Received: 30 June 2022 | Revised: 3 November 2022 | Accepted: 18 November 2022

DOI: 10.1111/cts.13463

REVIEW

Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review

Sunyang Fu[1] | Liwei Wang[1] | Sungrim Moon[1] | Nansu Zong[1] | Huan He[1] | Vikas Pejaver[2,3] | Rose Relevo[4] | Anita Walden[4] | Melissa Haendel[5] | Christopher G. Chute[6] | Hongfang Liu[1]

[1]Department of AI and Informatics Research, Mayo Clinic, Rochester, Minnesota, USA

[2]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[4]The National Center for Data to Health, Bethesda, Maryland, USA

[5]Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

[6]Department of Medicine, Johns Hopkins University, Baltimore, Maryland, USA

Correspondence

Hongfang Liu, Department of AI and Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.
Email: liu.hongfang@mayo.edu

Abstract

An increasing number of studies have reported using natural language processing (NLP) to assist observational research by extracting clinical information from electronic health records (EHRs). Currently, no standardized reporting guidelines for NLP-assisted observational studies exist. The absence of detailed reporting guidelines may create ambiguity in the use of NLP-derived content, knowledge gaps in the current research reporting practices, and reproducibility challenges. To address these issues, we conducted a scoping review of NLP-assisted observational clinical studies and examined their reporting practices, focusing on NLP methodology and evaluation. Through our investigation, we discovered a high variation regarding the reporting practices, such as inconsistent use of references for measurement studies, variation in the reporting location (reference, appendix, and manuscript), and different granularity of NLP methodology and evaluation details. To promote the wide adoption and utilization of NLP solutions in clinical research, we outline several perspectives that align with the six principles released by the World Health Organization (WHO) that guide the ethical use of artificial intelligence for health.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

An increasing number of natural language processing (NLP) applications have been leveraged to assist observational research.

WHAT QUESTION DID THIS STUDY ADDRESS?

To understand the reporting practices in NLP-assisted observational studies.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

The study reviews reporting practice of NLP-assisted observational studies and discusses the lessons learned from existing evaluation and reporting practices.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

The study emphasizes the importance of reproducibility and scientific rigor when using NLP for clinical research. It also defines recommendations to ensure translation into practice.

INTRODUCTION

Advancement in the digital transformation of health care has reshaped the documentation, representation, and management of electronic health records (EHRs). This transformation increases the need for innovative informatics solutions for facilitating and accelerating the secondary use of EHRs for clinical research defined as patient-oriented research, epidemiological and behavioral studies, or outcomes and health services research.[1,2] Observational research is a common type of clinical research that measures outcomes or observes individuals following cohort, cross-sectional, or case-control research study design.[3–5] Not all EHR data are represented in structured format, so it is necessary in many observational studies to leverage unstructured data to comprehensively capture the "complete" representation of patient profiles. Natural language processing (NLP) has been leveraged to assist chart review by automatically extracting clinical information from unstructured text.[6,7] For example, in our previous study, we assessed the agreement among International Classification of Diseases (ICD) codes, flowsheets, and clinical notes within the EHR documentation of patients with delirium. The unstructured clinical notes and semi-structured flowsheet data can improve phenotyping sensitivity by approximately 20% over a purely structured data (ICD-based) approach.[8–10] Similar results were achieved with incidental findings,[11] diseases with multifactorial causes,[9,12] diseases with no singular and conclusive diagnostic tests,[13] or surgical information.[13,14] Furthermore, to maximize the detection accuracy, studies have leveraged the combination of both structured (e.g., laboratory and medication) and unstructured data to determine patient's disease status, such as silent brain infarction,[15] type 2 diabetes mellitus,[16] and rheumatoid arthritis.[17] These findings strongly suggest that text information can improve the detection of conditions such as rare diseases, which are not routinely coded and/or are underdiagnosed in clinical practice.[18]

There are many approaches to the derivation of knowledge from EHRs for conducting observational research, one of which is chart review, a common, manual process of extracting or reviewing information from EHRs and assembling a data set for various research needs, such as case ascertainment,[6,15,19] data element extraction,[13,20,21] risk stratification,[22] and case matching.[23] Because a significant portion of clinical information is represented in textual format, execution of such a human-operated approach is time-consuming, non-standardized, and not very scalable.[7,23–25] The availability of digital information demands systematic solutions for high-throughput knowledge extraction. As illustrated in Figure 1, NLP can be integrated into the research workflow for observational studies by providing information extraction and knowledge conversion at multiple stages of study implementation.[6,7] Such a scalable solution promotes a new era of research that would not have been feasible decades ago. As a result, an increasing number of observational studies from disease study areas[26–30] to drug-related studies[22,31,32] benefit from state-of-the-art NLP solutions.

Despite the notable advantages of NLP-assisted observational research, there are remaining challenges – one of which is the inconsistency between community standards in observational research and the standards in clinical NLP communities. In the clinical NLP community, studies that

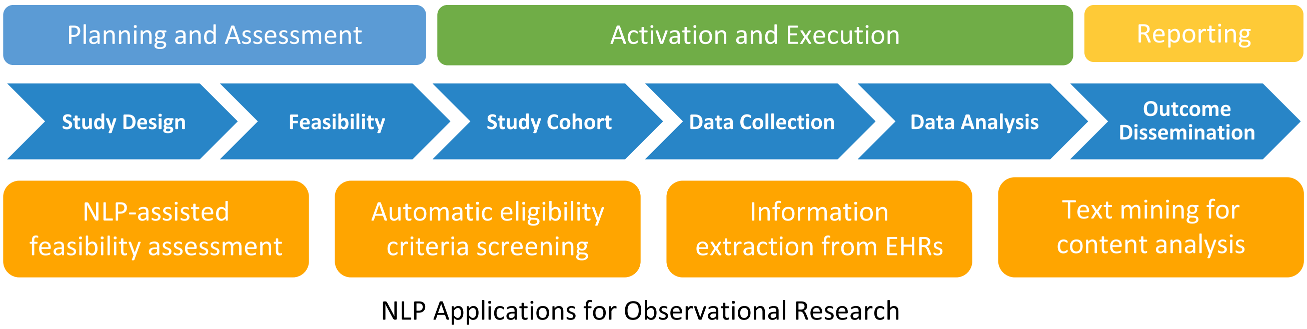

Planning and Assessment
Activation and Execution
Reporting
Study Design
Feasibility
Study Cohort
Data Collection
Data Analysis
Outcome Dissemination
NLP-assisted feasibility assessment
Automatic eligibility criteria screening
Information extraction from EHRs
Text mining for content analysis
NLP Applications for Observational Research

FIGURE 1 EHR-based observational research empowered by NLP. EHR, electronic health record; NLP, natural language processing.

focus on the development and evaluation of NLP solutions are commonly referred to as measurement studies, defined as research projects designed to develop and refine methods for determining how much error of the targeted attribute can be measured in a population of objects.[33–35] For example, Wyles et al.[20] conducted a measurement study to examine the accuracy of two NLP systems for the identification of common data elements in operative notes for total joint registry. Wang et al.[36] evaluated a comprehensive NLP algorithm for extracting lung cancer data elements, including stage, histology, tumor grade, and therapies. McCoy et al.[37] developed an NLP-based phenotype algorithm to extract five different cognitive and psychopathologic symptoms. Such study involved iterative processes to establish gold standard corpora – the benchmark language resources for the development, evaluation, and deployment of NLP methods. In observational studies, NLP is typically viewed as a specific measurement for patient characteristics through EHR data. For example, through applying the previously validated NLP algorithm for estimating cognitive symptomatology, McCoy et al.[38] conducted a survival analysis to examine the association of cognitive symptoms with incidence of dementia diagnosis. Similarly, Kent et al.[15] designed a study to estimate the incidence of future stroke in patients with incidentally discovered silent cerebrovascular disease (SCD). In this study, NLP was used to measure patients' status of SCD from neuroimaging reports.

