Journal of the American Medical Informatics Association, 28(12), 2021, 2716–2727
doi: 10.1093/jamia/ocab170
Advance Access Publication Date: 6 October 2021
Review

Review

Extracting social determinants of health from electronic health records using natural language processing: a systematic review

Braja G. Patra [1], Mohit M. Sharma [1], Veer Vekaria [1], Prakash Adekkanattu,[2] Olga V. Patterson [3,4], Benjamin Glicksberg [5], Lauren A. Lepow,[5] Euijung Ryu,[6] Joanna M. Biernacka,[6] Al'ona Furmanchuk,[7] Thomas J. George [8], William Hogan [9], Yonghui Wu,[8] Xi Yang,[8] Jiang Bian [8], Myrna Weissman,[10] Priya Wickramaratne,[10] J. John Mann,[10] Mark Olfson,[10] Thomas R. Campion Jr, [1,2] Mark Weiner [1], and Jyotishman Pathak [1]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA, [2]Information Technologies and Services, Weill Cornell Medicine, New York, New York, USA, [3]Department of Internal Medicine, Division of Epidemiology, University of Utah, Salt Lake City, Utah, USA, [4]US Department of Veterans Affairs, Salt Lake City, Utah, USA, [5]Icahn School of Medicine at Mount Sinai, New York, New York, USA, [6]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, USA, [7]Northwestern University, Chicago, Illinois, USA, [8]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA, [9]Division of Hematology & Oncology, Department of Medicine, College of Medicine, University of Florida, Gainesville, Florida, USA, and [10]Vagelos College of Physicians and Surgeons, Columbia University, New York, New York, USA

Corresponding Author: Jyotishman Pathak, PhD, Department of Population Health Sciences, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065, USA (jyp2001@med.cornell.edu)

Received 29 April 2021; Revised 9 July 2021; Editorial Decision 29 July 2021; Accepted 4 August 2021

ABSTRACT

Objective: Social determinants of health (SDoH) are nonclinical dispositions that impact patient health risks and clinical outcomes. Leveraging SDoH in clinical decision-making can potentially improve diagnosis, treatment planning, and patient outcomes. Despite increased interest in capturing SDoH in electronic health records (EHRs), such information is typically locked in unstructured clinical notes. Natural language processing (NLP) is the key technology to extract SDoH information from clinical text and expand its utility in patient care and research. This article presents a systematic review of the state-of-the-art NLP approaches and tools that focus on identifying and extracting SDoH data from unstructured clinical text in EHRs.

Materials and Methods: A broad literature search was conducted in February 2021 using 3 scholarly databases (ACL Anthology, PubMed, and Scopus) following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. A total of 6402 publications were initially identified, and after applying the study inclusion criteria, 82 publications were selected for the final review.

Results: Smoking status (n = 27), substance use (n = 21), homelessness (n = 20), and alcohol use (n = 15) are the most frequently studied SDoH categories. Homelessness (n = 7) and other less-studied SDoH (eg, education, financial problems, social isolation and support, family problems) are mostly identified using rule-based approaches. In contrast, machine learning approaches are popular for identifying smoking status (n = 13), substance use (n = 9), and alcohol use (n = 9).

Conclusion: NLP offers significant potential to extract SDoH data from narrative clinical notes, which in turn can aid in the development of screening tools, risk prediction models, and clinical decision support systems.

Key words: social determinants of health, population health outcomes, electronic health records, natural language processing, information extraction, machine learning

INTRODUCTION

Social determinants of health (SDoH) are circumstances in which people are born, live, learn, work, and age and are closely tied to individuals' health behaviors, lifestyle, and interpersonal relations. The upstream distribution of wealth, power, and resources at local, national, and global levels can trickle down to impact individual health outcomes and potentially lead to health disparities (https://www.who.int/gender-equity-rights/understanding/sdh-definition/en/). Several studies have investigated associations of SDoH with differential health outcomes, such as the effects of food insecurity on developing diabetes[1]; socioeconomic status, neighborhood, employment, race, and social support on breast cancer risk and survival[2]; and housing quality on mental health.[3] Not surprisingly, risky health behaviors and maldistribution of SDoH have been associated with increased financial burdens on patients and providers.[4]

The annual County Health Rankings (https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-healthrankings-model) gauge the impact of a wide range of health factors by ranking the health outcomes of the over 3000 counties across the United States. Figure 1 theorizes how upstream improvements in policies and programs impact health factors and ultimately ripple into downstream community health outcomes.[5] Social, economic, and physical environment factors contribute the most to health outcomes (50%), followed by health behaviors (30%). Just 20% of health outcomes are attributed to clinical care. Case in point, according to the Centers for Disease Control and Prevention (CDC), 39% of deaths from chronic lower respiratory disease resulted from social and environmental exposure to secondhand smoke, allergens, occupational agents, and other indoor and outdoor air pollutants. To a lesser degree, 33% of premature stroke deaths were attributed to risky health behaviors (tobacco use, alcohol use, and sedentary lifestyles) and their related clinical manifestations—high blood pressure, high cholesterol, heart disease, diabetes, obesity, and previous stroke (https://www.cdc.gov/media/releases/2014/p0501-preventable-deaths.html).

Such evidence suggests a strong association between nonclinical factors with clinical outcomes and has increased clinical and public health interest in incorporating SDoH into patient profiles on a much broader scale. Collecting and understanding SDoH information offers significant potential and can uncover important contextual information about patients' lifestyles to supplement clinical findings. Most US health systems and providers use electronic health records (EHRs) to document patient clinical information. In the last decade, adoption of EHRs has widely expanded, however qualitative information about patients' lifestyles is usually documented in unstructured clinical notes. Although SDoH information is often collected, the lack of standardized data elements, assessment tools, measurable inputs, and data collection practices in clinical notes greatly limits the access to this information. Attempts to improve standardization at the national level have been made by the Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE) (https://www.cdc.gov/media/releases/2014/p0501-preventable-deaths.html) and the National Academy of Medicine (https://nam.edu/social-determinants-of-health-101-for-health-care-five-plus-five/). Furthermore, the International Classification of Diseases, Ninth Revision (ICD-9) and Tenth Revision (ICD-10) V-codes (V-60-62) and Z-codes (Z55–Z65) have been implemented for diagnostic use; however, Truong et al[6] noted that these codes are ex-

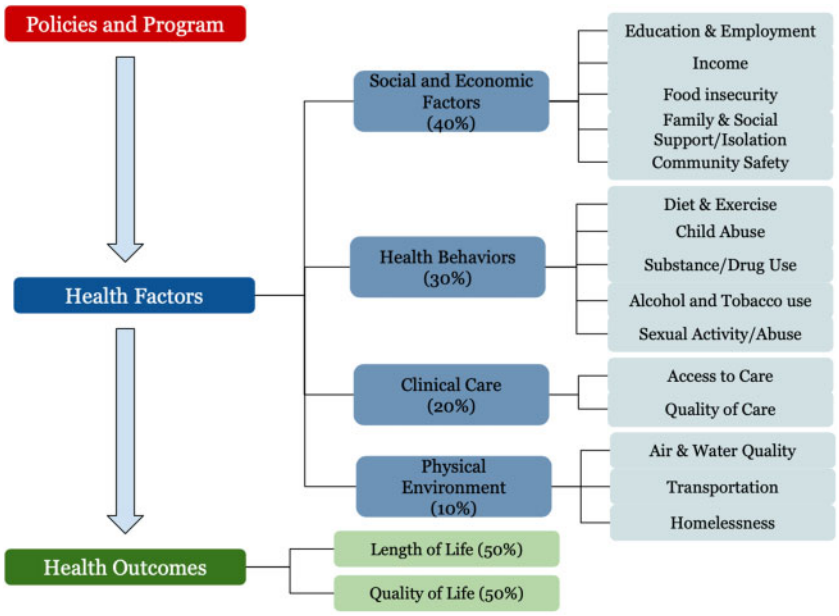

Figure 1. The County Health Rankings model of population health.

tremely underutilized, with less than 2% of inpatient hospitalizations having a Z-coded diagnosis during the first 2 years of their availability. A more comprehensive, coherent, and user-friendly set of codes could be developed to improve SDoH documentation and increase rates of adoption in healthcare settings.

