HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2024 May 31.

Published in final edited form as:

J Biomed Inform. 2024 May ; 153: 104642. doi:10.1016/j.jbi.2024.104642.

Identifying Social Determinants of Health from Clinical Narratives: A Study of Performance, Documentation Ratio, and Potential Bias

Zehao Yu, MS[1], **Cheng Peng, PhD**[1,2], **Xi Yang, PhD**[1,2], **Chong Dang, MS**[1], **Prakash Adekkanattu, PhD**[3], **Braja Gopal Patra, PhD**[4], **Yifan Peng, PhD**[4], **Jyotishman Pathak, PhD**[4], **Debbie L. Wilson, PhD, RN**[5], **Ching-Yuan Chang, PhD**[5], **Wei-Hsuan Lo-Ciganic, PhD**[5], **Thomas J. George**[6], **William R. Hogan, MD**[1], **Yi Guo, PhD**[1,2], **Jiang Bian, PhD**[1,2], **Yonghui Wu, PhD**[1,2]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA

[2]Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, Florida, USA

[3]Information Technologies and Services, Weill Cornell Medicine, New York, NY, USA.

[4]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.

[5]Department of Pharmaceutical Outcomes & Policy, College of Pharmacy, University of Florida, Gainesville, FL 32611, USA

[6]Division of Hematology & Oncology, Department of Medicine, College of Medicine, University of Florida, Gainesville, Florida, USA

Abstract

Objective—To develop a natural language processing (NLP) package to extract social determinants of health (SDoH) from clinical narratives, examine the bias among race and gender groups, test the generalizability of extracting SDoH for different disease groups, and examine population-level extraction ratio.

Methods—We developed SDoH corpora using clinical notes identified at the University of Florida (UF) Health. We systematically compared 7 transformer-based large language models (LLMs) and developed an open-source package – SODA (i.e., SOcial DeterminAnts) to facilitate SDoH extraction from clinical narratives. We examined the performance and potential bias of

Corresponding author: Yonghui Wu, PhD, Clinical and Translational Research Building, 2004 Mowry Road, PO Box 100177, Gainesville, FL, USA, 32610, yonghui.wu@ufl.edu.

CONTRIBUTORSHIP STATEMENT

ZY, XY, JB, and YW were responsible for the overall design, development, and evaluation of this study. ZY and CD annotated the SDoH corpus from cancer patients' notes. DLW, CYC, and WL annotated the SDoH corpus from opioid use patients. TJG, WRH served as domain expert created seed keywords for SDoH and solved the discrepancies in the annotation. YG performed power calculations to determine the number of notes to annotate. PA, BGP, YP, and JP participated in the development of annotation guidelines. ZY and CP conducted the experiments and data analysis. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

COMPETING INTERESTS STATEMENT

The authors have no conflicts of interest that are directly relevant to the content of this study.

SODA for different race and gender groups, tested the generalizability of SODA using two disease domains including cancer and opioid use, and explored strategies for improvement. We applied SODA to extract 19 categories of SDoH from the breast (n=7,971), lung (n=11,804), and colorectal cancer (n=6,240) cohorts to assess patient-level extraction ratio and examine the differences among race and gender groups.

Results—We developed an SDoH corpus using 629 clinical notes of cancer patients with annotations of 13,193 SDoH concepts/attributes from 19 categories of SDoH, and another cross-disease validation corpus using 200 notes from opioid use patients with 4,342 SDoH concepts/attributes. We compared 7 transformer models and the GatorTron model achieved the best mean average strict/lenient F1 scores of 0.9122 and 0.9367 for SDoH concept extraction and 0.9584 and 0.9593 for linking attributes to SDoH concepts. There is a small performance gap (~4%) between Males and Females, but a large performance gap (>16%) among race groups. The performance dropped when we applied the cancer SDoH model to the opioid cohort; fine-tuning using a smaller opioid SDoH corpus improved the performance. The extraction ratio varied in the three cancer cohorts, in which 10 SDoH could be extracted from over 70% of cancer patients, but 9 SDoH could be extracted from less than 70% of cancer patients. Individuals from the White and Black groups have a higher extraction ratio than other minority race groups.

Conclusions—Our SODA package achieved good performance in extracting 19 categories of SDoH from clinical narratives. The SODA package with pre-trained transformer models is available at https://github.com/uf-hobi-informatics-lab/SODA_Docker.

Keywords

Social determinants of health; Large language model; Transformer; Clinical concept extraction; Natural Language Processing; Cancer

1. INTRODUCTION

Social [e.g., education] and behavioral [e.g., smoking] determinants of health (hereafter SDoH for simplicity) are increasingly recognized as important factors affecting a wide range of health, functional, and quality of life outcomes, as well as healthcare fairness and disparities. For example, up to 75% of cancer occurrences are associated with SDoH, [1] which affect individual cancer risks and influence the likelihood of survival, early prevention, and health equity. [2–4] SDoH are associated with the frequency of opioid use and are important factors in preventing opioid misuse. [5–7] Various national and international organizations, such as the World Health Organization (WHO) [8], Healthy People 2030 [9], American Hospital Association (AHA) [10], National Institutes of Health (NIH), and Centers for Disease Control and Prevention (CDC) [11] have unanimously highlighted the importance of SDoH to people's health. There is an increasing interest in studying the role of SDoH in health outcomes and healthcare disparities, yet SDoH are not well-documented in electronic health records (EHRs). In February 2018, the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) Official Guidelines for Coding and Reporting approved that healthcare providers involved in the care of a patient can document SDOH using Z codes (Z55–Z65); however, current reporting of SDoH using ICD-10-CM Z codes is relatively low (2.03% at patient-level) [12] and most

individual-level SDoH are only documented in clinical narratives. [13] Natural language processing (NLP) systems that extract comprehensive SDoH information from clinical narratives are needed.

