Journal of the American Medical Informatics Association, 30(8), 2023, 1389–1397
https://doi.org/10.1093/jamia/ocad073
Advance Access Publication Date: 1 April 2023
Research and Applications

Research and Applications

Leveraging natural language processing to augment structured social determinants of health data in the electronic health record

Kevin Lybarger [1], Nicholas J. Dobbins [2,3], Ritche Long [3], Angad Singh [4], Patrick Wedgeworth [4], Özlem Uzuner [1], and Meliha Yetisgen [2]

[1]Department of Information Sciences and Technology, George Mason University, Fairfax, Virginia, USA, [2]Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, Washington, USA, [3]Department of Research IT, UW Medicine, University of Washington, Seattle, Washington, USA and [4]Department of Medicine University of Washington, Seattle, Washington, USA

Kevin Lybarger and Nicholas J. Dobbins contributed equally to this work.

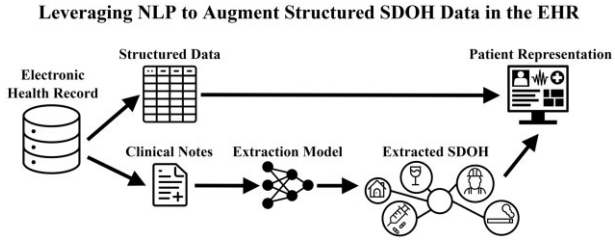
Corresponding Author: Kevin Lybarger, PhD, Department of Information Sciences and Technology, George Mason University, 4400 University Dr. MSN 1G8, Fairfax, VA 22030, USA; klybarge@gmu.edu

Received 4 December 2022; Revised 6 April 2023; Editorial Decision 11 April 2023; Accepted 12 April 2023

ABSTRACT

Objective: Social determinants of health (SDOH) impact health outcomes and are documented in the electronic health record (EHR) through structured data and unstructured clinical notes. However, clinical notes often contain more comprehensive SDOH information, detailing aspects such as status, severity, and temporality. This work has two primary objectives: (1) develop a natural language processing information extraction model to capture detailed SDOH information and (2) evaluate the information gain achieved by applying the SDOH extractor to clinical narratives and combining the extracted representations with existing structured data.
Materials and Methods: We developed a novel SDOH extractor using a deep learning entity and relation extraction architecture to characterize SDOH across various dimensions. In an EHR case study, we applied the SDOH extractor to a large clinical data set with 225 089 patients and 430 406 notes with social history sections and compared the extracted SDOH information with existing structured data.
Results: The SDOH extractor achieved 0.86 F1 on a withheld test set. In the EHR case study, we found extracted SDOH information complements existing structured data with 32% of homeless patients, 19% of current tobacco users, and 10% of drug users only having these health risk factors documented in the clinical narrative.
Conclusions: Utilizing EHR data to identify SDOH health risk factors and social needs may improve patient care and outcomes. Semantic representations of text-encoded SDOH information can augment existing structured data, and this more comprehensive SDOH representation can assist health systems in identifying and addressing these social needs.

GRAPHICAL ABSTRACT



Key words: social determinants of health, natural language processing, machine learning, electronic health records, data mining

INTRODUCTION

Social determinants of health (SDOH) are increasingly recognized for their influence on patient health, accounting for an estimated 40–90% of health outcomes.[1] SDOH include *protective factors* that reduce health risks (eg, family support) and *risk factors* that increase health risks (eg, housing instability).[2] SDOH interventions, such as initiating medication-assisted therapy in opioid patients, demonstrate a clear reduction in mortality.[3] Other studies have demonstrated the importance of SDOH data in improving the prediction of hospital readmissions, medication adherence, suicide attempts, and more.[4,5] Such studies reinforce the importance of screening patients for social needs, so clinical care teams can connect them with needed resources.

Patient SDOH information is captured in the Electronic Health Record (EHR) through structured data and unstructured clinical narrative text. The clinical narrative contains a more nuanced and detailed representation of many SDOH than is available through structured data. For example, substance use (alcohol, tobacco, and drug) is often documented through binary fields (yes/no) in structured data, while clinical narratives often document substance use frequency, amount, and history information. This information can be automatically extracted using natural language processing (NLP) information extraction techniques, which map the unstructured text to a structured SDOH representation. Combining extracted information from clinical narratives with existing structured data yields a more complete patient representation.[6,7] This more complete, automatically-derived patient representation can be used in large-scale secondary use applications, including clinical decision-support systems and retrospective studies. Utilizing already-collected data may reduce the workload and financial resources required for data collection.

