JAMIA Open, 6(2), 2023, ooad024
https://doi.org/10.1093/jamiaopen/ooad024
Research and Applications

Research and Applications

Natural language processing-driven state machines to extract social factors from unstructured clinical documentation

Katie S. Allen [1,2], Dan R. Hood[1], Jonathan Cummins[1], Suranga Kasturi[1], Eneida A. Mendonca [3,4], and Joshua R. Vest [1,2]

[1]Center for Biomedical Informatics, Regenstrief Institute, Inc., Indianapolis, Indiana, USA, [2]Department of Health Policy and Management, Richard M. Fairbanks School of Public Health, IUPUI, Indianapolis, Indiana, USA, [3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA and [4]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana, USA

Corresponding Author: Katie S. Allen, BS, Center for Biomedical Informatics, Regenstrief Institute, Inc., 1101 W. 10th Street, Indianapolis, IN 46202, USA; allenkat@regenstrief.org

Received 28 December 2022; Revised 8 March 2023; Editorial Decision 9 March 2023; Accepted 28 March 2023

ABSTRACT

Objective: This study sought to create natural language processing algorithms to extract the presence of social factors from clinical text in 3 areas: (1) housing, (2) financial, and (3) unemployment. For generalizability, finalized models were validated on data from a separate health system for generalizability.

Materials and Methods: Notes from 2 healthcare systems, representing a variety of note types, were utilized. To train models, the study utilized n-grams to identify keywords and implemented natural language processing (NLP) state machines across all note types. Manual review was conducted to determine performance. Sampling was based on a set percentage of notes, based on the prevalence of social need. Models were optimized over multiple training and evaluation cycles. Performance metrics were calculated using positive predictive value (PPV), negative predictive value, sensitivity, and specificity.

Results: PPV for housing rose from 0.71 to 0.95 over 3 training runs. PPV for financial rose from 0.83 to 0.89 over 2 training iterations, while PPV for unemployment rose from 0.78 to 0.88 over 3 iterations. The test data resulted in PPVs of 0.94, 0.97, and 0.95 for housing, financial, and unemployment, respectively. Final specificity scores were 0.95, 0.97, and 0.95 for housing, financial, and unemployment, respectively.

Discussion: We developed 3 rule-based NLP algorithms, trained across health systems. While this is a less sophisticated approach, the algorithms demonstrated a high degree of generalizability, maintaining >0.85 across all predictive performance metrics.

Conclusion: The rule-based NLP algorithms demonstrated consistent performance in identifying 3 social factors within clinical text. These methods may be a part of a strategy to measure social factors within an institution.

Key words: social factors, social determinants of health, natural language processing, clinical data

LAY SUMMARY

Social factors, such as an individual's housing, food, employment, and income situations, affect their overall health and well-being. As a result, data on patients' social factors aid in clinical decision making, planning by hospital administrators and policy-makers, and enrich research studies with data representative of more factors influencing the life of an individual. Data on social factors can be collected at the time of a healthcare visit through screening questionnaires or are often documented in the clinical text as part of the social narrative. This study examines the use of natural language processing—a machine method to identify certain text within a larger document—to identify housing instability, financial insecurity, and unemployment from within the clinical notes. Using a relatively unsophisticated methodology, this study demonstrates strong performance in identifying these social factors, which will enable stakeholders to utilize these details in support of improved clinical care.

INTRODUCTION

Social factors, that is, an individual's housing, food, employment, and income situations, affect health outcomes, utilization, and costs.[1–3] As a result, data on patients' social factors have application to clinical decision-making, organizational-level planning and decision-making, and research studies.[1,4] There is emerging interest in capturing social factor data via social screening questionnaires and structured diagnostic codes (ie, ICD-10 and LOINC).[5] However, a rich variety of social factor data may already be routinely documented in clinical free-text notes collected by clinical and nonclinical staff.[6,7] The goal of this project is to develop automated approaches to identify key social factors from multiple types of free-text notes collected from different patient populations. We created and validated natural language processing-driven state machines to detect 3 types of social factors: housing instability, financial insecurity, and unemployment.

Background and significance

Clinical care has a long tradition of recording social factors and risks within patients' health records as free-text data.[8,9] Social history is a standard portion of health records.[10] Social risk factor information appears in clinical notes from different medical specialties[11] and it could be in the form of products of services delivered by physicians, nurses, or social workers.[7] However, the unstructured nature of text creates well-known challenges in reuse for clinical decision-making, aggregate reporting, or as input for other informatics tools, such as referral systems.