The quality of such a measure can potentially impact the validity, reproducibility, and explainability of the research outcome. As indicated by the current version of STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) reporting criteria, there is, however, no standardized reporting guideline for NLP-assisted observational studies as of today. The absence of detailed reporting guidelines can create ambiguity and a lack of reproducibility in current research reporting practices (e.g., not defining the original cohort used in measurement study when using NLP in an observational study). This issue is particularly important for NLP applications due to their methodology complexity. For example, most NLP applications contain (i) a generic text processor (i.e., NLP framework) that is usually comprised of a set of information extraction pipelines, such as sector detector, sentence breaker, tokenizer, part-of-speech tagger, chunk annotator, information extractor, and context annotator and (ii) task-specific NLP algorithms (e.g., heart failure, stroke, and breast cancer).[6,7] In general, the text processor serves as a generic pipeline to pre-process clinical documents into a standard representation before disease-specific information extraction occurs. Well known generic text processors include MedLEE,[39] MetaMap,[40] KnowledgeMap,[41] cTAKES,[42] HiTEX,[43] and MedTaggerIE.[44,45] In addition to that, an NLP algorithm is often referred to as a specific set of rules (rule-based or statistical) that will be used to solve one or many defined problems. These algorithms are defined by specific NLP tasks and thus are closely dependent on various contextual factors, such as disease types, domain experts, EHR data, and evaluation environments. Because the final NLP output is dependent on the aggregated performance of all prior text processing components and algorithms, non-explicit and ambiguous reporting practices can cause unintended data reproducibility issues.

Considering NLP is becoming a vital part of EHR-based observational research and closely affects overall research reproducibility and scientific rigor, it is important to have a systematic understanding of the existing reporting practices such as the metadata, degree of granularity, and confidence of NLP methodology and evaluation in observational studies. In response to that, we conducted the scoping review to examine the manuscript reporting of NLP-specific methodological standards and evaluation processes in observational studies to provide potential evidence for strengthening the future reporting of NLP-assisted observational research. We reviewed all studies published from 2009 to 2021 that met the inclusion criteria. Based on our findings, we propose recommendations for the future development of NLP research standards for the clinical and translational science (CTS) community.

METHODS

Research questions of scoping review

The review was conducted following the scoping review guide proposed by Pham et al.[46] The initial process includes identification of key research questions and defining the scope (population) of research.

- *Why the review should be conducted*: existing studies focus on NLP methodologies and evaluations in measurement studies.[6,7] Little-to-no investigation on the methodologic standards and evaluation practices has been conducted when NLP is applied to support observational studies.
- *What the review will add to the reader's knowledge in the field*: understanding current reporting variability, lessons learned from NLP-assisted observational studies, and indications of what a more standardized NLP-reporting structure could or should look like.
- *What specifically is being investigated about the topic under investigation*:
    - What is the current reporting practice of the NLP component in the NLP-assisted observational research?
        - Are important NLP methodology and evaluation details reported?
        - Where are they reported?

  - What is the degree of granularity and confidence (e.g., data source definition, data abstraction and annotation process, NLP pre-processing pipeline, NLP method, evaluation statistics)?
- What are the potential barriers to adopting NLP for observational research?
- What are the lessons learned from the existing evaluation and reporting practices?

Data sources and search strategy

An experienced librarian (LP) retrieved the studies written in English and published from January 2009 through September 2021 from Ovid MEDLINE In-Process & Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, and Web of Science. The implementation of search patterns was consistent across different databases. Detailed search strategies are provided in Supplementary Material S1. A total of 1347 articles were retrieved from five libraries and after de-duplication, 471 articles remained for evaluation.

Eligibility criteria and screening

The title and abstract screenings were conducted by two reviewers (authors S.F. and R.R.) from the previously identified 471 articles. The main objective of the screening was to assess and identify observational studies that focused on the subsequent use of the NLP solutions to address clinical questions. In addition, the studies needed to comply with the standard design of the EHR-based observational research method.[3] A total of 390 articles were then removed during the screening process. Among them, we excluded 205 measurement studies of NLP that concentrated on developing and refining methods for making measurements.[33] We also removed 161 non-observational studies, 21 studies without NLP applications, and three studies with non-English EHR data. After that, 81 studies underwent full-text review, performed by five reviewers (authors S.F., R.R., N.Z., L.W., and S.M.). The full-text review further excluded 31 articles. Among them, seven studies were measurement design, five studies have no description of NLP in the full text, 15 studies did not comply with the standard design of the EHR-based observational research method, three studies had no full text or abstract, and one study was non-peer reviewed.

Data abstraction definition

To assess the reporting, methodologic, and evaluation standards of the identified articles, we considered Findable, Accessible, Interoperable, and Reusable (FAIR) and Reproducible, Implementable, Transparent, and Explainable (RITE) principles (Figure 2).[47,48] The FAIR data principles define the best practices of data stewardship and dissemination through emphasizing four criteria: findable, accessible, interoperable, and reusable.[47] After translating the FAIR principles into the use case of clinical NLP, the provenance of information resources for developing NLP applications include corpus location, corpus sharing criteria, metadata of corpus, definition of corpus (i.e., annotation guideline), and standard (semantic and syntactic). In addition to FAIR, the RITE implementation principles place an emphasis on process transparency and implementability, because the variability and explainability of the result are dependent on the process.[48] Even in a situation that does not require the process or results to be replicated, all important steps and details need to be documented and made available to ensure the traceability of the process and explainability of the result. This view implies the importance of the documentation and reporting of methodology (i.e., processes of NLP evaluation and rule refinement) and implementation (i.e., accessibility of key information resources for system deployment) details. We also considered the evaluation principles suggested by Friedman et al.[33] to guide the reporting practice: (1) the study team is obligated to identify the best possible way of measuring a standard given the research context and conduct measurement studies to estimate the error; and (2) the error estimated from the measurement studies need to be incorporated into analyses (observational study).

The key data elements that needed to be extracted include study metadata, NLP methodology (e.g., context, certainty, and normalization), gold standard development and evaluation (e.g., evaluation environment and sampling strategy), and overall reporting practice. Three categories (retrospective cohort, cross-sectional, and case-control) of study design were included based on Man's definition.[3–5] Based on the research question, study hypothesis, and clinical perspectives, we summarized the studies into five application domains: disease, drug, risk factors, social determinant of health, and other categories. The NLP methodology and evaluation-related data elements and definitions are provided in Table 1.

Quality control and synthesis

To ensure quality, we performed consensus development during the article screening and data abstraction. During the consensus development phase, two batches of articles

