ARTIFICIAL INTELLIGENCE

review articles

Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing

Liwei Wang, MD, PhD[1]; Sunyang Fu, PhD[1]; Andrew Wen, MS[1]; Xiaoyang Ruan, PhD[1]; Huan He, PhD[1]; Sijia Liu, PhD[1]; Sungrim Moon, PhD[1]; Michelle Mai[1]; Irbaz B. Riaz, MBBS, PhD[2]; Nan Wang, BS[3]; Ping Yang, MD, PhD[4]; Hua Xu, PhD[5]; Jeremy L. Warner, MD, MS[6,7]; and Hongfang Liu, PhD[1]

abstract

PURPOSE The advancement of natural language processing (NLP) has promoted the use of detailed textual data in electronic health records (EHRs) to support cancer research and to facilitate patient care. In this review, we aim to assess EHR for cancer research and patient care by using the Minimal Common Oncology Data Elements (mCODE), which is a community-driven effort to define a minimal set of data elements for cancer research and practice. Specifically, we aim to assess the alignment of NLP-extracted data elements with mCODE and review existing NLP methodologies for extracting said data elements.

METHODS Published literature studies were searched to retrieve cancer-related NLP articles that were written in English and published between January 2010 and September 2020 from main literature databases. After the retrieval, articles with EHRs as the data source were manually identified. A charting form was developed for relevant study analysis and used to categorize data including four main topics: metadata, EHR data and targeted cancer types, NLP methodology, and oncology data elements and standards.

RESULTS A total of 123 publications were selected finally and included in our analysis. We found that cancer research and patient care require some data elements beyond mCODE as expected. Transparency and reproductivity are not sufficient in NLP methods, and inconsistency in NLP evaluation exists.

CONCLUSION We conducted a comprehensive review of cancer NLP for research and patient care using EHRs data. Issues and barriers for wide adoption of cancer NLP were identified and discussed.

JCO Clin Cancer Inform 6:e2200006. © 2022 by American Society of Clinical Oncology

INTRODUCTION

As a real-world data source, electronic health records (EHRs) have the potential to provide the comprehensive and relatively timely clinical information necessary to facilitate cancer research and patient care. One of the major challenges associated with the use of EHR data for cancer research and patient care is data quality.[1] While ideally, all data elements necessary for cancer research and patient care purposes would be rendered accessible in a structured and standardized manner such that no additional efforts would be required to make use of the information contained therein, such is unfortunately not currently the case. For many current usages in so far as cancer research and patient care, the structured data provisioned as part of many popular EHR systems are considered to be incomplete[2] in that it is limited to specific subsets of clinical data, such as billing codes, and laboratory tests. Some data critical for cancer research and patient care may be recorded only in unstructured text, for example, whether and when a cancer improves or worsens after a given therapy.[3] Advancement in natural language processing (NLP) techniques has promoted the usage of clinical information extraction (IE) from unstructured texts to help supplement this information gap,[4,5] and consequently, the application of NLP in cancer domain has also been increasing.

To gain an understanding of the gaps and opportunities of NLP in EHR for cancer research and patient care, we conducted a scoping review of literature relevant to cancer NLP in EHR. We hypothesize that the need of NLP solutions to extract data elements reflects critical information not captured by structured EHR, and the readiness of NLP for extracting those data elements highly depends on the performance of NLP methodology involving many aspects including NLP tools, methods, evaluation, and reproducibility. Data elements defined as part of the Minimal Common Oncology Data Elements (mCODE) standard are used as a proxy for data elements

ASSOCIATED CONTENT

Appendix

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on June 15, 2022 and published at ascopubs.org/journal/cci on August 2, 2022: DOI https://doi.org/10.1200/CCI.22.00006

CONTEXT

Key Objective

To assess electronic health record (EHR) for cancer research and patient care through assessing the coverage of natural language processing (NLP)–derived data elements by the Minimal Common Oncology Data Elements and reviewing existing NLP methodologies for data extraction.

Knowledge Generated

A comprehensive review of cancer NLP for research and patient care using EHRs data extraction was conducted. Issues and barriers for wide adoption of cancer NLP were identified and discussed.

Relevance

Overcoming the identified issues and barriers will improve the readiness of EHRs for cancer research and patient care, thus propelling translational clinical research and care.

that would be important for cancer research and patient care.[6] The mCODE data standard was initiated from 2018 by ASCO, other founding collaborators, and a group of collaborators, including oncologists, informaticians, researchers, and experts in terminologies and standards, to develop and maintain standard computable oncology data formats in EHR for cancer research and practice. The final mCODE data standard (version 1.0) included six primary groups (domains): patient, disease, laboratory/vital, genomics, treatment, and outcome. Each domain is organized into several concepts, which then have associated data elements. These concepts are referred to as profiles. In total, 23 profiles exist across mCODE's six primary domains (Appendix Table A1). These data elements are linked to standard coding systems such as American Joint Committee on Cancer,[7] ClinVar,[8] International Classification of Diseases (10th revision), and Clinical Modification.[9]

Some prior work exists. In 2016, Yim et al conducted a similar literature review,[10] providing an introduction to NLP and its potential applications in oncology, describing specific tools available, and summarizing on the state of the current technology with respect to cancer case identification, staging, and outcomes quantification. Similarly, in 2019, Datta et al conducted a scoping review of clinical NLP literature extracting information from cancer-related EHR notes according to frame semantic principles.[11] They created frames from the reviewed articles pertaining to cancer information such as cancer diagnosis, tumor description, cancer procedure, breast cancer diagnosis, prostate cancer diagnosis, and pain in patients with prostate cancer. This review paper emphasized data model construction. Another relevant work reviewed the major NLP algorithmic advances and cancer NLP application developments over 3 years since 2016, summarizing the main trends of clinical cancer phenotype extraction from EHRs.[12]

In our scoping review, we focus on (1) presenting all the efforts using NLP in extracting cancer information as an end point or intermedium step and aligning them to mCODE, that is, the new data standard for oncology domain and (2) categorizing them based on NLP methodology.

METHODS

This scoping review was performed based on the following five stages of the framework from Arksey and O'Malley.[13]

Identifying the Research Question

In this scoping review, we aim to assess the alignment of NLP-extracted data elements with mCODE and review existing NLP methodologies for extracting said data elements.

Identifying Relevant Studies

We included articles to a 10-year period from January 1, 2010, to September 4, 2020. Only studies written in English were considered. Literature databases surveyed included Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, and Daily; Ovid Embase; Ovid Cochrane Central Register of Controlled Trials; Ovid Cochrane Database of Systematic Reviews; Scopus; and Web of Science. The search strategy for articles using NLP in cancer domain was designed and conducted by an experienced librarian (Larry J. Prokop). A detailed description of the search strategies used is provided in Appendix 1.

Study Selection

All the titles and abstracts after deduplication were screened, and the publications were included if

1. NLP was conducted for cancer, as defined below.
   a. The NLP involved could be used either as an end product or an inter-medium step for other downstream analytics.
   b. The NLP was cancer related.
2. EHR-sourced textual data in English were used as data source.

We excluded publications if they were

1. Not written in English.
2. Retrieved by irrelevant term matching.
3. Using non-EHR data sources such as literature, web resources, knowledge bases, clinical trials, and clinical guidelines.
4. Not using English EHR data.
5. Review papers/letters.

Charting the Relevant Studies

A standardized charting form was established to synthetize relevant publications. The information of interest can be categorized into four main sections: metadata, EHR data and targeted cancer types, oncology data elements and standards, and NLP methodology.

The Metadata section consisted of publication year; publication domain; major country of authors (first/senior author); types of author's organizations; the main study aim, defined as one of research and patient care, of the article; and the NLP study aim. The NLP study aim was classified into two groups: IE as an end point and IE as the input for machine learning. The research aim and patient care aim were further subcategorized, and more details are presented in Figure 1C.

The EHR Data and Targeted Cancer Types section aimed to summarize information including the targeted cancer types of the article, data time frame, and document types (eg, clinical notes, pathology report, and radiology reports). We defined the targeted cancer types using the ultimate cancer type around which the study was focused, for example, if the study focused on lung nodules but the aim of this study was to screen lung cancer, then we charted the targeted cancer type as lung cancer.