A potential solution to extracting person-level social factors from the clinical notes is natural language processing (NLP).[12] NLP has been successfully applied to clinical conditions including cancer,[13] cardiovascular issues,[14] and mental health.[15] Nevertheless, the use of NLP to identify social factors may pose greater challenges. For example, documentation of social factors in clinical text is highly variable between clinicians.[16] Also, different healthcare professionals may describe the same social concept using inconsistent language, or the concept may be inconsistently described in different types of clinical notes.[17] Additionally, social factors are complex, intertwined, and often recounted by patients as stories or narratives, which may be difficult to record within EHRs.[18] All these factors complicate extraction.

Social factors are not a completely new domain for NLP research. However, previous NLP work has tended more towards the identification of socio-behavioral risks than social factors, per se. For example, a recent systematic review concluded that the identification of tobacco use, alcohol use, substance abuse, physical activity, or sexual activity constituted the majority of publications on NLP.[12] Specific to social factors, numerous publications have used NLP to identify homelessness,[7,19–23] which is the most acute manifestation of housing instability.[24] However, efforts to identify factors such as financial insecurity and employment are much less frequent.[12] Moreover, previous work has focused primarily on NLP within a specific context. For example, the work will focus on one condition/cohort, such as substance abuse.[25,26] Other work relies heavily on existing, freely available clinical text (eg, the Medical Information Mart for Intensive Care dataset [MIMIC]),[27–29] certain note types or specific sections of notes (eg, social work/social history),[27,29] or utilizes surveys collected for nonclinical care purposes.[30] Studies looking at the generalizability across health systems are rare, with limited evidence to date. However, recent work has examined the ability to identify housing insecurity across multiple healthcare systems.[23]

This study is motivated by the current state of applied NLP research as outlined above and seeks to contribute to the literature in 3 ways: (1) identifies a methodology applied to multiple social factors, with a focus on understudied factors related to economic stability, (2) is applied to a broad population representing multiple clinical settings and note types, and (3) focuses on the generalizability of the NLP algorithms to outside health systems.

MATERIALS AND METHODS

We developed a series of NLP-driven state machines[31] to identify social risk factors present in clinical notes obtained from 2 different health systems based in Indiana, USA. Development included a multistep, systematic, iterative process with keyword and rule identification, state machine development, and expert validation.

Social factors

We sought to identify the following highly prevalent social risk factors: (1) housing instability, (2) financial insecurity, and (3) unemployment. All 3 risk factors have been associated with economic stability, which is a central concern of Healthy People 2030.[32] Housing instability is estimated to affect 20% of households in the United States.[33] The connection between housing instability and clinical care is extremely critical as issues related to housing have been shown to share risk factors with physical trauma[34] as well as health impacts from unstable or inadequate housing.[35] Evidence shows that income shapes health behaviors[36] as well as the ability to seek appropriate medical care.[37] Economic stability, or its inverse of instability, may change behaviors in a myriad of ways and clinician awareness of these social factors may help shape the care provided. For this study, the 3 constructs were defined as follows:

- *Housing instability*: Any housing disruption or housing-related problem such as frequent moves, difficulty paying rent, eviction, or homelessness.[38]

- *Financial insecurity*: A situation in which a person cannot fully meet current and ongoing financial obligations without fear of not having enough money to cover necessary expenses.[39,40]
- *Unemployment*: Individuals who are not working, but eligible to participate in the workforce. This excludes individuals who are students, homemakers, or disabled.[41]

Sample and data

The primary data source was the Indiana Network for Patient Care (INPC).[42] INPC is a regional health information exchange with multiple participating health systems. For the purposes of this study, clinical notes were isolated from 2 Indianapolis, IN area health systems that represented high-volume, diverse clinical populations: one is a public safety-net provider, and the second is a nonprofit system serving largely privately insured individuals. We purposefully selected these different sources to support model generalizability. Clinical notes included all clinical documentation shared with INPC, which includes a variety of note types (see Supplementary Appendix S1 for most common note types). The training corpus of notes was obtained from a multihospital health system representing urban and rural settings. Notes associated with any type of clinical encounter were included if they were created between January 1, 2019, and December 31, 2019 ($n = 1\,710\,124$ notes, 581 205 unique patients). The test corpus of clinical notes was extracted from any clinical encounter for a second, separate health system. These notes were obtained from a safety-net hospital with multiple health clinics that were documented between September 1, 2020, and March 31, 2021 ($n = 724\,308$ notes, 74 239 unique patients). Notes did not undergo any preparation processes and were utilized in their raw form. Clinical notes were linked via patient identifiers to clinical and demographic information from INPC. These measures included age, gender, race, ethnicity, rural/urban status, Modified Townsend Index (social deprivation based on residential zip code),[43] and Charlson Comorbidity score.[44]

