JAMIA Open, 2023, **6(4)**, ooad085
https://doi.org/10.1093/jamiaopen/ooad085
Research and Applications

Research and Applications

Application of natural language processing to identify social needs from patient medical notes: development and assessment of a scalable, performant, and rule-based model in an integrated healthcare delivery system

Geoffrey M. Gray, PhD[1], Ayah Zirikly, PhD[2], Luis M. Ahumada, PhD[1], Masoud Rouhizadeh, PhD[3], Thomas Richards, MS[4], Christopher Kitchen, MS[4], Iman Foroughmand, MD[4], Elham Hatef, MD, MPH*[,4,5]

[1]Center for Pediatric Data Science and Analytic Methodology, Johns Hopkins All Children's Hospital, St. Petersburg, FL, United States, [2]Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, United States, [3]Department of Pharmaceutical Outcomes and Policy, University of Florida College of Pharmacy, Gainesville, FL, United States, [4]Department of Health Policy and Management, Center for Population Health Information Technology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, [5]Division of General Internal Medicine, Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States

*Corresponding author: Elham Hatef, MD, MPH, Department of Health Policy and Management, Center for Population Health Information Technology, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Room 502, Baltimore, MD 21205, USA (ehatef1@jhu.edu)

Author Contributions: Geoffrey M. Gray and Ayah Zirikly are dual first authors.

Abstract

Objectives: To develop and test a scalable, performant, and rule-based model for identifying 3 major domains of social needs (residential instability, food insecurity, and transportation issues) from the unstructured data in electronic health records (EHRs).

Materials and Methods: We included patients aged 18 years or older who received care at the Johns Hopkins Health System (JHHS) between July 2016 and June 2021 and had at least 1 unstructured (free-text) note in their EHR during the study period. We used a combination of manual lexicon curation and semiautomated lexicon creation for feature development. We developed an initial rules-based pipeline (Match Pipeline) using 2 keyword sets for each social needs domain. We performed rule-based keyword matching for distinct lexicons and tested the algorithm using an annotated dataset comprising 192 patients. Starting with a set of expert-identified keywords, we tested the adjustments by evaluating false positives and negatives identified in the labeled dataset. We assessed the performance of the algorithm using measures of precision, recall, and *F*1 score.

Results: The algorithm for identifying residential instability had the best overall performance, with a weighted average for precision, recall, and *F*1 score of 0.92, 0.84, and 0.92 for identifying patients with homelessness and 0.84, 0.82, and 0.79 for identifying patients with housing insecurity. Metrics for the food insecurity algorithm were high but the transportation issues algorithm was the lowest overall performing metric.

Discussion: The NLP algorithm in identifying social needs at JHHS performed relatively well and would provide the opportunity for implementation in a healthcare system.

Conclusion: The NLP approach developed in this project could be adapted and potentially operationalized in the routine data processes of a healthcare system.

Lay Summary

We developed and tested an algorithm for identifying 3 major domains of social needs (residential instability, food insecurity, and transportation issues) from the free-text notes in electronic health records (EHRs). Thus, we included patients aged 18 years or older who received care at the Johns Hopkins Health System between July 2016 and June 2021 and had at least 1 note in their EHR during the study period. We developed keywords and phrases, which described the social needs, and developed natural language processing (NLP) algorithms that used those keywords to identify different social needs in free-text EHR. We assessed the performance of these algorithms and compared what they identified in the notes with what a human identified through a direct review of the notes. The algorithm for identifying residential instability had the best overall performance, the algorithm for identifying food insecurity performed relatively well but the transportation issues algorithm was the lowest overall performing metric. The NLP algorithms developed in this study would provide the opportunity for implementation in different healthcare systems and could be adapted and potentially operationalized in the routine data processes of the healthcare systems.

Key words: social needs; residential instability; food insecurity; transportation; natural language processing; free text.

Received: May 31, 2023; **Revised:** September 13, 2023; **Editorial Decision:** September 20, 2023; **Accepted:** September 21, 2023
© The Author(s) 2023. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Background and significance

Social needs and social determinants of health (SDOH) represent important indicators of clinical outcomes but they are infrequently documented in the electronic health records (EHRs) of healthcare systems.[1–6] Most of the available data in EHRs on social needs and SDOH challenges are currently documented as unstructured medical notes as opposed to structured data.[7] Some of the challenges to routine assessment and documentation of social needs and SDOH, particularly in a structured format in the EHR, include difficulty in recognition of social needs as part of disease etiology, assessment, and documentation burden on the healthcare providers, lack of reimbursement mechanisms by payers, and lack of identifiable coding systems such as the International Classification of Diseases 10th revision (ICD-10) for all social needs/SDOH.

Despite these limitations, there has been a recent effort across healthcare systems to assess, document, and design interventions that address social needs and SDOH challenges as part of the standard of the care process.[8] The impact of social needs and SDOH challenges on health outcomes coupled with the barriers to capturing them in structured EHRs highlights the need for developing methods to identify social needs and SDOH challenges in other forms of EHR documentation such as in the EHR unstructured data (ie, providers' free-text notes).

Machine learning (ML) techniques represent a promising approach for information retrieval from EHR clinical notes. However, traditionally unstructured EHR data have been less amicable to such ML techniques, due to the high level of variability in the encoding of the information compared to structured data.[9,10] To overcome these barriers, natural language processing (NLP) has been used to convert unstructured EHR data into ML insights.[11,12] Recent developments in NLP of unstructured EHR data have allowed for reliable, low-cost, and rapid extraction of information from EHRs. However, most of these efforts have been limited to research settings.[13–15] Developing NLP algorithms that could be operationalized in routine data processes of a healthcare system would improve the application of such methods in extracting social needs from the EHR's unstructured data.