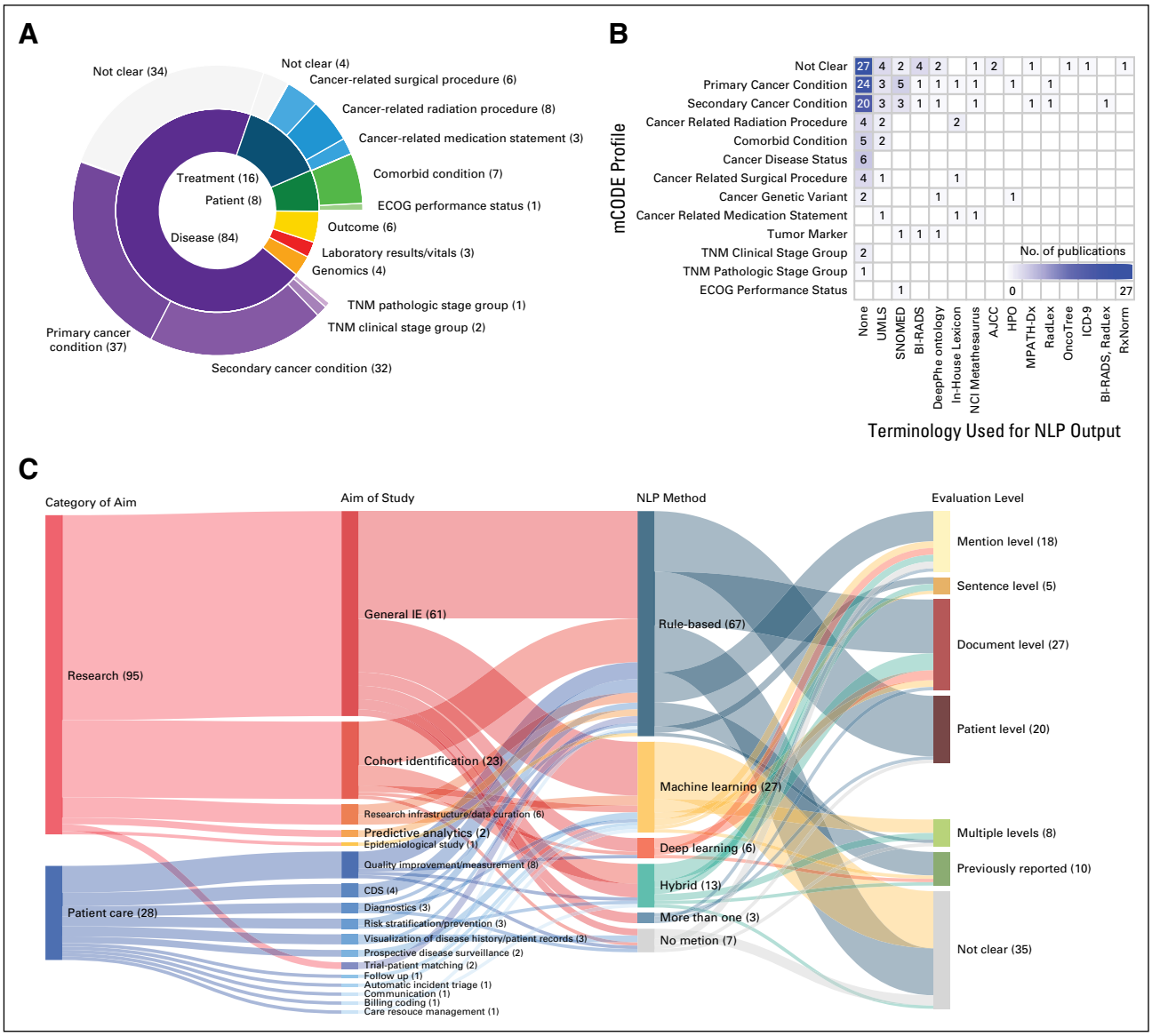

FIG 1. Synthetic analysis for mCODE and NLP methodology. (A) Distribution of data elements covered by mCODE. (B) Clustering visualization of the mCODE profiles and standardized terminologies. (C) Synthetic analysis for NLP methods, study aim, and evaluation level. AJCC, American Joint Committee on Cancer; BI-RADS, Breast Imaging Reporting and Data System; CDS, clinical decision support; ECOG, Eastern Cooperative Oncology Group; HPO, Human Phenotype Ontology; ICD-9, International Classification of Diseases (9th revision); IE, information extraction; mCODE, Minimal Common Oncology Data Elements; MPATH-Dx, Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis; NCI, National Cancer Institute; NLP, natural language processing; RadLex, Radiology Lexicon; RxNorm, no full name; UMLS, Unified Medical Language System.

In summarizing oncology data elements and standards, we aggregated NLP-extracted data elements based on the 23 profiles presented in Appendix Table A1. Additionally, we also examined the standardized terminology used for any normalization done of the NLP output for these data elements.

In the NLP methodology section, we described the most frequently used NLP tools, frameworks, or toolkits and cancer-specific NLP tools. In addition, we also analyzed NLP methods, the evaluation environment, performance metrics used, NLP methods evaluation granularity, and the study's reproducibility and rigor of evaluation.

NLP methods were categorized into one of six groups: rule-based, machine learning, deep learning, hybrid (one model with different approaches, eg, rules and machine learning), more than one (multiple models), and unspecified. Evaluation granularities included mention level, sentence level, document level, patient level, multiple levels, previously reported, and unknown.

To understand the study's evaluation scope (internal *v* external), we categorized the evaluation environment into one of the following groups: single center, multiple centers, Veterans Affairs, benchmark data set, single center and benchmark, no evaluation, and unspecified.

In many NLP studies, trust and adoption of any study outcomes are dependent on the validity and reproducibility of the NLP methods used. As such, in this review, we assessed the reporting patterns for NLP methods and evaluation methodologies. We first examine the reproducibility of NLP methods by evaluating code sharing, a crucial component of transparent and reproducible NLP research[14] (Table 1). To assess the rigor of evaluation, we considered four major evaluation best practices[5] (Table 1). Specifically, each publication is defaulted to have a score of 4, from which 1 point is subtracted whenever one of the criteria is met. Therefore, the maximum score for rigor of evaluation is 4 while the minimum is 0.

Each reviewer is responsible for charting 2-4 aspects and checking the charting quality of other reviewers by randomly sampling 10% of the data in the easily charted case or otherwise fully reviewing all data. When charting results disagreed between individual reviewers, reviewers met to resolve uncertainties.

TABLE 1. Definitions for Assessing Reporting and Evaluation Methods

Reproducibility of NLP Methods	Rigor of Evaluation Methods
Level 1: Source code not available, not replicable Level 2: Source code not available, experimental procedure replicable but individual components may differ (eg, tokenization and sentence parsing were done but method used not specified) Level 3: Code available	1. Evaluation description unclear/not present, for example, no description of study cohort, data collection, and annotation process, definition of eligibility criteria and data sources 2. Evaluation narrow/not comprehensive, small n 3. Evaluation design unclear (training, test) 4. Lack of evaluation rigor, for example, no mention of quality control (training, measuring inter-rater reliability, adjudication, etc)

Abbreviation: NLP, natural language processing.

Collating, Summarizing, and Reporting the Results

The results from the data charting were summarized, analyzed, and visualized both within and across the sections to present an overview of the scope of the application of NLP in cancer domain.

RESULTS

Figure 2 shows the article selection process. Finally, a comprehensive full-text review of the resulting 123 studies was performed by the study team. All data extractions from the articles (charting items) are presented in the Data Supplement.

Metadata

Figure 3A presents the distribution of articles, stratified by study aim and year, showing an increase in research interest for NLP in cancer. As shown in the author-country distribution depicted in Figure 3C, 111 articles (90%) had major authors from the United States while six (5%) had major authors from Australia. Figure 3B shows the distribution of articles according to the organization categories of the respective authors. In terms of publication venue, medical informatics or medical journals were the main venues for related studies, as shown in Figure 3D. In addition, NLP was used for IE as an end point in 101 articles and as machine learning input in 22 articles.

EHR Data and Targeted Cancer Types

Document types. Of the reviewed studies (Fig 3G), 58 (47%) studies extracted information from pathology reports,[15-72] 43 (35%) studies from clinical notes,[3,20,28,30,35,36,46,61,64,65,70,72-103] and 41 (33%) studies from radiology reports.[20,36-38,43,46,48,52,61,62,70,76,89,104-131] As various document types could be used for one study, document type numbers in each article were further analyzed (Fig 3H).

Targeted cancer types. The cancer types of interest for our reviewed studies were scattered across a wide spectrum, and significant variability was present in the number of articles for each cancer type of interest (Fig 3E). Of note, a single study may involve multiple cancer types. Among the 22 cancer types specified across all reviewed articles, breast cancer was the most intensively studied, being the primary cancer type of interest for 29 articles (24%), followed by prostate, colorectal, and lung cancers at 19 (15%), 15 (12%), and 15 (12%), respectively.

A clustering can be visualized for the document types and targeted cancer types (Appendix Fig A1). Pathology reports, clinical notes, and radiology reports were the major data sources for extracting information related to breast,

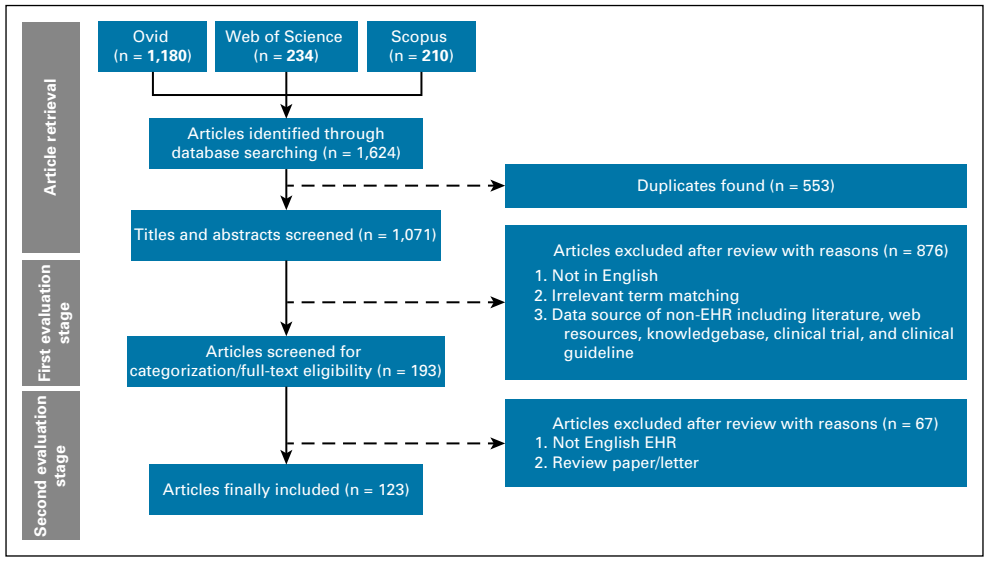

FIG 2. Overview of article selection process. EHR, electronic health record.

lung, liver, prostate, colorectal, brain, pancreatic, melanoma, head, and neck cancers.

Data time frame. In terms of the time frame of the data used for NLP, we calculated the age in years of the data used after excluding 33 studies that did not specify a data time frame. Before the 10-year mark, the number of studies increased as data age increased. Conversely, past the 10-year mark, the number of studies decreased as data age increased. We hypothesize that this trend may reflect the availability of EHR data (Fig 3J).

Oncology Data Elements and Standards

We first analyzed the 106 articles extracting data elements that can be mapped to a mCODE (Fig 1A). The distribution of data elements extracted was imbalanced across all six mCODE groups, with coverage extending to 12 of the 23 mCODE profiles (Appendix Table A1). Note that there are 31 articles extracting data elements corresponding to more than one mCODE group.

The most studied mCODE group was disease with 84 unique studies, associated with mCODE profiles of primary cancer condition, secondary cancer condition, TNM clinical stage group, and TNM pathologic stage group (Fig 1A). Our review revealed that most studies recorded no obvious differentiation between clinical and pathologic staging, and it was therefore difficult to categorize studies under these profiles as mCODE requires. As such, TNM staging was labeled as not clear under the disease group. In addition, those data elements without clear indications as to the primary cancer or secondary cancer conditions involved were similarly labeled as not clear. The second most studied mCODE group was treatment with 16 unique studies, followed by the patient, outcome, genomics, and laboratory/vital groups with 8, 6, 4, and 3 corresponding studies, respectively.