Traditionally, extracting valuable information from unstructured data is performed manually through chart review, which can be time-consuming. However, recent advances in natural language processing (NLP) offer more efficient, automated approaches to unlock and analyze insightful information from existing EHR data. Currently, about 80% of medical data are unstructured and do not fit into easily actionable categories, including clinician encounter notes, discharge summaries, patient-reported information, and radiology/pathology reports,[7] but they can be indexed and leveraged to guide more informed clinical decision-making. Various clinical decision support systems are being developed using EHRs, and many healthcare organizations are allocating substantial resources to support the integration of NLP technologies in an effort to expand the amount of usable data, enhance analytic insights, and improve patient outcomes.[8,9]

Recent studies specified the scope and importance of NLP or information retrieval (IR) methods to extract SDoH information from clinical notes.[10–12] However, there are several different designs of SDoH identification/extraction tools in recent literature, and the utility of each tool depends largely on the type of SDoH in question. Furthermore, the lack of a comprehensive review that delineates the tools available and their most suitable purposes may hinder efficient progress on this research problem in terms of deciphering what has and has not been explored. In this review, we investigated various NLP techniques for SDoH lexicon curation and implementation of SDoH extraction systems that have been developed for extracting SDoH data from unstructured clinical notes in EHR systems. We list EHR systems and categories of SDoH concepts extracted using NLP techniques and tools.

We also identified 2 relevant works that study the implementation of SDoH in EHR databases.[13,14] Chen et al[13] studied the integration of SDoH in EHRs, their impact on risk prediction, and the specific outcomes. In addition, Bompelli et al[14] studied artificial intelligence (AI) methods to extract SDoH from EHRs. They briefly discussed different NLP methods used for identifying SDoH from EHRs and surveyed the papers that studied the healthcare outcomes using SDoH. These studies, while informative, do not describe the details of NLP methodologies used for SDoH from clinical text which is the major focus of this systematic review .

MATERIALS AND METHODS

Searching methods and screening

An initial literature survey was conducted to identify SDoH-related keywords that could be used for searching relevant publications. We also identified keywords related to multiple categories of SDoH from the County Health Rankings model. A total of 73 SDoH-related keywords were identified in addition to variants of "natural language processing" and "electronic health records" (see Supplementary Table S1).

We searched 3 scholarly databases—ACL Anthology, PubMed, and Scopus—between October 2020 and February 2021 with the goal of identifying all relevant articles related to SDoH extraction from EHRs using NLP methods. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[15] For PubMed and Scopus, we created queries for each SDoH keyword. A total of 3301 publications were identified from PubMed (n = 2672) and Scopus (n = 629) using our search strategy. Search-related observations are reported in the Supplementary File. After removing duplicates, we screened 1874 publications from PubMed and Scopus, during which 1746 publications were further excluded as not relevant, or not having the full text available. In the case of ACL Anthology, 3101 publications were identified. Although the search results could not be saved, all 3101 publications were screened during this stage. The screening resulted in 128 articles from PubMed and Scopus, and 11 articles from ACL Anthology for the full text review. The PRISMA workflow is shown in Figure 2 and example queries for each database are provided in Supplementary Table S2.

Full text review

To be eligible for inclusion in the review, the articles needed to focus on the description, evaluation, or use of NLP or text mining algorithm/pipeline to identify or extract SDoH information from EHRs. Two authors (BGP and MMS) independently reviewed each full text article and discussed to reach consensus if there were any discrepancies. There were several reasons for exclusion, for example, if the article was not relevant to NLP, EHRs, or any SDoH categories that we intended to study. SDoH categories that were not evaluated in the County Health Rankings (such as race, ethnicity, mental health conditions, or stress) were not included in our study. Other reasons for exclusion included if the SDoH extraction was performed on non-EHR data (eg, survey results, homelessness assistance program applications, or social media data), or if the publication involved non-English EHR systems. Finally, we excluded studies that were not peer-reviewed articles, such as abstracts and commentary, perspective, or opinion pieces.

Data

A final total of 82 publications were included in this study at the end of the full text review process. We also searched references of all 82 articles, and most of these publications overlap with our search results. The included articles were published between 2005 and 2021, as illustrated in Figure 3. Authors BGP and MMS performed the data extraction and analysis of all 82 publications.

RESULTS

We list SDoH categories studied in the collected publications and describe the steps associated with extraction of SDoH in EHRs. In general, we observed 2 major steps associated with SDoH extraction systems from literature. The first step is gathering SDoH-related keywords to create lexicons for each SDoH category, and the second step is developing rule-based or supervised systems to locate clinical notes associated with SDoH categories or extract SDoH concepts. The SDoH lexicons can be created using manual chart review of clinical notes/medical dictionaries or using semisupervised or supervised algorithms. Many keyword/lexicon matching, supervised (classification using semantic and syntactic features), and unsupervised (topic modeling) approaches were developed in the past for SDoH identification.

SDoH categories

Figure 4a shows counts of publications that studied different SDoH categories, whereas Figure 4b presents the combinations of SDoH

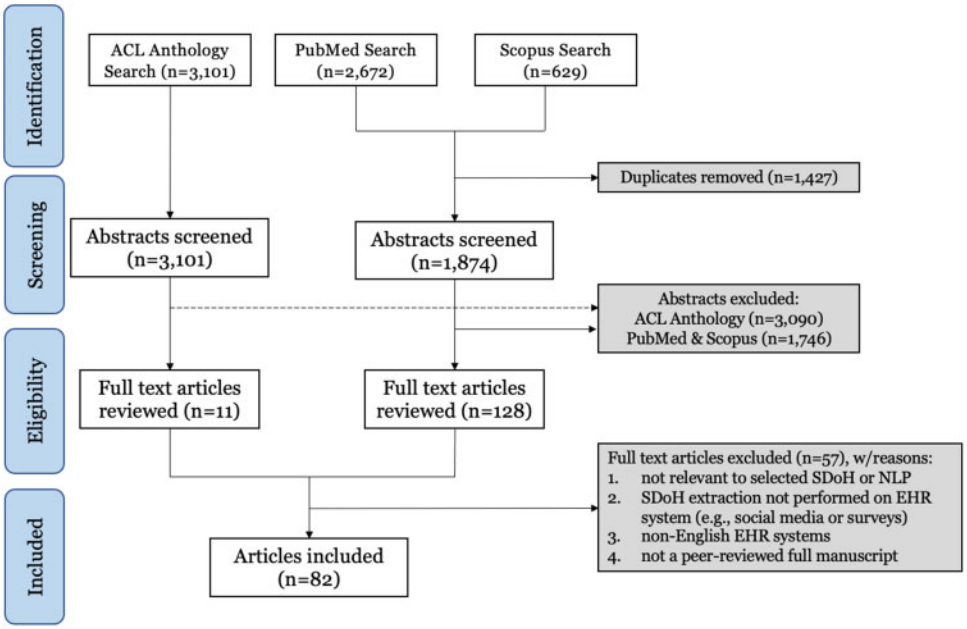

Figure 2. PRISMA workflow of included articles.