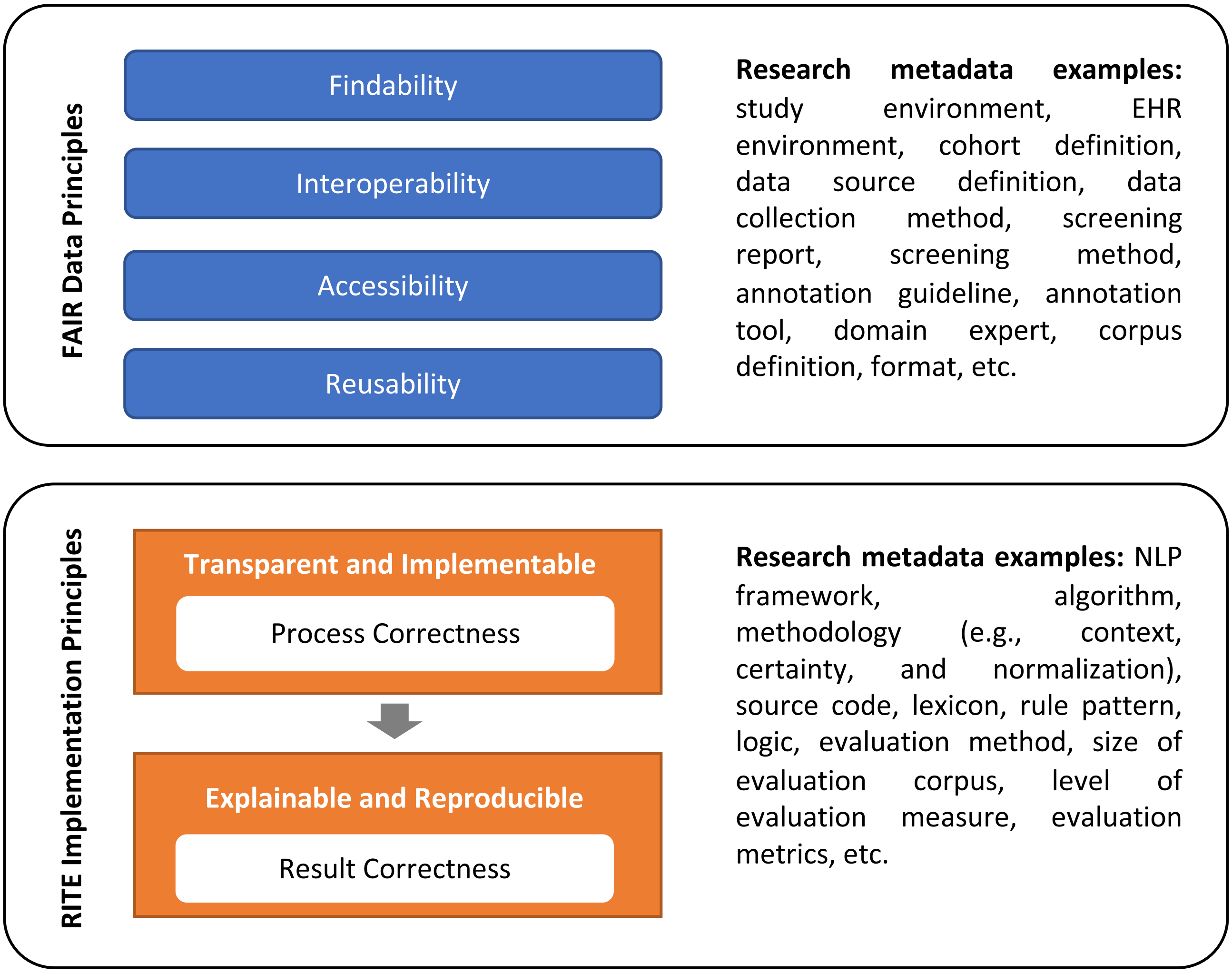

FAIR Data Principles
Findability
Interoperability
Accessibility
Reusability
Research metadata examples: study environment, EHR environment, cohort definition, data source definition, data collection method, screening report, screening method, annotation guideline, annotation tool, domain expert, corpus definition, format, etc.
RITE Implementation Principles
Transparent and Implementable
Process Correctness
Explainable and Reproducible
Result Correctness
Research metadata examples: NLP framework, algorithm, methodology (e.g., context, certainty, and normalization), source code, lexicon, rule pattern, logic, evaluation method, size of evaluation corpus, level of evaluation measure, evaluation metrics, etc.

FIGURE 2 FAIR data principles and RITE implementation principles. FAIR, Findable, Accessible, Interoperable, and Reusable; RITE, Reproducible, Implementable, Transparent, and Explainable.

were randomly sampled from the screening pool. Each batch contains 20 articles. Two reviewers (authors S.F. and R.R.) annotated the same set of articles. After each round, disagreements were measured using kappa statistics. The agreement scores in the two rounds were 0.573 (confidence interval [CI] 0.210–0.936) and 0.700 (CI 0.389–1.000), respectively. Two consensus meetings were organized to resolve the disagreements. During the meeting, the same reviewers (authors S.F. and R.R.) present the cases with disagreements to two adjudicators (authors H.L. and A.W.). Each case was discussed for reaching a consensus. In a rare situation when a consensus cannot be reached, the adjudicators provided the final decision. The full-text data abstraction was performed by five reviewers (authors S.F., R.R., N.Z., L.W., and S.M.). During the data abstraction, two reviewers (authors S.F. and L.W.) verified the abstraction quality of other reviewers by randomly sampling 40% of the data. Conflicting results were discussed and adjudicated between the primary reviewer and the validation reviewer. The abstracted data from the included studies were charted using frequencies of the following variables: year of publication, study period, geographic region of conduct, and study design. Descriptive statistics were used in the analysis of the data collected in this study.

RESULTS

Characteristics of included scoping reviews

Among the total 471 articles after the de-duplication, we identified 205 (44%) NLP-related measurement studies and 161 (34%) non-observational studies without NLP methodology to be excluded based on title and abstract screening. After the screening process, 81 articles were considered for full text review, of which we identified 50 articles for a comprehensive full-text review and subsequent data abstraction. A flow chart of this article selection process is shown in Figure 3.

TABLE 1 Definition of NLP methodology and evaluation-related data elements

Data elements	Definitions
Data source	Type of clinical documents used for NLP development and evaluation (e.g., clinical notes, radiology reports, pathology reports, and surgical operative reports)
EHR Environment	Single center, multiple centers, benchmark dataset, EHR and benchmark, no evaluation
Primary purpose of NLP	Risk factor extraction, outcome extraction, other
NLP method	Rule-based, statistical machine learning, hybrid, and deep learning
Source code, keyword, and definition	Release source code for machine learning models or lexicons and rule-patterns for rule-based models
Medical concept Normalization	A task aims at mapping medical mentions to a corresponding medical concept in controlled vocabularies such as the Unified Medical Language System (UMLS) or Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)
Has context and certainty level definition	Negated; status; hypothetical; experienced by someone other than the patient
Methods of evaluation	Corpus annotation design (splitting corpus into training, validation and test sets that contain both cases and controls), chart review on positive results only (apply NLP first and only review the returned results), compare with existing literature (EHR environment), compare with publicly available benchmark dataset, 5: Criteria validity
Level of evaluation measure	Concept (i.e., mentions or named entities), sentences, document, patient
Evaluation metrics	F1-score, precision, recall, positive predictive value (PPV), negative predictive value (NPV), kappa
Iteration and refinement	A process of applying the algorithm on the data. Incorrect cases were manually reviewed by domain experts and iteratively refined until all issues were resolved

Abbreviations: EHR, electronic health record; NLP, natural language processing.

Based on the comparison of our previous review on clinical NLP methodologies (Figure 4a),[6] we observed an upward trend in published research on NLP use for observational research, especially within the past 5 years. This finding not only affirms the need for leveraging NLP to extract important clinical information from EHRs, but also suggests an increasing number of translational efforts in applying NLP solutions in the context of clinical research. Among these methods being translated into clinical research, rule-based methodology was the most reported in the demonstration studies. On the other hand, although we observed an upward trend in the published NLP-assisted observational studies, compared with the distribution of overall EHR-based observational research (Figure 4b), utilization of NLP is still low. Since the HITECH Act of 2008, the growth of observational studies leveraging EHR data was exponential. However, such a pattern was not observed for NLP-assisted studies, despite the plethora of NLP methods papers. This number suggested an implementation gap between traditional NLP research and NLP-assisted observational research.

The summaries of primary study design, application domain, and disease classification are provided in Figure 5. Among the 50 studies, the primary research design was retrospective cohort (66%), followed by cross-sectional (28%) and case-control (6%; Figure 5a). We identified that 24 (48%) studies used NLP to extract risk factors, 21 (42%) studies extracted outcome, two (4%) studies extracted exposure, and three (6%) studies applied NLP for extracting multiple types of research variables. The study distribution of disease, drug, other categories, risk factors, and social determinant of health was 41 (82%), 4 (8%), 2(4%), 2 (4%), and 1 (2%), respectively (Figure 5b). The "Diseases" sub-domain was further classified based on the ICD-10 classification system. Top classification systems include Mental, Behavioral, and Neurodevelopmental disorders (20%), Diseases of the genitourinary system (8%), Diseases of the respiratory system (8%), and Neoplasms (8%; Figure 5c).