There are 20 articles extracting data elements not covered by mCODE. Table 2 presents the statistics of these data elements. Certain cancer screening criteria and social determinants of health were the two areas with data elements outside of mCODE's scope. For those 20 articles, only five used standard terminologies to code NLP output, including the Thyroid Imaging Reporting & Data System [131] and the Unified Medical Language System.[54,90,100,102]

About one third of the studies (43 of 123) adopted standard terminologies to normalize NLP output. Appendix Figure A2 shows a comparison between the reviewed articles and mCODE regarding adopted standard terminologies.

Figure 1B shows the clustering visualization of the mCODE profiles and standardized terminologies. A significant portion of reviewed studies failed to adopt standards for NLP output, while those profiles under the disease group were the major mCODE profiles normalized. The Unified Medical Language System covered the most profiles, followed by SNOMED.

NLP Methodology

NLP tools. Table 3 presents cancer-specific NLP tools, and the most frequently used general NLP tools, frameworks, or toolkits, which are consistent with our previous review.[4,5]

NLP methods and synthetic analysis. Appendix Figure A3 shows the changing trend of various NLP methods over time and the total number for each method. Rule-based methods are currently predominant. There is, however, an increasing trend in adoption of machine learning and hybrid methods. Despite the recently increasing adoption of deep learning methods[132] for NLP in the general domain, a delay of such application in the cancer domain can be observed.

To delineate the association among the charted items, we conducted a synthetic analysis for NLP methods, study aim of the article, and evaluation level (Fig 1C).

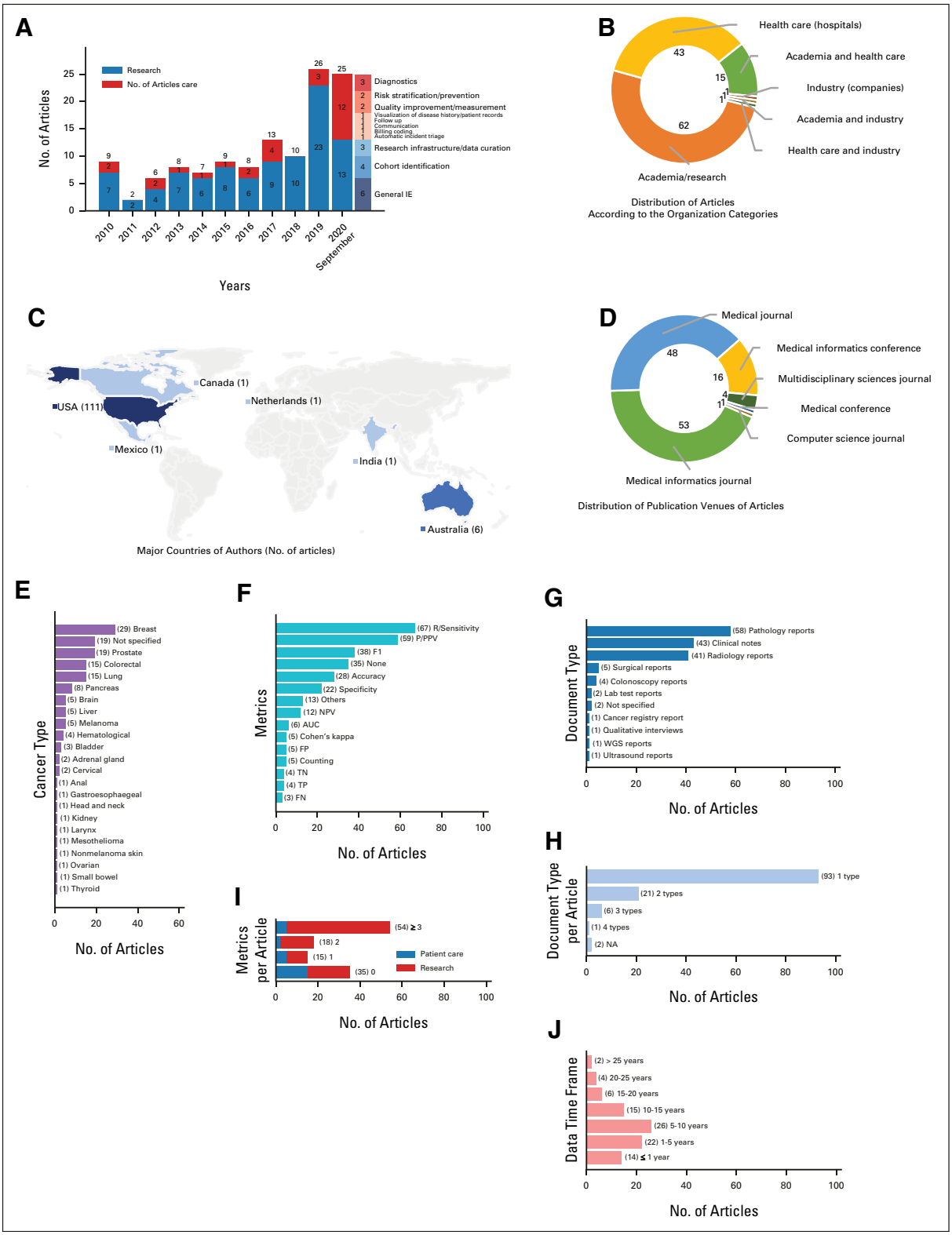

FIG 3. Analysis of metadata, EHR data scope, and evaluation metrics of included articles. (A) Distribution of articles over years. (B) Distribution of articles according to the organization categories. (C) Countries of major authors (No. of articles). (D) Distribution of publication venues of articles. (E) Distribution of cancer type. (F) Distribution of metrics. (G) Distribution of document type. (H) Histogram of document type number. (I) Histogram of metric number. (J) Histogram of data time frame. AUC, area under the curve; EHR, electronic health record; F1, no full name; FN, false negative; FP, false positive; IE, information extraction; NA, not applicable; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive; WGS, Whole Genome Sequencing.

Evaluation setting. Among the total 123 articles, 86 (70%) articles conducted the NLP evaluation in a single-site environment, 14 (11%) articles conducted a multisite evaluation, two (2%) studies used an external benchmark data set (eg, i2b2 shared task, Medical Information Mart for Intensive Care database), and four (3%) studies were based on VA data. In addition, 10 (8%) studies did not specify an evaluation environment, and seven (6%) studies did not report evaluation details.

Performance metrics. A variety of metrics were used to evaluate NLP methods (Fig 3F). The lack of consistency in performance evaluation makes cross-comparison of NLP systems difficult. Among the 35 articles reporting no evaluation metrics of NLP methods, 20 articles had the study aim for research, accounting for 21% of articles for research, and 15 were for patient care, accounting for 56% of this category (Fig 3I). In general, articles for research purpose reported more NLP evaluation metrics compared with those for patient care (Fig 3I).

Reproducibility and evaluation rigor. For the ability to replicate NLP methods, there were 78 articles (63%) in level 1. Level 2 contained 35 articles (28%), and level 3 had only 10 articles (8%). For the evaluation rigor of NLP methods in each study (defined in Table 1), 21 (17%) articles were rated as 0, 24 (20%) articles rated as 1, 32 (26%) rated as 2, 41 (34%) rated as 3, and five (4%) rated as 4.

DISCUSSION

The rapid growth of dense longitudinal EHR data sets provides substantial opportunities for the application of NLP in the cancer domain in recent years, as evidenced by the increased article count. It is extremely encouraging to see a jump of the NLP applications with a patient care (as opposed to research) focus in 2020 (12 articles). This growth reflects the value of NLP for clinical practice as NLP becomes more accessible. In the meanwhile, issues and barriers for wide adoption of cancer NLP were identified and discussed as follows.

NLP-targeted cancer types have covered most of the common cancer types collected by Cancer Stat Facts of the National Cancer Institute.[133] Cancer is highly complex and diverse; consequently, cancer research and patient care require diverse types of data, which can be reflected in our document type summary (Fig 3G).

The high utilization of NLP to extract information from pathology reports, clinical notes, and radiology reports demonstrates that important data elements for cancer research and patient care were embedded in text. Even with the ongoing efforts of standardizing pathology and radiology reporting,[134,135] the actual implementations seem to be insufficient.

mCODE is a current standard for EHR data to represent essential clinical elements for cancer patients, benefiting both oncology researchers and providers. We observed an imbalanced distribution across the six mCODE groups. Some groups, that is, the genomics group (four studies) and the laboratory/vital group (three studies), primarily come from structured data, thus less involved. Nevertheless, from the perspective of precision oncology and given that genomic data created through molecular diagnostics and treatment have been increasingly accumulated in EHR, tackling genomic data extraction from text could greatly advance the efficient secondary use of EHRs. In the meanwhile, aligning the standardized terminologies for data elements to mCODEs, for example, International Classification of Diseases for Oncology, American Joint Committee on Cancer, and ClinVar, is also important.

NLP for the outcome group (six studies) was under explored probably because of the intrinsic challenge of cancer disease status assessment as it mostly relies on clinicians' qualitative judgment on the current trend of the cancer. The judgment can be based on a single type or multiple kinds of evidence, such as imaging data, assessment of symptoms, tumor markers, and laboratory data, at a given time. Among those six studies, a study by Lee et al[104] proposed a scalable NLP pipeline that was capable of inferring Brain Tumor Reporting and Data System report scores. In the study by Sevenster et al, measurements used to synthesize treatment response status were extracted and paired across consecutive free-text computed tomography reports.[125] Clinically relevant outcomes can be extracted by NLP methods based on rules and machine learning[130] as well as deep NLP models[113] from radiologic reports. The codependent effects of NLP and machine learning in categorizing cancer disease status were investigated in computed tomography and magnetic resonance imaging reports.[117] Those studies focused on radiology reports which are ubiquitous and central to ascertainment of cancer disease status. However, additional information relevant to cancer status cannot be captured in radiology reports, such as laboratory results. The only study that did not use radiology reports applied deep NLP models to extract meaningful outcomes from clinical progress notes.[3]

Owing to the complexity of cancer disease status assessment, current EHR recording practice does not favor a seamless outcome assessment. For example, there is generally not a mechanism to input the staging information into the radiation oncology EHR or link metastatic sites to the original diagnosis, which are usually of interest for outcome analyses.[136] Moreover, oncologists sometimes have irreducible uncertainty about whether the cancer is responding or progressing when a clinical note is filed.[3] In addition, recorded cancer disease statuses may be time varying resulting in multiple instances of outcomes, which poses additional challenges for outcome extraction. For such mCODE data element as cancer disease status without a discrete data field in EHR, we believe NLP could play the most important role in extracting and structuring it.