Keyword identification

Utilizing 1 month of notes from the training set, we produced a list of continuous sequence words (n-grams)[31] for review. This list was filtered utilizing Term Frequency-Inverse Document Frequency (TF-IDF) measurements, to identify and remove stop words not contributing value (eg, "this", "what"). Team members reviewed 1-, 2-, and 3-grams to determine any connection to housing instability, financial insecurity, or unemployment. Each n-gram was reviewed independently by 2 research assistants with differences adjudicated by the authors. We also identified potential keywords from published social risk factor screening questionnaires.[45,46] The result of this step was an initial set of keywords for inclusion in each state machine.

State machine development

We applied a deterministic finite state machine methodology,[47] which has been previously proven to work with clinical data,[48] using a multistep development and validation processes (see Fig. 1). These processes identified clinical notes with mentions of housing instability, financial insecurity, or unemployment. We constructed and validated the machines independently from one another, utilizing the Regenstrief natural language processing environment nDepth[49–51] to implement the state machines. nDepth is a platform that enables logical keyword-based searches, extraction of structured information such as n-grams and processing pipelines for text segmentation and tokenization (state machines), and manual validation of pipeline performance. nDepth indexes and searches the corpus of clinical notes using the keyword dictionary and associated rules (eg, negation).[49–51] Figure 1 provides an overview of our approach.

First, we constructed individual initial state machines for each of the 3 social factors utilizing the terms generated in the keyword identification phase. For each machine, 2 reviewers (KA and DH) manually reviewed 200 randomly selected notes from the training corpus to assess machine performance using positive and negative predictive values (PPV and NPV). This was utilized to determine the initial performance of each machine and to identify keyword adjustments and any issues related to language (eg, template language containing keyword or alternate meaning to keyword).

Next, we retrained the state machine on a new, larger set of training corpora of notes, derived from the original set of notes identified previously. This iterative step included adjustments to each state machine based on how each term performed and reviewer comments. The state machines were again reassessed via manual review. We continued this process, gradually updating each state machine based on expert validation results. Updates to the state machines included the addition of new keywords as well as the creation of new rules such as (1) the definition of formal search spaces (eg, search for the keyword "sleep" within 5 words of the word 'car', (2) grammatical rules to identify negation (eg, the patient is able "to afford" vs unable "to afford"), and (3) to identify context (eg, differentiate between the use of the phrase "does not work" as it applies to a clinical concept such as 'medication does not work' vs how it applies to an individual who is not employed).

Given the low prevalence of reported social factors in the notes, we focused on a subset of note types where the social factors were most commonly identified (see Supplementary Appendix S1). For

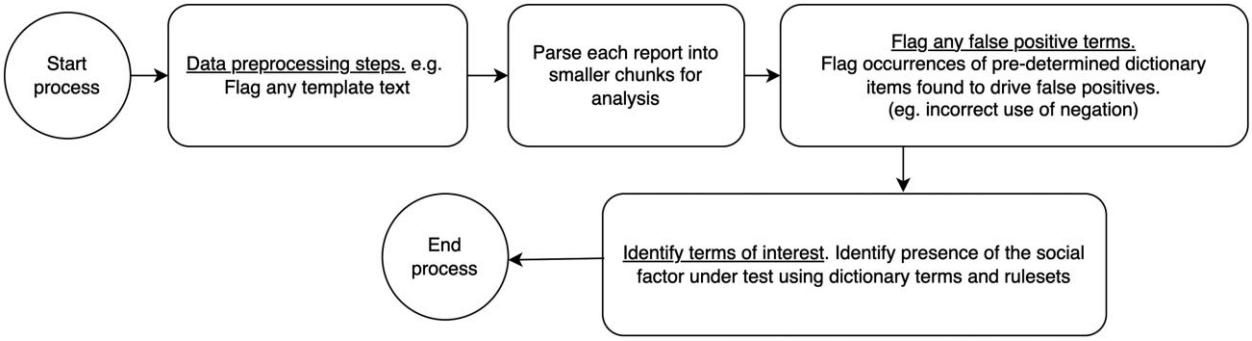

Figure 1. State machine process diagram.