SDoH are often referred to as factors related to the conditions and status where people are born, live, and work, and are distinct from medical determinants of health (MDoH, e.g., diseases, medical procedures) from healthcare. [9] The definition of SDoH varies across different organizations. Still, common SDoH categories usually include economic stability, education access and quality, social and community context, neighborhood and built environment, and healthcare access and quality. [9] There is growing evidence on the significant association of SDoH with healthcare outcomes such as mortality [14], morbidity [15], mental health status [16], functional limitations [17], and substance use including opioid crisis [7]. For example, Galea *et al.* [18] estimated the number of cancer deaths attributable to SDoH in the United States and reported that low education, racial segregation, low social support, poverty, and income inequality attributed to cancer deaths were comparable to pathophysiological and behavioral causes. Albright *et al.* [5] identified education, housing stability, and employment status significantly associated with the frequency of opioid abuse.[5,6] Cantu *et al.* [7] examined three counties with opioid misuse in Ohio and identified social and economic instability such as unemployment, criminalization of substance use, limited access to healthcare, poverty, and social isolation among the root causes. As SDoH are not well-documented in structured EHRs, many studies [8,19,20] have explored SDoH collected using surveys.

Extracting SDoH from clinical narratives is a typical task of clinical concept extraction or named entity recognition (NER), which identifies phrases of interest (represented using the beginning position and ending position in the text) and determines their semantic categories (e.g., homelessness, smoking). While SDoH were more frequently captured in clinical narratives than structured codes, they were captured only in a subset of notes. Therefore, researchers typically identify the subset of notes with SDoH mentions using note types or key words, to facilitate the developing of corpora. Previous studies [13] have applied NLP methods to extract a single SDoH category from clinical narratives such as homelessness and housing insecurity [21,22], employment status [23], suicide detection [24], marital status [25], and substance use [26,27]. Rule-based and traditional machine learning models have been applied. Recent studies developed corpora with multiple common SDoH categories and applied deep learning-based NLP models. Yetisgen *et al.* [28] developed a corpus of 13 SDoH categories using notes from the publicly available MTSample dataset; Lybarger *et al.* [29] developed a corpus of 12 SDoH using clinical notes from the University of Washington and applied deep learning models including bidirectional long short-term memory (bi-LSTM) and BERT; Feller *et al.* [30] developed a corpus of 5 SDoH categories using notes from Columbia University Medical Center and applied traditional machine learning models; Stemerman *et al.* [16] developed a corpus of 6 SDoH categories and applied the BI-LSTM model; Gehrmann *et al.* [31] and Han *et al.* [32] explored SDoH using clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset; Feller *et al.* [33] developed a corpus of 6 SDOH categories using notes from Columbia University Irving Medical Center. We also have developed an SDoH corpus and transformer-based NLP methods [34], examined the extraction ratio for a lung cancer cohort

[35], and identified potential disparity for treatment options in a type 2 diabetes cohort [36]. As SDoH is not routinely collected in EHRs, previous studies used “key words” or “section names” to identify the notes or sections potentially with mentions of SDoH for annotation. For example, Gundlapalli *et al.* [21] used key words “homeless” to identify notes related with homeless; Feller *et al.* [30] used distributional semantic distance to identify sections contains SDoH; the n2c2 NLP challenge [37] identified social history sections for development of corpora.

Most recent studies for SDoH often applied deep learning models [38]. The 2022 n2c2 organized an NLP challenge focusing on SDoH, which greatly improved the adoption of transformer-based large language models (LLMs)[37]. Recent studies have explored transformer architectures such as BERT and RoBERTa [39,40]. Most NLP methods for SDoH were developed without a disease domain, yet researchers must apply these methods to a disease-specific cohort to study the role of SDoH in EHR-based retrospective cohorts. It is unclear how well current NLP systems can be used to extract SDoH for retrospective patient cohorts and across different disease domains. Until now, there is no off-the-shelf NLP package to facilitate the use of SDoH for EHR-based studies. It is important to develop not only accurate but fair and inclusive NLP methods to prevent potential disparities caused by medical AI systems. The research community has become increasingly aware of potential bias of LLM-based NLP methods for healthcare. Recent studies have discovered the potential bias of LLMs and NLP for healthcare. [41,42] For example, a study reported an NLP method to detect Opioid misuse had systematic bias in false negative rate (32%) for the Black population compared to the White population (17%). [43] Yet, there is no study to examine the bias in extracting SDoH for different race and gender groups.

The goals of this study were (1) to develop an SDoH corpus and an open-source NLP package, SODA (i.e., SOcial DeterminAnts), with pre-trained state-of-the-art transformer models for SDoH extraction from clinical narratives, (2) examine potential bias of SODA for different race and gender groups and test the generalizability of SDoH extraction across two disease domains including cancer and opioid use, and (3) examine extraction rates for various SDoH categories in 3 cancer-specific (breast, lung, colorectal) cohorts, and variations of extraction rates among race and gender groups. We developed a SDoH corpus using clinical notes of cancer and a smaller cross-disease validation corpus using opioid use patients identified at the University of Florida (UF) Health and compared transformer models including Bidirectional Encoder Representations from Transformers (BERT) [44] and RoBERTa [45], DeBERTa[46], Longformer[47], and GatorTron[48]. Then, we explored strategies to customize the cancer-specific NLP model to an opioid user cohort. We integrated SODA with pre-trained clinical models into an open-source software package.

2. METHODS

2.1 Data

This study used clinical narratives from UF Health Integrated Data Repository (IDR). The UF Health IDR is a clinical data warehouse that aggregates data from the university’s various clinical and administrative information systems, including the Epic (Epic Systems

Corporation) system. This study was approved by the UF Institutional Review Board (IRB #IRB201902362).

General cancer cohort: We identified a general cancer cohort between 2012 and 2020 in UF Health IDR using ICD-9 and ICD-10 cancer diagnoses codes, and randomly selected 20,000 cancer patients using stratified random sampling (by cancer types). Using this general cancer cohort, we identified and collected a total number of ~1.5 million clinical notes.

Opioid use cohort: We identified an opioid use cohort between 2016 and 2020 in UF Health IDR. Adult patients aged ≥18 who had at least one outpatient visit and at least one eligible opioid prescribing order (excluding injectable and buprenorphine approved for opioid use disorder). We excluded patients who had non-malignant cancers and who had their first opioid prescription order after Oct 1, 2019.

Identify SDoH keywords: We created a list of keywords to identify clinical notes that contained SDoH using a snowball strategy. We first collected seed keywords indicating SDoH from domain experts (TJG, WRH), healthcare representatives in stakeholders' panel meetings, as well as the biomedical literature. Then, we iteratively reviewed notes to identify new SDoH keywords and extend the seed SDoH keywords until there were no new keywords identified.

Training and test datasets from the cancer cohort: We identified clinical notes containing SDoH by searching the SDoH keywords in the general cancer cohort's clinical notes. Then, we identified clinical notes with at least three unique mentions of SDoH keywords and randomly sampled a subset for annotation. After annotation, we divided the annotated notes into a training set and a testing set with an 8:2 ratio and held out 10% of the training sample which we used as a validation set.