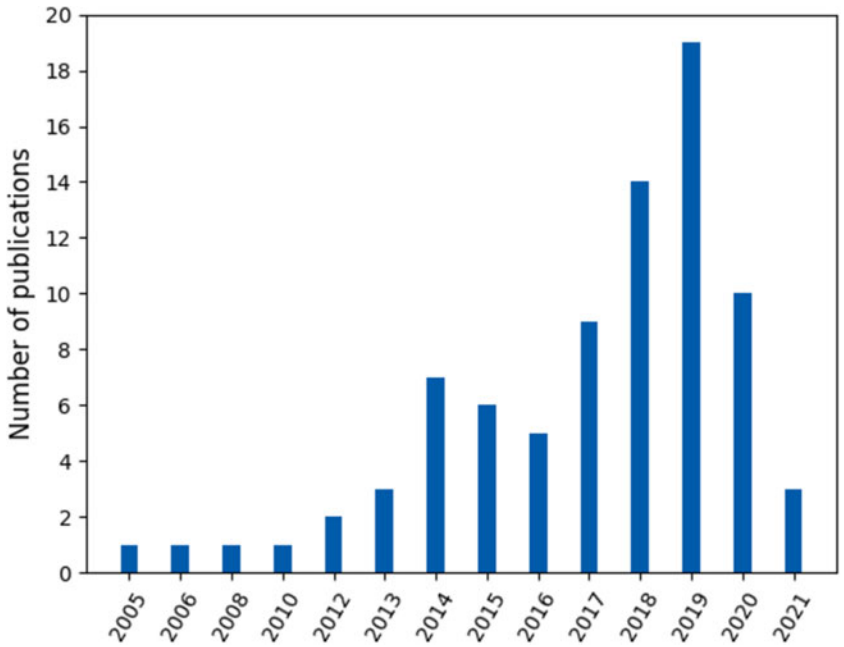

Figure 3. Publication years of included articles.

categories that were studied in publications. Factors related to smoking status are some of the most studied SDoH using NLP algorithms, followed by drug abuse and homelessness. Figure 4c shows historical counts of different SDoH categories that were extracted using NLP methods (here, we grouped the SDoH factors into 4 categories). NLP studies in health behaviors focused primarily on extracting smoking status in the early 2000s. The first set of studies in extracting physical environment (eg, housing issues/homelessness) and social and economic factors were conducted in 2012 and 2013, respectively.

SDoH lexicon development

Manual lexicon curation

Initial approaches to SDoH lexicon development were based on manually reviewing the literature and filtering clinical notes with the help of domain experts.[10,11,16–19] Figure 5a highlights the frequencies of different techniques to develop lexicons and suggests that manual lexicon development techniques were more commonly employed than semiautomated techniques. This finding could be explained by data that are too noisy or limited in availability to support development of semiautomated approaches. SDoH catego-

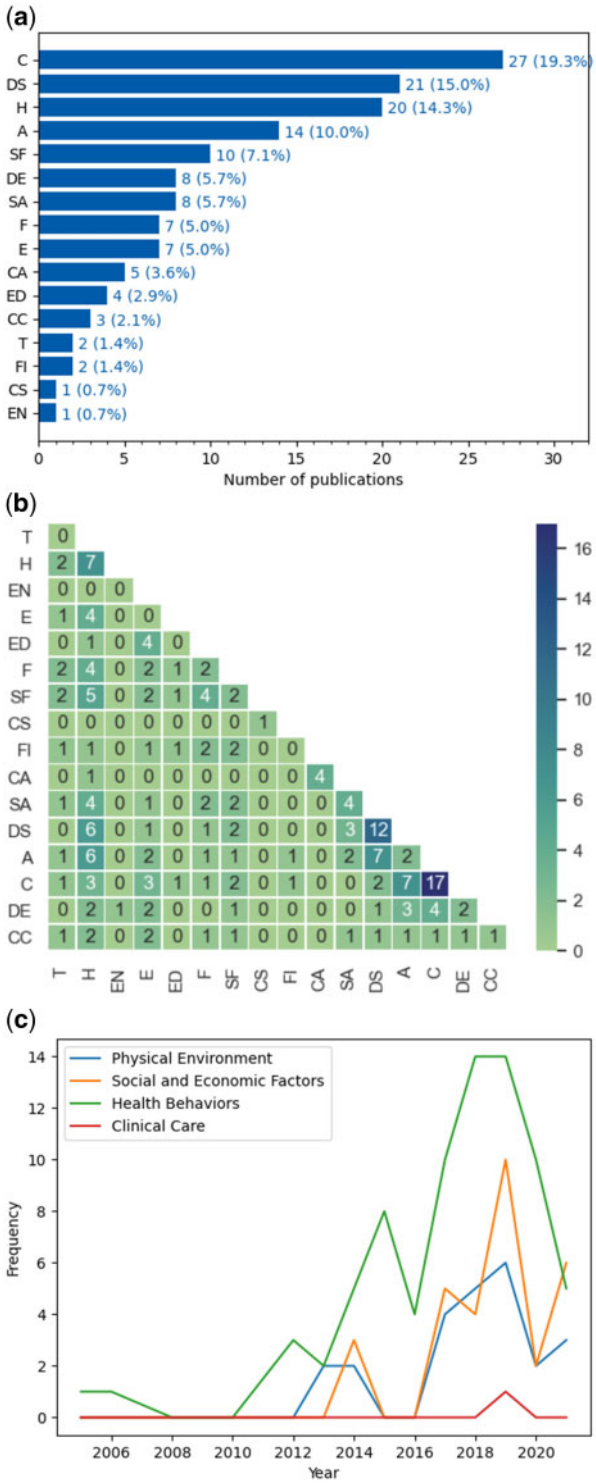

Figure 4. (a) Frequency of SDoH categories in the collected publications. (b) Heatmap of different SDoH categories combinations implemented in publications. (c) Year-wise frequencies of SDoH categories that were extracted using NLP.

Abbreviations: A, alcohol abuse/use; C, cigarettes/smoking status; CA, child abuse/adverse childhood experiences; CC, clinical care (access to care/quality of care); CS, community safety; DE, diet & exercise; DS, drug/substance abuse; E, employment; ED, education; EN, environment (water/air quality); F, financial/income issues; FI, food insecurity; H, housing issues; SA, sexual activity/abuse; SF, social connection/isolation or family problem; T, Transportation.