NLP methodology

Among the 50 surveyed articles, the rule-based method (72%) was the most reported NLP method, followed by statistical machine learning (12%), and hybrid (6%; Figure 6a). Five (10%) studies did not specify the NLP method used. Compared with the previous study reviewing NLP measurement studies that reported using the rule-based approach,[6] utilization of this approach increased from 48% to 72% among the demonstration studies. Figure 6b presents the reporting adherence of NLP methodologic definitions. Fifty-eight percent of the studies did not report model definition (e.g., lexicon, dictionary, code, etc.), 74% of the studies did not report normalization techniques (e.g., UMLS, SNOMED CT),

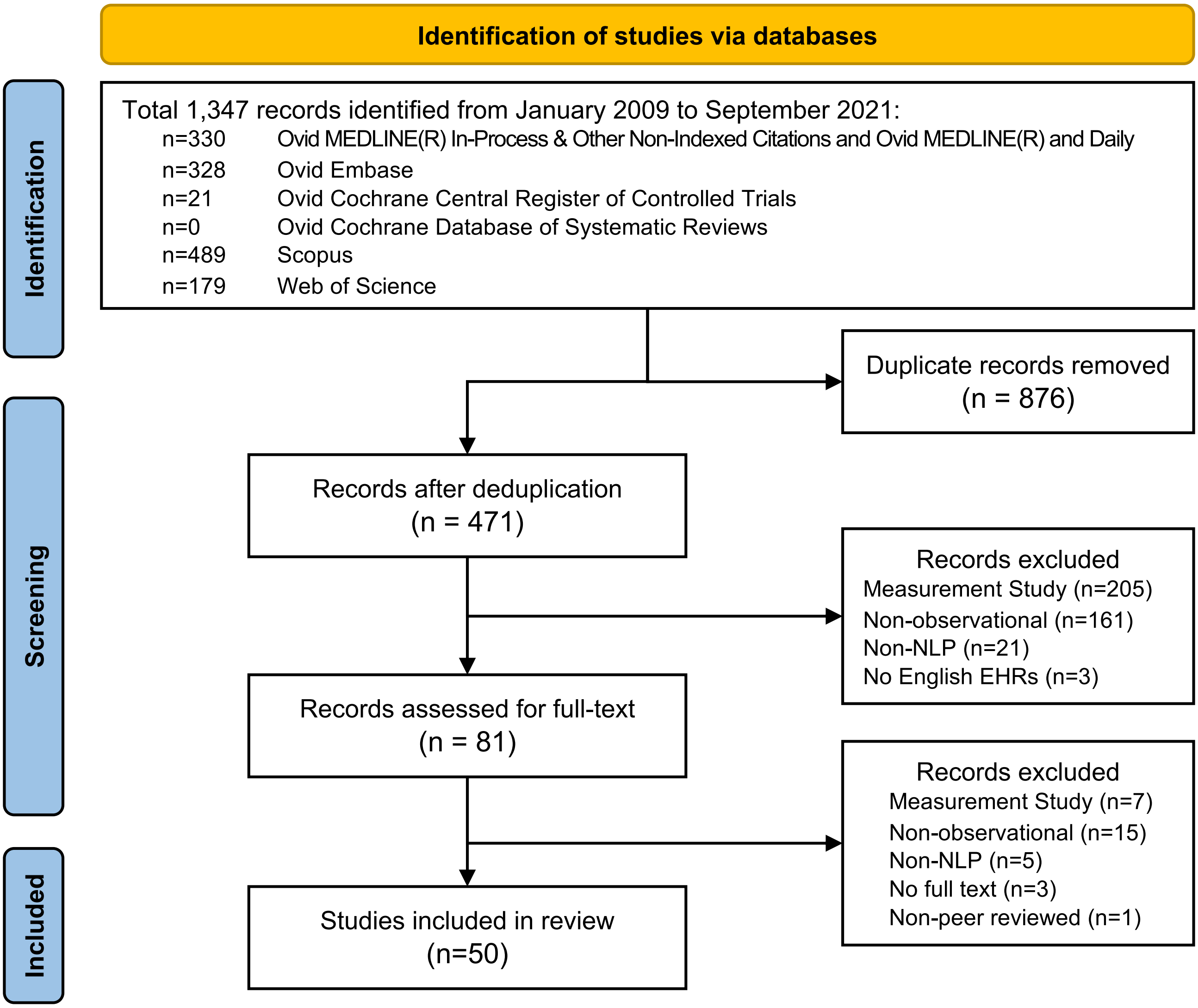

Identification of studies via databases
Identification
Screening
Included
Total 1,347 records identified from January 2009 to September 2021:
n=330 Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) and Daily
n=328 Ovid Embase
n=21 Ovid Cochrane Central Register of Controlled Trials
n=0 Ovid Cochrane Database of Systematic Reviews
n=489 Scopus
n=179 Web of Science
Duplicate records removed (n = 876)
Records after deduplication (n = 471)
Records excluded
Measurement Study (n=205)
Non-observational (n=161)
Non-NLP (n=21)
No English EHRs (n=3)
Records assessed for full-text (n = 81)
Records excluded
Measurement Study (n=7)
Non-observational (n=15)
Non-NLP (n=5)
No full text (n=3)
Non-peer reviewed (n=1)
Studies included in review (n=50)

FIGURE 3 Overview of article selection process. NLP, natural language processing.

and 58% of the studies did not report context definition (e.g., certainty, patient status [historical vs. present], experiencer [patient vs. family member]). Only six (12%) of the studies reported all three definitions. In summary, the reporting of NLP methodology needs to be enhanced for demonstrating method reproducibility and confidence in technical soundness.

Gold standard development and evaluation

In assessing the utilization and reporting of NLP-related evaluation methodologic standards (Figure 6c–f), most studies (66%) conducted the evaluation in a single EHR environment. Eleven (22%) of the studies involved multiple EHRs, one (2%) study utilized benchmark datasets, one (2%) study utilized both EHRs and benchmark datasets, three (6%) studies utilized the VA environment, and one (2%) study did not report evaluation environment. In addition, we discovered that 17 (34%) studies reported the use of sampling methods for establishing the evaluation cohort and preparing gold standard datasets. Furthermore, 25 (50%) studies reported the definition of the level of evaluation. Among these studies, 12 (24%) reported that evaluations were conducted at patient level, 12 (24%) at document level, and one (2%) at concept level. The median, minimal, and maximal evaluation size was 40, 716, and 2000 at document level and 18, 245, and 1042 at patient level, respectively (Figure 6d). Evaluation metrics, such as sensitivity and specificity, quantify the performance of NLP and can provide a direct estimate of the validity models. Based on the review, 14 (28%) studies did not report evaluation metrics, 11 (22%) studies reported one metric, 11 (22%) reported two metrics, and 14 (28%) reported more than two. The most-reported metric was sensitivity, followed by precision, specificity, and f1-score (Figure 6e). Post evaluation refinement is an important process for optimizing the NLP before using it in an