2.2 SDoH annotation

We reviewed SDoH categories defined by healthcare organizations and national agencies including the WHO, Healthy People 2030, and CDC and identified all SDoH categories and their attributes. We developed initial annotation guidelines according to the SDoH definitions from different resources and iteratively fine-tuned the guidelines in training sessions of annotation. During the training sessions, the study team met routinely to identify and review the discrepancies in annotation. Our domain experts served as judges when the two annotators could not reach an agreement. We monitored the annotation agreement using Cohen's Kappa. When a good agreement score (>0.8) was achieved, the two annotators started annotation independently. We used the Brat Rapid annotation tool in this study.

2.3 Assess performance bias and extraction ratio among different race and gender groups

We assessed the performance of SODA for different race groups including White, Black, and Other, as well as gender groups including Female and Male. We also assessed the extraction ratio among different race and gender groups. The extraction ratio for a specific

SDoH category and a specific patient group is defined as the number of unique patients from the specific patient group who have at least one SDoH concept from the specific SDoH category divided by the total number of patients in that specific patient group.

2.4 Cross-disease evaluation dataset from an opioid use cohort

We sought to examine how well the SDoH NLP models developed using cancer patients performed in a different cohort representing opioid use. We adopted the same procedure to identify clinical notes with at least three mentions of unique SDoH from the opioid use cohort and sampled a subset for annotation following the same annotation guidelines. We excluded cancer patients when sampling opioid notes for annotation to avoid any overlap between the two SDoH corpora. After annotation, we split the annotated notes into an additional fine-tuning set – used to fine-tune the cancer SDoH model, and a test set – used to evaluate the cross-disease performance on the opioid population.

2.5 NLP methods to extract SDoH

We approached SDoH extraction as a two-stage NLP task, including (1) a concept extraction step to identify SDoH concepts and attributes and (2) a relation extraction step to link the attribute to the targeted SDoH concepts. Table S1 (in the Supplement) provides the attributes identified for SDoH categories. For example, “attend religious service” is a concept for “social cohesion” where “1 to 4 times per year” is an attribute indicating the frequency of attending religious service; “every day smoker” is an SDoH concept for “tobacco use” where “cigarettes”, “1 packs/day”, and “46 years” are the attributes indicating the smoking type, pack per day, and years of smoking, respectively. We explored pre-trained models from two state-of-the-art transformer architectures, BERT and RoBERTa. Our previous study showed that BERT and RoBERTa consistently outperformed other transformer models for clinical concept extraction [49]. Following our previous studies on clinical transformers, we examined pre-trained transformers from general English corpus (denoted as ‘_general’, e.g., ‘BERT_general’) and clinical transformers pre-trained using clinical notes from the MIMIC-III database (denoted as ‘_mimic’, e.g., ‘BERT_mimic’). We adopted the default parameters optimized in our clinical transformer package [49]. We also explored new transformer architectures including DeBERTa [46], Longformer [47], and GatorTron [48]. GatorTron is developed using BERT architecture with 82 billion words from over 290 million clinical notes at UF Health, 6 billion words from PubMed, and 2.5 billion words from Wikipedia, which is the largest encoder-only LLM in the clinical domain. We used the GatorTron-base model with 345 million parameters.

Identification of SDoH concepts and attributes using concept extraction—We approached clinical concept extraction as a sequence labeling problem and adopted the ‘BIO’ labeling schema, where ‘B-’ and ‘I-’ are label prefixes indicating words at the beginning and inside of a concept, and ‘O’ stands for words located outside of any concepts of interests. We solved the task as a classification – for each word in a sentence, we determined a label in [‘B’, ‘I’, ‘O’]. In this study, we used the pre-trained transformer models to generate distributed word-level and sentence-level representations, then added a classification layer with Softmax activation to calculate a probability for each category. The cross-entropy loss was used for fine-tuning.

Linking attributes to core SDoH concepts using relation classification—The goal was to link attribute concepts (e.g., smoking frequency) to the core SDoH concept (e.g., tobacco use). Following our previous experience in relation classification, we approached attribute linking as a classification task – we generated candidate pairs of concepts and trained machine learning classifiers to classify them into predefined relation classes. We adopted a heuristic method developed in our previous studies [50,51] to identify candidate pairs of clinical concepts. Specifically, two concepts can be considered as a candidate pair if there is a relation defined between the semantic categories of the two concepts. For example, an "Employment" concept and an "Occupation" concept can be a candidate pair, but an "Education" concept and an "Occupation" concept cannot. The Supplement Table S1 provides detailed information about the heuristic rules. Then, pre-trained transformer models were used to generate a distributed representation. To distinguish concepts, we introduced two sets of entity markers, i.e., [S1] and [E1] for the first concept, and [S2] and [E2] for the second concept. To determine the relation type, we concatenated the contextual representations of the model special [CLS] token and all four entity markers and added a classification layer (a linear layer with Softmax activation) to calculate a score for each relation category. The cross-entropy loss was used in fine-tuning.

2.6 Evaluation and experiments design

Evaluation methods: We first evaluated SODA using a standard setting where both the training and test data were from a cancer cohort. We evaluated SODA on three subtasks including (1) a concept extraction task to extract SDoH concepts and attributes, (2) a relation extraction task to link attributes to the target SDoH concept (given ground-truth SDoH concepts), (3) an end-to-end task to extract SDoH concepts and link attributes to SDoH concepts. Then, we conducted a cross-disease evaluation to evaluate the NLP models using clinical notes from an opioid use cohort. We compared three application scenarios to evaluate SODA in cross-disease settings including (1) directly applying the NLP models developed for cancer patients to patients of opioid use, (2) merging the cancer corpus with the opioid fine-tuning corpus and training a model from scratch, and (3) fine-tuning the cancer SDoH model using the opioid fine-tuning set.

Evaluation metrics: Cohen's Kappa: We evaluated annotator agreement using Cohen's Kappa, κ, coefficient, where higher κ denotes better annotator agreement. We used the strict and lenient micro-averaged precision, recall, and F1-score aggregated from all classes to evaluate the concept extraction and relation extraction. The strict evaluation requires machine learning models to precisely detect the same span of concepts as in the gold standard annotations. Whereas in the lenient evaluation, it is sufficient if the model detected span of concepts overlaps with the gold standard annotation. The official evaluation scripts provided by the 2018 n2c2 challenge [52] were used to calculate these scores.