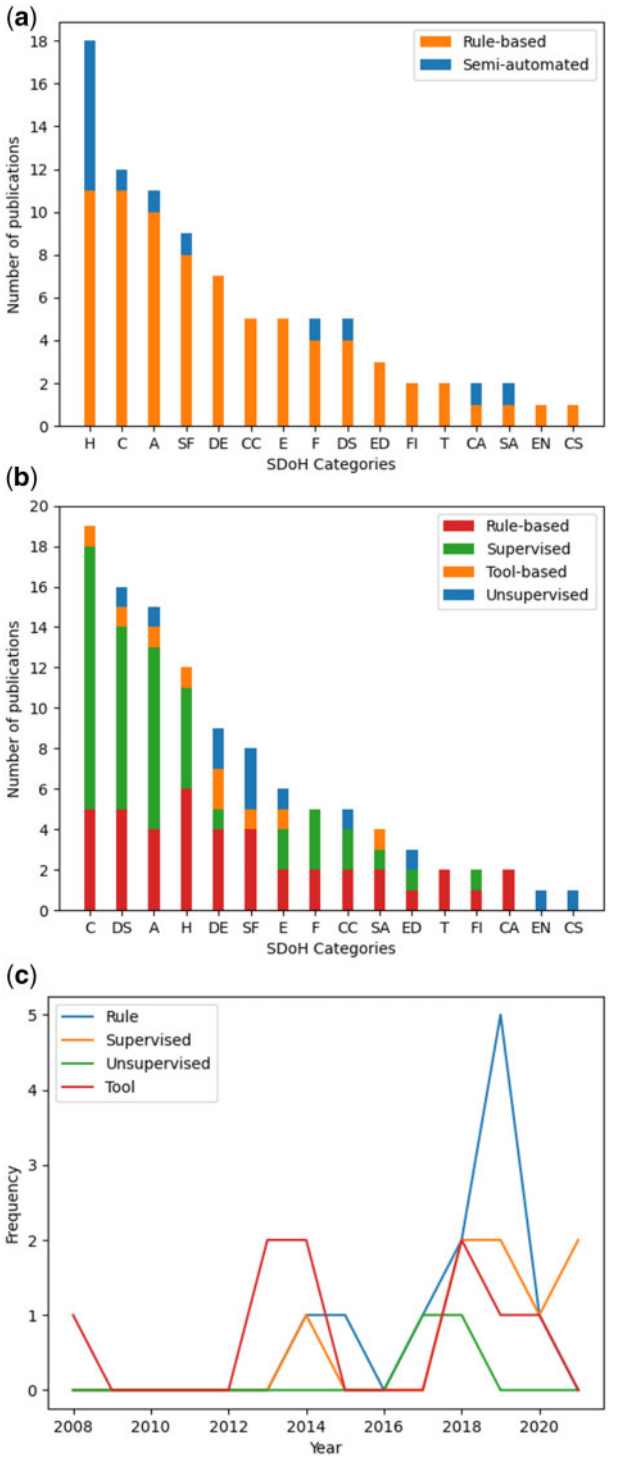

Figure 5. (a) Frequencies of rule-based and semiautomated methods for SDoH lexicon creation. (b) Frequencies of existing tools and systems (rule-based, supervised, and unsupervised) for SDoH identification/extraction. (c) Year-wise frequencies of different NLP methods that were used to extract different SDoH categories.

Abbreviations: A, alcohol abuse/use; C, cigarettes/smoking status; CA, child abuse/adverse childhood experiences; CC, clinical care (access to care/quality of care); CS, community safety; DE, diet & exercise; DS, drug/substance abuse; E, employment; ED, education; EN, environment (water/air quality); F, financial/income issues; FI, food insecurity; H, housing issues; SA, sexual activity/abuse; SF, social connection/isolation or family problem; T, Transportation.

ries that used manual lexicon development techniques are listed in Supplementary Table S4.

Gundlapalli et al[18] created a homelessness lexicon after manual chart review and used the National Library of Medicine (NLM)'s lexical generation tool for extending lexicons. Greenwald et al[20] created lexicons for housing, financial issues, substance use, social and psychosocial support, substance abuse, and physical abuse from multiple focus group conversations with front-line clinical staffs (nurses, case managers, physical therapists, occupational therapists, physicians, and social workers). Hatef et al[10] created a homelessness lexicon using various public health surveys and instruments such as the American Community Survey (ACS); the American Housing Survey (AHS); the Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE); and the Accountable Health Communities Model from the Center for Medicare and Medicaid Innovation. Hatef et al[10] also used ICD-10 codes, Current Procedural Terminology (CPT) codes, Logical Observation Identifiers Names and Codes (LOINC), and Systematized NOmenclature of MEDicine (SNOMED) codes to create lexicons for social connection/isolation, housing issues, and income/financial resource strain. Blosnich et al[21] used ICD-10 codes, Veterans Health Administration (VHA) stop codes, and VHA health factors to create lexicons for housing issues, financial issues, adult violence, and military sexual trauma. Winden et al[22] initially reviewed SNOMED codes for homelessness keywords and then used these to identify additional terms in the EHR flowsheet.

Semiautomated lexicon creation

Figure 5a suggests that fewer efforts were made to create lexicons automatically. Bejan et al[12] proposed a semisupervised method to develop SDoH lexicons related to housing and adverse childhood experiences. Their work used the seed words (single or multi-words) at first, followed by lexical association and word2vec in clinical notes to find similar SDoH keywords to further increase the inventory of lexicons. The initial set of lexicons were modified iteratively using clinical notes retrieved by similarity calculation. In a similar approach, Bettencourt-Silva et al[23] used the skip-gram model and word embedding on EHRs and Wikipedia. They collected relevant keywords related to 5 SDoH categories (housing, criminal justice, financial services, health behaviors, and homelessness) using semisupervised methods. Initially, they provided 57 related terms to a word2vec model trained on Wikipedia pages and then collected 2000 similar words as candidate lexicons for further annotation by domain experts.

Topaz et al[24] used word2vec to expand the seed word list for alcohol use and substance-related disorders and then manually reviewed this list to generate the final lexicon. Velupillai et al[25] used WordNet and word2vec to develop a substance use lexicon. Conway et al[9] collected a lexical dictionary of words and phrases from the Unified Medical Language System (UMLS) metathesaurus. Each SDoH category involving semiautomated lexicon development techniques and the corresponding citations are shown in Supplementary Table S4.

SDoH identification/extraction

Most studies used rule-based (keyword searching/keyword similarity), traditional supervised machine learning (ML), and deep learning (DL) approaches to identify SDoH at the keyword or document level. Many SDoH identification systems leveraged existing NLP systems, terminologies, and infrastructures. Figure 5b shows the frequencies of existing tools and systems (using rules, supervised, and unsupervised methods) for identifying SDoH categories. Different techniques for SDoH extraction for each category and the corresponding citations are provided in Supplementary Table S4.

Rule-based methods

In total, 22 out of 82 publications (about 27%) used rule-based methods to identify SDoH in clinical notes. (Note: There may be multiple SDoH/methods in a single publication. However, we considered each SDoH/method separately while counting the numbers.) Figure 5b shows how frequently different methods were used to identify SDoH categories.

Rule-based systems were used more frequently for housing, transport, and social isolation and less frequently for smoking, alcohol, and substance use. Insufficient volumes of annotated or structured data for homelessness, social support, and other socioeconomic factors may explain why rule-based systems were more common for these variables. The rule-based systems were developed either using keyword matching/counts or regular expression or similarity matching to identify the presence of SDoH in any clinical document.[10,12,20–22,26–28] We provided the corresponding citations for SDoH extraction methods in Supplementary Table S5. Figure 5c shows frequencies of different NLP methods that were used to extract SDoH data.

Supervised methods

In supervised classification approaches, the features can be broadly classified into 2 categories. First, a category of features that were based on developed lexicon/keywords and the others that were based on embeddings. A wide range of embeddings (bag-of-words/term frequency and inverse document frequency [TF-IDF], n-grams, and word2vec) served as features for classifying SDoH categories.[29–35] In addition to these 2 feature categories, several other features such as part-of-speech (POS), POS unigram, POS bigrams, lexicon counts, and concepts (identified using UMLS, MetaMap, and the clinical Text Analysis and Knowledge Extraction System [cTAKES]) were also used for classification in several studies.[36,37] Our investigation found that diagnosis codes from ICD-9 and ICD-10 and concepts identified using cTAKES and UMLS, were extensively used as features for substance use, alcohol use, and smoking status.[17,26,38,39]

In 24 out of 82 publications (about 32%), classification methods were used for identifying SDoH. Several studies that identified homelessness (n = 5), alcohol use (n = 9), substance use (n = 9), and tobacco/smoking (n = 13) used supervised techniques. Support Vector Machine (SVM), random forests, and logistic regression classifiers were commonly used for classification tasks.[24,29,30,32–35,37,38,40–44]

Furthermore, we observed 7 studies that investigated deep learning algorithms for SDoH extraction.[29,32,33,42,45–47] Convolutional neural network (CNN) and feedforward neural network (FNN) performed poorly compared to traditional ML algorithms, such as SVM, random forests, Classification and Regression Tree (CaRT), and AdaBoost due to scarcity of annotated data for identifying SDoH such as alcohol abuse, substance abuse, homelessness, and sexual orientation.[33] In more recent work, Lybarger et al[42] used Bidirectional Encoder Representations from Transformers (BERT) for embedding, Bidirectional Long Short-Term Memory (BiLSTM) for trigger and label identification, and Conditional Random Field (CRF) for SDoH span identification. This study identified living status and employment status, as well as drug, alcohol, and tobacco

use. A comprehensive list of traditional ML and DL algorithms used in publications with their citations is provided in Supplementary Table S6.