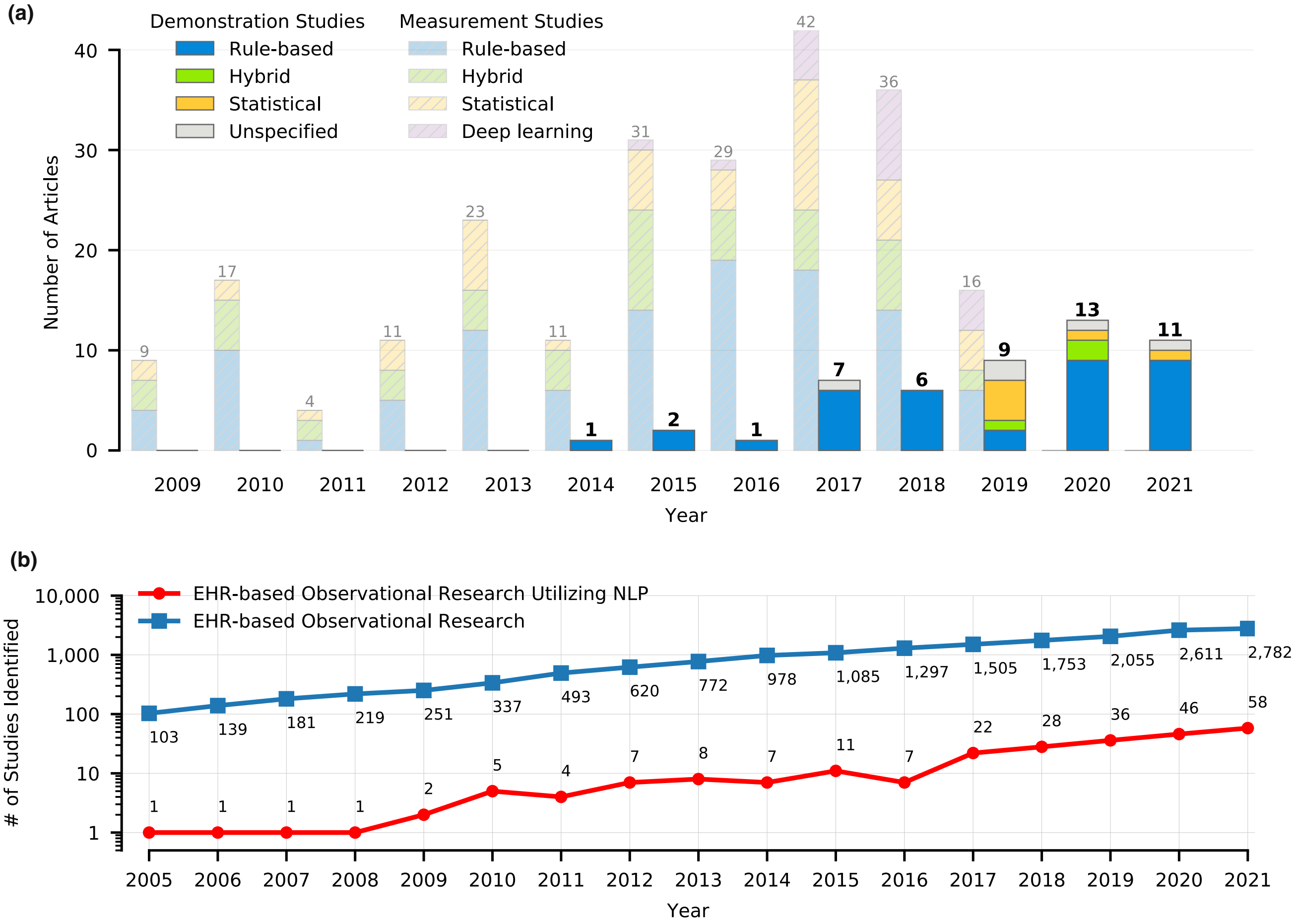


FIGURE 4 Trend view comparison of the number of EHR-based measurement studies and the number of observational studies with and without utilizing NLP. EHR, electronic health record; NLP, natural language processing.

observational study. We observed that 12 (24%) studies reported that NLP rules and keywords were iteratively revised prior to being used for research analyses.

Overall reporting practice

Among the total 50 studies, 18 (36%) studies reported NLP evaluation in the body of the manuscript, nine (18%) provided the reference of a prior study, two (4%) reported in the supplementary appendix, nine (18%) reported in the body of the manuscript and provided reference, four (8%) reported in the body of the manuscript and supplementary appendix, one (2%) provided reference and appendix, and seven (14%) did not report NLP evaluation (Figure 6g).

DISCUSSION

To better understand the current translation effort of NLP applications to clinical research, we conducted a scoping review on NLP-assisted observational studies and examined (1) whether important NLP methodology and evaluation details are reported; (2) where they are reported, and (3) the degree of granularity and confidence with which they are reported. Through our investigation, we discovered a high variation in reporting practices, such as missing details about measurement studies and inconsistent reporting of the location and granularity of NLP methodology and evaluation details. Fourteen percent of studies did not report NLP methodology and evaluation, 22% of studies did not report evaluation design, and 10% of studies did not report NLP methodology. A few studies claimed that "'NLP has been evaluated prior to release," however, no additional details could be found to justify the validity of the NLP results and clinical findings. In addition, among the 14 studies with both measurement and demonstration studies reported, we found no study evaluated the difference between the two environments. As suggested by our previous investigation,[49] if there is a substantial variation in the study setting, EHR environment, or cohort definition between measurement study and observational study, NLP may suffer from a portability issue. This issue could also occur within a single institution

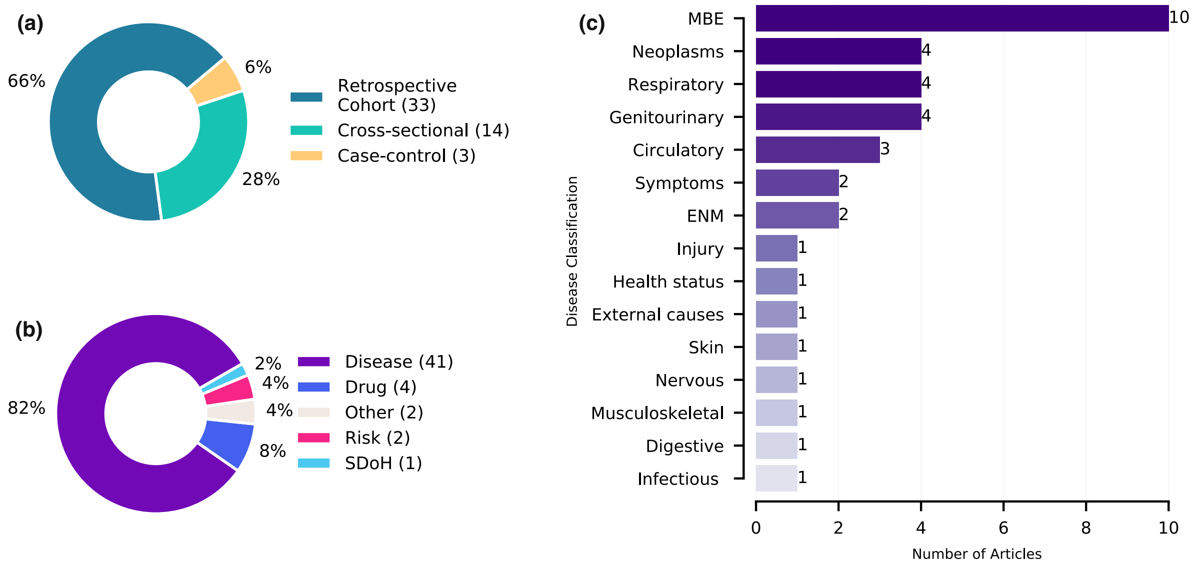
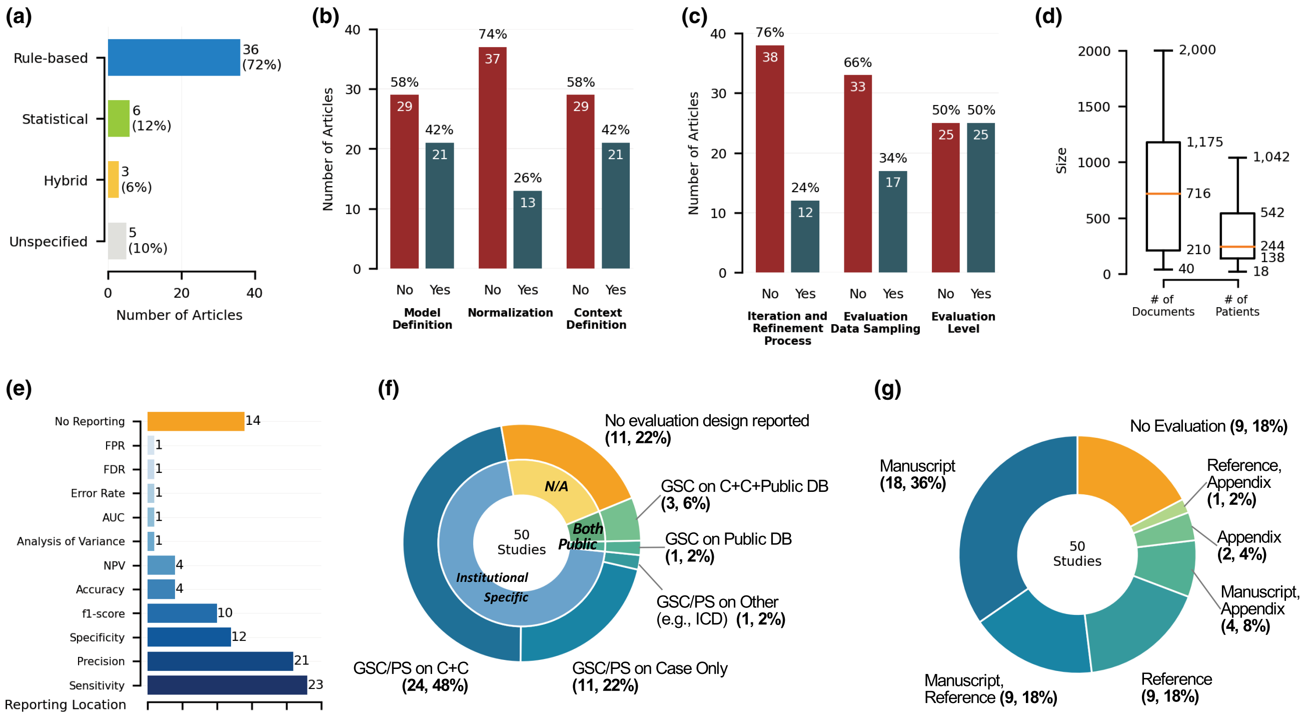