This article presents the application of NLP techniques and text mining to identify patients' social needs in EPIC-based EHR of a multilevel academic healthcare system that provides both inpatient and outpatient care to patients with varying social needs and SDOH challenges across Maryland.[5–7] Here we developed and tested a scalable, performant, and rule-based model for the identification of 3 major domains of social needs namely residential instability (ie, homelessness and housing insecurity), food insecurity, and transportation issues. These domains were identified as the top priorities for social needs screening and referral initiatives at JHHS as they were among the most common social needs in the JHHS patient population.

Methods

Study design

We included patients who were 18 years of age and older, who received care at the Johns Hopkins Health System (JHHS) from July 2016 to June 2021, and who had at least 1 free-text note in their EHR during the study period. Table S1 defines the selected social needs assessed in the study and provides examples of how these needs were documented in the EHR. The study protocol was reviewed and approved by the Institutional Review Board at the Johns Hopkins University School of Public Health.

Keyword development

For feature development, we employed both manual lexicon curation and semiautomated lexicon creation methods.[15] To develop hand-crafted linguistic patterns, a team of subject matter experts (eg, preventive medicine, primary care, and geriatric physicians, former social workers/care managers, and health disparities experts) at JHHS reviewed codes and phrases related to each social needs domain in ICD-10, Current Procedural Terminology, Logical Observation Identifiers Names and Codes, and Systematized Nomenclature of Medicine—Clinical Terms terminologies.[16,17] We used a comprehensive list of codes related to different domains of social needs, which was extracted in a systematic search of the medical vocabularies by the Social Interventions Research and Evaluation Network at the University of California, San Francisco to help interested stakeholders more easily identify existing related codes.[16] The expert team also reviewed the description of selected social needs in public health surveys and instruments such as the Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences and similar surveys.[18–21] Furthermore, our expert team reviewed phrases derived from a literature review and the results of a manual annotation process from a past study.[22,23] To finalize the linguistic patterns, the expert team developed a comprehensive list of all available codes (in the absence of ICD-10 codes for transportation issues, our expert team relied on other medical vocabularies, public health surveys, and expert input), specific content areas, and phrases for each social needs domain. Then matched them across different coding systems and developed several phrases and synonyms to describe each content area.

Following the development of a curated manual lexicon, we applied a semiautomated approach using ngram, keyword matching, and statistical analysis to refine the curated keywords. We extracted the ngrams containing the matched phrases in an annotated dataset to further understand the context of the keywords and to reduce false positives and negatives. We moved the ambiguous keywords to the dataset comprised of keywords that potentially indicated social needs and used the median count from ICD-10/questionnaire-identified patients to determine the cut-off for the keyword (the details of the process are described in the following).

Developing the validation dataset

We identified 1 879 626 patients with an encounter from July 1, 2016, through June 30, 2021, and aged 18+ at the time of encounter. This was reduced to 1 317 335 patients who had a valid Maryland jurisdiction code required for matching. From these, we identified 5502 patients with an ICD-10 diagnosis code for residential instability (homelessness/housing insecurity) or food insecurity. From the remaining patients with no ICD-10 codes for residential instability or food insecurity, we produced a matched (on age, gender, and place of living) control sample of 5502 who also had 1+ free-text notes. We assigned a positive label (1) to those with a relevant ICD-10 code for residential instability or food insecurity and a

negative label (0) to patients without relevant ICD-10 codes for the social need of interest.

Since there is no specific ICD-10 code related to transportation issues, we assessed over 362 million EHR questionnaire and flowsheet data points (responses to questions in different questionnaires and flowsheets) available in the EHR structured data by searching *transportation* keyword. There were 3156 flowsheet templates with a total of 93 102 questions and we identified 7 questions related to transportation issues and 8024 relevant responses addressing transportation issues. We identified 1790 patients with positive responses and 4364 patients with only negative responses to questions related to transportation issues. These patients were all added to the dataset. We assigned a positive label (1) to patients with a response indicating an unmet transportation issue (ie, an indication of an existing transportation issue identified during the encounter) and a negative label (0) to patients with a response indicating no current transportation issues.

We extracted provider notes occurring ±45 days of the encounter with relevant ICD-10 codes or questionnaire's date and linked them to each relevant ICD-10 or questionnaire result. If multiple ICD-10 codes or questionnaires were available within 45 days of each other, we assigned the overlapping provider notes to the latest ICD-10 or questionnaire date. If any text of the selected questionnaires was identified in the provider's notes (some providers may copy/paste a questionnaire into a note), we excluded the text to ensure it would not impact the performance of our NLP algorithms. We did not have any limitations in selecting the provider notes and only excluded lab results and radiology and pathology reports.

We performed manual annotation on the notes for a small sample of patients. Thus, we randomly selected 150 patients, 50 patients with any ICD-10 diagnosis code of residential instability, 50 with any ICD-10 diagnosis code of food insecurity, and 50 with transportation issues indicated on the flowsheets. We also selected a matched sample of 42 patients with no ICD-10 codes or flowsheets of interest. This process resulted in 192 patients, and we extracted providers' free-text notes for all those patients. A team of trained and experienced researchers performed manual annotation independently to determine the status of social needs of interest for the selected study sample and highlighted mentions of social needs in the notes using a customized version of SynWrite, a freely available text editing tool.[24]

The experienced researchers used the definitions of the selected social needs and examples of how these needs were documented in the EHR (Table S1) as a guide to identifying social needs of interest and used their expert judgment to assign a positive or negative label. Each note was reviewed by 1 experienced researcher. While in doubt about assigning a social needs label, the researchers reviewed the case with the study senior author/social needs expert and the team made a final decision. Patients with a provider note indicating an unmet social need (ie, an indication of an existing social need identified during the encounter) were assigned a positive label (1), while patients with any provider note indicating no current social need were assigned a negative label (0). The study team reviewed and resolved any discrepancies between social needs labels generated from ICD-10 codes and the results of the manual annotation.