Table 1. Overall cohort characteristics (persons)

	Total cohort characteristics ($N = 9907$)	Training corpus ($n = 3621$)	Testing corpus ($n = 6286$)
Age, mean (SD)	48.72 (17.21)	51.47 (18.58)	47.14 (16.42)
Female, *n* (%)	5316 (53.7)	1871 (51.7)	3445 (54.8)
Race, *n* (%)			
White	5519 (55.7)	2922 (80.7)	2597 (41.3)
Black/African American	3320 (33.5)	487 (13.4)	2833 (45.1)
AAPI	177 (1.8)	25 (0.7)	152 (2.4)
Multiracial	132 (1.3)		132 (2.1)
Other	759 (7.7)	187 (5.2)	572 (9.1)
Hispanic	1131 (11.4)	81 (2.2)	1050 (16.7)
Geographic characteristics			
Urban, *n* (%)	8933 (90.2)	2799 (77.3)	6134 (97.6)
Modified townsend, mean (SD)	1.78 (4.33)	0.11 (4.51)	2.68 (3.90)
Clinical characteristics, mean (SD)			
Encounters PP[a]	19.2 (20.69)	13.8 (1.30)	22.32 (24.33)
Notes PP[b]	30.67 (35.45)	15.36 (13.23)	39.48 (42.22)
Charlson comorbidity score	1.84 (2.51)	1.63 (2.44)	1.93 (2.53)

[a]Encounters PP: Statistics related to the number of clinical encounters per person in the data set.
[b]Notes PP: Statistics related to the number of clinical notes per person in the data set.

example, the more common note types included progress and visit notes and we ultimately excluded documents related to pathological reports, which do not typically have social history. For subsequent training iterations, we identified the top 10 most common note types, and 4 reviewers manually examined an equal number of positive and negative results, matched on note type. This is similar to methods deployed in other studies requiring manual review.[52] We completed the iteration process if the state machine reported PPV and sensitivity scores $\geq$0.85.

To ensure consistency, the reviewers annotated an overlapping subset of 50 notes from each of the 3 state machines. Reviewers exhibited a high level of agreement (see Supplementary Appendix S2).

Finally, we applied each of the finalized state machines to the test corpus of notes from the second health system. Manual review was completed per the same methodology as the second step. The state machine was considered finalized if PPV and sensitivity were $\geq$0.90.

Analysis

Frequencies, percent, means, and standard deviation described the study samples. PPV, NPV, sensitivity, and specificity described the performance of each state machine. The Indiana University Institutional Review Board approved this study.

RESULTS

Cohort characteristics

The corpora of notes validated in this study included a total of 9907 unique patients across the 2 health systems (Table 1). Consistent in an urban US healthcare setting, the majority of patients were from urban zip codes and were female. Patients in the test cohort tended to be more diverse and from more socially deprived areas, which is reflective of the safety-net role of the health system.

State machine performance

The state machines identified a prevalence of social factors which ranged from 0.06% to 1.54%. This prevalence rate suggests a low baseline of documentation but should be interpreted cautiously as the state machine was run across all notes, including imaging, which is less likely to contain mentions of social factors. Table 2 shows all results for the initial training run, final training run, and the performance of each machine on the final holdout test set of notes. Initial state machine results, based solely on the identified keywords, were promising with moderate PPV and excellent NPV. Repetitive iterations with the training data resulted in increased PPVs, meeting our criteria for movement to test data, and excellent sensitivity results.

Overall, the housing instability state machine required 3 training iterations, financial insecurity 2 training iterations, and unemployment 3 training iterations to reach the threshold sensitivity threshold score (>0.85) we defined. Specificity was lower but remained over the 0.85 threshold. Following slight adjustments, the final state machine performance, conducted on the test corpus of notes, was

Table 2. Validation results by state machine (note level)

	Housing	Financial	Unemployment
Training run 1[a]	$n = 200$	$n = 200$	$n = 200$
PPV	0.71	0.83	0.78
NPV	1.00	1.00	1.00
Training run 2	$n = 200$	$n = 1025$	$n = 1980$
PPV	0.89	0.89	0.83
NPV	1.00	0.99	0.99
Sensitivity	1.00[a]	0.99	0.99
Specificity	0.90[a]	0.9	0.85
Training run 3	$n = 997$	n/a	$n = 200$
PPV	0.95	—	0.88
NPV	1.00	—	0.98
Sensitivity	1.00	—	0.94[b]
Specificity	0.95		0.96[b]
Test run (final)	$n = 997$	$n = 2136$	$n = 4110$
PPV	0.94	0.97	0.95
NPV	1.00	0.98	0.99
Sensitivity	1.00	0.98	0.99
Specificity	0.95	0.97	0.95

[a]Represents the first training iteration with all keywords and no modifications.
[b]Should be interpreted cautiously given the artificially constrained sample size.

strong. PPV, sensitivity, and specificity were above 0.90 for all state machines.