FIGURE 5 The Summaries of the Research Trend, Primary Study design, Application Domain and Disease Classification. MBE, Mental, Behavioral and Neurodevelopmental disorders; Genitourinary, Diseases of the genitourinary system; Respiratory, Diseases of the respiratory system; Neoplasms, Neoplasms; Circulatory, Diseases of the circulatory system; ENM, Endocrine, nutritional and metabolic diseases; Symptoms, Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; Infectious, Certain infectious and parasitic diseases; Digestive, Diseases of the digestive system; Musculoskeletal, Diseases of the musculoskeletal system and connective tissue; Nervous, Diseases of the nervous system; Skin, Diseases of the skin and subcutaneous tissue; External causes, External causes of morbidity; Health status, Factors influencing health status and contact with health services; Injury, Injury, poisoning and certain other consequences of external causes. EHR, electronic health record; NLP, natural language processing.



GSC/PS on C+C: Gold standard corpora or patient status is established on chart review of both case and control
GSC/PS on Case Only: Gold standard corpora or patient status label is established on chart review of case only
GSC/PS on Other: Gold standard patient status is based on other clinical measurements (e.g., ICD)
GSC on Public DB: Gold standard corpora is utilized based on publicly available database (e.g., MIMIC, i2b2 shared task)
GSC on C+C+Public DB: Gold standard corpora is established on chart review of both case and control and external database

FIGURE 6 Summary distribution of NLP methods (a, b), reporting practices of model definition, normalization, and context definition (certainty, status, and experiencer), evaluation methods (c – f), and overall reporting (6g).

due to EHR system migration, cohort definition, change of definition, and variation in the extract, transform, and load (ETL) process.

We further observed ambiguity and inconsistency when authors used the terms "framework," "system," "tool," and "algorithm." We observed that only 42% of the studies provided definitions of the NLP algorithm, and 18% of studies provided the reference of a prior measurement study. The lack of direct access to the original text processors, machine learning models, and pattern definitions may create substantial reproducibility challenges and may impact clinical validity. These findings highlight the strong need for more community efforts, such as partnering with international initiatives such as EQUATOR Network to establish reporting standards for observational studies empowered by NLP methods. As an initial effort, we emphasize several crucial points that need to be carefully considered when conducting NLP-assisted observational studies (Table 2).

One potential barrier to the wide adoption of NLP may be due to the high-stake nature of clinical research because findings could be translated into routine care delivery through practice-based knowledge discovery and evidence-based medicine.[62,63] Based on our review, we discovered that most data elements extracted by NLP applications are either risk factors (48%) or outcomes (42%), which are the primary variables used for the statistical analyses for discovering associations. Consequently, invalid NLP results will cause systematic bias, measurement error, and misclassification, which can ultimately impact the overall validity of any produced research and subsequent clinical guidelines derived from these studies. Therefore, when

TABLE 2 Towards a reporting standard for NLP-assisted observational research

Category	Recommendations	Examples
Overall recommendation	• Conducting an NLP measurement (evaluation) study prior to the clinical application • If a measurement study has been disseminated, (1) providing the reference, and (2) reporting evaluation and methodological details in either body of the manuscript or supplemental appendix • If a measurement study has not been disseminated, (1) reporting evaluation and methodological details in supplemental appendix, (2) reporting the abstract version in the body of the manuscript	19,28,29,32,38,50–55
Study environment, data source, and cohort	• Assessing and reporting potential variability between evaluation environment and application environment • Providing detailed cohort definition (e.g., cohort duration, clinical criteria, validity of clinical criteria) • Providing detailed methods for constructing the cohort (e.g., EHR data sources, query methods) • Providing evaluation method for cohort screening (e.g., chart review) • Reporting validity of cohort screening methods	50,56
Corpus annotation	• Identifying annotators with (clinical) domain expertise • Determining the appropriate number of annotators (minimal two) and the size of corpus for annotation • Conducting training and consensus development processes • Defining a minimum agreement score (e.g., Kappa, F1-score) that must be reached before proceeding with production	21,57
NLP algorithm and framework definition	• Specifying context definitions: certainty (e.g., negated, hypothetical, possible/probable concepts); status (future, present, history); experiencer (patient, family) • Specifying normalization techniques and resources (e.g., UMLS) • Reporting rule-based lexicons, patterns, and logic in either body of the manuscript or supplemental appendix • Providing reference and code access (e.g., Git repository) to any generic text processors and statistical models	21,58–61
NLP evaluation and refinement	• Defining the measurement level of evaluation (e.g., concept level, document level, and patient level) • Defining NLP evaluation duration • Conducting evaluation if the application environment is substantially different from the evaluation environment • Refining the system based on error analyses	58,60,61

Abbreviations: EHR, electronic health record; NLP, natural language processing.

translating and applying NLP solutions to observational studies, the process must require necessary assessment and evaluation efforts for mitigating various quality issues during the real-world implementation within the EHR environment to ensure research reproducibility and scientific rigor. To align with the six guiding principles (human autonomy, human safety, transparency, explainability and intelligibility, responsibility and accountability, inclusiveness and equity, and responsive and sustainable) released by the World Health Organization (WHO) to guide the ethical use of artificial intelligence for health, we outlined two key perspectives to promote the wide adoption and utilization of NLP solutions in clinical research in the following discussion sections.

Ensuring process transparency and scientific rigor

Reproducibility is crucial to NLP-assisted observational research because the validity of NLP models is dependent on the data from which they are derived. Solid data understanding and documentation can promote a good data curation plan and solutions for mitigating potential biases or confounders prior to model development and re-deployment. From the implementation perspective (RITE principle), deploying NLP for a real-world research application would require substantial translational efforts (e.g., rigorous evaluation, optimization, and dissemination) to ensure the model (1) is successfully deployed and executed and (2) can produce valid and explainable outcomes. According to the review, the utilization of the rule-based method (72%) for observational research was much higher than all other reported methods combined. The unique characteristic of the rule-based approach is that the lexicons and rule patterns are interpretable and able to be modified easily. On the other hand, no deep learning methods were reported in the observational research studies. Similarly, despite the recently increasing adoption of deep learning methods for NLP in the general domain, we discovered a delay of such application in the cancer domain in another scoping review for cancer NLP.[64] Such findings point out potential misaligned research objectives between translational research and informatics research – translational research emphasizes quality-driven objectives (e.g., practicality, customizability, and explainability) and informatics research focuses more on methodological innovation (i.e., sophisticated deep learning methods). To better bridge this gap, it is important for us to have a stronger engagement in translational research as well as community-wide interdisciplinary collaboration.

