision.

### ACKNOWLEDGMENTS

We express our gratitude to Dr. James P. d'Etienne and Ms. Alejandra Powers for their review and valuable editorial comments on the manuscript.

### FUNDING INFORMATION

Usha Sambamoorthi received funds from the National Institute on Minority Health and Health Disparities through the Texas Center for Health Disparities (NIMHD) 5S21MD012472-05 and the NIH Common Fund Grant 1OT2OD032581-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

### ORCID

*Hao Wang* https://orcid.org/0000-0002-5105-0951

### REFERENCES

1. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216-1219.
2. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347-1358.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
4. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6:26094.
5. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46:547-553.
6. Gao Y, Cai GY, Fang W, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun*. 2020;11:5033.
7. Taheriyan M, Ayyoubzadeh SM, Ebrahimi M, et al. Prediction of COVID-19 Patients' survival by deep learning approaches. *Med J Islam Repub Iran*. 2022;36:144.
8. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38:500-507.

9. Phakhounthong K, Chaovalit P, Jittamala P, et al. Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis. *BMC Pediatr*. 2018;18:109.
10. Hughes HE, Edeghere O, O'Brien SJ, Vivancos R, Elliot AJ. Emergency department syndromic surveillance systems: a systematic review. *BMC Public Health*. 2020;20:1891.
11. Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA Annu Symp Proc*. 2008;2008:172-176.
12. Mueller B, Kinoshita T, Peebles A, Graber MA, Lee S. Artificial intelligence and machine learning in emergency medicine: a narrative review. *Acute Med Surg*. 2022;9:e740.
13. Burns A, Donnelly B, Feyi-Waboso J, et al. How do electronic risk assessment tools affect the communication and understanding of diagnostic uncertainty in the primary care consultation? A systematic review and thematic synthesis. *BMJ Open*. 2022;12:e060101.
14. CASP Checklists. Critical Appraisal Skills Program. 2023. Accessed Aug 1, 2023. https://casp-uk.net/casp-tools-checklists/.
15. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006;333:597-600.
16. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58:882-893.
17. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. 2006;25:3443-3457.
18. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*. 2006;295:676-680.
19. Safari S, Baratloo A, Elfil M, Negida A. Evidence based emergency medicine; part 4: pre-test and post-test probabilities and Fagan's nomogram. *Emerg (Tehran)*. 2016;4:48-51.
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
21. Aronis JM, Ye Y, Espino J, Hochheiser H, Michaels MG, Cooper GF. A Bayesian system to track outbreaks of influenza-like illnesses including novel diseases. *medRxiv*. 2023.
22. Desai A, Zumbo A, Giordano M, et al. Word2vec word embedding-based artificial intelligence model in the triage of patients with suspected diagnosis of major ischemic stroke: a feasibility study. *Int J Environ Res Public Health*. 2022;19(22):15295.
23. Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int J Med Inform*. 2019;129:184-188.
24. Ye Y, Tsui FR, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *J Am Med Inform Assoc*. 2014;21:815-823.
25. Goss FR, Plasek JM, Lau JJ, Seger DL, Chang FY, Zhou L. An evaluation of a natural language processing tool for identifying and encoding allergy information in emergency department clinical notes. *AMIA Annu Symp Proc*. 2014;2014:580-588.
26. Afshar M, Phillips A, Karnik N, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc*. 2019;26:254-261.
27. Travers D, Wu S, Scholer M, Westlake M, Waller A, McCalla AL. Evaluation of a chief complaint pre-processor for biosurveillance. *AMIA Annu Symp Proc*. 2007;2007:736-740.
28. Raja AS, Ip IK, Prevedello LM, et al. Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. *Radiology*. 2012;262:468-474.
29. Travers D, Haas SW, Waller AE, et al. Implementation of emergency medical text classifier for syndromic surveillance. *AMIA Annu Symp Proc*. 2013;2013:1365-1374.
30. Yadav K, Sarioglu E, Smith M, Choi HA. Automated outcome classification of emergency department computed tomography imaging reports. *Acad Emerg Med*. 2013;20:848-854.
31. Dutta S, Long WJ, Brown DF, Reisner AT. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. *Ann Emerg Med*. 2013;62:162-169.
32. Wagholikar A, Zuccon G, Nguyen A, et al. Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *Australas Med J*. 2013;6:301-307.
33. Haas SW, Travers D, Waller A, et al. Emergency medical text classifier: new system improves processing and classification of triage notes. *Online J Public Health Inform*. 2014;6:e178.
34. Koopman B, Zuccon G, Wagholikar A, et al. Automated reconciliation of radiology reports and discharge summaries. *AMIA Annu Symp Proc*. 2015;2015:775-784.
35. Yadav K, Sarioglu E, Choi HA, Cartwright WB, Hinds PS, Chamberlain JM. Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. *Acad Emerg Med*. 2016;23:171-178.
36. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, et al. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide Life Threat Behav*. 2016;46:154-159.
37. Jones BE, South BR, Shao Y, et al. Development and validation of a natural language processing tool to identify patients treated for pneumonia across VA emergency departments. *Appl Clin Inform*. 2018;9:122-128.
38. Patterson BW, Jacobsohn GC, Shah MN, et al. Development and validation of a pragmatic natural language processing approach to identifying falls in older adults in the emergency department. *BMC Med Inform Decis Mak*. 2019;19:138.
39. Fernandes M, Mendes R, Vieira SM, et al. Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing. *PLoS One*. 2020;15:e0230876.
40. Osborne JD, Booth JS, O'Leary T, et al. Identification of gout flares in chief complaint text using natural language processing. *AMIA Annu Symp Proc*. 2020;2020:973-982.
41. Bouchouar E, Hetman BM, Hanley B. Development and validation of an automated emergency department-based syndromic surveillance system to enhance public health surveillance in Yukon: a lower-resourced and remote setting. *BMC Public Health*. 2021;21:1247.
42. Chartash D, Sharifi M, Emerson B, et al. Documentation of shared decision making in the emergency department. *Ann Emerg Med*. 2021;78:637-649.
43. Shung D, Tsay C, Laine L, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. *J Gastroenterol Hepatol*. 2021;36:1590-1597.
44. Sax DR, Mark DG, Huang J, et al. Use of machine learning to develop a risk-stratification tool for emergency department patients with acute heart failure. *Ann Emerg Med*. 2021;77:237-248.
45. Cohen J, Wright-Berryman J, Rohlfs L, Trocinski D, Daniel L, Klatt TW. Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. *Front Digit Health*. 2022;4:818705.
46. Irvin JA, Pareek A, Long J, et al. CheXED: comparison of a deep learning model to a clinical decision support system for pneumonia in the emergency department. *J Thorac Imaging*. 2022;37:162-167.