Note processing

We performed the following note-processing steps to clean and normalize data: (1) cleaning special and nonword or digital characters (eg, removing the *dot-phrase* or segments with extraneous formatting characters that may interfere with model performance), (2) spell checking and correction for mistyped, misspelled, or concatenated words detected during the NLP development process in previous studies, (3) sentence separation, and (4) tokenization (ie, segmenting text into linguistic units such as words and punctuation).[25] We did not use any section identification, left the note sections undivided, and searched the entire provider note for NLP-ML model development as our clinical experts recommended not to do so. The rationale was that social needs can be identified in any part of the notes and focusing on specific sections (eg, social history) might result in missing some information.

Rules-based matching algorithm

We developed an initial rules-based pipeline (Match Pipeline) utilizing 2 sets of keywords for each social needs domain. The first set comprised keywords that definitively indicated the presence of a social need, while the second set comprised keywords that potentially indicated social needs. We assigned a social need flag if any keyword from the first set was present, while a positive flag from the second set required exceeding a predetermined cut-off value. Because of the possible presence of keywords from the second set in patients without social needs, we determined a cut-off value from patients with ICD-10 or questionnaire identifiers for social needs. We obtained a cut-off value using the median count of the keywords in the notes of patients with ICD-10 identifiers indicating residential instability (ie, homelessness and housing insecurity) and food insecurity, as well as the notes of patients with positive indicators of transportation issues obtained from the questionnaires. We matched the potential keywords to the notes of patients with an ICD-10 code or positive indicator in a questionnaire for each social need and used the median of the total number of matched keywords as the cut-off. For example, we matched the list of potential keywords describing housing insecurity to the subset of patients with ICD-10 identifiers for residential instability-housing insecurity (Table S2). We calculated the median number of matches per patient for the total subset and used it as the cut-off value. This allowed for a data-driven approach to determining the appropriate frequency of these keywords in patients with social needs. We did not assess the temporality of social needs occurrence and did not assess the text for negations.

Full pipeline architecture

Preliminary analysis of the patient notes indicated the presence of some semistructured text of selected questionnaires in the *Social History* section of some providers' notes as some providers copy/pasted a questionnaire into a note (eg, questions containing phrases such as: *Food insecurity Worry:* and *Transportation needs Medical:* and responses set to *not on file* indicating that no social needs were present). However, the relative simplicity of a basic matching scheme resulted in false positives. To address this issue, we built a more complex rule-based pipeline and subdivided the corpus into 2 components, 1 containing the *Social History* section (semistructured) and the other containing all remaining notes (unstructured). We structured the rules-based keyword pipeline as described

above and processed the semistructured notes by extracting the values recorded for each of the false positive phrases identified above. We then combined the results, with any positive indicator (either keywords from the text or values recorded in the semistructured text) acting as a flag for social needs of interest. Figure 1 presents the full pipeline architecture.

Algorithm development

We performed rule-based keyword matching for distinct lexicons for each social needs domain and validated the models using the annotated dataset of 192 patients. Starting with a set of expert-identified keywords, we modified the algorithm and tested the adjustments by evaluating false positives and negatives identified in the annotated dataset. The original keyword dataset was used since performance gains were not sufficient to offset concerns about overfitting to the annotated dataset. This process resulted in 9 definite and 17 probable keywords for housing insecurity; 101 definite and 21 probable keywords for homelessness; 52 definite and 33 probable keywords for food insecurity; and 270 definite and 125 probable keywords for transportation. After final adjustments in the algorithm, we applied the keywords to the dataset of the study population at large containing 63 352 328 notes to identify patients with social needs of interest. We performed the analysis at the patient level, using Microsoft Azure Databricks and Spark NLP.[26] Full development of the pipeline extraction is shown in Figure 2.

We used Spark, a powerful tool for highly parallel processing and effective distribution of large datasets across multiple processors to design the pipeline.[27] This was achieved using an extension of the MapReduce algorithm. This algorithm consisted of 3 steps, (1) applying a map function to local data and storing the output keys; (2) shuffling the data with the same keys to the same worker node; (3) reducing the data by processing it in parallel on worker nodes.[28] Spark extended the MapReduce algorithm using resilient distributed datasets, which better captured a wide variety of data sources.

Machine learning model development

While the proposed rules-based approaches could be effective, more advanced ML techniques may be required depending on the complexity of the patient notes. The modular structure of our pipeline allowed the processing of various sections of the free-text notes using different approaches. Thus, we developed and tested ML models to identify patients with social needs by leveraging the validation dataset of manually annotated notes of 192 patients and compared the performance of the relatively simple yet highly efficient rules-based keyword-matching algorithms with the more sophisticated ML models. Because of the scale of the data to be processed and the research questions of interest, we focused on developing 2 logistic regression models and explored the use of context-based transformer models, namely ClinicalBERT,[29] and deferred exploration of more recent models, such as GPT-3, to future studies.