Feature selection plays an important role in improving the performance of supervised systems, and we observed these methods in SDoH identification. Wang et al[35] used latent Dirichlet allocation (LDA) and information gain feature selection techniques to identify drug, alcohol, and nicotine use. Regression models and random forest classifiers were used to evaluate the importance of features on different tasks.[24,34,38,45] Feller et al[32] used chi-squared goodness of fit tests to perform feature selection for housing status, sexual history, substance use, alcohol use, sexual orientation, and gender documentation identification.

Unsupervised analysis

Three studies investigated unsupervised approaches for SDoH extraction from clinical notes. Lindemann et al[48] used topic modeling to perform a detailed analysis of social history topic variation, which was followed by validation through a separate manual analysis of 1400 clinical documents from Fairview Health Services (FHS). Similarly, Afshar et al[49] employed LDA to detect the subtypes of opioid misuse. In this study, latent class analysis (LCA) was used to cluster documents into 20 categories. The authors concluded that the 4-class latent model was the most parsimonious model to define clinically relevant subtypes for opioid misuse. In another related study, Wang et al[50] used LDA to identify topics for nutrition (swallow function, artificial feeding, and nutritional status) and social support (family support, spiritual support, caregiver support, and social history) in patients with Alzheimer's disease and related dementias (ADRD).

Other methods

Multiple publications used previously developed NLP systems, terminologies, and infrastructure to accomplish SDoH extraction/identification tasks. Table 1 lists experiments that used NLP tools for different SDoH extraction, and most of these systems were developed using lexicons and rules.[9,17,51,63] MTERMS was also used for deidentification in addition to SDoH identification.[50]

Evaluation

One of the steps in the development of NLP systems is the manual review of relevant documents or, similarly, the creation of a gold standard through chart review. UMLS, SNOMED, LOINC, CPT, ICD-9, and ICD-10 codes were used for both lexicon development and NLP systems evaluation.[10,17,21,38,55] Many supervised NLP systems used these codes to identify notes for training and testing (eg, if a patient has an ICD-10 Z-code for homelessness, then their notes were used for training/testing homelessness models). ICD-9 and ICD-10 codes, VHA stop codes, and VHA health factors in EHR systems were used to validate systems for homelessness.[21,40] Several publications described chart review of clinical notes for evaluation.[48,60,64] In addition, Wang et al[29] used weak labeling to generate labels for raw data. The authors manually reviewed the initial labeled data from a rule-based system and developed a supervised ML system using embeddings to identify unseen notes containing smoking status.

DISCUSSION

The majority of lexicons were created using manual curation. The manual curation requires more efforts to develop and manually evaluate a lexicon, however, it helps to understand the characteristics of EHRs. One of the advantages of using semiautomated techniques is that these greatly reduce the human efforts and resources needed to create lexicons.

Keyword matching and classification were the 2 most common techniques in the literature used to identify notes with SDoH. Selecting a technique depends on the availability of an SDoH annotated gold standard corpus. Researchers used a keyword matching-based method if there was no gold standard corpus, followed by manual evaluation of the matching algorithms' performances. Researchers used supervised techniques if a gold standard SDoH annotated dataset was available. Furthermore, concepts identified using UMLS, cTAKES or any dictionaries were used as features for identifying behavioral determinants. It can be observed that these terminology dictionaries are more developed for behavioral determinants than others.

Table 1. Tools used for SDoH identifications and the corresponding citations

NLP systems, terminologies, and infrastructure	Task (citations)
cTAKES	alcohol use status, tobacco cessation, diet and exercise,[17] opioid misuse[49]
Moonstone NLP	housing and social issues,[9] lifestyle modification,[17] lived alone, marginal housing, alcohol use, substance use[51]
ARC	homelessness,[16] adverse childhood experiences[52]
V3NLP	homelessness,[18,19] sexual trauma[53,54]
MediClass	opioid related overdose,[38,39] opioid use[55]
I2E	social isolation[56]
UMLS	lifestyle modification[17]
HITEx	smoking status[57,58]
MTERMS	homelessness, social support, and drug abuse[59]
MedTagger and MedTime	smoking status[60]
VINCI	MST[8]
VISA	adverse childhood experiences[52]
CRIS-IE	smoking[61]
TextHunter	cannabis use,[43] neighborhood characteristics, and physical violence[62]

Abbreviations: ARC, automated retrieval console; CRIS, clinical record interactive search; cTAKES, clinical Text Analysis and Knowledge Extraction System; HITEx, health information text extraction; MST, military sexual trauma ; VISA, veterans indexed search for analytics; VINCI, VA Informatics and Computing Infrastructure.

Many studies used the NegEx and ConText algorithms to identify negations, experiences, and temporal status in the clinical notes and reported improved accuracy.[12,19,32,33,37,47,59,65] BRAT (https://brat.nlplab.org) was used for annotating SDoH in clinical texts.[52,66] Knowtator was also used for annotation in VA EHRs for annotating homelessness,[18] education, and employment.[67] From the publications included in this study, we compiled all SDoH classes extracted from EHR systems using NLP and summarized this information in Table S3. We included papers with at least 1 SDoH category in our study. We mentioned additional categories in the "Other" column of Table S3. We collected a list of EHR data sources and associated papers from the literature, and these are listed in Table 2. Many experiments on SDoH identification were carried out on VHA patient populations in the United States and are mostly related to homelessness and military sexual trauma.

Potential benefits may emerge in the healthcare delivery landscape from integrating SDoH extraction tools to EHR systems. In clinical settings, providers report spending less time on patient care and more time on administrative burdens that are byproducts of data management in the EHR.[86] Manual screening of SDoH could potentially further complicate and delay the process for healthcare staff. Furthermore, as SDoH categories grow in number and complexity, storing SDoH in a structured framework could potentially become inefficient and require frequent maintenance. In light of these scenarios, we believe that the NLP-based SDoH identification and the developed outcome analysis tools may offer an optimal solution that may minimize impact on current documentation routines while guiding providers to make better, informed and holistic clinical decisions.

SDoH outcome analysis

Many publications included in this survey focused on developing decision/intervention systems for SDoH categories. The outcome analysis helps in the diagnosis, treatment, and clinical outcomes. We

Table 2. EHR data sources used for SDoH experiments. Here * represents different databases from the same source

Datasets	Citations
100 synthetic data sets using Monte Carlo methods	68
Academic Health Center Information Exchange (AHC-IE), Academic Health Center (AHC)	22
Brigham and Women's Hospital or Massachusetts General Hospital	57,58
Centre Clinical Record Interactive Search (CRIS)	62
Cerner Corporation, Kansas, MO	69,70
Child Health Department Netherlands	30
Columbia University Irving Medical Center (CUIMC), Columbia University Medical Center (CUMC)	32,33
Epic EHR systems*	17,71–73
Fairview Health System	22,48
four HMOs	74
Group Health, Washington State	11,55
Informatics for Integrating Biology and the Bedside (i2b2) smoking database	29,36,46,75,76
Kaiser Permanente*	4,38,39,77
Level I Trauma Center	78
Loyola University Medical Center	45
Marshfield Clinic's Enterprise Data Warehouse (MC-EDW)*	34,79
Mayo Clinic	60
Medical University of South Carolina (MUSC) Research Data Warehouse	56
Midwestern academic medical center	80
MIMIC-II	24,25
MIMIC-III	42,81
Minnesota Disability Determination Services	37
MTSamples	35,66
Multilevel academic health care system	10
National Homeless Registry	40
Loyola University Medical Center	41
Partners Healthcare System	59
SLaM Case Register	61
South London and Maudsley (SLaM) Biomedical Research Centre (BRC) Case Register	43
State child welfare agencies	31
UK Clinical Record Interactive Search (UK-CRIS)	82
University of Pittsburgh Medical Center (UPMC)	66
University of Minnesota*	22,35,66
University of Vermont Medical Center (UVMMC)	35
University Hospital, University of Utah, Salt Lake City	25
University of Massachusetts Memorial Health Care	44
University of Utah Health Sciences Center	83
University of Washington (UW) and Harborview Medical Centers	42
Urban tertiary academic center 18	49
US academic health system	84
Vanderbilt HER, Vanderbilt Synthetic Derivative, Vanderbilt University Medical Center (VUMC)	12,63,76
VeteransHealthAdministration, VA'sCorporate Data Warehouse (CDW)*	8,9,16,18,19,21,26,27,40,52–54,65,68,85