Our review does indicate that some critical information elements covered by mCODE and needed for cancer

TABLE 2. Distribution of Data Elements Not Covered by mCODE

Application Areas	Data Elements	No. of Articles	No. of Articles (in total)
Cancer screening	Colonoscopy/polyps[16,40,53,54,60,66,100,102]	8	15
	Pancreatic cysts[46,99]	2	
	Digital rectal examination for prostate cancer[83,91]	2	
	Follow-up recommendations[105]	1	
	Lung nodule[129]	1	
	Thyroid nodule[131]	1	
Social determinants of health	Smoking[94]	1	3
	Social isolation[90]	1	
	Family history[96]	1	
Others	Test[101]	1	2
	Age of onset and death in family history[88]	1	

Abbreviation: mCODE, Minimal Common Oncology Data Elements.

research and patient care are not currently captured as part of structured EHR data. Since mCODE is used only for minimal critical oncology specialty information, it is as expected that those data elements extracted by NLP but not covered by mCODE are also important for cancer research and patient care, ie, cancer screening such as lung nodule screening and social determinants of health such as family history and smoking. This implies a great opportunity for improving structured data capture in EHRs to improve the readiness of EHRs for cancer research and patient care.

Recently, US Food and Drug Administration published a guidance for assessing EHRs and medical claims data to support regulatory decision making for drug and biological products, which also provides insightful and sharable recommendations for real-world data applications in other domains. For AI methods extracting data elements from unstructured data, it recommended to specify methods, tools, data sources, and the metrics associated with validation of the methods. Although the ability to replicate NLP methods was not mentioned in the US Food and Drug Administration guidance, we consider it a crucial factor that affects the adoption of NLP methods in the cancer domain. Unfortunately, our review revealed that more than half of the surveyed articles failed to provide either the source code or sufficient detail in the methodology to fully replicate the developed NLP systems, indicating a poor reporting practice in the domain.

TABLE 3. General and Specific NLP Tools, Frameworks, and Toolkits

Category	Name	No. of Papers
General NLP tools, frameworks, and toolkits	cTAKES[40,43,44,53-55,86,97-99,131]	11
	NLTK[74,77,104,106,110-112,124]	8
	UIMA[22,42,46,96,121]	5
	MedTagger[28,72,91,94]	4
	MetaMap[34,48,93,126]	4
Cancer-specific NLP tools	TIES NLP system[18,41,63]	3
	DeepPhe[38,82]	2
	PEP[49]	1
	BROK[128]	1
	The MOTTE[37]	1
	Clamp cancer module[26]	1

Abbreviations: BI-RADS, Breast Imaging Reporting and Data System; BROK, BI-RADS observation kit; MOTTE, methodist hospital text teaser; NLP, natural language processing; NLTK, natural language toolkit; PEP, pathology extraction pipeline; TIES, text information extraction system; UIMA, Unstructured Information Management Architecture.

Transparency and sharing contribute to assessment of research reproducibility, robustness, and replicability. Guidance on code sharing aligned with a specific study aim would be helpful to support transparency and reproducibility that would then strengthen the credibility of NLP results and promote downstream patient care leveraging NLP results.

For studies reporting no evaluation metrics for NLP methods, reasons that justify no assessment include that NLP evaluation has been reported in a previous publication or that NLP was simply used as an intermediate feature extraction step that feeds into a downstream machine learning algorithm that was itself evaluated. Although most studies reported metrics for NLP methods, reported metrics were not consistent across all studies, but rather study dependent, even for those with the same study aim.

Most studies inherited traditional and typical NLP study paradigms, focusing on mention-level, sentence-level, or document-level evaluations using a benchmark data set. Patient-level evaluation was relatively sparsely used (20 articles), and such an alignment to real-world clinical settings was very seldom seen. A considerable number of studies assigned with not clear evaluation level were extracting data elements from pathology reports. Without pondering that potential conflicting information could be reported for a single patient through multiple text reports, the evaluation level was failed to be clarified in these studies. There is certainly a need to evaluate NLP methods to an extent where the true level of complexity of clinical EHR data could be reflected following a scientific and rigorous evaluation process,[137] thus propelling the translation to clinical application.

Although NLP solutions can be potentially leveraged for large-scale IE, single-site studies are still predominant among current cancer NLP research. We acknowledge the potential barriers of multisite cancer NLP research such as complex concept definitions requiring extensive effort for participating sites to reach consensus and variations in clinical documentation patterns and data infrastructures

(eg, different extract, transform, and load processes) hampering cross-institutional experimentation. Regardless, we observed that 11% of studies surveyed involved more than one EHR in the study, demonstrating the feasibility of multi-institutional NLP efforts in cancer research and care.

There exist some limitations in our study. First, this review may be biased due to the potential of missing relevant articles caused by search strings and databases selected. Second, we only included articles written in English with the focus on using NLP in cancer EHR. Articles written in other languages would also provide valuable information. Third, our review did not include methods based on non-English EHRs. Finally, our study may also suffer the inherent ambiguity associated with data element collection, normalization, and analysis due to subjectivity introduced in the review process.

AFFILIATIONS

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN
[2]Department of Hematology/Oncology, Mayo Clinic, Scottsdale, AZ
[3]Department of Computer Science and Engineering, College of Science and Engineering, University of Minnesota, Minneapolis, MN
[4]Department of Quantitative Health Sciences, Mayo Clinic, Scottsdale, AZ
[5]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
[6]Departments of Medicine (Hematology/Oncology), Vanderbilt University, Nashville, TN
[7]Department Biomedical Informatics, Vanderbilt University, Nashville, TN

CORRESPONDING AUTHOR

Hongfang Liu, PhD, Mayo Clinic, 200 1st St SW, Rochester, MN 55905; e-mail: liu.hongfang@mayo.edu.

DISCLAIMER

The views expressed in the submitted article are authors' own and not an official position of the institution or funder.

SUPPORT

Supported by the National Institutes of Health (NIH) Grant Number 1U01TR002062-01 and U24CA194215-01A1.

AUTHOR CONTRIBUTIONS

Conception and design: Liwei Wang, Sunyang Fu, Andrew Wen, Hongfang Liu
Financial support: Ping Yang, Hua Xu, Jeremy L. Warner, Hongfang Liu
Administrative support: Hongfang Liu
Collection and assembly of data: Liwei Wang, Sunyang Fu, Andrew Wen, Xiaoyang Ruan, Huan He, Sijia Liu, Sungrim Moon, Michelle Mai, Irbaz B. Riaz, Nan Wang
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Irbaz B. Riaz

This author is a member of the *JCO Clinical Cancer Informatics* Editorial Board. Journal policy recused the author from having any role in the peer review of this manuscript.

Hua Xu

Employment: Melax Technologies Inc
Stock and Other Ownership Interests: Melax Technologies Inc
Consulting or Advisory Role: More Health Inc, Hebta LLC, Melax Technologies Inc
Patents, Royalties, Other Intellectual Property: Receive royalties from software license from UTHealth

Jeremy L. Warner

This author is an Associate Editor for *JCO Clinical Cancer Informatics*. Journal policy recused the author from having any role in the peer review of this manuscript.
Stock and Other Ownership Interests: HemOnc.org
Consulting or Advisory Role: Westat, Roche, Flatiron Health, Melax Tech

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We gratefully acknowledge Larry J. Prokop for implementing search strategies.

REFERENCES

1. Tayefi M, Ngo P, Chomutare T, et al: Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdiscip Rev Comput Stat 13:e1549, 2021
2. Bernstam EV, Warner JL, Krauss JC, et al: Quantitating and assessing interoperability between electronic health records. J Am Med Inform Assoc 29:753-760, 2022
3. Kehl KL, Xu W, Lepisto E, et al: Natural language processing to ascertain cancer outcomes from medical oncologist notes. JCO Clin Cancer Inform 4:680-690, 2020
4. Wang Y, Wang L, Rastegar-Mojarad M, et al: Clinical information extraction applications: A literature review. J Biomed Inform 77:34-49, 2018
5. Fu S, Chen D, He H, et al: Clinical concept extraction: A methodology review. J Biomed Inform 17:103526, 2020
6. Osterman TJ, Terry M, Miller RS: Improving cancer data interoperability: The promise of the Minimal Common Oncology Data Elements (mCODE) initiative. JCO Clin Cancer Inform 4:993-1001, 2020