We used 3-fold cross-validation on our dataset to account for its size and reduce potential bias that may exist in any sample, with a split of 70/10/20 for training, development, and testing datasets. The dataset showed a class imbalance, with majority-class-to-minority class ratios ranging from 1.56 to 5.62. To lessen its impact, we applied class weight balancing. The classification was applied at the patient level, as opposed to the note level, where patients with any annotated social needs were grouped in the positive class. Two models were built: the notes-based (NB) model and the terminologies-based (TB) model. For the NB model, we calculated the term frequency-inverse document frequency (TF-IDF) of the normalized text in its surface, tokenized, and lemmatized form. Then, we developed a multilabel logistic regression model with an L2 regularizer to classify patients with social needs.

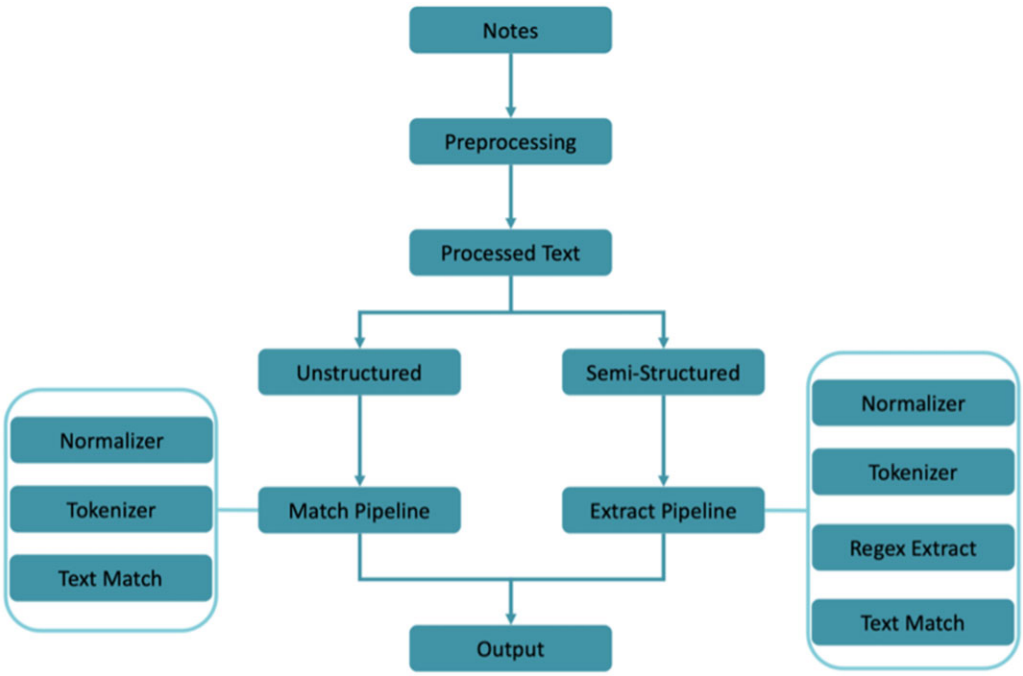

Figure 1. Pipeline architecture for processing of semistructured and unstructured notes.

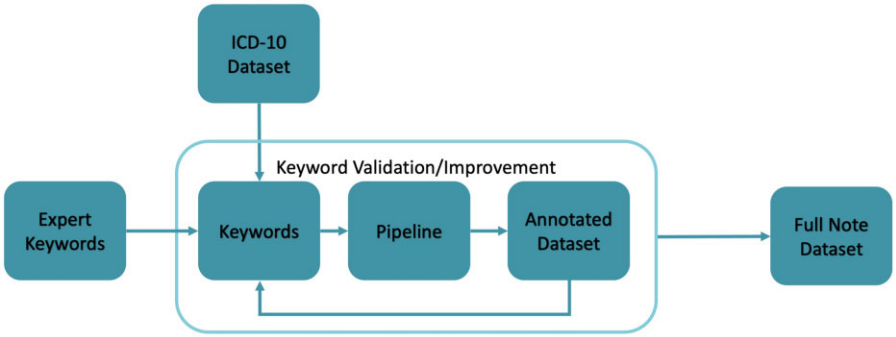

Figure 2. Keyword development process for each pipeline.

We assigned social needs labels on the patient level in the following mechanism: if the patient had a note with a social need (eg, homelessness or food insecurity), then the label for that social need was 1/enabled. Otherwise, if the patient did not have any notes with a social need then the label for the patient was set to 0. This approach did not consider the temporal aspect of the label where a patient's social need may not exist within a different timeframe. We selected patient-level assignment as such an approach would be most beneficial if the model was integrated into the clinical settings. For the TB model, we developed a multilabel logistic regression model with the same parameters as the NB model while using the context window around expert-identified keywords as input. We used the same list of keywords as the rule-based model and experimented with multiple different window sizes between 3 and 10 tokens before and after the term and found that 4 yielded the best performance. For our BERT-based model, we fine-tuned ClinicalBERT29 on our task and followed the same approach and hyperparameters proposed by experts in similar studies.[30]

Table 1. Performance metrics for identifying social needs using a rules-based approach and the annotated dataset for validation.

	Precision	Recall	*F*1 score	No.[a]
Residential instability—homelessness				
Not homeless	0.99	0.82	0.90	163
Homeless	0.49	0.97	0.65	29
Weighted average	0.92	0.84	0.92	192
Residential instability—housing insecurity				
No housing insecurity	0.81	0.99	0.89	141
Housing insecurity	0.90	0.37	0.53	51
Weighted average	0.84	0.82	0.79	192
Food insecurity				
No food insecurity	0.93	0.74	0.83	148
Food insecurity	0.49	0.82	0.61	44
Weighted average	0.83	0.76	0.78	192
Transportation issues				
No transportation issues	0.97	0.56	0.71	117
Transportation issues	0.59	0.97	0.73	75
Weighted average	0.82	0.72	0.72	192

[a] Number of patients in the annotated dataset with and without the social need of interest.