Table 3. SDoH classes and the corresponding healthcare outcomes

SDoH categories	Outcome
Transportation	Multiple social and behavioral factors,[10] suicide[21]
Housing issues	30-day readmission,[20] suicide,[21] acute myocardial infarction (AMI), mental and behavioral disorders and multiple SDOH,[47] heart failure (HF), or pneumonia[51]
Environment (water/air quality)	Food and drug allergies[63]
Employment	Suicide,[21] mental and behavioral disorders and multiple SDOH,[47] post-deployment rehabilitation (mild traumatic brain injury)[67]
Education	Mental and behavioral disorders and multiple SDOH,[47] postdeployment rehabilitation (mild traumatic brain injury)[67]
Financial issues/income	Multiple Social and Behavioral factors,[10] 30-day readmission,[20] suicide,[21] cost/financial considerations[84]
Social connection/isolation or family problem	Multiple social and behavioral factors,[10] 30-day readmission,[20] mental and behavioral disorders and multiple SDOH,[47] dementia,[50] cardiovascular diseases,[59] geriatric syndrome[87]
Community safety	Mental illness[62]
Food insecurity	Multiple social and behavioral factors,[10] mental and behavioral disorders and multiple SDOH[47]
Child abuse/adverse childhood experiences	Childhood abuse,[12,69,70] suicide[52]
Sexual activity/abuse	Suicide,[21] sexual trauma[54]
Drug/substance abuse	Chronic opioid therapy,[11] suicide attempt and depression,[38] hospitalization[49]
Alcohol abuse/use	Multiple social and behavioral factors,[10] suicide,[26] myocardial infarction (AMI), heart failure (HF), or pneumonia,[51] mental health/social and behavioral factors,[82] emergency admission[88]
Cigarettes/smoking status	Multiple social and behavioral factors,[10] asthma or chronic obstructive pulmonary disease (COPD),[57] mood disorders (depression, anxiety, or bipolar disorder),[58] smoking status,[60] tobacco use/smoking,[71,85] smoking behavior,[75] smoking status (including vape, electronic cigarette, pen),[77] tobacco use status[79]
Diet & exercise	dementia,[50] weight management[64]
Clinical care (access to care/quality of care)	Mental and behavioral disorders and multiple SDOH,[47] breast cancer[80]

found 33 publications performed SDoH identification only and 49 publications performed outcome analyses in addition to SDoH identification. Table 3 describes relationships between SDoH and corresponding health outcomes. Mental health (n = 12) is a notable outcome that is associated with all SDoH categories except environmental factors. Emergency hospitalization or readmission (n = 3) was another major SDoH associated with several SDoH categories, such as housing, financial issues, social connection/isolation or family problems, and drug and alcohol abuse.

Limitations

The findings of this review should be understood within the context of a few methodological limitations. First, the number of search results in the ACL Anthology was higher compared to PubMed and Scopus; however, the search resulted in many nonrelevant publications. Second, we chose to include existing NLP systems, terminologies, and infrastructures used for SDoH extraction, however a comprehensive study is required to capture details of these facets, and future review would benefit from further elaboration. Third, we did not include non-English EHR systems, however, it will be interesting to see the SDoH extraction in these EHRs. Despite these limitations, this study presents a compelling overview of the most recent and reliable NLP approaches that can be applied to identify SDoH in EHR systems.

Future work

A number of future directions can be derived from this literature review. Several studies were performed on identifying smoking, substance abuse, housing, and alcohol status in the EHR systems using NLP techniques. In the future, the NLP research community might focus on less-studied SDoH such as child and sexual abuse, financial issues, transportation, neighborhood, social isolation, family problems, employment, education, food insecurity, and healthcare access. Another interesting study would be to compare different aspects of NLP algorithms, such as system performance, amount of annotated data, type of NLP systems, and so forth with the difficulty of SDoH extraction.

Analyzing longitudinal aspects of the data may be helpful for developing outcome-based systems. Bejan et al[12] analyzed longitudinal data on homelessness and found that homelessness status changes with time. Temporal information (temporal information [current and past], amount/quantity/frequency, and type) extraction systems were developed for alcohol abuse, smoking status, and substance use/abuse.[35,55] However, temporal extraction was rare in homelessness (temporal)[16] and employment (status, duration, history, and type).[42] Also, Lybarger et al[42] identified the span of SDoH concepts related to living status type, employment status, drug, alcohol, and tobacco from clinical notes. Fewer experiments have identified the span of SDoH keywords in clinical texts. Furthermore, relation extraction/identifying relationships between medical concepts are semantic tasks that have been popular among the NLP community. Thus, developing comprehensive systems for 1) capturing longitudinal information; 2) extracting temporal information; 3) extracting SDoH concepts; and 4) extracting relations among SDoH concepts will be valuable future goals in this area of interest.

Next, the implementation of DL in clinical text has increased in the last decade, and DL-based NLP systems obtain better performances than other state-of-the-art NLP systems. Surprisingly, DL algorithms were rare for identifying SDoH. This may be attributed to insufficient amounts of annotated data whereas DL models necessitate large volumes of annotated data. More experiments implementing DL algorithms for SDoH identification and relation extraction can be performed in future.

Finally, NLP algorithms developed for SDoH extraction heavily rely on structural and linguistic information available in the data. Significant progress has been made in terms of enhancing portability of NLP systems across clinical specialties. In the future, developing a cross-site NLP algorithm will be helpful to delineate individual noting styles of providers and to generalize NLP systems for SDoH.

CONCLUSION

This review presented a qualitative analysis of 82 publications focused on the extraction of SDoH concepts in EHR systems using NLP techniques. With increasing recognition of nonclinical factors that define patients' health risks, needs, and outcomes, it becomes equally imperative that social and behavioral concepts are captured in order to be leveraged during clinical decision-making related to diagnosis and therapy planning. Devising novel ways in which such data can be extracted and leveraged with as little impact on current documentation routines of providers is an ideal solution. With the valuable knowledge of the relatively new literature in this area, researchers can leverage such reviews to steer their study in innovative ways.

FUNDING

This research was supported in part by NIH R01MH119177, R01MH121907, and R01MH121922.