7. Edge SB, Compton CC: The American Joint Committee on Cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 17:1471-1474, 2010
8. Landrum MJ, Lee JM, Benson M, et al: ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46:D1062-D1067, 2018
9. Wu P, Gifford A, Meng X, et al: Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and initial evaluation. JMIR Med Inform 7:e14325, 2019
10. Yim W-W, Yetisgen M, Harris WP, Kwan SW: Natural Language processing in oncology: A review. JAMA Oncol 2:797-804, 2016
11. Datta S, Bernstam EV, Roberts K: A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 100:103301, 2019
12. Savova GK, Danciu I, Alamudun F, et al: Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 79:5463-5470, 2019
13. Arksey H, O'Malley L: Scoping studies: Towards a methodological framework. Int J Soc Res Methodol 8:19-32, 2005
14. Errington TM, Denis A, Perfito N, et al: Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. eLife 10:e67995, 2021
15. Deshmukh PR, Phalnikar R: Anatomic stage extraction from medical reports of breast cancer patients using natural language processing. Health Technol 10:1555-1570, 2020
16. Fevrier HB, Liu L, Herrinton LJ, Li D: A transparent and adaptable method to extract colonoscopy and pathology data using natural language processing. J Med Syst 44:1-10, 2020
17. Oliveira CR, Niccolai P, Ortiz A, et al: Development and validation of a natural language processing algorithm for surveillance of cervical and anal cancer and precancer: A split-validation study. JMIR Med Inform 8:e20826, 2020
18. Jacobson RS, Becich MJ, Bollag RJ, et al: A federated network for translational cancer research using clinical data and biospecimens. Cancer Res 75:5194-5201, 2015
19. Parthasarathy G, Lopez R, McMichael J, Burke CA: A natural language-based tool for diagnosis of serrated polyposis syndrome. Gastrointest Endosc 92:886-890, 2020
20. Banerjee I, Bozkurt S, Caswell-Jin JL, et al: Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. JCO Clin Cancer Inform 3:1-12, 2019
21. Goulart BHL, Silgard ET, Baik CS, et al: Validity of natural language processing for ascertainment of EGFR and ALK test results in SEER cases of stage IV non-small-cell lung cancer. JCO Clin Cancer Inform 3:1-15, 2019
22. Malke JC, Jin S, Camp SP, et al: Enhancing case capture, quality, and completeness of primary melanoma pathology records via natural language processing. JCO Clin Cancer Inform 3:1-11, 2019
23. Odisho AY, Bridge M, Webb M, et al: Automating the capture of structured pathology data for prostate cancer clinical care and research. JCO Clin Cancer Inform 3:1-8, 2019
24. Oliwa T, Maron SB, Chase LM, et al: Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. JCO Clin Cancer Inform 3:1-8, 2019
25. Santus E, Li C, Yala A, et al: Do neural information extraction algorithms generalize across institutions? JCO Clin Cancer Inform 3:1-8, 2019
26. Soysal E, Warner JL, Wang J, et al: Developing customizable cancer information extraction modules for pathology reports using CLAMP. Stud Health Technol Inform 264:1041-1045, 2019
27. Thompson J, Hu J, Mudaranthakam DP, et al: Relevant word order vectorization for improved natural language processing in electronic health records. Sci Rep 9:9253, 2019
28. Wang L, Wampfler J, Dispenzieri A, et al: Achievability to extract specific date information for cancer research. AMIA Annu Symp Proc 2019:893-902, 2019
29. AalAbdulsalam AK, Garvin JH, Redd A, et al: Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. AMIA Jt Summits Transl Sci Proc 2017:16-25, 2018
30. Breitenstein MK, Liu H, Maxwell KN, et al: Electronic health record phenotypes for precision medicine: Perspectives and caveats from treatment of breast cancer at a single institution. Clin Transl Sci 11:85-92, 2018
31. Glaser AP, Jordan BJ, Cohen J, et al: Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform 2:1-8, 2018
32. Lott JP, Boudreau DM, Barnhill RL, et al: Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. JAMA Dermatol 154:24-29, 2018
33. Qiu JX, Yoon H-J, Srivastava K, et al: Scalable deep text comprehension for cancer surveillance on high-performance computing. BMC Bioinform 19:488, 2018 (suppl 18)
34. Zeng Z, Espino S, Roy A, et al: Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics 19:498, 2018 (suppl 17)
35. Gregg JR, Lang M, Wang LL, et al: Automating the determination of prostate cancer risk strata from electronic medical records. JCO Clin Cancer Inform 1:CCI.16.00045, 2017
36. Lin FP-Y, Pokorny A, Teng C, Epstein RJ: TEPAPA: A novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. Sci Rep 7:6918, 2017
37. Patel TA, Puppala M, Ogunti RO, et al: Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer 123:114-121, 2017
38. Savova GK, Tseytlin E, Finan S, et al: DeepPhe: A natural language processing system for extracting cancer phenotypes from clinical records. Cancer Res 77:e115-e118, 2017
39. Schroeck FR, Patterson OV, Alba PR, et al: Development of a natural language processing engine to generate bladder cancer pathology data for health services research. Urology 110:84-91, 2017
40. Wadia R, Shifman M, Levin FL, et al: A clinical decision support system for monitoring post-colonoscopy patient follow-up and scheduling. AMIA Jt Summits Transl Sci Proc 2017:295-301, 2017
41. Xie F, Lee J, Munoz-Plaza CE, et al: Application of text information extraction system for real-time cancer case identification in an integrated healthcare organization. J Pathol Inform 8:48, 2017
42. Osborne JD, Wyatt M, Westfall AO, et al: Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc 23:1077-1084, 2016

43. Sada Y, Hou J, Richardson P, et al: Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care 54:e9-e14, 2016
44. Imler TD, Morea J, Kahi C, et al: Multi-center colonoscopy quality measurement utilizing natural language processing. Am J Gastroenterol 110:543-552, 2015
45. Nguyen AN, Moore J, O'Dwyer J, Philpot S: Assessing the utility of automatic Cancer Registry notifications data extraction from free-text pathology reports. AMIA Annu Symp Proc 2015:953-962, 2015
46. Roch AM, Mehrabi S, Krishnan A, et al: Automated pancreatic cyst screening using natural language processing: A new tool in the early detection of pancreatic cancer. HPB 17:447-453, 2015
47. Wieneke AE, Bowles EJA, Cronkite D, et al: Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform 6:38, 2015
48. Wagholikar AS, Nguyen A, Fung M: A method for matching patients to advanced prostate cancer clinical trials. Electron J Health Inf 8:e61-e66, 2014
49. Ashish N, Dahm L, Boicey C: University of California, Irvine-Pathology Extraction Pipeline: The pathology extraction pipeline for information extraction from pathology reports. Health Inform J 20:288-305, 2014
50. Kim BJ, Merchant M, Zheng C, et al: A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports. J Endourol 28:1474-1478, 2014
51. Thomas AA, Zheng C, Jung H, et al: Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results. World J Urol 32:99-103, 2014
52. Heintzelman NH, Taylor RJ, Simonsen L, et al: Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. J Am Med Inform Assoc 20:898-905, 2013
53. Hou JK, Chang M, Nguyen T, et al: Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. Dig Dis Sci 58:936-941, 2013
54. Imler TD, Morea J, Imperiale TF: Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. Clin Gastroenterol Hepatol 12:1130-1136, 2014
55. Imler TD, Morea J, Kahi C, Imperiale TF: Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol 11:689-694, 2013
56. Strauss JA, Chao CR, Kwan ML, et al: Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Inform Assoc 20:349-355, 2013
57. Buckley JM, Coopey SB, Sharko J, et al: The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 3:23, 2012
58. Coopey SB, Mazzola E, Buckley JM, et al: The role of chemoprevention in modifying the risk of breast cancer in women with atypical breast lesions. Breast Cancer Res Treat 136:627-633, 2012
59. Eide MJ, Tuthill JM, Krajenta RJ, et al: Validation of claims data algorithms to identify nonmelanoma skin cancer. J Invest Dermatol 132:2005-2009, 2012
60. Mehrotra A, Dellon ES, Schoen RE, et al: Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. Gastrointest Endosc 75:1233-1239.e14, 2012
61. Xu H, Fu Z, Shah A, et al: Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011:1564-1572, 2011
62. Al-Haddad MA, Friedlin J, Kesterson J, et al: natural Language processing for the development of a clinical registry: A validation study in intraductal papillary mucinous neoplasms. HPB 12:688-695, 2010
63. Crowley RS, Castine M, Mitchell K, et al: caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 17:253-264, 2010
64. Friedlin J, Overhage M, Al-Haddad MA, et al: Comparing methods for identifying pancreatic cancer patients using electronic data sources. AMIA Annu Symp Proc 2010:237-241, 2010
65. Hsu W, Taira RK: Tools for improving the characterization and visualization of changes in neuro-oncology patients. AMIA Annu Symp Proc 2010:316-320, 2010
66. Nayor J, Borges LF, Goryachev S, et al: Natural language processing accurately calculates adenoma and sessile serrated polyp detection rates. Dig Dis Sci 63:1794-1800, 2018
67. Wilson RA, Chapman WW, Defries SJ, et al: Automated ancillary cancer history classification for mesothelioma patients from free-text clinical reports. J Pathol Inform 1:24, 2010
68. Ou Y, Patrick J: Automatic population of structured reports from narrative pathology reports. Presented at: 7th Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2014), Auckland, New Zealand, January 20-23, 2014
69. Yala A, Barzilay R, Salama L, et al: Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 161:203-211, 2016
70. Dexter PR, He J, Mark L, et al: A comparison of structured data query methods versus natural language processing to identify metastatic melanoma cases from electronic health records. Pharmacoepidemiol Drug Saf 25:253-254, 2016 (suppl 3)
71. Napolitano G, Fox C, Middleton R, Connolly D: Pattern-based information extraction from pathology reports for cancer registration. Cancer Causes Control 21:1887-1894, 2010
72. Wang L, Luo L, Wang Y, et al: Natural language processing for populating lung cancer clinical research data. BMC Med Inf Decis Mak 19:239, 2019
73. Montelongo González EE, Reyes Ortiz JA, González Beltrán BA: Machine learning models for cancer type classification with unstructured data. Comput Sist 24:403-411, 2020
74. Bozkurt S, Paul R, Coquet J, et al: Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. Learn Health Syst 4:e10237, 2020
75. Agaronnik ND, Lindvall C, El-Jawahri A, et al: Challenges of developing a natural language processing method with electronic health records to identify persons with chronic mobility disability. Arch Phys Med Rehabil 21:21, 2020
76. Brizzi K, Zupanc SN, Udelsman BV, et al: Natural language processing to assess palliative care and end-of-life process measures in patients with breast cancer with leptomeningeal disease. Am J Hosp Palliat Care 37:371-376, 2020
77. Hernandez-Boussard T, Blayney DW, Brooks JD: Leveraging digital data to inform and improve quality cancer care. Cancer Epidemiol Biomarkers Prev 29:816-822, 2020
78. Li K, Banerjee I, Magnani CJ, et al: Clinical documentation to predict factors associated with urinary incontinence following prostatectomy for prostate cancer. Res 12:7-14, 2020
79. Narayanan A, Topaloglu U, Laurini JA, Diaz-Garelli F: Building cancer diagnosis text to OncoTree mapping pipelines for clinical sequencing data integration and Curation. AMIA Jt Summits Transl Sci Proc 2020:440-448, 2020