Experimental setup: We used pretrained transformer models developed in our previous study [49], where the transformer architecture was implemented in PyTorch. We fine-tuned transformer models using the training set. The best model was selected according to the validation performance measured by strict F1-scores on the validation set. We adopted

an early stop strategy to stop the training when no improvements were observed in 5 consecutive epochs. We conducted all experiments using two Nvidia A100 GPUs.

2.7 SODA package

We implemented SODA in Python based on the Transformers library developed by HuggingFace. We used a pipeline-based architecture with multiple components including preprocessing for tokenization and sentence boundary, SDoH concepts and attributes extraction, relation extraction to link SDoH concepts to attributes, and postprocessing to combine extracted SDoH concepts and relations into an output following the format used by the Brat rapid annotation tool. The fine-tuned GatorTron models were integrated into this package. We also created a Docker image to facilitate the study of SDoH using SODA. Using one NVIDIA A100 GPU, SODA could process 1 million notes in about 6 days.

3. RESULTS

Domain experts identified a total of 44 keywords (provided in the Supplement) suggesting SDoH in the snowball sampling procedure. We identified a total of 225,441 clinical notes containing at least three unique SDoH mentions from cancer patients and randomly sampled 700 for annotation. After de-duplicating and removing notes without valid SDoH annotations, there remained 629 notes in the cancer SDoH corpus. Two annotators (ZY and CD) annotated a total of 13,193 SDoH concepts in these notes. There were 19 categories of SDoH identified from the annotation. Table S1 (in the Supplement) provides the attributes identified for the 19 subclasses of SDoH. The inter-annotator agreement between the two annotators calculated by kappa score (using 20 overlapped notes) was low at 0.47 in the first training session, which was improved to 0.68 in the second round and eventually reached 0.89 after 5 iterative rounds of training followed by meetings to discuss and solve discrepancies. From the opioid cohort, we identified ~13 million clinical notes from 98,074 patients. We followed the same annotation guidelines and annotated an SDoH corpus of 200 notes for cross-disease evaluation. Table 1 shows detailed numbers of concepts annotated for each SDoH category. Table 2 shows the distribution of notes and SDoH concepts for training, validation, and test set of the two disease domains.

Table 3 compares 7 transformer-based NLP models for SDoH concept/attribute extraction and attribute linking. We run each transformer model 10 times using different random initializations to calculate the mean average scores and the standard deviations. For SDoH concept extraction, the GatorTron model achieved the best mean average strict/lenient F1 scores of 0.9122 and 0.9367, respectively. Table S2 (in the Supplement) provides detailed scores for each SDoH subclass. For attribute linking using relation classification, the GatorTron again achieved the best mean average strict/lenient F1 scores of 0.9584 and 0.9593, respectively. The end-to-end system using the best GatorTron model achieved strict/lenient F1 scores of 0.8963 and 0.9133, respectively. Statistical test results showed that GatorTron is significantly better ($p<0.05$) than the second-best model for concept extraction, but not significantly ($p>0.05$) for attribute linking.

Table 4 compares performance differences of SODA for different gender and race groups using the end-to-end model based on GatorTron. There are large differences among race

groups where SODA has the highest performance in extracting SDoH for White group (F1 scores of 0.9038 and 0.9160) and the lowest performance (F1 scores of 0.7465 and 0.7960) for the Other group with a performance gap over 15% in strict F1-score and 12% in relaxed F1-score. The performance difference among Male and Female is relatively small about 4%.

Table 5 shows the results of the cross-disease evaluation for concept extraction. When directly applying GatorTron trained using cancer data to the opioid cohort, we observed a performance drop from strict/lenient F1 scores of 0.9122 and 0.9367 to 0.8244 and 0.8565, respectively. Both customization strategies improved the F1-score of SDoH extraction for opioid use patients. The best strict/lenient F1 score of 0.8444 and 0.8760 was achieved by fine-tuning the cancer SDoH model using the opioid fine-tuning data.

Table 6 reports for three cancer cohorts the total number of SDoH concepts and the population-level extraction ratio – defined as the total number of patients with at least one specific SDoH category divided by the total number of patients. For lung cancer, we identified a total of 11,804 patients with 1,796,131 notes. For breast cancer, we identified 7,971 patients with 1,143,304 clinical notes. For colorectal cancer, we identified 6,240 patients with 1,021,405 clinical notes. We applied the end-to-end NLP model to extract 19 SDoH categories and aggregated the SDoH to the patient level to examine the extraction ratio.

As shown in Table 7, the extraction ratio (same definition as in Table 6) varied among different race and gender groups. Among race groups, individuals from the “Other” group have a lower extraction ratio than individuals from the White and the Black group. For gender groups, individuals from the Male group have a lower extraction ratio than individuals from the Female group.

4. DISCUSSION AND CONCLUSION

NLP is the key technology to extract SDoH from clinical narratives. This study examined transformer-based NLP models for SDoH extraction from clinical narratives. We developed SDoH corpora from two disease domains (cancer and opioid use patients) with 19 SDoH categories and compared seven transformer-based NLP models for extraction. The end-to-end NLP system using the GatorTron model achieved the best strict mean average F1-score of 0.8963 and the best lenient mean average F1-score of 0.9133, demonstrating the effectiveness of LLMs for SDoH extraction from clinical narratives. We examined the performance differences among gender and race groups and assessed patient-level extraction ratio using 3 real-world cancer cohorts including breast, lung, and colorectal. We integrate our transformer models and pipelines into an open-source package to facilitate the extraction of SDoH from clinical narratives.

This study systematically compared 7 transformer models using 19 categories of SDoH and developed an open-source package to facilitate SDoH extraction from clinical narratives. We performed statistical tests and found that the best GatorTron model is significantly better ($p<0.05$) than other transformer models for concept extraction. Our previous studies [49,51] showed that a clinical fine-tuned BERT model (BERT_mimic) outperformed a

general BERT model (BERT_general) on extracting clinical concepts. This study showed that BERT_general outperformed BERT_mimic for SDoH extraction. One potential reason is that most SDoH concepts are composed of general English words rather than medical words. This could be the potential reason that there is limited benefit from GatorTron [48], an LLM trained using a larger clinical corpus.