47. Yaeger JP, Lu J, Jones J, Ertefaie A, Fiscella K, Gildea D. Derivation of a natural language processing algorithm to identify febrile infants. *J Hosp Med*. 2022;17:11-18.
48. Rozova V, Witt K, Robinson J, Li Y, Verspoor K. Detection of self-harm and suicidal ideation in emergency department triage notes. *J Am Med Inform Assoc*. 2022;29:472-480.
49. Kooragayala K, Crudeli C, Kalola A, et al. Utilization of natural language processing software to identify worrisome pancreatic lesions. *Ann Surg Oncol*. 2022;29:8513-8519.
50. Gordon AJ, Banerjee I, Block J, et al. Natural language processing of head CT reports to identify intracranial mass effect: CTIME algorithm. *Am J Emerg Med*. 2022;51:388-392.
51. Evans CS, Dorris HD, Kane MT, et al. A natural language processing and machine learning approach to identification of incidental radiology findings in trauma patients discharged from the emergency department. *Ann Emerg Med*. 2023;81:262-269.
52. Dotson TL, Gasimova A, Watkins J, Chometon Q, Bellinger CR. Identifying patients with pulmonary nodules from CT radiology reports using natural language processing (NLP). *Am J Respir Crit Care Med*. 2023;207:A6516.
53. MacPhaul E, Zhou L, Mooney SJ, et al. Classifying firearm injury intent in electronic hospital records using natural language processing. *JAMA Netw Open*. 2023;6:e235870.
54. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof*. 2013;10:12.
55. Nagurney JT, Brown DF, Sane S, Weiner JB, Wang AC, Chang Y. The accuracy and completeness of data collected by prospective and retrospective methods. *Acad Emerg Med*. 2005;12:884-895.
56. Yawn BP, Wollan P. Interrater reliability: completing the methods description in medical records review studies. *Am J Epidemiol*. 2005;161:974-977.
57. Lu Z, Sim JA, Wang JX, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. *J Med Internet Res*. 2021;23:e26777.
58. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform*. 2019;7:e12239.
59. Gholipour M, Khajouei R, Amiri P, Hajesmaeel Gohari S, Ahmadian L. Extracting cancer concepts from clinical notes using natural language processing: a systematic review. *BMC Bioinformatics*. 2023;24:405.
60. Shi J, Liu S, Pruitt LCC, et al. Using natural language processing to improve EHR structured data-based surgical site infection surveillance. *AMIA Annu Symp Proc*. 2019;2019:794-803.
61. Berge GT, Granmo OC, Tveit TO, Munkvold BE, Ruthjersen AL, Sharma J. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. *BMC Med Inform Decis Mak*. 2023;23:5.
62. Mellia JA, Basta MN, Toyoda Y, et al. Natural language processing in surgery: a systematic review and meta-analysis. *Ann Surg*. 2021;273:900-908.
63. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Ling Compass*. 2021;15:e12432.
64. Chen CH, Hsieh JG, Cheng SL, Lin YL, Lin PH, Jeng JH. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. *Int J Med Inform*. 2020;139:104146.
65. Su D, Li Q, Zhang T, et al. Prediction of acute appendicitis among patients with undifferentiated abdominal pain at emergency department. *BMC Med Res Methodol*. 2022;22:18.
66. Lee J, Kim W, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers – part II. Statistical methods of meta-analysis. *Korean J Radiol*. 2015;16:1188-1196.
67. Leeflang MMG. Systematic review and meta-analysis of diagnostic test accuracy. *Clin Microbiol Infect*. 2014;20:105-113.
68. Klang E, Kummer BR, Dangayach NS, et al. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. *Sci Rep*. 2021;11:1381.
69. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inf Med*. 2017;56:377-389.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Wang H, Alanis N, Haygood L, et al. Using natural language processing in emergency medicine health service research: A systematic review and meta-analysis. *Acad Emerg Med*. 2024;31:696-706. doi:10.1111/acem.14937