Performance evaluation

To assess the performance of our rule-based and the ML models, we calculated the performance metrics for each class (those with and without a social need) and reported precision (positive predictive value), recall (sensitivity), and *F*1 score (the harmonic mean of precision and recall) for each social needs domain using the annotated dataset of 192 patients. We also reported the overall weighted average for each parameter (the average of the metrics for each class weighted by the fraction of patients in that class). The overall metrics indicated the performance of the model in the larger population when encountering patients without social needs, and how well the models correctly sorted patients into having a social need or not (how well the keywords filtered patients into each group). Metrics for the positive social needs group demonstrated performance within that specific group.

A public GitHub repository contains all the details including the rule-based pipeline at https://github.com/ggray15jh/sdoh_pipeline.git.

Results

Rules-based model performance

Table 1 presents the performance metrics of the final rule-based algorithm using the annotated dataset. Among social needs domains, the residential instability-homelessness algorithm had the best overall performance, with a weighted average for precision, recall, and *F*1 score of 0.92, 0.84, and 0.92, respectively. The algorithm for patients classified as with residential instability-housing insecurity had a weighted average for precision, recall, and *F*1 score of 0.84, 0.82, and 0.79, respectively. Metrics for the food insecurity algorithm remained high with a weighted average for precision, recall, and *F*1 score of 0.83, 0.76, and 0.78, respectively. The transportation issues algorithm was the lowest overall performing metric, with a weighted average for precision, recall, and *F*1 score of 0.82, 0.72, and 0.72, respectively. Using the *F*1 score, the overall ranking of algorithm performance for the positive social needs group was: (1) transportation issues; (2) residential instability-homelessness; (3) food insecurity; and (4) residential instability-housing insecurity.

Machine learning model performance

Table 2 presents the performance metrics of the NB, TB, and BERT-based models using the annotated dataset. For NB and TB, we only reported the results on the lemmatized text since it yielded the best performance across all social needs domains. The residential instability-homelessness NB, TB, and BERT-based models had the best overall performance, with a weighted average for precision, recall, and *F*1 score of

Table 2. Performance metrics for identifying social needs using a machine learning approach and the annotated dataset for validation.

	Precision	Recall	*F*1 Score	No.[a]
Note-based model				
Residential instability—homelessness				
Not homeless	0.88	0.93	0.91	163
Homeless	0.81	0.62	0.68	29
Weighted average	0.86	0.86	0.85	192
Residential instability—housing insecurity				
No housing insecurity	0.71	0.58	0.63	141
Housing insecurity	0.59	0.70	0.64	51
Weighted average	0.65	0.64	0.63	192
Food insecurity				
No food insecurity	0.75	0.84	0.79	148
Food insecurity	0.70	0.56	0.62	44
Weighted average	0.73	0.73	0.72	192
Transportation issues				
No transportation issues	0.68	0.43	0.52	117
Transportation issues	0.71	0.86	0.77	75
Weighted average	0.69	0.69	0.68	192
Terminology-based model				
Residential instability—homelessness				
Not homeless	0.81	0.97	0.88	163
Homeless	0.76	0.41	0.53	29
Weighted average	0.81	0.80	0.78	192
Residential instability—housing insecurity				
No housing insecurity	0.63	1.00	0.77	141
Housing insecurity	1.00	0.30	0.46	51
Weighted average	0.80	0.68	0.63	192
Food insecurity				
No food insecurity	0.65	0.96	0.77	148
Food insecurity	0.31	0.10	0.15	44
Weighted average	0.66	0.65	0.58	192
Transportation issues				
No transportation issues	0.42	0.96	0.58	117
Transportation issues	0.88	0.20	0.31	75
Weighted average	0.70	0.48	0.41	192
BERT-based model				
Residential instability—homelessness				
Not homeless	0.88	0.93	0.91	163
Homeless	0.81	0.62	0.68	29
Weighted average	0.87	0.88	0.88	192
Residential instability—housing insecurity				
No housing insecurity	0.72	0.63	0.64	141
Housing insecurity	0.64	0.74	0.67	51
Weighted average	0.7	0.66	0.65	192
Food insecurity				
No food insecurity	0.78	0.84	0.8	148
Food insecurity	0.73	0.58	0.64	44
Weighted average	0.77	0.78	0.76	192
Transportation issues				
No transportation issues	0.71	0.46	0.55	117
Transportation issues	0.72	0.88	0.79	75
Weighted average	0.71	0.62	0.64	192

[a] Number of patients in the annotated dataset with and without the social need of interest.

0.86, 0.86, 0.85 and 0.81, 0.80, and 0.78 for NB and TB models, respectively, and 0.87, 0.88, and 0.88 for the BERT-based model. For both NB and BERT-based models using the *F*1 score, the overall ranking of model performance for the positive social needs group was: (1) transportation issues; (2) residential instability-homelessness; (3) residential instability-housing insecurity; and (4) food insecurity. For TB models, the overall ranking was: (1) residential instability-homelessness; (2) residential instability-housing insecurity; (3) transportation issues; and (4) food insecurity.

Cohort analysis using provider notes for patient population at large

We applied the rule-based model to provider notes for the patient population at large, assessing the social needs in 5 separate cohorts of the study population (Table 3). We identified 3.75% of the population with residential instability, 28.97% with food insecurity, and 1.33% with transportation issues.

Discussion

Our work represents the development of a novel rules-based NLP algorithm for identifying different domains of social needs from the EHR free-text notes with the possibility of operationalizing and deploying it in a healthcare system. The performance of the algorithm was validated on a human-labeled dataset comprising 192 patients. Subsequently, the tool was applied to a population-level dataset obtained from the EHR to determine the prevalence of selected social needs in the larger population.