80. Udelsman BV, Lee KC, Lilley EJ, et al: Variation in serious illness communication among surgical patients receiving palliative care. J Palliat Med 23:411-414, 2020
81. Venkataraman GR, Pineda AL, Bear Don't Walk Iv OJ, et al: FasTag: Automatic text classification of unstructured medical narratives. PLoS One 15:e0234647, 2020
82. Yuan Z, Finan S, Warner J, et al: Interactive exploration of longitudinal cancer patient histories extracted from clinical text. JCO Clin Cancer Inform 4:412-420, 2020
83. Bozkurt S, Kan KM, Ferrari MK, et al: Is it possible to automatically assess pretreatment digital rectal examination documentation using natural language processing? A single-centre retrospective study. BMJ Open 9:e027182, 2019
84. Coquet J, Bozkurt S, Kan KM, et al: Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients. J Biomed Inform 94:103184, 2019
85. Guan M, Cho S, Petro R, et al: Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes. JAMIA Open 2:139-149, 2019
86. Lindvall C, Lilley EJ, Zupanc SN, et al: Natural language processing to assess end-of-life quality indicators in cancer patients receiving palliative surgery. J Palliat Med 22:183-187, 2019
87. Ling AY, Kurian AW, Caswell-Jin JL, et al: Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open 2:528-537, 2019
88. Mowery DL, Kawamoto K, Bradshaw R, et al: Determining onset for familial breast and colorectal cancer from family history Comments in the electronic health record. AMIA Jt Summits Transl Sci Proc 2019:173-181, 2019
89. Udelsman B, Chien I, Ouchi K, et al: Needle in a haystack: Natural language processing to identify serious illness. J Palliat Med 22:179-182, 2019
90. Zhu VJ, Lenert LA, Bunnell BE, et al: Automatically identifying social isolation from clinical narratives for patients with prostate Cancer. BMC Med Inf Decis Mak 19:43, 2019
91. Bozkurt S, Park JI, Kan KM, et al: An automated feature engineering for digital rectal examination documentation using natural language processing. AMIA Annu Symp Proc 2018:288-294, 2018
92. Si Y, Roberts K: A frame-based NLP system for cancer-related information extraction. AMIA Annu Symp Proc 2018:1524-1533, 2018
93. Hernandez-Boussard T, Kourdis PD, Seto T, et al: Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. AMIA Annu Symp Proc 2017:876-882, 2017
94. Wang L, Ruan X, Yang P, Liu H: Comparison of three information sources for smoking information in electronic health records. Cancer Inform 15:237-242, 2016
95. Joffe E, Pettigrew EJ, Herskovic JR, et al: Expert guided natural language processing using one-class classification. J Am Med Inform Assoc 22:962-966, 2015
96. Mehrabi S, Krishnan A, Roch AM, et al: Identification of patients with family history of pancreatic cancer-Investigation of an NLP system portability. Stud Health Technol Inform 216:604-608, 2015
97. Ni Y, Wright J, Perentesis J, et al: Increasing the efficiency of trial-patient matching: Automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inf Decis Mak 15:28, 2015
98. Carrell DS, Halgrim S, Tran D-T, et al: Using natural Language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. Am J Epidemiol 179:749-758, 2014
99. Mehrabi S, Schmidt CM, Waters JA, et al: An efficient pancreatic cyst identification methodology using natural language processing. Stud Health Technol Inform 192:822-826, 2013
100. Denny JC, Choma NN, Peterson JF, et al: Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Med Decis Mak 32:188-197, 2012
101. Warner JL, Anick P, Hong P, Xue N: Natural Language processing and the oncologic history: Is there a match?. J Oncol Pract 7:e15-e19, 2011
102. Denny JC, Peterson JF, Choma NN, et al: Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 17:383-388, 2010
103. Mull HJ, Stolzmann KL, Shin MH, et al: Novel method to flag cardiac implantable device infections by integrating text mining with structured data in the Veterans Health Administration's electronic medical record. JAMA Netw 3:e2012264-e2012264, 2020
104. Lee SJ, Weinberg BD, Gore A, Banerjee I: A scalable natural language processing for inferring BT-RADS categorization from unstructured brain magnetic resonance reports. J Digit Imaging 33:1393-1400, 2020
105. Lou R, Lalevic D, Chambers C, et al: Automated detection of radiology reports that require follow-up imaging using natural language processing feature engineering and machine learning classification. J Digit Imaging 33:131-136, 2020
106. Senders JT, Cho LD, Calvachi P, et al: Automating clinical chart review: An open-source natural language processing pipeline developed on free-text radiology reports from patients with glioblastoma. JCO Clin Cancer Inform 4:25-34, 2020
107. Syed K, Iv WS, Ivey K, et al: Integrated natural language processing and machine learning models for standardizing radiotherapy structure Names. Healthcare (Basel) 8:30, 2020
108. Syed K, Sleeman Wt, Hagan M, et al: Automatic incident triage in radiation oncology incident learning system. Healthcare (Basel) 8:14, 2020
109. Tan JR, Cheong EHT, Chan LP, Tham WP: Implementation of an artificial intelligence-based double read system in capturing pulmonary nodule discrepancy in CT studies. Curr Probl Diagn Radiol 50:119-122, 2021
110. Bozkurt S, Alkim E, Banerjee I, Rubin DL: Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. J Digit Imaging 32:544-553, 2019
111. Brown AD, Kachura JR: Natural language processing of radiology reports in patients with hepatocellular carcinoma to predict radiology resource utilization. J Am Coll Radiol 16:840-844, 2019
112. Chen W, Butler RK, Zhou Y, et al: Prediction of pancreatic cancer based on imaging features in patients with duct abnormalities. Pancreas 49:413-419, 2020
113. Kehl KL, Elmarakeby H, Nishino M, et al: Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol 25:25, 2019
114. Senders JT, Karhade AV, Cote DJ, et al: Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. JCO Clin Cancer Inform 3:1-9, 2019
115. Van Haren RM, Correa AM, Sepesi B, et al: Ground glass lesions on chest imaging: Evaluation of reported incidence in cancer patients using natural language processing. Ann Thorac Surg 107:936-940, 2019
116. Walker G, Soysal E, Xu H: Development of a natural language processing tool to extract radiation treatment sites. Cureus 11:e6010, 2019
117. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging 31:178-184, 2018
118. Wadia R, Akgun K, Brandt C, et al: Comparison of natural language processing and manual coding for the identification of cross-sectional imaging reports suspicious for lung cancer. JCO Clin Cancer Inform 2:1-7, 2018

119. Ananda-Rajah MR, Bergmeir C, Petitjean F, et al: Toward electronic surveillance of invasive mold diseases in hematology-oncology patients: An expert system combining natural language processing of chest computed tomography reports, microbiology, and antifungal drug data. JCO Clin Cancer Inform 1:1-10, 2017
120. Beyer SE, McKee BJ, Regis SM, et al: Automatic Lung-RADS™ classification with a natural language processing system. J Thorac Dis 9:3114-3122, 2017
121. Castro SM, Tseytlin E, Medvedeva O, et al: Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 69:177-187, 2017
122. Moore CR, Farrag A, Ashkin E: Using natural language processing to extract abnormal results from cancer screening reports. J Patient Saf 13:138-143, 2017
123. Bozkurt S, Gimenez F, Burnside ES, et al: Using automatically extracted information from mammography reports for decision-support. J Biomed Inform 62:224-231, 2016
124. Yim W-W, Denman T, Kwan SW, Yetisgen M: Tumor information extraction in radiology reports for hepatocellular carcinoma patients. AMIA Jt Summits Transl Sci Proc 2016:455-464, 2016
125. Sevenster M, Bozeman J, Cowhy A, Trost W: A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. J Biomed Inform 53:36-48, 2015
126. Ananda-Rajah MR, Martinez D, Slavin MA, et al: Facilitating surveillance of pulmonary invasive mold diseases in patients with haematological malignancies by screening computed tomography reports using natural language processing. PLoS One 9:e107797, 2014
127. Sevenster M, Bozeman J, Cowhy A, Trost W: Automatically pairing measured findings across narrative abdomen CT reports. AMIA Annu Symp Proc 2013:1262-1271, 2013
128. Sippo DA, Warden GI, Andriole KP, et al: Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. J Digit Imaging 26:989-994, 2013
129. Danforth KN, Early MI, Ngan S, et al: Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 7:1257-1262, 2012
130. Cheng LT, Zheng J, Savova GK, Erickson BJ: Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 23:119-132, 2010
131. Chen KJ, Dedhia PH, Imbus JR, Schneider DF: Thyroid ultrasound reports: Will TI-RADS improve natural language processing capture of critical thyroid nodule features?. J Surg Res 256:557-563, 2020
132. Young T, Hazarika D, Poria S, Cambria E: Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag 13:55-75, 2018
133. Facts CS. Cancer Stat Facts: Common Cancer Sites, 2021. https://seer.cancer.gov/statfacts/html/common.html
134. Sluijter CE, van Lonkhuijzen LR, van Slooten H-J, et al: The effects of implementing synoptic pathology reporting in cancer diagnosis: A systematic review. Virchows Archiv 468:639-649, 2016
135. Cramer JA, Eisenmenger LB, Pierson NS, et al: Structured and templated reporting: An overview. Appl Radiol 43:18-21, 2014
136. Matuszak MM, Fuller CD, Yock TI, et al.: Performance/outcomes data and physician process challenges for practical big data efforts in radiation oncology. Med Phys 45:e811-e819, 2018
137. Fu S: TRUST: Clinical Text Retrieval and Use towards Scientific Rigor and Transparent Process. Minneapolis, MN, University of Minnesota, 2021

APPENDIX 1. SEARCH STRATEGIES

Ovid

Database(s): EBM Reviews—Cochrane Central Register of Controlled Trials August 2020, EBM Reviews—Cochrane Database of Systematic Reviews 2005 to September 3, 2020, Embase 1974 to 2020 September 4, Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily 1946 to September 4, 2020.