DISCUSSION

We developed 3 NLP-driven state machines using free-text notes captured from a health system that served a commercially insured population and demonstrated their generalizability to a health system serving an underserved patient population. By developing and calibrating models across health systems that serve different patient populations, we demonstrated high degrees of model generalizability, which is critical in working with multi-institutional data, such as health information exchange systems.

Methodologically, this study highlights 2 challenges often identified in applied healthcare analytics. The first is generalizability. The ability of models, NLP or otherwise, developed using data from one population to deliver satisfactory predictive performance across other previously unseen, statistically different patient populations, is an ongoing challenge.[53] For example, some projects targeting social factors identification have focused on vulnerable populations (eg, substance abuse, HIV patients).[25,26] The advantage of using these targeted cohorts is that the higher likelihood social factors are frequently documented due to their known associations with health outcomes. While this may make statistical model or algorithm development more feasible, generalizability to a less vulnerable population is uncertain. For this study, we were able to leverage 2 healthcare institutions, that, while in the same geographic area, generally serve 2 different patient populations, that is, the commercially insured and the underserved. This is a more rigorous approach towards establishing generalizability.

The second methodological issue was our choice of a state machine-based approach. A majority of analytical efforts involving free-text datasets have shifted towards complex, resource-intensive approaches such as neural networks and deep learning[29] to identify and classify various social factors.[29,54] While these models may yield superior performance under test settings, they are complex approaches that require a high degree of technical expertise, greater computing resources, and present scalability issues. In any machine learning application, researchers and practitioners must make choices between the tradeoffs of performance, implementability, and maintainability.[7,26,29,54] While these simple, finite-state machine methods are not as sophisticated as neural network-based approaches, they have several advantages. They are less complex, and thereby easier to develop, interpret, implement, and maintain.[55] In many cases, rules-based systems are more transparent, easier to communicate to nonexperts, and therefore more easily implemented in other health systems. Past studies have shown that rule-based methods are more sensitive, which may be important for phenomena that are more rarely documented, such as social factors.[25] Due to their simplicity, rule-based methods may be more generalizable, and thus, report more consistent predictive performance across health systems. Our choice of a state-machine approach may address some of the limitations facing advanced analytical methods including the potential for bias in artificial intelligence which particularly impact more complex, black box approaches such as deep learning methods. Learning algorithms run the risk of incorporating underlying societal biases contained within medical datasets.[56] This potential for bias is increased with the perpetuation of 2 particular practices, namely the utilization of the off-the-shelf NLP programs that may not have been trained to utilize clinical data and the reliance on shared/public annotated note sets (eg, MIMIC).[27–29] The methods discussed in this study limit bias by utilizing human intervention in the development of state machines and by leveraging a unique set of clinical notes.

The demand from payers, policy-makers, and advocates for information on patients' social factors and needs is substantial[57] and multiple approaches are requested to obtain this information. In recent years, coding standards for recording social risks as structured data within EHRs using ICD-10 or LOINC codes have advanced substantially. Nevertheless, these structured data are very underutilized in practice.[58] Additionally, health systems and researchers have integrated social screening surveys into EHR environments. However, evidence suggests that social screening is not an exceptionally common practice.[59] Our findings indicate that the use of NLP to analyze existing, routinely collected free-text reports may be a feasible approach for healthcare organizations and researchers to identify patients with documented social factors. As clinical notes represent a different data generation process than coding or screening surveys, NLP could be applied as part of an overall social health measurement strategy. It is important to not discard clinical text in favor of screening or other structured methods for data collection. However, social factors extracted via NLP could be utilized to impute missing survey results, augment survey data, or—given the ability to apply retrospectively—provide a longitudinal description of social factors. As products of a clinical encounter, these patient interactions and the information within clinical notes are important. However, it is also critical to remember that the text is, by nature, selective, filtered, and containing omissions (either left unrecorded by the provider or never volunteered by the patient).[60–62] A comprehensive health measurement strategy will include formalized screening as well as information garnered from clinical documentation.