Dataset performance and comparison with previous studies

Our rule-based algorithm had satisfactory performance across the social domains (Table 1). The algorithm for residential instability-homelessness was the best performing overall, while the transportation needs algorithm performed the best in the positive social group. Residential instability-housing insecurity had an overall comparable precision and recall, suggesting a similar number of false positives and false negatives. This finding indicated that the algorithm was unlikely to overestimate or underestimate the population with residential instability, and the obtained population values were not biased. Overall, food insecurity, residential instability-homelessness, and transportation issues algorithms all had higher precision than recall. This indicated that these models tended to produce more false negatives than false positives. Thus, the values obtained in the population may indicate the lower bound for social needs.

We also applied the rule-based algorithm to provider notes for the patient population at large (Table 3). Values for residential instability obtained from the overall population were likely to be close to the true values. However, due to the higher number of false positives and false negatives for food insecurity and transportation issues, respectively, these values likely represented upper and lower bounds for the total JHHS population. Although the true prevalence of the social needs in the JHHS population was not available, we identified social needs in the structured EHR for our study population using available ICD-10 codes (ie, 0.40% for residential instability, 0.10% for food insecurity, and 1.9% for transportation). The comparison of the findings from free-text EHR with the identified social needs in the structured data presented the free-text notes as a rich source of information on patient's social needs.

Like other studies, we used a manual approach for the development of key phrases and a rule-based approach for the identification of those phrases in free-text notes.[15] The manual development of key phrases for the rule-based algorithm introduced potential subjectivity and bias. To mitigate these issues, we applied the semiautomated approach using ngram, keyword matching, and statistical analysis to refine the keywords. The performance of our algorithm was

Table 3. Frequency of social needs in the patient population at large using the rule-based model to identify provider notes with social needs of interest.

2016-2017 ($N = 657\,234$)	2017-2018 ($N = 661\,934$)	2018-2019 ($N = 666\,647$)	2019-2020 ($N = 656{,}211$)	2020-2021 ($N = 679\,018$)
Residential instability—homelessness				
2.35%	3.19%	3.47%	4.19%	4.94%
Residential instability—housing insecurity				
0.19%	0.18%	0.18%	0.20%	0.29%
Food insecurity				
0.17%	0.28%	0.27%	0.60%	14.08%
Transportation issues				
6.58%	7.37%	7.48%	7.69%	18.86%

comparable to previous studies. For example, Conway et al.[13] tested the performance of Moonstone, a new, highly configurable rule-based clinical NLP system for extraction of information requiring inferencing from clinical notes derived from the Veterans Health Administration. Their system achieved a precision of 0.66 (comparable with the precision of 0.49-0.90 for identifying patients with homelessness and housing insecurity in our study) and a recall of 0.87 (comparable with the recall of 0.37-0.97 for identifying patients with homelessness and housing insecurity in our study) for phrases related to homeless and marginally housed. Dorr et al.[14] extracted the phenotypic profiles for 4 key psychosocial vital signs, including housing insecurity or homelessness from EHR data. Their system achieved a precision of >0.90 in all psychosocial vital signs except for social isolation. Lastly, our team tested the performance of a rules-based NLP to extract mentions of residential instability (ie, homelessness and housing insecurity) from EHRs at JHHS, Kaiser Permanente Mid-Atlantic States, and Kaiser Permanente Southern California. The NLP algorithm demonstrated moderate precision (0.45, 0.73, and 1.0) at 3 sites, and the sensitivity and specificity of the algorithm varied across 3 sites (sensitivity: 0.68, 0.85, and 0.96; specificity: 0.69, 0.89, and 1.0).[7]

Pipeline scalability

Because of the size of the intended data processing, we designed our pipeline from the beginning to be scalable and efficient when handling large datasets. This involved identifying suitable technologies such as Spark[27] that were capable of processing large amounts of information effectively. Spark[27] facilitated the locality of data, which reduced communication overhead and facilitated better scaling. Spark extension of the MapReduce algorithm helped to better capture a wide variety of data sources. This extension had the further benefit of allowing multiple operations to be performed over the same dataset, whereas previous architectures required transfer between engines, which increased computational overhead. Spark NLP, an enterprise-grade NLP tool, leveraged Spark processing for highly scalable, efficient applications. One advantage of both of these technologies was the highly efficient scaling, which allowed more processors to be added to a task to handle changes in data volume. Our approach of initially designing the pipeline to assure sufficient performance allowed for deployment without substantial modifications and enabled the processing of larger datasets that would otherwise not be achievable.

Moreover, the performance of this pipeline on provider notes for the patient population at large indicated that it was appropriate and sufficient for deployment as a clinical decision-support tool to identify patients with social needs in a healthcare system. While our previously designed pipelines required several weeks of processing time (data not shown), our new pipeline processed 733 notes/second requiring approximately 24 h to complete the task and could conceivably handle ~32 000 000 notes in a 12-h period, which should be sufficient for most clinical settings. Thus, our pipeline architecture would be sufficiently performant for direct clinical deployment and scalable for the retrospective analysis of historical notes.