Search Strategy

No.	Searches	Results
1	exp Natural Language Processing/	9,794
2	("coreference resolution" or "co-reference resolution" or "information extraction" or "named entity extraction" or "named entity recognition" or "natural language processing" or NLP or "relation extraction" or "text mining").ti,ab,hw,kw.	14,976
3	1 or 2	14,976
4	exp neoplasms/	7,921,733
5	exp Medical Oncology/	191,672
6	((hodgkin* adj1 disease) or adenocarcinoma* or adenoma* or anticarcinogen* or Astrocytoma* or blastoma* or burkitt* or cancer* or carcinogen* or carcinoid* or carcinom* or carcinosarcoma* or chordoma* or "Chronic Myeloproliferative Disorder*" or craniopharyngioma* or ependymoma* or Esthesioneuroblastoma* or germinoma* or "gestational trophoblastic disease*" or Glioblastoma* or glioma* or gonadoblastoma* or hepatoblastoma* or histeocytoma* or histiocytoma* or histiocytos* or leukaemi* or leukemi* or lymphangioma* or lymphangiomyoma* or lymphangiosarcoma* or lymphom* or Macroglobulinemia* or malignan* or melanom* or meningioma* or mesenchymoma* or mesonephroma* or Mesothelioma* or metasta* or "multiple myeloma*" or "Mycosis Fungoide*" or neoplas* or neuroblastoma* or neuroma* or nonmelanoma* or nsclc or oncogen* or oncolog* or ostesarcoma* or Papillomatos* or paraganglioma* or paraneoplas* or pheochromocytoma* or plasmacytoma* or precancerous or retinoblastoma* or Rhabdomyosarcoma* or sarcoma* or "section 16" or "Szary Syndrome*" or teratocarcinoma* or teratoma* or tumor* or tumour*).ti,ab,hw,kw.	10,960,738
7	4 or 5 or 6	11,269,350
8	3 and 7	1801
9	(exp animals/or exp nonhuman/) not exp humans/	11,156,069
10	((alpaca or alpacas or amphibian or amphibians or animal or animals or antelope or armadillo or armadillos or avian or baboon or baboons or beagle or beagles or bee or bees or bird or birds or bison or bovine or buffalo or buffaloes or buffalos or "c elegans" or "Caenorhabditis elegans" or camel or camels or canine or canines or carp or cats or cattle or chick or chicken or chickens or chicks or chimp or chimpanze or chimpanzees or chimps or cow or cows or "D melanogaster" or "dairy calf" or "dairy calves" or deer or dog or dogs or donkey or donkeys or drosophila or "Drosophila melanogaster" or duck or duckling or ducklings or ducks or equid or equids or equine or equines or feline or felines or ferret or ferrets or finch or finches or fish or flatworm or flatworms or fox or foxes or frog or frogs or "fruit flies" or "fruit fly" or "G mellonella" or "Galleria mellonella" or geese or gerbil or gerbils or goat or goats or goose or gorilla or gorillas or hamster or hamsters or hare or hares or heifer or heifers or horse or horses or insect or insects or jellyfish or kangaroo or kangaroos or kitten or kittens or lagomorph or lagomorphs or lamb or lambs or llama or llamas or macaque or macaques or macaw or macaws or marmoset or marmosets or mice or minipig or minipigs or mink or minks or monkey or monkeys or mouse or mule or mules or nematode or nematodes or octopus or octopuses or orangutan or "orang-utan" or orangutans or "orang-utans" or oxen or parrot or parrots or pig or pigeon or pigeons or piglet or piglets or pigs or porcine or primate or primates or quail or rabbit or rabbits or rat or rats or reptile or reptiles or rodent or rodents or ruminant or ruminants or salmon or sheep or shrimp or slug or slugs or swine or tamarin or tamarins or toad or toads or trout or urchin or urchins or vole or voles or waxworm or waxworms or worm or worms or xenopus or "zebra fish" or zebrafish) not (human or humans or patient or patients)).ti,ab,hw,kw.	9,608,843
11	8 not (9 or 10)	1,748
12	limit 11 to (editorial or erratum or note or addresses or autobiography or bibliography or biography or blogs or comment or dictionary or directory or interactive tutorial or interview or lectures or legal cases or legislation or news or newspaper article or overall or patient education handout or periodical index or portraits or published erratum or video-audio media or webcasts) [Limit not valid in CCTR, CDSR, Embase, Ovid MEDLINE(R), Ovid MEDLINE(R) Daily Update, Ovid MEDLINE(R) In-Process,Ovid MEDLINE(R) Publisher; records were retained]	32
13	11 not 12	1,716
14	remove duplicates from 13	1,180

Scopus

1. TITLE-ABS-KEY("coreference resolution" OR "co-reference resolution" OR "information extraction" OR "named entity extraction" OR "named entity recognition" OR "natural language processing" OR NLP OR "relation extraction" OR "text mining")
2. TITLE-ABS-KEY((hodgkin* W/1 disease) or adenocarcinoma* or adenoma* or anticarcinogen* or Astrocytoma* or blastoma* or burkitt* or cancer* or carcinogen* or carcinoid* or carcinom* or carcinosarcoma* or chordoma* or "Chronic Myeloproliferative Disorder*" or craniopharyngioma* or ependymoma* or Esthesioneuroblastoma* or germinoma* or "gestational trophoblastic disease*" or Glioblastoma* or glioma* or gonadoblastoma* or hepatoblastoma* or histeocytoma* or histiocytoma* or histiocytos* or leukaemi* or leukemi* or lymphangioma* or lymphangiomyoma* or lymphangiosarcoma* or lymphom* or Macroglobulinemia* or malignan* or melanom* or meningioma* or mesenchymoma* or mesonephroma* or Mesothelioma* or metasta* or "multiple myeloma*" or "Mycosis Fungoide*" or neoplas* or neuroblastoma* or neuroma* or nonmelanoma* or nsclc or oncogen* or oncolog* or ostesarcoma* or Papillomatos* or paraganglioma* or paraneoplas* or pheochromocytoma* or plasmacytoma* or precancerous or retinoblastoma* or Rhabdomyosarcoma* or Sarcoma* or "section 16" or "Szary Syndrome*" or teratocarcinoma* or teratoma* or tumor* or tumor*)
3. LANGUAGE(english)
4. 1 and 2 and 3
5. TITLE-ABS-KEY((alpaca OR alpacas OR amphibian OR amphibians OR animal OR animals OR antelope OR armadillo OR armadillos OR avian OR baboon OR baboons OR beagle OR beagles OR bee OR bees OR bird OR birds OR bison OR bovine OR buffalo OR buffaloes OR buffalos OR "c elegans" OR "Caenorhabditis elegans" OR camel OR camels OR canine OR canines OR carp OR cats OR cattle OR chick OR chicken OR chickens OR chicks OR chimp OR chimpanze OR chimpanzees OR chimps OR cow OR cows OR "D melanogaster" OR "dairy calf" OR "dairy calves" OR deer OR dog OR dogs OR donkey OR donkeys OR drosophila OR "Drosophila melanogaster" OR duck OR duckling OR ducklings OR ducks OR equid OR equids OR equine OR equines OR feline OR felines OR ferret OR ferrets OR finch OR finches OR fish OR flatworm OR flatworms OR fox OR foxes OR frog OR frogs OR "fruit flies" OR "fruit fly" OR "G mellonella" OR "Galleria mellonella" OR geese OR gerbil OR gerbils OR goat OR goats OR goose OR gorilla OR gorillas OR hamster OR hamsters OR hare OR hares OR heifer OR heifers OR horse OR horses OR insect OR insects OR jellyfish OR kangaroo OR kangaroos OR kitten OR kittens OR lagomorph OR lagomorphs OR lamb OR lambs OR llama OR llamas OR macaque OR macaques OR macaw OR macaws OR marmoset OR marmosets OR mice OR minipig OR minipigs OR mink OR minks OR monkey OR monkeys OR mouse OR mule OR mules OR nematode OR nematodes OR octopus OR octopuses OR orangutan OR "orang-utan" OR orangutans OR "orang-utans" OR oxen OR parrot OR parrots OR pig OR pigeon OR pigeons OR piglet OR piglets OR pigs OR porcine OR primate OR primates OR quail OR rabbit OR rabbits OR rat OR rats OR reptile OR reptiles OR rodent OR rodents OR ruminant OR ruminants OR salmon OR sheep OR shrimp OR slug OR slugs OR swine OR tamarin OR tamarins OR toad OR toads OR trout OR urchin OR urchins OR vole OR voles OR waxworm OR waxworms OR worm OR worms OR xenopus OR "zebra fish" OR zebrafish) AND NOT (human OR humans or patient or patients))
6. 4 and not 5
7. DOCTYPE(ed) OR DOCTYPE(bk) OR DOCTYPE(er) OR DOCTYPE(no) OR DOCTYPE(sh)
8. 6 and not 7
9. INDEX(embase) OR INDEX(medline) OR PMID(0* OR 1* OR 2* OR 3* OR 4* OR 5* OR 6* OR 7* OR 8* OR 9*)
10. 8 and not 9