Once identified, social factor data have wide application. These data can improve risk prediction models,[63–66] match patients to appropriate social services,[67,68] and illuminate underlying disparities in population health.[69] Free-text data constitute an important source for feature creation[11] and are one that generally contribute to better prediction models.[70] Additionally, unlike structured codes or patient surveys that cannot be collected retroactively, NLP algorithms can be applied to historical data, for times prior to the implementation of formalized screening processes. While still acknowledging the caveats and the limitations of clinical notes, NLP could provide both a more reasonable measure of longitudinal social risk as well as data for retrospective cohorts.

Limitations

While these NLP algorithms were developed and validated across 2 distinct health systems, performance may not be similar in other statistically diverse patient populations. Further analyses could push generalizability in the area of settings by validating performance in a different geographic location. Additionally, while the state machines achieved acceptable performance for the 3 target social factors, such an approach may not be feasible for other social factors such as transportation barriers, legal issues, or food insecurity. More sophisticated methods may be necessary for these social factors or to achieve even better performance than our existing models. In addition, as social factor screening becomes more common in clinical practice, it is possible that documentation practices or language will change, which will necessitate the revision and updating of the state machines. Finally, our efforts were susceptible to limitations caused by misspellings, errors present in the free-text data, and challenges caused by the use of abbreviations specific to the medical domain.

CONCLUSION

NLP algorithms demonstrated strong performance in identifying cases of financial insecurity, housing instability, and unemployment among adult patients. NLP could be a part of an overall social health measurement strategy.

FUNDING

This work was supported by the funding from the Indiana University Addictions Grand Challenge (PI: Dr. Peter Embi), a University initiative to respond to the addictions crisis.

AUTHOR CONTRIBUTIONS

JRV and KSA conceived of this study and were responsible for primary writing activities. KSA and DRH provided manual validation as well as keyword identification guidance. KSA also completed the analyses. JC provided technical guidance and programming requirements. SK and EAM provided subject matter expertise, overall guidance, and manuscript editing.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

The study team would like to acknowledge the indispensable assistance provided by the Regenstrief Institute, Inc. Research Data Services team. Additionally, the study team would like to thank Harold Kooreman (HK) and Amber Blackmon (AB) for their assistance with annotations.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Due to the ethical considerations, data and detailed analyses generated during the course of this study are available upon request. Access is subject to review and approval by the applicable Privacy Officers of the participating institutions. Contact corresponding author for assistance with request.