For relation extraction, the statistical test results showed that GatorTron is not significantly ($p>0.05$) better than the second-best model RoBERTa_general and the third-best model BERT_general. This study approached relation extraction as a classification task, therefore, our results may indicate that smaller transformer models could achieve performance comparable to large transformer-based LLMs for classification tasks. Future studies should examine this finding using more text classification datasets and exploring more transformer architectures.

Previously, we applied GatorTron in the 2022 n2c2/UW challenge [37] for the extraction of 5 categories of SDoH, 9 SDoH attributes, and 28 categories of relations. For concept extraction, our GatorTron model achieved a strict F1 score of 0.8341 and a lenient F1 score of 0.9318, respectively. The end-to-end system based on GatorTron achieved a strict F1 score of 0.6395 and a lenient F1 score of 0.7913. (Scores were calculated using the official evaluation script developed by the 2018 n2c2 challenge) During the 2022 n2c2 challenge, we fine-tuned GatorTron models for both SDoH concept/relation extraction and argument subtype classification and our system achieved the second best F1-score of 0.8903 in extracting 5 categories of SDoH and their attributes according to the 2022 n2c2 official evaluation results.

The research community is increasingly aware of the potential bias and fairness of applying AI in healthcare. We assessed SODA for different race and gender groups and found that there are small performance gaps between Male and Female groups (~4%) but large performance gaps among race groups (>16% in strict F1-score). One potential reason is the relative smaller number of individuals randomly sampled from the Black and Other race groups. Further studies should examine potential causes such as the documentation variations for different race groups. In addition to the standard training/test evaluation using data from the same disease domain, we conducted a cross disease evaluation to examine how the cancer SDoH models perform on an opioid use cohort. We observed a performance drop when directly applying the cancer SDoH models to opioid use patients, indicating that the documentation of SDoH varied among different disease domains. We explored two strategies to customize the NLP model and the fine-tuning strategy achieved the best strict F1 score. The experimental results from the cross-disease evaluation showed that it is necessary to fine-tune the NLP module by annotating corpora from a new disease domain.

We also examined the patient-level extraction ratio for the 19 SDoH categories. The patient-level extraction ratio was largely consistent among three cancer cohorts with some variations. For example, the lung cancer cohort had a higher extraction ratio for tobacco use than the breast and colorectal cancer cohorts. This result is expected given the association between smoking and lung cancer and the strong emphasis on smoking cessation as a component of lung cancer therapy. There are 10 categories of SDoH extracted from

> 70% population of the cancer patients, including gender, race, tobacco use, alcohol use, drug use, education, living supply, marital status, occupation, and sexual activity; 9 other categories had a relatively low extraction ratio (< 70% population), indicating a potential gap of documenting SDoH in EHRs. The extraction ratios vary among different gender and race groups. For example, the documentation ratios of the Black group for "Abuse", "Financial constraint", "Living condition", "Physical activity", "Social cohesion", and "Transportation" are remarkably higher than other groups (over 10%). We searched existing studies examining these SDoH among different race groups for potential insights. It has been reported that Black children were more than twice as likely to be referred for abused victims[53]; Black older adults are at heightened risk of overall mistreatment and financial abuse [54]; the discrimination towards Black individuals in rental and housing markets remains pervasive[55]; there is higher work-related physical activity among Black compared with White[56]; and African American has more transportation burden than other groups[57,58]. The documentation of these SDoH categories might be affected by existing findings from the healthcare research community. However, future studies need to examine if the variations in documentation ratios reflect real-world incidences.

We identified a total of 38 SDoH categories (shown in Supplement Figure S1) from WHO, Healthy People 2030, and CDC, yet only found 19 categories from the randomly sampled 629 notes, indicating the current reporting of SDoH in EHRs needs to be improved. In a previous study [59], we conducted a focused interview of stakeholders including oncologists, data analysts, citizen scientists, and patient navigators, and identified potential challenges and barriers to the low documentation ratio of SDoH in EHRs, including lack of integration into clinical workflow, lack of incentives for SDoH data collection, and lack of training and tools for clinicians to derive actionable insights for decision making. Future studies should explore strategies to reduce these barriers and improve the documentation of SDoH in EHRs.

This study has limitations. First, a limited number of instances were annotated for some SDoH categories (e.g., language) due to low documentation rates. As most notes don't have SDoH documented, we sampled from clinical notes with at least three SDoH concepts identified by keywords, which may cause bias for annotation. We plan to annotate more notes from the minority race groups to increase the sample size and mitigate the performance differences. Similarly, the cross-disease performance of the NLP models could be further improved. Second, we may miss some keywords in the snowball procedure used to identify the seed SDoH keywords using domain experts. The NLP models were developed using notes from patients with cancer and opioid use. Customization through fine-tuning is needed when applying SODA to other disease domains. Our future work will investigate how person-level SDoH affect cancer risks, treatment outcomes, and disparities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation and NVIDIA AI Technology Center with the donation of the GPUs and the computing resources used for this research. We acknowledge the support from the Cancer Informatics Shared Resource in the UF Health Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

FUNDING STATEMENT

This study was partially supported by grants from the Patient-Centered Outcomes Research Institute® (PCORI®) (ME-2018C3-14754), the National Institute on Aging (1R56AG069880, 1R01AG080624-01, R21AG068717, R01AG080991), Ed and Ethel Moore Alzheimer's Disease Research Program (23A09), the National Cancer Institute (1R01CA246418, 3R01CA246418-02S1, 1R21CA245858-01A1, R21CA245858-01A1S1, R21CA253394-01A1, R21CA253394-01A1), National Institute of Allergy and Infectious Diseases (R01AI172875), National Heart, Lung, and Blood Institute (1R01HL169277), National Institute on Drug Abuse (1R01DA050676), National Institute of Mental Health (1R01MH121907, 5R21MH129682-02), National Library of Medicine (4R00LM013001), NSF Career (2145640), and Centers for Disease Control and Prevention (1U18DP006512).