AUTHOR CONTRIBUTTIONS

BGP conceived of the idea and worked on planning, data collection, data extraction, writing, and editing. MMS helped in data collection, data extraction, and writing. VV helped in planning, writing, and editing. PA helped in planning, advising, and editing. OVP, BG, LAL, ER, JMB, AF, TJG, WH, YW, XY, JB, MW, PW, JJM, MO, TC, and MW helped in advising and editing. JP provided critical direction and helped in planning, advising, and editing. JP also gathered funding.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY

The extracted data are available in the supplementary file.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Gucciardi E, Vahabi M, Norris N, Del Monte JP, Farnum C. The intersection between food insecurity and diabetes: a review. *Curr Nutr Rep* 2014; 3 (4): 324–332.
2. Coughlin SS. Social determinants of breast cancer risk, stage, and survival. *Breast Cancer Res Treat* 2019; 177 (3): 537–548.
3. Suglia SF, Duarte CS, Sandel MT. Housing quality, housing instability, and maternal mental health. *J Urban Health* 2011; 88 (6): 1105–1116.
4. Masters ET, Ramaprasan A, Mardekian J, *et al*. Natural language processing—identified problem opioid use and its associated health care costs. *J Pain Palliat Care Pharmacother* 2018; 32 (2-3): 106–115.
5. Magnan S. Social determinants of health 101 for health care: five plus five. *NAM Perspectives* 2017: 1–9.
6. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health icd-10 z-codes among hospitalized patients in the United States, 2016–2017. *Med Care* 2020; 58 (12): 1037–1043.
7. Kong H-J. Managing unstructured big data in healthcare system. *Healthc Inform Res* 2019; 25 (1): 1–2.
8. Gundlapalli AV, Brignone E, Divita G, *et al*. Using structured and unstructured data to refine estimates of military sexual trauma status among US military veterans. *Stud Health Technol Inform* 2017; 238: 128–131.
9. Conway M, Keyhani S, Christensen L, *et al*. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *J Biomed Semantics* 2019; 10 (1): 1–10.
10. Hatef E, Rouhizadeh M, Tia I, *et al*. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform* 2019; 7 (3): e13802.
11. Palmer RE, Carrell DS, Cronkite D, *et al*. The prevalence of problem opioid use in patients receiving chronic opioid therapy: computer-assisted review of electronic health record clinical notes. *Pain* 2015; 156 (7): 1208–1214.
12. Bejan CA, Angiolillo J, Conway D, *et al*. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc* 2018; 25 (1): 61–71.
13. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *J Am Med Inform Assoc* 2020; 27 (11): 1764–1773.
14. Bompelli A, Wang Y, Wan R, *et al*. Social determinants of health in the era of artificial intelligence with electronic health records: a systematic review. arXiv preprint arXiv:2102.04216, 2021, preprint: not peer-reviewed.
15. Hutton B, Catala-Lopez F, Moher D. The PRISMA statement extension for systematic reviews incorporating network meta-analysis: PRISMA-NMA. *Med Clin (Barc)* 2016; 147 (6): 262–266.
16. Gundlapalli AV, Carter ME, Palmer, M, *et al*. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among us veterans. *AMIA Annu Symp Proc* 2013 2013: 537–546.
17. Shoenbill K, Song Y, Gress L, Johnson H, Smith M, Mendonca EA. Natural language processing of lifestyle modification documentation. *Health Informatics J* 2020; 26 (1): 388–405.
18. Gundlapalli AV, Carter ME, Divita G, *et al*. Extracting concepts related to homelessness from the free text of VA electronic medical records. *AMIA Annu Symp Proc 2014;* 2014: 589; Washington, DC.
19. Redd A, Carter M, Divita G, *et al*. Detecting earlier indicators of homelessness in the free text of medical records. In: *International Conference on Informatics, Management and Technology in Healthcare*. IOS Press; 2014: 153–156; Washington, DC.
20. Greenwald JL, Cronin PR, Carballo V, Danaei G, Choy G. A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial support using natural language processing. *Med Care* 2017; 55 (3): 261–266.
21. Blosnich JR, Montgomery AE, Dichter ME, *et al*. Social determinants and military veterans' suicide ideation and attempt: a cross-sectional analysis of electronic health record data. *J Gen Intern Med* 2020; 35 (6): 1759–1767.
22. Winden TJ, Chen ES, Monsen KA, Wang Y, Melton GB. Evaluation of flowsheet documentation in the electronic health record for residence, living situation, and living conditions. *AMIA Jt Summits Transl Sci Proc* 2018; 2018: 236–245.
23. Bettencourt-Silva JH, Mulligan N, Sbodio M, *et al*. Discovering new social determinants of health concepts from unstructured data: framework and evaluation. *Stud Health Technol Inform* 2020; 270: 173–177.
24. Topaz M, Murga L, Bar-Bachar O, Cato K, Collins S. Extracting alcohol and substance abuse status from clinical notes: the added value of nursing data. *Stud Health Technol Inform* 2019; 264: 1056–1060.

25. Velupillai S, Mowery DL, Conway M, Hurdle J, Kious B. Vocabulary development to support information extraction of substance abuse from psychiatry notes. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics; 2016: 92–101.
26. Myra Kim H, Smith EG, Ganoczy D, *et al.* Predictors of suicide in patient charts among patients with depression in the veterans health administration health system: importance of prescription drug and alcohol abuse. *J Clin Psychiatry* 2012; 73 (10): 1269–1275.
27. Mowery DL, South B, Patterson O, Zhu S-H, Conway M. Investigating the documentation of electronic cigarette use in the veteran affairs electronic health record: a pilot study. In: *BioNLP 2017*. Association for Computational Linguistics; 2017: 282–286.
28. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from deidentified electronic health records. In: proceedings of the 2017 *Pacific Symposium on Biocomputing*; World Scientific; January 3–7, 2017; Island of Hawaii.
29. Wang Y, Sohn S, Liu S, *et al.* A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019; 19 (1): 1–13.
30. Amrit C, Paauw T, Aly R, Lavric M. Identifying child abuse through text mining and machine learning. *Expert Syst Appl* 2017; 88: 402–418.
31. Perron BE, Victor BG, Bushman G, *et al.* Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child Abuse Negl* 2019; 98: 104180.
32. Feller DJ, Zucker J, Bear Don't Walk IV O, *et al.* Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. *AMIA Annu Symp Proc* 2018; 2018: 422–429.
33. Feller DJ, Zucker J, Bear Don't Walk OJ IV, Yin MT, Gordon P, Elhadad N. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Appl Clin Inform* 2020; 11 (1): 172–181.
34. Badger J, LaRose E, Mayer J, Bashiri F, Page D, Peissig P. Machine learning for phenotyping opioid overdose events. *J Biomed Inform* 2019; 94: 103185.
35. Wang Y, Chen ES, Pakhomov S, *et al.* Automated extraction of substance use information from clinical texts. *AMIA Annu Symp Proc* 2015; 2015: 2121–2130; San Francisco, CA.
36. Jonnagaddala J, Dai H-J, Ray P, Liaw S-T. A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. In: *Proceedings of BioNLP 15*. NLP; 2015: 147–51.
37. Erickson J, Abbott K, Susienka L. Automatic address validation and health record review to identify homeless social security disability applicants. *J Biomed Inform* 2018; 82: 41–46.
38. Green CA, Perrin NA, Hazlehurst B, *et al.* Identifying and classifying opioid-related overdoses: a validation study. *Pharmacoepidemiol Drug Saf* 2019; 28 (8): 1127–1137.
39. Hazlehurst B, Green CA, Perrin NA, *et al.* Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data. *Pharmacoepidemiol Drug Saf* 2019; 28 (8): 1143–1151.
40. Byrne T, Montgomery AE, Fargo JD. Predictive modeling of housing instability and homelessness in the Veterans Health Administration. *Health Serv Res* 2019; 54 (1): 75–85.
41. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. *Alcohol* 2020; 84: 49–55.
42. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform* 2021; 113: 103631.
43. Patel R, Wilson R, Jackson R, *et al.* Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: an observational study. *BMJ Open* 2016; 6 (3): e009888.
44. Lingeman JM, Wang P, Becker W, Yu H. Detecting opioid-related aberrant behavior using natural language processing. *AMIA Ann Symp Proc* 2017; 2017: 1179–1185.
45. Sharma B, Dligach D, Swope K, *et al.* Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. *BMC Med Inform Decis Mak* 2020; 20 (1): 1–11.
46. Rajendran S, Topaloglu U. Extracting smoking status from electronic health records using NLP and deep learning. *AMIA Jt Summits Transl Sci Proc* 2020; 2020: 507–516.
47. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open* 2021.
48. Lindemann EA, Chen ES, Wang Y, Skube SJ, Genevieve BM. Representation of social history factors across age groups: a topic analysis of freetext social documentation. *AMIA Annu Symp Proc* 2017; 2017: 1169–1178; Washington, DC.
49. Afshar M, Joyce C, Dligach D, *et al.* Subtypes in patients with opioid misuse: a prognostic enrichment strategy using electronic health record data in hospitalized patients. *PLoS One* 2019; 14 (7): e0219717.
50. Wang L, Lakin J, Riley C, Korach Z, Frain LN, Zhou L. Disease trajectories and end-of-life care for dementias: latent topic modeling and trend analysis using clinical notes. *AMIA Annu Symp Proc* 2018; 2018: 1056–1065.
51. Wray CM, Vali M, Walter LC, *et al.* Examining the interfacility variation of social determinants of health in the Veterans Health Administration. *Fed Pract* 2021; 38 (1): 15.
52. Hammond KW, Laundry RJ. Application of a hybrid text mining approach to the study of suicidal behavior in a large population. In: *2014 47th Hawaii International Conference on System Sciences*, IEEE, 2014: 2555–2561; Waikoloa, HI.
53. Divitaa G, Brignonea E, Carter ME, *et al.* Extracting sexual trauma mentions from electronic medical notes using natural language processing. In: *MEDINFO 2017: Precision Healthcare Through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics*. Vol. 245. IOS Press, 2018: 351–355; Hangzhou, China.
54. Jones AL, Pettey WB, Carte, ME, *et al.* Regional variations in documentation of sexual trauma concepts in electronic medical records in the United States Veterans Health Administration. *AMIA Annu Symp Proc* 2019; 2019: 514–522; Washington, DC.
55. Carrell DS, Cronkite D, Palmer RE, *et al.* Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inform* 2015; 84 (12): 1057–1064.
56. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate cancer. *BMC Med Inform Decis Mak* 2019; 19 (1): 89.
57. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006; 6 (1): 1–9.
58. Regan S, Meigs JB, Grinspoon SK, Triant VA. Determinants of smoking and quitting in hiv-infected individuals. *PLoS One* 2016; 11 (4): e0153103.
59. Navathe AS, Zhong F, Lei VJ, *et al.* Hospital readmission and social risk factors identified from physician notes. *Health Serv Res* 2018; 53 (2): 1110–1136.
60. Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. *Cancer Inform* 2016; 15: 237–242.
61. Wu C-Y, Chang C-K, Robson D, *et al.* Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PLoS One* 2013; 8 (9): e74262.
62. Bhavsar V, Sanyal J, Patel R, *et al.* The association between neighbourhood characteristics and physical victimisation in men and women with mental disorders. *BJPsych Open* 2020; 6 (4): e73.
63. Epstein RH, St Jacques P, Stockin M, Rothman B, Ehrenfeld JM, Denny JC. Automated identification of drug and food allergies entered using non-standard terminology. *J Am Med Inform Assoc* 2013; 20 (5): 962–968.