REFERENCES

1. Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. *Thorax* 2014; 69 (9): 876–8.
2. Centers for Disease Control. Social Determinants of Health: Know What Affects Health. 2022. https://www.cdc.gov/socialdeterminants/index.htm. Accessed February 15, 2022.
3. Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. *Popul Health Manag* 2021; 24 (2): 222–30.
4. Kreuter MW, Thompson T, McQueen A, Garg R. Addressing social needs in health care settings: evidence, challenges, and opportunities for public health. *Annu Rev Public Health* 2021; 42 (1): 329–44.
5. HL7 International. Gravity Project. https://www.hl7.org/gravity/. Accessed December 1, 2020.
6. 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.
7. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, 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): 172–81.
8. Weed LL. Medical records that guide and teach. *N Engl J Med* 1968; 278 (12): 652–7.
9. Zander LI. Recording family and social history. *J R Coll Gen Pract* 1977; 27 (182): 518–20.
10. Podder V, Lew V, Ghassemzadeh S. SOAP notes. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2022: 1–5. http://www.ncbi.nlm.nih.gov/books/NBK482263/. Accessed July 1, 2022.
11. Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. *AMIA Annu Symp Proc* 2011; 2011: 227–36.
12. Patra BG, Sharma MM, Vekaria V, *et al.* Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc* 2021; 28 (12): 2716–27.
13. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. *JAMA Oncol* 2016; 2 (6): 797–804.
14. Reading Turchioe M, Volodarskiy A, Pathak J, Wright DN, Tcheng JE, Slotwiner D. Systematic review of current natural language processing methods and applications in cardiology. *Heart* 2022; 108 (12): 909–16.
15. Le Glaz A, Haralambous Y, Kim-Dufor DH, *et al.* Machine learning and natural language processing in mental health: systematic review. *J Med Internet Res* 2021; 23 (5): e15708.
16. Cohen GR, Friedman CP, Ryan AM, Richardson CR, Adler-Milstein J. Variation in physicians' electronic health record documentation and potential patient harm from that variation. *J Gen Intern Med* 2019; 34 (11): 2355–67.
17. Walsh C, Elhadad N. Modeling clinical context: rediscovering the social history and evaluating language from the clinic to the wards. *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 224–31.
18. Kotay A, Huang JL, Jordan WB, Korin E. Exploring family and social context through the electronic health record: physicians' experiences. *Fam Syst Health* 2016; 34 (2): 92–103.
19. 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: 537–46.
20. 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.
21. Zech J, Husk G, Moore T, Kuperman GJ, Shapiro JS. Identifying homelessness using health information exchange data. *J Am Med Inform Assoc* 2015; 22 (3): 682–7.
22. Hatef E, Singh Deol G, Rouhizadeh M, *et al.* Measuring the value of a practical text mining approach to identify patients with housing issues in the free-text notes in electronic health record: findings of a retrospective cohort study. *Front Public Health* 2021; 9: 697501.
23. Hatef E, Rouhizadeh M, Nau C, *et al.* Development and assessment of a natural language processing model to identify residential instability in electronic health records' unstructured data: a comparison of 3 integrated healthcare delivery systems. *JAMIA Open* 2022; 5 (1): ooac006.
24. Frederick TJ, Chwalek M, Hughes J, Karabanow J, Kidd S. How stable is stable? Defining and measuring housing stability: defining and measuring housing stability. *J Community Psychol* 2014; 42 (8): 964–79.
25. 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.
26. 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; 4 (3): ooaa069.
27. 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.
28. 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.
29. Han S, Zhang RF, Shi L, *et al*. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J Biomed Inform* 2022; 127: 103984.
30. Rouillard CJ, Nasser MA, Hu H, Roblin DW. Evaluation of a natural language processing approach to identify social determinants of health in electronic health records in a diverse community cohort. *Med Care* 2022; 60 (3): 248–55.
31. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011; 18 (5): 544–51.
32. Office of Disease Prevention and Health Promotion. *Healthy People 2030*. Washington, DC: US Department of Health and Human Services; 2022. https://health.gov/healthypeople.
33. National Alliance to End Homelessness. State of Homelessness: 2022 Edition. 2022. https://endhomelessness.org/homelessness-in-america/homelessness-statistics/state-of-homelessness/. Accessed February 13, 2023.
34. Vera L, Reed KK, Rose E, *et al*. Prevalence of housing insecurity in survivors of traumatic injury. *Am Surg* 2022; 88 (9): 2274–9.
35. D'Alessandro D, Appolloni L. Housing and health: an overview. *Ann Ig* 2020; 32(5 Supple 1): 17–26.
36. Stringhini S, Sabia S, Shipley M, *et al*. Association of socioeconomic position with health behaviors and mortality. *JAMA* 2010; 303 (12): 1159–66.
37. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep* 2014; 129 (Suppl 2): 19–31.
38. Burgard SA, Seefeldt KS, Zelner S. Housing instability and health: findings from the Michigan Recession and Recovery Study. *Soc Sci Med* 2012; 75 (12): 2215–24.
39. Sinclair RR, Cheung JH. Money matters: recommendations for financial stress research in occupational health psychology. *Stress Health* 2016; 32 (3): 181–93.