Web of Science

1. TOPIC: ((("coreference resolution" OR "co-reference resolution" OR "information extraction"OR "named entity extraction" OR "named entity recognition" OR "natural language processing" OR NLP OR "relation extraction" OR "text mining")) *AND* **TOPIC:** (((hodgkin* NEAR/1 disease) or adenocarcinoma* or adenoma* or anticarcinogen* or Astrocytoma* or blastoma* or burkitt* or cancer* or carcinogen* or carcinoid* or carcinom* or carcinosarcoma* or chordoma* or "Chronic Myeloproliferative Disorder*" or craniopharyngioma* or ependymoma* or Esthesioneuroblastoma* or germinoma* or "gestational trophoblastic disease*" or Glioblastoma* or glioma* or gonadoblastoma* or hepatoblastoma* or histeocytoma* or histiocytoma* or histiocytos* or leukaemi* or leukemi* or lymphangioma* or lymphangiomyoma* or lymphangiosarcoma* or lymphom* or Macroglobulinemia* or malignan* or melanom* or meningioma* or mesenchymoma* or mesonephroma* or Mesothelioma* or metasta* or "multiple myeloma*" or "Mycosis Fungoide*" or neoplas* or neuroblastoma* or neuroma* or nonmelanoma* or nsclc or oncogen* or oncolog* or ostesarcoma* or Papillomatos* or paraganglioma* or paraneoplas* or pheochromocytoma* or plasmacytoma* or precancerous or retinoblastoma* or Rhabdomyosarcoma* or Sarcoma* or "section 16" or "Szary Syndrome*" or teratocarcinoma* or teratoma* or tumor* or tumour*)) *AND* **LANGUAGE:** (English) *AND* **DOCUMENT TYPES:** (Article OR Abstract of Published Item OR Data Paper OR Letter OR Meeting Abstract OR Proceedings Paper OR Review OR Software Review)Indexes = SCI-EXPANDED Timespan = All years
2. TS=((alpaca OR alpacas OR amphibian OR amphibians OR animal OR animals OR antelope OR armadillo OR armadillos OR avian OR baboon OR baboons OR beagle OR beagles OR bee OR bees OR bird OR birds OR bison OR bovine OR buffalo OR buffaloes OR buffalos OR "c elegans" OR "Caenorhabditis elegans" OR camel OR camels OR canine OR canines OR carp OR cats OR cattle OR chick OR chicken OR chickens OR chicks OR chimp OR chimpanze OR chimpanzees OR chimps OR cow OR cows OR "D melanogaster" OR "dairy calf" OR "dairy calves" OR deer OR dog OR dogs OR donkey OR donkeys OR drosophila OR "Drosophila melanogaster" OR duck OR duckling OR ducklings OR ducks OR equid OR equids OR equine OR equines OR feline OR felines OR ferret OR ferrets OR finch OR finches OR fish OR flatworm OR flatworms OR fox OR foxes OR frog OR frogs OR "fruit flies" OR "fruit fly" OR "G mellonella" OR "Galleria mellonella" OR geese OR gerbil OR gerbils OR goat OR goats OR goose OR gorilla OR gorillas OR hamster OR hamsters OR hare OR hares OR heifer OR heifers OR horse OR horses OR insect OR insects OR jellyfish OR kangaroo OR kangaroos OR kitten OR kittens OR lagomorph OR lagomorphs OR lamb OR lambs OR llama OR llamas OR macaque OR macaques OR macaw OR macaws OR marmoset OR marmosets OR mice OR minipig OR minipigs OR mink OR minks OR monkey OR monkeys OR mouse OR mule OR mules OR nematode OR nematodes OR octopus OR octopuses OR orangutan OR "orang-utan" OR orangutans OR "orang-utans" OR oxen OR parrot OR parrots OR pig OR pigeon OR pigeons OR piglet OR piglets OR pigs OR porcine OR primate OR primates OR quail OR rabbit OR rabbits OR rat OR rats OR reptile OR reptiles OR rodent OR rodents OR ruminant OR ruminants OR salmon OR sheep OR shrimp OR slug OR slugs OR swine OR tamarin OR tamarins OR toad OR toads OR trout OR urchin OR urchins OR vole OR voles OR waxworm OR waxworms OR worm OR worms OR xenopus OR "zebra fish" OR zebrafish) NOT (human OR humans or patient or patients))
3. 1 NOT 2
4. PMID=(0* or 1* or 2* or 3* or 4* or 5* or 6* or 7* or 8* or 9*)
5. 3 NOT 4

TABLE A1. mCODE Groups and Profiles

mCODE Group	mCODE Profile
Disease	Primary cancer condition
Disease	Secondary cancer condition
Disease	TNM clinical distant metastases category
Disease	TNM clinical primary tumor category
Disease	TNM clinical regional nodes category
Disease	TNM clinical stage group
Disease	TNM pathologic distant metastases category
Disease	TNM pathologic primary tumor category
Disease	TNM pathologic regional nodes category
Disease	TNM pathologic stage group
Genomics	Genetic specimen
Genomics	Genomic region studied
Genomics	Cancer genetic variant
Genomics	Cancer genomics report
Laboratory/vital	Tumor marker
Outcome	Cancer disease status
Patient	Comorbid condition
Patient	ECOG performance status
Patient	Karnofsky performance status
Patient	Cancer patient
Treatment	Cancer-related radiation procedure
Treatment	Cancer-related surgical procedure
Treatment	Cancer-related medication statement

Abbreviations: ECOG, Eastern Cooperative Oncology Group; mCODE, Minimal Common Oncology Data Elements.

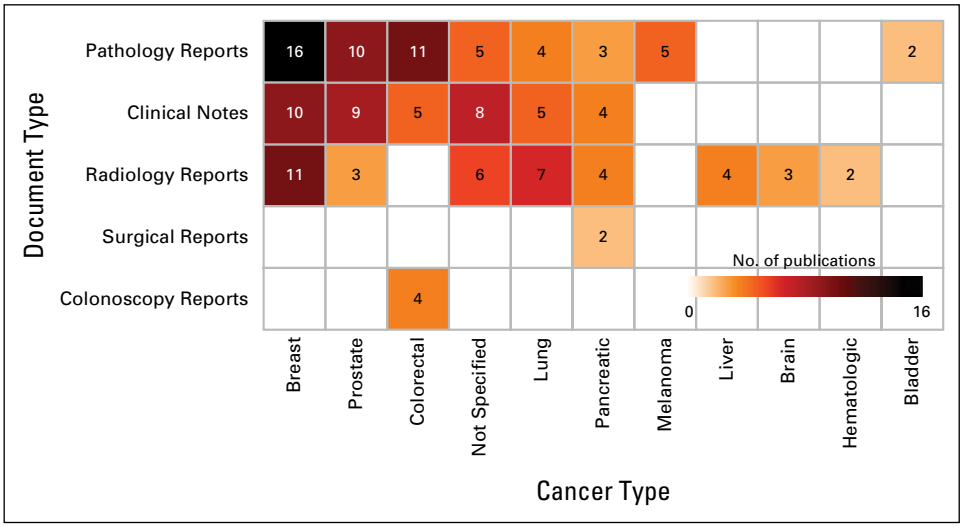

FIG A1. A heatmap of document types and targeted cancer types. Those with < 2 publications not shown.

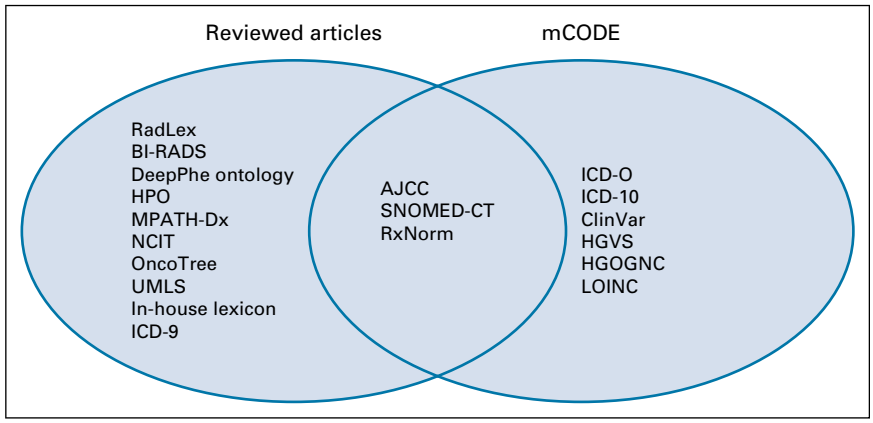

FIG A2. Comparison of standardized terminologies for data elements between the reviewed articles and mCODE. AJCC, American Joint Committee on Cancer; BI-RADS, Breast Imaging Reporting and Data System; HGOGNC, Human Genome Organization Gene Nomenclature Committee; HGVS, Human Genome Variation Society; HPO, Human Phenotype Ontology; ICD-9, International Classification of Diseases (9th revision); ICD-10, International Classification of Diseases (10th revision); ICD-O, International Classification of Diseases for Oncology; LOINC, Logical Observation Identifiers Names and Codes; MPATH-Dx, Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis; NCIT, National Cancer Institute Thesaurus; RadLex, Radiology Lexicon; RxNorm, no full name; SNOMED-CT, SNOMED Clinical Terms; UMLS, Unified Medical Language System.

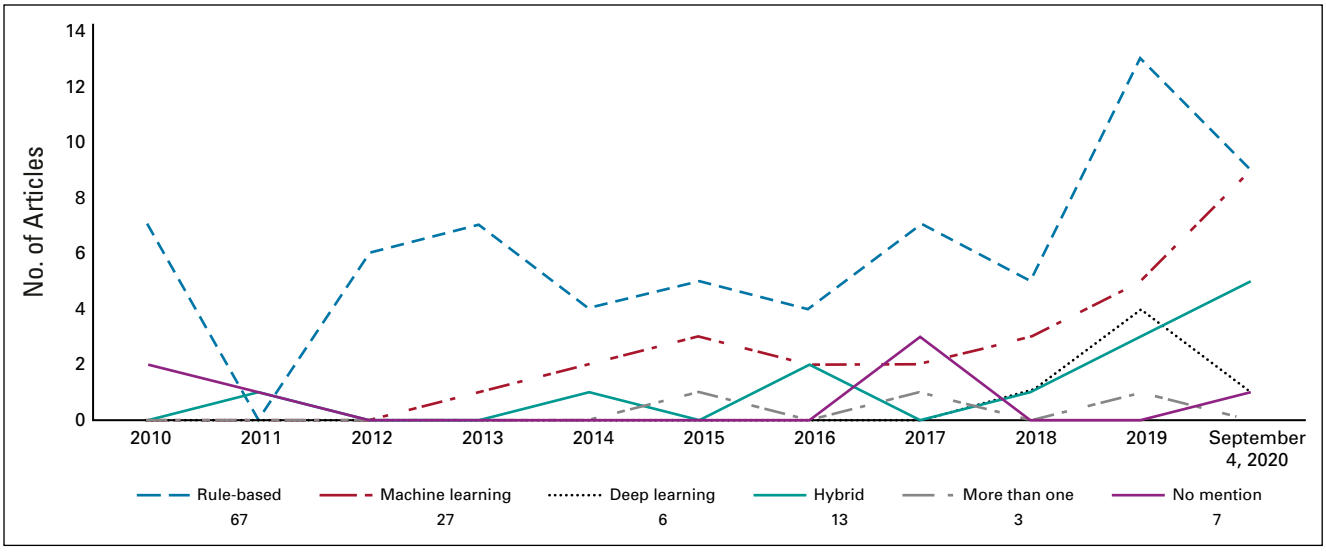

FIG A3. Analysis of NLP methods. NLP, natural language processing.