Promoting open NLP communities through team science collaboration

Open science carries the criteria of transparency, openness, and reproducibility, which are recognized as vital features of science.[65] NLP technologies, on the other hand, have a strong presence in both commercial industry and academic research. Consequently, the protection of intellectual property may prevent disclosing enough details about NLP solutions (e.g., methodology and evaluation) and become a potential barrier to open-science collaboration and knowledge sharing. When non-open and non-cross-validated NLP solutions are leveraged in clinical research, outcomes could potentially face limited utilization, external validation, and result in generalizability and trust issues. In contrast, open solutions can promote collaborative, and trustworthy environments for NLP development, evaluation, and deployment, foster partnership with diverse communities such as the National COVID Cohort Collaborative (N3C),[66] PCORnet,[67] Observational Health Data Sciences and Informatics (OHDSI),[68] and Precision Medicine Initiative (PMI),[69] and bring positive impacts to future workforce development. Therefore, there is a strong need for open collaboration across different disciplines and among the CTSA community. There are some early efforts for promoting open NLP collaboration including the National Center for Advancing Translational Sciences (NCATS) funded national consortium Open Health Natural Language Processing (OHNLP) consortium[70] and National Center for Data to Health (CD2H) funded federated evaluation platform NLP Sandbox.[71]

In 2010, the National Cancer Institute (NCI) published its first field guide entitled *Collaboration and Team Science* to foster the education, practice, and professional development of team science and interdisciplinary research.[72] Due to the increasing specialization and complexity of research methods and domain expertise, there is a strong need for adopting collaborative approaches among scientists across discipline-based fields. The development, evaluation, and application of NLP for observational research is an example of team science collaboration (TSC) across the disciplines of biomedical informatics, medicine, statistics, and epidemiology. For example, the development and evaluation of a research cohort is a process of operationalizing clinical definition to EHR data. This process requires extensive collaboration and communication among clinical research scientists and informaticians to ensure the validity and reliability of EHR-based measurements (e.g., ICD, CPT, and LOINC). Furthermore, developing gold standard resources requires expertise in clinical domain knowledge, data, infrastructure, and model. Failure to adhere to the TSC approach can affect the robustness of NLP models and the validity of research outcomes. This issue can be illustrated by not engaging domain experts and scientists when drawing sample populations, matching control cases, estimating sample size, conducting corpus annotation, etc.

The study does have some limitations. First, the review is limited by the defined scope. Studies which had relevant content but did not meet the inclusion and exclusion criteria were not included (e.g., written in non-English). In addition, the study could miss relevant articles due to the search strings and databases selected in the review. Finally, abstraction tasks themselves are often subjective in nature, and manual data abstraction may therefore introduce data quality issues. Supporting this is the fact that the first round of inter-abstractor agreement only had a kappa of 0.57. This moderate agreement suggests that several challenges were encountered when differentiating NLP related measurement studies and demonstration studies during the article screening process. Despite adjudication of any disagreements, misclassification is still possible. Our results should therefore be better viewed as an exploratory investigation, and we hope that this study can motivate future studies in a similar scope to provide additional evidence.

CONCLUSION

As NLP solutions are increasingly integrated into the research workflow for observational studies, the validity and reproducibility of these solutions take on high importance. Our study reviewed the reporting of NLP-specific methodological standards and evaluation processes in EHR-based observational studies. Our assessment reveals a high variation in the reporting practices, such as inconsistent use of references for measurement studies, variable reporting locations, and differing granularity of NLP methodology and evaluation details, which entails the need for continuously improving research reproducibility and scientific rigor for EHR-based observational studies. These critical issues must be addressed through fostering partnerships with clinical NLP and clinical and translational science communities through promotion of collaborative and transparent environments for the development, evaluation, and deployment of NLP solutions.

AUTHOR CONTRIBUTIONS

S.F. wrote the manuscript. H.L., S.F., M.H., A.W., and C.G.C. designed the research. S.F., R.R., L.W., S.M., N.Z., H.H., V.P., R.R., A.W., M.H., C.G.C., and H.L. performed the research. S.F., H.H., R.R., S.M., L.W., and N.Z. analyzed the data.

ACKNOWLEDGMENTS

This work was made possible by National Institute of Health (NIH) grant number U01TR002062. We also want to acknowledge the OHNLP Consortium, National Center for Data to Health (CD2H), and iEC text analytics community.

FUNDING INFORMATION

This study was funded by NIH-NCATS-U01TR002062.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

ORCID

Sunyang Fu https://orcid.org/0000-0003-1691-5179
Liwei Wang https://orcid.org/0000-0001-9970-8604
Sungrim Moon https://orcid.org/0000-0002-9191-3897
Nansu Zong https://orcid.org/0000-0003-0066-9524
Huan He https://orcid.org/0000-0003-1312-4195
Vikas Pejaver https://orcid.org/0000-0002-1943-0284
Anita Walden https://orcid.org/0000-0002-3327-7423
Melissa Haendel https://orcid.org/0000-0001-9114-8737
Christopher G. Chute https://orcid.org/0000-0001-5437-2545
Hongfang Liu https://orcid.org/0000-0003-2570-3741

REFERENCES

1. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med*. 2010;363(6):501-504.
2. Health NIo. Mission, Vision and Goals of the NIH Office of Clinical Research 2022 https://ocr.od.nih.gov/
3. Mann C. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J*. 2003;20(1):54-60.
4. Chute CG. Invited commentary: observational research in the age of the electronic health record. *Am J Epidemiol*. 2014;179(6):759-761.
5. Callahan A, Shah NH, Chen JH. Research and reporting considerations for observational studies using electronic health record data. *Ann Intern Med*. 2020;172(11_Supplement):S79-S84.
6. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. *J Biomed Inform*. 2020;109:103526.
7. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49.
8. Fu S, Wen A, Pagali S, St Sauver J, Sohn S, Fan J, et al. The Implication of Latent Information Quality to the Reproducibility of Secondary Use of Electronic Health Records. MEDINFO2021.
9. Fu S, Lopes GS, Pagali SR, et al. Ascertainment of delirium status using natural language processing from electronic health records. *J Gerontol*. 2022;77(3):524-530.
10. Pagali S, Fu S, Lindroth H, Sohn S, Burton MC, Lapid M. Delirium occurrence and association with outcomes in hospitalized COVID-19 patients. *Int Psychogeriatr*. 2021;33(10):1105-1109.