REFERENCES

1. Akushevich I, Kravchenko J, Akushevich L, et al. Cancer Risk and Behavioral Factors, Comorbidities, and Functional Status in the US Elderly Population. ISRN Oncol. 2011;2011. doi: 10.5402/2011/415790
2. Hiatt RA, Breen N. The social determinants of cancer: a challenge for transdisciplinary science. Am J Prev Med. 2008;35:S141–150. [PubMed: 18619394]
3. Matthews AK, Breen E, Kittiteerasack P. Social Determinants of LGBT Cancer Health Inequities. Semin Oncol Nurs. 2018;34:12–20. [PubMed: 29373163]
4. Gerend MA, Pai M. Social determinants of Black-White disparities in breast cancer mortality: a review. Cancer Epidemiol Biomarkers Prev. 2008;17:2913–23. [PubMed: 18990731]
5. Albright DL, Johnson K, Laha-Walsh K, et al. Social Determinants of Opioid Use among Patients in Rural Primary Care Settings. Soc Work Public Health. 2021;36:723–31. [PubMed: 34167439]
6. Rangachari P, Govindarajan A, Mehta R, et al. The relationship between Social Determinants of Health (SDoH) and death from cardiovascular disease or opioid use in counties across the United States (2009–2018). BMC Public Health. 2022;22:236. [PubMed: 35120479]
7. Cantu R, Fields-Johnson D, Savannah S. Applying a Social Determinants of Health Approach to the Opioid Epidemic. Health Promotion Practice. 2023;24:16–9. [PubMed: 32713219]
8. Singh GK, Daus GP, Allender M, et al. Social Determinants of Health in the United States: Addressing Major Health Inequality Trends for the Nation, 1935–2016. Int J MCH AIDS. 2017;6:139–64. [PubMed: 29367890]
9. Social Determinants of Health - Healthy People 2030 | health.gov. https://health.gov/healthypeople/objectives-and-data/social-determinants-health (accessed 14 September 2021)
10. American Hospital Association| ICD-10-CM Coding for Social Determinants of Health. https://www.aha.org/system/files/2018-04/value-initiative-icd-10-code-social-determinants-of-health.pdf (accessed 2 December 2022)
11. Hillemeier M, Lynch J, Harper S, et al. Data Set Directory of Social Determinants of Health at the Local Level.; 75.
12. Guo Y, Chen Z, Xu K, et al. International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. Medicine (Baltimore). 2020;99:e23818.
13. 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;28:2716–27. [PubMed: 34613399]
14. Sterling MR, Ringel JB, Pinheiro LC, et al. Social Determinants of Health and 90-Day Mortality After Hospitalization for Heart Failure in the REGARDS Study. J Am Heart Assoc. 2020;9:e014836.

15. Eppes C, Salahuddin M, Ramsey PS, et al. Social Determinants of Health and Severe Maternal Morbidity During Delivery Hospitalizations in Texas [36L]. Obstetrics & Gynecology. 2020;135:133S.
16. Stemerman R, Arguello J, Brice J, et al. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. JAMIA Open. Published Online First: 9 February 2021. doi: 10.1093/jamiaopen/ooaa069
17. May 10 EHP, 2018. Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity. KFF. 2018 https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ (accessed 12 November 2021)
18. Galea S, Tracy M, Hoggatt KJ, et al. Estimated Deaths Attributable to Social Factors in the United States. Am J Public Health. 2011;101:1456–65. [PubMed: 21680937]
19. 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.
20. Chen X, Jacques-Tiura AJ. Smoking Initiation Associated With Specific Periods in the Life Course From Birth to Young Adulthood: Data From the National Longitudinal Survey of Youth 1997. Am J Public Health. 2014;104:e119–26.
21. 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. [PubMed: 24551356]
22. 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:ooac006.
23. Dillahunt-Aspillaga C, Finch D, Massengale J, et al. Using Information from the Electronic Health Record to Improve Measurement of Unemployment in Service Members and Veterans with mTBI and Post-Deployment Stress. PLOS ONE. 2014;9:e115873.
24. Carson NJ, Mullin B, Sanchez MJ, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLOS ONE. 2019;14:e0211116.
25. Bucher BT, Shi J, Pettit RJ, et al. Determination of Marital Status of Patients from Structured and Unstructured Electronic Healthcare Data. AMIA Annu Symp Proc. 2020;2019:267–74. [PubMed: 32308819]
26. Wang Y, Chen ES, Pakhomov S, et al. Automated Extraction of Substance Use Information from Clinical Texts. AMIA Annu Symp Proc. 2015;2015:2121–30. [PubMed: 26958312]
27. Rajendran S, Topaloglu U. Extracting Smoking Status from Electronic Health Records Using NLP and Deep Learning. AMIA Jt Summits Transl Sci Proc. 2020;2020:507–16. [PubMed: 32477672]
28. Yetisgen M, Vanderwende L. Automatic Identification of Substance Abuse from Social History in Clinical Text. In: ten Teije A, Popow C, Holmes JH, et al., eds. Artificial Intelligence in Medicine. Cham: Springer International Publishing 2017:171–81. 10.1007/978-3-319-59758-4_18
29. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. Journal of Biomedical Informatics. 2021;113:103631.
30. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, et al. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. Appl Clin Inform. 2020;11:172–81. [PubMed: 32131117]
31. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLOS ONE. 2018;13:e0192360.
32. 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.
33. Feller DJ, Zucker J, Don't Walk OB, et al. Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning. AMIA Annu Symp Proc. 2018;2018:422–9. [PubMed: 30815082]