Clinical impact

While some new EHR-based questionnaires and flowsheets have facilitated documentation of social needs and SDOH challenges in the structured EHRs, their application only has contributed to the growing documentation burden of healthcare providers. An automated NLP tool could support better identification of patients with social needs without substantially increasing the documentation burden of clinicians or the need for manual chart review. Rule-based keyword matching using subject matter expert-defined phrases is a promising approach to identifying patients with social needs. While several well-developed NLP systems have presented variable levels of success in extracting social needs information from EHR free-text notes, all these attempts have been mostly limited to selected datasets for research purposes. To the best of our knowledge, our study is one of the first attempts to develop and test a scalable, performant, and rule-based model for the identification of social needs that could be operationalized in routine data processes of a multilevel academic healthcare system. The development of NLP techniques that extract data from unstructured EHRs and can be operationalized in daily activities of healthcare systems, would result in the identification of patients at risk and assist providers in focusing their resources on addressing the needs of medically underserved patients.

Machine learning models

Our ML models presented promising results. More specifically, comparing the results between the NB and TB models revealed better performance for the model that used the full free-text notes (ie, NB model) as the input for all domains of social needs. Furthermore, using context-based transformer models yielded the best results among all 3 models. The generalizability of these findings may be limited considering the small dataset and the potential bias in the findings. However, the findings implied that the rest of the content in the free-text notes may be beneficial to improve the model performance.

To further analyze the most salient features, we applied local interpretable model-agnostic explanations (LIME) to interpret our model and identify the most predictive features in our NB model.[31] We picked NB since logistic regression models tend to be more interpretable than neural network models. For residential instability-homelessness, the model used tokens such as *homeless*, *shelter*, and *schizophrenia*. Although *schizophrenia* seemed, at first, less relevant, the association between mental health and the presence of social needs has been documented in the literature.[32] Analyzing the frequency of the tokens for the terms produced by LIME revealed that the frequency for terms such as *shelter* in free-text notes of patients with social needs was significantly higher than those without a social need. For instance, the term *shelter* occurred 5 times and *schizophrenia* occurred 21 times more often in free-text notes of patients with *homelessness*. For residential instability-housing insecurity, the model was weighing on medication-related information such as *tablet* and *oral* which may seem irrelevant to residential instability. However, it is worth noting that the association of medication adherence with social needs such as residential instability and food insecurity has been reported in other studies,[33] which may explain higher weight assignment to these features in the model.

We compared the results of our NB model to the rule-based model, noting that the comparison was not exact since the validation data for the NB model was a subset of the data for the rule-based model. When comparing the *F1* score for each of the social needs domains with label enabled (existence of the social need of interest), we noted that the NB model performed better. And the difference in the precision and recall for different social needs domains in the NB model was not as large as the ones in the rule-based models. To further analyze these findings, we generated the confusion matrices of each social needs domain for the NB model (Figure S1). The model robustly generated false negatives, as an important aspect of a classifier focusing on identifying patients with social needs which are less frequently documented events in the medical records. The results of the rule-based and ML models presented the opportunity to integrate these 2 systems into an ensemble model to achieve higher performance.

Limitations

Our study had several limitations. Our rule-based and ML models were validated using a human-labeled dataset comprising 192 patients, which may not fully represent the diversity of social needs in the larger population. While we applied our rule-based model to a population-level dataset obtained from the EHR to determine the prevalence of selected social needs, the true prevalence of social needs in the population was not available, and our findings may represent upper and lower bounds due to the potential for false positives and false negatives. Also, our models did not include negation identification, which may have impacted their ability to distinguish between true positive and false positive instances. However, we believe this limitation did not significantly impact the scalability and utility of the algorithms. In our manual annotation of free-text notes, we found negation as a rare approach to the documentation of social needs in the free-text EHR. The screening and documentation of social needs was not a common or standard process in our healthcare system (or other healthcare systems) at the time of performing this study. Thus, very few providers would ask for or document the existence of a social need, and documenting the absence of a social need was a rare practice.

Furthermore, our rule-based algorithm had higher precision than recall for food insecurity, residential instability-homelessness, and transportation issues algorithms, indicating that false negatives may be more common. This finding suggested that our algorithm may underestimate the true prevalence of social needs, and the values obtained may indicate a lower bound. Additionally, we utilized a manual approach for the development of key phrases and a rule-based approach for the identification of those phrases in free-text notes, which may introduce subjectivity and potential bias in the results.

Finally, social needs and SDOH challenges are often poorly documented in free-text notes, which may result in incomplete or inaccurate identification of social needs using our models. Moreover, our dataset included the first one and a half years of the COVID-19 pandemic, where social distancing protocols (eg, stay-at-home orders) unprecedently limited transportation, and healthcare access, among other factors, which significantly impacted the documentation of such information in the EHRs. These limitations may highlight the need for improved documentation of social needs in healthcare settings to enhance the accuracy and validity of rule-based and ML models.

Conclusion

Our work presents the development of a novel model for identifying social needs from free-text notes with the possibility of operationalizing and deploying it in a healthcare system. Our rule-based algorithm showed satisfactory performance across selected social domains and our NB ML model performed robustly in generating false negatives and could be utilized as a proper classifier focusing on identifying patients with social needs. Our approach to the design of a scalable and efficient pipeline using suitable technologies could enable the effective processing of large amounts of information and the potential for deployment without substantial modifications. Our rule-based and ML models could be adapted and potentially integrated into an ensemble model for higher performance as a tool to efficiently identify social needs in EHR free-text notes for targeted interventions among patients with such social needs. Future research should further investigate the generalizability of these models using larger and more diverse datasets to ensure their effectiveness across different patient populations. External validation of the models, using data from different healthcare systems, would also enhance the reliability of the models and their potential applicability in diverse settings.