11. Fu S, Leung LY, Wang Y, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Med Inform*. 2019;7(2):e12109.
12. Inouye SK, Westendorp RG, Saczynski JS, Kimchi EY, Cleinman AA. Delirium in elderly people–Authors' reply. *Lancet*. 2014;383(9934):2045.
13. Fu S, Wyles CC, Osmon DR, et al. Automated detection of periprosthetic joint infections and data elements using natural language processing. *J Arthroplast*. 2021;36(2):688-692.
14. Tibbo ME, Wyles CC, Fu S, et al. Use of natural language processing tools to identify and classify periprosthetic femur fractures. *J Arthroplast*. 2019;34(10):2216-2219.
15. Kent DM, Leung LY, Zhou Y, et al. Association of silent cerebrovascular disease identified using natural language processing and future ischemic stroke. *Neurology*. 2021;97(13):e1313-e1321.
16. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc*. 2012;19(2):212-218.
17. Wang L, Olson JE, Bielinski SJ, et al. Impact of diverse data sources on computational phenotyping. *Front Genet*. 2020;11:556.
18. Shen F, Wang S, Wang Y, Afzal L, Liu N. Leveraging collaborative filtering to accelerate rare disease diagnosis. *AMIA Annu Symp Proc*. 2017;2017:1554-1563.
19. Lynch KEG, Whitcomb KC, Alba BW, DuVall PR, Blosnich SL, John R. Evaluation of suicide mortality among sexual minority us veterans from 2000 to 2017. *JAMA Netw*. 2020;3(12):e2031357.
20. Wyles CC, Tibbo ME, Fu S, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *J Bone Joint Surg Am*. 2019;101(21):1931.
21. Lott JP, Boudreau DM, Barnhill RL, et al. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. *JAMA Dermatol*. 2018;154(1):24-29.
22. Hylan TR, Von Korff M, Saunders K, et al. Automated prediction of risk for problem opioid use in a primary care setting. *J Pain*. 2015;16(4):380-387.
23. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: Where are the methods? *Ann Emerg Med*. 1996;27(3):305-308. doi:10.1016/s0196-0644(96)70264-0
24. Kaur H, Sohn S, Wi CI, et al. Automated chart review utilizing natural language processing algorithm for asthma predictive index. *BMC Pulm Med*. 2018;18(1):34. doi:10.1186/s12890-018-0593-9
25. Fu S, Peterson AL, Wang KJ, et al. Natural language processing for the evaluation of methodological standards and best practices of EHR-based clinical research. *AMIA Summits Transl Sci Proc*. 2020;2020:171-180.
26. Zhu H, Fan J, Wang H, Wang Q, Gao W. Prevalence and risk factors of suicidal ideation in adult inpatients with five different types of mental disorders. *J Affect Disord*. 2021;291:344-351. doi:10.1016/j.jad.2021.05.033
27. Hoogenboom WS, Perlis RH, Smoller JW, et al. Limbic system white matter microstructure and long-term treatment outcome in major depressive disorder: a diffusion tensor imaging study using legacy data. *World J Biol Psychiatry*. 2014;15(2).
28. Wu J, Morrison F, Zhao Z, et al. Reasons for discontinuing insulin and factors associated with insulin discontinuation in patients with type 2 diabetes mellitus: a real-world evidence study. *Clin Dermatol*. 2021;7(1):1. doi:10.1186/s40842-020-00115-2
29. Anthony MSR, Reed SD, Armstrong MA, et al. Design of the association of uterine perforation and expulsion of intrauterine device study: a multisite retrospective cohort study. *Am J Obstet Gynecol*. 2021;224(6):599.e1-e18. doi:10.1016/j.ajog.2021.01.003
30. Wang L, Fu S, Wen A, et al. Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. *JCO Clin Cancer Informat*. 2022;6:e2200006.
31. Tyczynski JE, Potluri R, Kilpatrick R, Mazumder D, Ghosh A, Liede A. Incidence and risk factors of pneumonitis in patients with non-small cell lung cancer: an observational analysis of real-world data. *Oncologia*. 2021;28:28.
32. Blumenthal KGL, Lai KH, Huang M, Wallace ZS, Wickner PG, Zhou L. Adverse and hypersensitivity reactions to prescription nonsteroidal anti-inflammatory agents in a large health care system. *J Allergy Clin Immunol Pract*. 2017;5(3):737-743.e3.
33. Friedman CP, Wyatt J. *Evaluation methods in biomedical informatics*. Springer Science & Business Media; 2005.
34. Michaelis J, Wellek S, Willems J. Reference standards for software evaluation. *Methods Inf Med*. 1990;29(04):289-297.
35. Quaglini S, Stefanelli M, Barosi G, Berzuini A. A performance evaluation of the expert system ANEMIA. *Comput Biomed Res*. 1988;21(4):307-323.
36. Wang L, Luo L, Wang Y, Wampfler J, Yang P, Liu H. Natural language processing for populating lung cancer clinical research data. *BMC Med Inform Decis Mak*. 2019;19(5):1-10.
37. McCoy TH, Yu S, Hart KL, et al. High throughput phenotyping for dimensional psychopathology in electronic health records. *Biol Psychiatry*. 2018;83(12):997-1004. doi:10.1016/j.biopsych.2018.01.011
38. McCoy TH, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study. *Alzheimers Dement*. 2020;16(3):531-540. doi:10.1016/j.jalz.2019.09.084
39. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*. 1994;1(2):161-174.
40. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229-236.
41. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A III. *The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2003.
42. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513.
43. Goryachev S, Sordo M, Zeng QT. *A suite of natural language processing tools developed for the I2B2 project. AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2006.
44. Liu H, Bielinski SJ, Sohn S, et al. An information extraction framework for cohort identification using electronic health records. *AMIA Summits Transl Sci Proc*. 2013;2013:149.

45. Wen A, Fu S, Moon S, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digital Med*. 2019;2(1):1-7.

46. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Res Synth Methods*. 2014;5(4):371-385.

47. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):1-9.

48. Fu S. *TRUST: Clinical Text Retrieval and Use towards Scientific Rigor and Transparent Process*. University of Minnesota; 2021.

49. Fu S, Leung LY, Raulli A-O, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med Inform Decis Mak*. 2020;20(1):1-12.

50. Zhang H, Plutzky J, Shubina M, Turchin A. Continued statin prescriptions after adverse reactions and patient outcomes: a cohort study. *Ann Intern Med*. 2017;167(4):221-227. doi:10.7326/M16-0838

51. Giess CSP, Pourjabbar S, Ip IK, Lacson R, Alper E, Khorasani R. Comparing diagnostic performance of digital breast tomosynthesis and full-field digital mammography in a hybrid screening environment. *AJR Am J Roentgenol*. 2017;209(4):929-934. doi:10.2214/AJR.17.17983

52. Li K, Banerjee I, Magnani CJ, Blayney DW, Brooks JD, Hernandez-Boussard T. Clinical documentation to predict factors associated with urinary incontinence following prostatectomy for prostate cancer. *Res Reports Urol*. 2020;12:7.

53. D Sullivan M, Boudreau D, Ichikawa L, et al. Primary care opioid taper plans are associated with sustained opioid dose reduction. *J Gen Intern Med*. 2020;35(3):687-695. doi:10.1007/s11606-019-05445-1

54. Bellows BK, DuVall SL, Kamauu AW, et al. Characteristics and use of treatment modalities of patients with binge-eating disorder in the Department of Veterans Affairs. *Eat Behav*. 2016;21:161-167.

55. Downs J, HarryLechler SS, Patel N, et al. Negative symptoms in early-onset psychosis and their association with antipsychotic treatment failure. *Schizophr Bull*. 2019;45(1):69-79. doi:10.1093/schbul/sbx197

56. McCoy TH Jr, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: a retrospective cohort study. *Alzheimers Dement*. 2019. (in press).

57. Colling C, Evans L, Broadbent M, et al. Identification of the delivery of cognitive behavioural therapy for psychosis (CBTp) using a cross-sectional sample from electronic health records and open-text information in a large UK-based mental health case register. *BMJ Open*. 2017;7(7):e015297.

58. Blumenthal DM, Singal G, Mangla SS, Macklin EA, Chung DC. Predicting non-adherence with outpatient colonoscopy using a novel electronic tool that measures prior non-adherence. *J Gen Intern Med*. 2015;30(6):724-731.

59. Irving JC, Shetty C, Pritchard H, et al. Gender differences in clinical presentation and illicit substance use during first episode psychosis: A natural language processing, electronic case register study. *BMJ Open*. 2021;11(4). doi:10.1136/bmjopen-2020-042949

60. Anzaldi LJD, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr*. 2017;17(1):248. doi:10.1186/s12877-017-0645-7

61. Jeon CYC, Chen Q, Yu W, et al. Identification of individuals at increased risk for pancreatic cancer in a community-based cohort of patients with suspected chronic pancreatitis. *Clin Transl Gastroenterol*. 2020;11(4):e00147. doi:10.14309/ctg.0000000000000147

62. Gelijns AC, Gabriel SE. Looking beyond translation – integrating clinical research with medical practice. *N Engl J Med*. 2012;366(18):1659-1661. doi:10.1056/NEJMp1201850

63. Milstein A. Code red and blue--safely limiting health care's GDP footprint. *N Engl J Med*. 2013;368(1):1-3. doi:10.1056/NEJMp1211374

64. Wang L, Fu S, Wen A, et al. Assessment of EHR for cancer research and patient care through a scoping review of cancer natural language processing. *JCO Clin Cancer Informat*. 2022;348(6242).

65. Nosek BA, Alter G, Banks GC, et al. Promoting an open research culture. *Science*. 2015;348(6242):1422-1425.

66. Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28(3):427-443.

67. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578-582.

68. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574.

69. Ashley EA. The precision medicine initiative: a new national effort. *JAMA*. 2015;313(21):2119-2120.

70. Consortium O. OHNLP Consortium 2022 http://ohnlp.org/

71. Health NCfDt. Cloud-based Sandbox for Analytics (Natural Language Processing) 2022 https://cd2h.org/node/152

72. Bennett LM, Gadlin H. Collaboration and team science: from theory to practice. *J Investig Med*. 2012;60(5):768-775.

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: Fu S, Wang L, Moon S, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review. *Clin Transl Sci*. 2022;00:1-14. doi:10.1111/cts.13463