34. Yu Z, Yang X, Dang C, et al. A Study of Social and Behavioral Determinants of Health in Lung Cancer Patients Using Transformers-based Natural Language Processing Models. arXiv:210804949 [cs]. Published Online First: 10 August 2021.
35. Yu Z, Yang X, Guo Y, et al. Assessing the Documentation of Social Determinants of Health for Lung Cancer Patients in Clinical Narratives. Front Public Health. 2022;10:778463.
36. Guo J, Wu Y, Guo Y, et al. Abstract P108: Natural Language Processing Extracted Social And Behavioral Determinants Of Health And Newer Glucose-lowering Drug Initiation Among Real-world Patients With Type 2 Diabetes. Circulation. 2022;145:AP108–AP108.
37. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. Journal of the American Medical Informatics Association. 2023;ocad012.
38. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. [PubMed: 26017442]
39. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:190711692 [cs]. Published Online First: 26 July 2019.
40. Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics 2016:260–70. 10.18653/v1/N16-1030
41. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356:183–6. [PubMed: 28408601]
42. Schramowski P, Turan C, Andersen N, et al. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat Mach Intell. 2022;4:258–68. doi: 10.1038/s42256-022-00458-8
43. Thompson HM, Sharma B, Bhalla S, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. J Am Med Inform Assoc. 2021;28:2393–403. [PubMed: 34383925]
44. Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
45. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv. 2019;abs/1907.11692.
46. He P, Liu X, Gao J, et al. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. 2021. 10.48550/arXiv.2006.03654
47. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. 2020. 10.48550/arXiv.2004.05150
48. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. npj Digit Med. 2022;5:1–9. doi: 10.1038/s41746-022-00742-2 [PubMed: 35013539]
49. Yang X, Bian J, Hogan WR, et al. Clinical concept extraction using transformers. Journal of the American Medical Informatics Association. 2020;27:1935–42. [PubMed: 33120431]
50. Yang X, Bian J, Gong Y, et al. MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. Drug Saf. Published Online First: 2 January 2019. doi: 10.1007/s40264-018-0761-0
51. Yang X, Bian J, Fang R, et al. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc. 2020;27:65–72. [PubMed: 31504605]
52. Henry S, Buchan K, Filannino M, et al. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020;27:3–12. [PubMed: 31584655]
53. Putnam-Hornstein E, Needell B, King B, et al. Racial and ethnic disparities: A population-based examination of risk factors for involvement with child protective services. Child Abuse & Neglect. 2013;37:33–46. [PubMed: 23317921]
54. Burnes D, Hancock DW, Eckenrode J, et al. Estimated Incidence and Factors Associated With Risk of Elder Mistreatment in New York State. JAMA Network Open. 2021;4:e2117758.
55. Pager D, Shepherd H. The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. Annual Review of Sociology. 2008;34:181–209.

56. Saffer H, Dave D, Grossman M, et al. Racial, Ethnic, and Gender Differences in Physical Activity. Journal of Human Capital. 2013;7:378–410. [PubMed: 25632311]

57. Wolfe MK, McDonald NC, Holmes GM. Transportation Barriers to Health Care in the United States: Findings From the National Health Interview Survey, 1997–2017. Am J Public Health. 2020;110:815–22. [PubMed: 32298170]

58. Probst JC, Laditka SB, Wang J-Y, et al. Effects of residence and race on burden of travel for care: cross sectional analysis of the 2001 US National Household Travel Survey. BMC Health Serv Res. 2007;7:40. [PubMed: 17349050]

59. Alpert J, Kim H (Julia), McDonnell C, et al. Barriers and Facilitators of Obtaining Social Determinants of Health of Patients With Cancer Through the Electronic Health Record Using Natural Language Processing Technology: Qualitative Feasibility Study With Stakeholder Interviews. JMIR Form Res. 2022;6:e43059.

Table 1.

Annotation results for the Cancer cohort and the Opioid cohort.

SDoH Class	#Concepts Cancer	#Concepts Opioid	SDoH Subclasses	#Concepts Cancer	#Concepts Opioid
			Financial constraint	97	42
Economic Stability	596	282	Employment	499	240
			Language	25	2
Education	602	210	Education	577	208
			Physical activity	223	78
			SDoH ICD	61	0
			Sexual activity	637	159
			Drug use	577	210
			Tobacco use	1,998	425
Health and Health care	4,370	1,194	Alcohol use	874	**322**
			Marital status	488	177
Social and community context	908	397	Social cohesion	420	220
			Abuse (physical or mental)	412	183
			Transportation	193	75
			Living supply	523	214
Neighborhood and physical environment	1,257	499	Living condition	129	27
			Gender	846	283
			Race	110	44
Gender, Race, and Ethnicity	990	332	Ethnicity	34	5

Table 2.

Distribution of notes and SDoH in training, validation/fine-tuning, and test sets of the cancer cohort; and the opioid cohort.

Disease domain		Total #	Training/Fine-tuning	Validation	Test
Cancer	Total notes	629	452	51	126
	Total entities	13,193	9,497	1,009	2,687
	Entity/note	20	21	20	21
Opioid	Total notes	200	90	10	100
	Total entities	4,342	1,952	173	2,217
	Entity/note	21	22	17	22

Table 3.

Comparison of transformer models to identify SDoH concepts and link attributes on the cancer cohort.

Task	Model	Strict			Lenient		
		Prec.	**Rec.**	**F(b=1)**	**Prec.**	**Rec.**	**F(b=1)**
Concept extraction to identify SDoH concepts and attributes	BERT_general	0.9024 (± 0.006)	0.9074 (± 0.005)	0.9048 (± 0.003)	0.9335 (± 0.010)	0.9339 (± 0.004)	0.9336 (± 0.004)
	BERT_mimic	0.9057 (± 0.006)	0.9094 (± 0.005)	0.9080 (± 0.004)	0.9333 (± 0.005)	0.9344 (± 0.005)	0.9338 (± 0.002)
	Roberta_general	0.9097 (± 0.006)	**0.9137** (± 0.004)	0.9116 (± 0.003)	**0.9335** (± 0.007)	0.9347 (± 0.003)	0.9341 (± 0.003)
	Roberta_mimic	0.9022 (± 0.008)	0.9063 (± 0.003)	0.9042 (± 0.004)	0.9311 (± 0.007)	0.9321 (± 0.003)	0.9316 (± 0.003)
	DeBERTa	**0.9179** (± 0.007)	0.9037 (± 0.009)	0.9107 (± 0.003)	0.9439 (± 0.007)	0.9272 (± 0.008)	0.9354 (± 0.002)
	GatorTron	0.9114 (± 0.006)	0.9130 (± 0.006)	**0.9122** (± 0.003)	0.9373 (± 0.006)	**0.9363** (± 0.005)	**0.9367** (± 0.002)
	Longformer	0.9136 (± 0.008)	0.9069 (± 0.004)	0.9102 (± 0.003)	0.9373 (± 0.006)	0.9284 (± 0.005)	0.9327 (± 0.003)
Relation classification to link attributes to core SDoH concepts	BERT_general	0.9608 (± 0.002)	0.9540 (± 0.006)	0.9574 (± 0.004)	0.9615 (± 0.002)	0.9548 (± 0.006)	0.9582 (± 0.004)
	BERT_mimic	0.9439 (± 0.007)	**0.9659** (± 0.005)	0.9548 (± 0.003)	0.9447 (± 0.007)	**0.9667** (± 0.005)	0.9555 (± 0.004)
	Roberta_general	0.9621 (± 0.004)	0.9544 (± 0.011)	0.9582 (± 0.005)	0.9631 (± 0.004)	0.9554 (± 0.011)	0.9592 (± 0.005)
	Roberta_mimic	0.9557 (± 0.004)	0.9494 (± 0.017)	0.9525 (± 0.008)	0.9567 (± 0.004)	0.9503 (± 0.017)	0.9535 (± 0.008)
	DeBERTa	0.9603 (± 0.004)	0.9431 (± 0.012)	0.9516 (± 0.007)	0.9611 (± 0.004)	0.9439 (± 0.011)	0.9524 (± 0.007)
	GatorTron	0.9601 (± 0.003)	0.9569 (± 0.006)	**0.9584** (± 0.004)	0.9609 (± 0.003)	0.9579 (± 0.007)	**0.9593** (± 0.004)
	Longformer	**0.9625** (± 0.005)	0.9498 (± 0.007)	0.9561 (± 0.004)	**0.9635** (± 0.005)	0.9507 (± 0.007)	0.9571 (± 0.004)
End-to-end	GatorTron	0.9183	0.8754	**0.8963**	0.9356	0.8919	**0.9133**