Author contributions

Study concept and design: G.M.G., A.Z., L.M.A., and E.H. NLP and ML algorithm development and testing: G.M.G., A.Z., L.M.A., and E.H. Interpretation of Results: G.M.G., A.Z., L.M.A., M.R., T.R., C.K., and E.H. Drafting of the manuscript: G.M.G., A.Z., M.H., T.R., C.K., I.F., and E.H. Critical revision of the manuscript for important intellectual content: G.M.G., A.Z., and E.H. Administrative, technical, and material support: L.M.A., M.R., T.R., C.K., I.F., and E.H. Study supervision: E.H.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by a grant from National Institute on Minority Health and Health Disparities (NIMHD, Grant Number R01MD015844-01). Its contents are solely the responsibility of the authors and do not necessarily represent the official NIMHD views.

Conflicts of interest

None declared.

Data availability

The data underlying this article were extracted from the electronic health record at the study site and cannot be shared publicly for the privacy of individuals that participated in the study. A public GitHub repository contains all the details of the rule-based pipeline developed in this study. https://github.com/ggray15jh/sdoh_pipeline.git.

References

1. Hatef E, Ma XM, 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):222-230.
2. Hatef E, Kharrazi H, Nelson K, *et al.* The association between neighborhood socioeconomic and housing characteristics with hospitalization: results of a national study of veterans. *J Am Board Fam Med*. 2019;32(6):890-903.
3. Hatef E, Searle KM, Predmore Z, *et al.* The impact of social determinants of health on hospitalization in the veterans health administration. *Am J Prev Med*. 2019;56(6):811-818.
4. Hatef E, Rouhizadeh M, Lasser TI, Hill-Briggs E, Marsteller F, Kharrazi JH. 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 Inf*. 2019;7(3):14.
5. Berkowitz SA, Basu S, Meigs JB, Seligman HK. Food insecurity and health care expenditures in the United States, 2011-2013. *Health Serv. Res*. 2018;53(3):1600-1620.
6. Kushel MB, Gupta R, Gee L, Haas JS. Housing instability and food insecurity as barriers to health care among low-income Americans. *J Gen Intern Med*. 2006;21(1):71-77.
7. Hatef ER, Nau C, Xie F, *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.
8. Byhoff E, Gottlieb LM. When there is value in asking: an argument for social risk screening in clinical practice. *Ann Intern Med*. 2022;175(8):1181-1182.
9. Xiao C, Choi E, Sun JM. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inf Assoc*. 2018;25(10):1419-1428.
10. Li I, Pan J, Goldwasser J, *et al.* Neural natural language processing for unstructured data in electronic health records: a review. *Comput Sci Rev*. 2022;46:29.
11. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. *Trends Anaesth Crit Care*. 2021;38:4-9.
12. Juhn Y, Liu HF. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol*. 2020;145(2):463-469.
13. Conway M, Keyhani S, Christensen L, *et al.* Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *J Biomed Semant*. 2019;10(1):10.
14. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying patients with significant problems related to social determinants of health with natural language processing. *Stud Health Technol Inform*. 2019;264:1456-1457.
15. 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 Inf Assoc*. 2021;28(12):2716-2727.
16. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Compendium of medical terminology codes for social risk factor, 2018. Accessed January 30, 2023. https://sirenetwork.ucsf.edu/tools-resources/resources/compendium-medical-terminology-codes-social-risk-factors
17. Richard M, Krebs AX, Charlet MO. Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts. *Stud Health Technol Inform*. 2015;210:221-223.
18. Bureau TUSC. American Housing Survey (AHS). Accessed January 30, 2023. https://www.census.gov/programs-surveys/ahs.html
19. Bureau TUSC. American Community Survey (ACS). Accessed January 30, 2023. https://www.census.gov/programs-surveys/acs/
20. Centers. NAoCH. The protocol for responding to and assessing patients' assets, risks, and experiences (PRAPARE). Accessed January 30, 2023. http://www.nachc.org/research-and-data/prapare/
21. Alley DE, Asomugha CN, Conway PH, Sanghavi DM. Accountable health Communities - addressing social needs through Medicare and Medicaid. *N Engl J Med*. 2016;374(1):8-11.
22. 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.
23. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr*. 2017;17(1):248.
24. SynWrite A free text and source code editor. Accessed September 28, 2023. https://cudatext.github.io/synwrite/
25. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Research Repository; 2017. Accessed September 28, 2023. https://sborms.github.io/econometrics-meets-sentiment/
26. John Snow Labs State of the Art Natural Language Processing in Python. Accessed September 28, 2023. https://nlp.johnsnowlabs.com
27. Zaharia M, Xin RS, Wendell P, *et al.* Apache spark: a unified engine for big data processing. *Commun ACM*. 2016;59(11):56-65.
28. Dean J, Ghemawat S. Mapreduce: simplified data processing on large clusters. *Commun ACM*. 2008;51(1):107-113.
29. Alsentzer EM, Boag W, Weng WH, Jin D, Naumann T, McDermott M. 2019. Publicly available clinical BERT embeddings, arXiv, arXiv preprint arXiv:190403323, preprint: not peer reviewed.
30. Kexin HA, Ranganath R. 2019. Clinicalbert: modeling clinical notes and predicting hospital readmission, arXiv, arXiv preprint arXiv:190405342, preprint: not peer reviewed.
31. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California: Association for Computing Machinery; 2016: 1135-1144.
32. Lin D, Kim H, Wada K, *et al.* Unemployment, homelessness, and other societal outcomes among US veterans with schizophrenia relapse: a retrospective cohort study. *Prim Care Companion CNS Disord*. 2022;24(5):21m03173.
33. Wilder ME, Kulie P, Jensen C, *et al.* The impact of social determinants of health on medication adherence: a systematic review and meta-analysis. *J Gen Intern Med*. 2021;36(5):1359-1370.