64. Hazlehurst BL, Lawrence JM, Donahoo WT, *et al.* Automating assessment of lifestyle counseling in electronic health records. *Am J Prev Med* 2014; 46 (5): 457–464.
65. Bellows BK, LaFleur J, Kamauu AW, *et al.* Automated identification of patients with a diagnosis of binge eating disorder from narrative electronic health records. *J Am Med Inform Assoc* 2014; 21 (e1): e163–e168.
66. Winden TJ, Chen ES, Wang Y, Lindemann E, Melton GB. Residence, living situation, and living conditions information documentation in clinical practice. *AMIA Annu Symp Proc* 2017; 2017: 1783–1792.
67. Dillahunt-Aspillaga C, Finch D, Massengale J, Kretzmer T, Luther SL, McCart JA. Using information from the electronic health record to improve measurement of unemployment in service members and veterans with mTBI and postdeployment stress. *PLoS One* 2014; 9 (12): e115873.
68. Lynch KE, Whitcomb BW, DuVall SL. How confounder strength can affect allocation of resources in electronic health records. *Perspect Health Inf Manag* 2018; 15 (Winter): 1d.
69. Rosenthal B, Skrbin J, Fromkin J, *et al.* Integration of physical abuse clinical decision support at 2 general emergency departments. *J Am Med Inform Assoc* 2019; 26 (10): 1020–1029.
70. Suresh S, Saladino RA, Fromkin J, *et al.* Integration of physical abuse clinical decision support into the electronic health record at a tertiary care children's hospital. *J Am Med Inform Assoc* 2018; 25 (7): 833–840.
71. Chen ES, Carter EW, Sarkar IN, Winden TJ, Melton GB. Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. In: *AMIA Annu Symp Proc* 2014; 2014: 366–374.
72. Wang Y, Chen ES, Pakhomov S, Lindemann E, Melton GB. Investigating longitudinal tobacco use information from social history and clinical notes in the electronic health record. In: *AMIA Annu Symp Proc* 2016; 2016: 1209–1218.
73. 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.
74. Hazlehurst B, Sittig DF, Stevens VJ, *et al.* Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. *Am J Prev Med* 2005; 29 (5): 434–439.
75. Palmer EL, Hassanpour S, Higgins J, Doherty JA, Onega T. Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC Med Inform Decis Mak* 2019; 19 (1): 141.
76. Liu M, Shah A, Jiang M, *et al.* A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu Symp Proc* 2012; 2012: 577–586.
77. Young-Wolff KC, Klebaner D, Folck B, *et al.* Do you vape? Leveraging electronic health records to assess clinician documentation of electronic nicotine delivery system use among adolescents and adults. *Prev Med* 2017; 105: 32–36.
78. Afshar M, Phillips A, Karnik N, *et al.* Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc* 2019; 26 (3): 254–261.
79. Hegde H, Shimpi N, Glurich I, Acharya A. Tobacco use status from clinical notes using natural language processing and rule based algorithm. *Technol Health Care* 2018; 26 (3): 445–456.
80. Brandt Baldwin K. Evaluating healthcare quality using natural language processing. *J Healthc Qual* 2008; 30 (4): 24–29.
81. Gordon DD, Patel I, Pellegrini AM, Perlis RH. Prevalence and nature of financial considerations documented in narrative clinical records in intensive care units. *JAMA Netw Open* 2018; 1 (7): e184178.
82. Goodday SM, Kormilitzin A, Vaci N, *et al.* Maximizing the use of social and behavioural information from secondary care mental health electronic health records. *J Biomed Inform* 2020; 107: 103429.
83. Bucher BT, Shi J, John Pettit R, Ferraro J, Chapman WW, Gundlapalli A. Determination of marital status of patients from structured and unstructured electronic healthcare data. *AMIA Annu Symp Proc* 2019; 2019: 267–274.
84. Skaljic M, Patel IH, Pellegrini AM, Castro VM, Perlis RH, Gordon DD. Prevalence of financial considerations documented in primary care encounters as identified by natural language processing methods. *JAMA Netw Open* 2019; 2 (8): e1910399.
85. Bellows BK, DuVall SL, Kamauu AW, Supina D, Babcock T, LaFleur J. Healthcare costs and resource utilization of patients with binge-eating disorder and eating disorder not otherwise specified in the department of veterans affairs. *Int J Eat Disord* 2015; 48 (8): 1082–1091.
86. Gottschalk A, Flocke SA. Time spent in face-to-face patient care and work outside the examination room. *Ann Fam Med* 2005; 3 (6): 488–493.
87. Kharrazi H, Anzaldi LJ, Hernandez L, *et al.* The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc* 2018; 66 (8): 1499–1507.
88. Rahimian F, Salimi-Khorshidi G, Payberah AH, *et al.* Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med* 2018; 15 (11): e1002695.