Mean average precision, recall, and F1 score with standard deviation are reported. Best precision, recall, and F1-score are highlighted in bold. The best models for GatorTron and BERT_general was used in the end-to-end evaluation. The official evaluation script developed by the 2018 n2c2 challenge was used to calculate the evaluation scores.

Table 4.

Comparison of performance among different race and gender groups using end-to-end model based on the best GatorTron model.

	Group	# Patients	Strict			Lenient		
			Prec.	Rec.	F(b=1)	Prec.	Rec.	F(b=1)
Race	White	92	0.9342	0.8753	0.9038	0.9468	0.8871	0.9160
	Black	41	0.8269	0.9085	0.8658	0.8489	0.9296	0.8874
	Other	16	0.7922	0.7042	**0.7456**	0.8458	0.7518	**0.7960**
Gender	Male	37	0.9482	0.8960	0.9214	0.9644	0.9113	0.9371
	Female	85	0.9045	0.8657	0.8847	0.9224	0.8829	0.9022

Lowest precision, recall, and F1-score are highlighted in bold. Other race: Include Asian or Pacific Islander, multi-racial, American Indian or Alaska Native, Indian, Native Hawaiian or Other Pacific Islander.

Table 5.

Cross-disease evaluation of concept extraction on the opioid test data set.

	Strict			Lenient		
	Prec.	**Rec.**	**F(b=1)**	**Prec.**	**Rec.**	**F(b=1)**
Direct evaluation	0.8295	0.8192	0.8244	0.8653	0.8479	0.8565
Fine-tuning	**0.8350**	**0.8541**	**0.8444**	**0.8688**	**0.8833**	**0.8760**
Merge and retrain	0.8329	0.8484	0.8406	**0.8688**	0.8804	0.8746

Direct evaluation: directly evaluating the cancer SDoH model using opioid test set; Fine-tuning: fine-tuning the cancer model using the Opioid fine-tuning set; Merge and retrain: merging the Cancer training set and opioid fine-tuning set and retraining the model.

Table 6.

Number of SDoH instances and population-level extraction ratio from lung, breast, and colorectal cancers.

	Breast cancer N=7,971		Colorectal cancer N=6,240		Lung cancer N=11,804	
SDoH	**# Concepts**	**Ratio**	**# Concepts**	**Ratio**	**# Concepts**	**Ratio**
Abuse (physical or mental)	3,077	0.47	1,378	0.36	4,145	0.43
Alcohol use	6,179	0.94	3,598	0.95	9,195	0.95
Drug use	6,055	0.92	3,521	0.93	8,756	0.91
Education	5,825	0.88	3,370	0.89	8,463	0.87
Ethnicity	5,173	0.79	2,509	0.66	5,231	0.54
Financial constraint	2,485	0.38	981	0.26	2,766	0.29
Gender	6,486	0.99	3,731	0.99	9,552	0.99
Language	5,158	0.78	2,466	0.65	5,173	0.53
Living condition	3,192	0.48	1,866	0.49	5,359	0.55
Living supply	5,853	0.89	3,285	0.87	7,861	0.81
Marital status	6,015	0.91	3,472	0.92	8,655	0.89
Occupation/Employment	5,882	0.89	3,324	0.88	8,345	0.86
Physical activity	2,992	0.45	1,136	0.30	3,092	0.32
Race	5,709	0.87	3,087	0.82	7,376	0.76
SDoH ICD	562	0.09	345	0.09	1,239	0.13
Sexual activity	5,606	0.85	3,173	0.84	8,124	0.84
Social cohesion	2,458	0.37	981	0.26	2,727	0.28
Tobacco use	4,940	0.75	2,669	0.71	7,639	0.79
Transportation	2,524	0.38	1,018	0.27	2,877	0.30

SDoH: social determinants of health; ICD: International Classification of Diseases; Ratio: population-level extraction ratio.

Table 7.

Comparison of extraction ratio for different race and gender groups over all individuals from lung, breast, and colorectal cancers.

SDoH	White N=15,543	Black N=3,289	Other N=994	Male N=8,216	Female N=12,972
Abuse (physical or mental)	0.42	0.55	0.34	0.40	0.46
Alcohol use	0.95	0.98	0.91	0.94	0.95
Drug use	0.92	0.96	0.88	0.91	0.92
Education	0.88	0.95	0.83	0.87	0.89
Ethnicity	0.65	0.70	0.60	0.56	0.69
Financial constraint	0.30	0.42	0.28	0.27	0.35
Gender	0.99	0.99	0.97	0.98	0.99
Language	0.64	0.70	0.58	0.55	0.69
Living condition	0.50	0.64	0.46	0.52	0.51
Living supply	0.84	0.91	0.79	0.82	0.86
Marital status	0.91	0.96	0.89	0.89	0.91
Occupation/Employment	0.88	0.92	0.86	0.86	0.89
Physical activity	0.35	0.47	0.31	0.30	0.41
Race	0.82	0.88	0.56	0.76	0.83
SDoH ICD	0.09	0.19	0.06	0.12	0.10
Sexual activity	0.85	0.92	0.78	0.83	0.85
Social cohesion	0.30	0.43	0.26	0.27	0.35
Tobacco use	0.75	0.84	0.69	0.77	0.75
Transportation	0.31	0.43	0.28	0.28	0.36

Other: Include Asian or Pacific Islander, multi-racial, American Indian or Alaska Native, Indian, Native Hawaiian or Other Pacific Islander.