40. Consumer Financial Protection Bureau. Financial Well-Being Scale: Scale Development Technical Report. 2017. https://www.consumerfinance.gov/data-research/research-reports/financial-well-being-technical-report/. Accessed June 15, 2022.
41. Dooley D. Unemployment, underemployment, and mental health: conceptualizing employment status as a continuum. *Am J Community Psychol* 2003; 32 (1–2): 9–20.
42. McDonald CJ, Overhage JM, Barnes M, *et al*.; INPC Management Committee. The Indiana network for patient care: a working local health information infrastructure. *Health Affairs* 2005; 24 (5): 1214–20.
43. Schwartz BS, Stewart WF, Godby S, *et al*. Body mass index and the built and social environments in children and adolescents using electronic health records. *Am J Prev Med* 2011; 41 (4): e17–28.
44. Charlson ME, Charlson RE, Peterson JC, Marinopoulos SS, Briggs WM, Hollenberg JP. The Charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. *J Clin Epidemiol* 2008; 61 (12): 1234–40.
45. Health Leads. The Health Leads Screening Toolkit. 2018. https://healthleadsusa.org/resources/the-health-leads-screening-toolkit/. Accessed February 5, 2020.
46. National Association of Community Health Centers. PRAPARE Implementation and Action Toolkit. 2019. http://www.nachc.org/research-and-data/prapare/toolkit/. Accessed February 5, 2020.
47. Karttunen L, Chanod JP, Grefenstette G, Schille A. Regular expressions for language engineering. *Nat Lang Eng* 1996; 2 (4): 305–28.
48. Sai Prashanthi G, Deva A, Vadapalli R, Das AV. Automated categorization of systemic disease and duration from electronic medical record system data using finite-state machine modeling: prospective validation study. *JMIR Form Res* 2020; 4 (12): e24490.
49. Weiner M, Dexter PR, Heithoff K, *et al*. Identifying and characterizing a chronic cough cohort through electronic health records. *Chest* 2021; 159 (6): 2346–55.
50. Duke J, Chase M, Poznanski-Ring N, Martin J, Fuhr R, Chatterjee A. Natural language processing to improve identification of peripheral arterial disease in electronic health data. *J Am Coll Cardiol* 2016; 67 (13): 2280.
51. Regenstrief Institute. What is nDepth? 2022. https://www.regenstrief.org/real-world-solutions/ndepth/what-is-ndepth/. Accessed August 31, 2022.
52. Weerahandi HM, Horwitz LI, Blecker SB. Diabetes phenotyping using the electronic health record. *J Gen Intern Med* 2020; 35 (12): 3716–8.
53. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. *Pharmacotherapy* 2018; 38 (8): 822–41.
54. Feller DJ, Zucker J, Srikishan B, *et al*. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. *Appl Clin Inform* 2020; 11 (01): 172–81.
55. 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): 6.
56. Chang KW, Prabhakaran V, Ordonez V. Bias and fairness in natural language processing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019.
57. Institute of Medicine (U.S.), editor. Capturing Social and Behavioral Domains in Electronic Health Records. Phase 1. Washington, DC: The National Academies Press; 2014: 123p.
58. 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): 1037–43.
59. Fraze TK, Brewster AL, Lewis VA, Beidler LB, Murray GF, Colla CH. Prevalence of screening for food insecurity, housing instability, utility needs, transportation needs, and interpersonal violence by US physician practices and hospitals. *JAMA Netw Open* 2019; 2 (9): e1911514.
60. Berg M. Practices of reading and writing: the constitutive role of the patient record in medical work. *Sociol Health Illness* 1996; 18 (4): 499–524.
61. Bansler J, Havn E, Mønsted T, Schmidt K, Svendsen JH. Physicians' progress notes. In: Bertelsen OW, Ciolfi L, Grasso MA, Papadopoulos GA, eds. *ECSCW 2013: Proceedings of the 13th European Conference on Computer Supported Cooperative Work, 21–25 September 2013, Paphos, Cyprus*. London: Springer, London; 2013: 123–42. http://link.springer.com/10.1007/978-1-4471-5346-7_7. Accessed August 31, 2022.
62. Weiner SJ, Wang S, Kelly B, Sharma G, Schwartz A. How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *J Am Med Inform Assoc* 2020; 27 (5): 770–5.
63. Bardsley M, Billings J, Dixon J, Georghiou T, Lewis GH, Steventon A. Predicting who will use intensive social care: case finding tools based on linked health and social care data. *Age Ageing* 2011; 40 (2): 265–70.
64. Nijhawan AE, Clark C, Kaplan R, Moore B, Halm EA, Amarasingham R. An electronic medical record-based model to predict 30-day risk of readmission and death among HIV-infected inpatients. *J Acquir Immune Defic Syndr* 2012; 61 (3): 349–58.
65. Hao S, Jin B, Shin AY, *et al*. Risk prediction of emergency department revisit 30 days post discharge: a prospective study. *PLoS One* 2014; 9 (11): e112944.
66. Vest JR, Ben-Assuli O. Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *Int J Med Inform* 2019; 129: 205–10.
67. Vest JR, Menachemi N, Grannis SJ, *et al*. Impact of risk stratification on referrals and uptake of wraparound services that address social determinants: a stepped wedged trial. *Am J Prev Med* 2019; 56 (4): e125–33–e133.

68. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med* 2015; 48 (2): 215–8.
69. Bazemore AW, Cottrell EK, Gold R, *et al*. "Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc* 2016; 23 (2): 407–12.
70. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak* 2020; 20 (1): 280.