BACKGROUND AND SIGNIFICANCE

Secondary use of SDOH information from clinical narratives requires extraction of relevant information and conversion of the SDOH descriptions to structured semantic representations. The extraction of SDOH information from clinical text is increasingly explored; however, the nature and granularity of the target SDOH varies across the research space.[8] Several studies treated SDOH extraction as a text classification task, where labels are assigned at the sentence or note-level.[9–14] Narrative SDOH information has also been the target of relation or event extraction, where SDOH are characterized across multiple dimensions related to status, temporality, and other attributes.[15–17] SDOH information extraction techniques include rule-based[7,18,19] and supervised learning approaches such as Support Vector Machines, random forest, and logistic regression.[8] More recent supervised extraction approaches utilized deep learning architectures, such as convolutional neural networks, recurrent neural networks, and pre-trained transformers, including Bidirectional Encoder Representations from Transformers (BERT)[20] and Text-To-Text Transfer Transformer (T5).[8,14,21,22] Pre-trained transformers, such as BERT and T5, allow pre-training on large quantities of unlabeled text and fine-tuning model parameters to specific classification tasks.[20,21] The fine-tuning of pre-trained transformers is an effective transfer learning strategy that has achieved state-of-the-art performance in several SDOH information extraction tasks.[13,14,22]

Prior studies have applied SDOH extractors to clinical data sets to understand the prevalence of SDOH information within clinical narratives and assess information gained relative to existing structured data. Hatef et al developed hand-crafted linguistic patterns for social isolation, housing insecurity, and financial strain, which were applied to a large clinical data set. [7] Navathe et al used a rule-based system[23] to extract SDOH from notes and demonstrated a more complete representation of patient substance use, depression, housing instability, fall risk, and poor social support can be obtained when combined with diagnosis codes.[6] Zhang et al similarly combined narrative text and structured data to predict patient outcomes using deep learning.[24] Focusing on lung cancer patients, Yu et al utilized BERT and RoBERTa[25] to identify SDOH concepts at the document-level and compared the extracted results with structured EHR data.[13,26]

Contributions

This article presents two main contributions. First, we present a state-of-the-art event-based deep learning extractor for SDOH, the multi-label span-based entity and relation transformer (mSpERT). mSpERT was trained on the Social History Annotated Corpus (SHAC),[17] the benchmark gold standard dataset from the 2022 National NLP Clinical Challenges SDOH extraction task (n2c2/UW SDOH Challenge).[22] In prior work, we developed SHAC to address the limitations of published studies in terms of SDOH representation and normalization. For example, Han et al used BERT for sentence-level SDOH classification but did not extract granular information related to substance types, duration, and frequency.[14] Yu et al extracted text spans referring to smoking but did not identify specific entity types or normalize the spans to SDOH concepts (eg, the phrase "smoked 2 packs per day until 5 years ago" would not be

labeled as a past habit or as containing specific frequency or amount information). SHAC is novel in the granularity of the annotations, size of the corpus, and inclusion of multi-institution data. The granular SDOH annotations enable a broader set of secondary downstream applications. If compared to n2c2/UW SDOH Challenge shared task systems trained and evaluated on SHAC, the mSpERT performance of 0.86 overall F1 would only be surpassed by two teams, which achieved 0.89 and 0.88 overall F1.[22] We provide the code, trained extraction model, and annotated data (https://github.com/Lybarger/sdoh_extraction). To our knowledge, this is the first publicly available SDOH extractor trained on SHAC to the research community.

Our second contribution is a large-scale EHR case study that demonstrates the utility of NLP for SDOH extraction. We measured the prevalence of substance use, living situation, and employment information in the clinical narrative and structured SDOH data. Previous studies exploring the prevalence of SDOH information in the clinical narrative are limited by the EHR dataset size, patient population scope, and extraction methods. These studies have either applied extractors to relatively small, often disease-specific cohorts[6,10,13,15,27] or used rule-based approaches.[6,7,27] In contrast, we applied mSpERT to a large clinical dataset of 225 089 patents and 430 406 notes spanning all patient populations from the University of Washington Medicine. We compared extracted SDOH information from clinical narratives with the structured EHR data. The results show that combining the narrative SDOH information with the existing structured data yields a more comprehensive patient representation, which can help guide patient care, assess health risks, and identify social needs.

MATERIALS AND METHODS

To extract detailed representations of SDOH from the clinical narrative, we developed a high-performing event-based SDOH extractor, mSpERT, using SHAC. mSpERT can extract multiple SDOH events in the patient timeline, including past and current SDOH, and characterize SDOH events through detailed arguments related to status, severity, type, and temporality. Through an EHR case study, we applied mSpERT to a University of Washington (UW) dataset that includes 430 406 notes with social history sections for 225 089 patients and compared the extracted information with existing structured data to quantify differences in SDOH coverage. The structured data captures SDOH information through coarse encounter-level labels. To facilitate a direct comparison between the extracted information and existing structured data, we mapped a subset of the extracted SDOH information to note-level labels. To validate mSpERT on the UW dataset, we randomly sampled and annotated a subset of the notes in the UW dataset with note-level labels that can be directly compared with structured fields. All parts of this work were approved by our institution's IRB. This section presents the: (1) data used, (2) information extraction methodology, and (3) EHR case-study design.

Data

We used SHAC to develop and evaluate mSpERT.[17,22] SHAC includes 4405 annotated social history sections from clinical notes from MIMIC-III[28] and UW. It includes train, development, and test partitions for both sources. SHAC uses an event-based schema, where each event includes a trigger that identifies the event type and arguments that characterize the event. The SHAC event annotation schema characterizes each SDOH event in the patient timeline across multiple dimensions. Figure 1 presents an annotated social history section from SHAC with a slot filling interpretation of the events. In this example, the event schema can differentiate between the current use of heroin and the past use of methamphetamines and crack. It can also resolve the method of use, intravenous, for all substances. The slot-filling representation in Figure 1 illustrates how the SDOH annotation scheme can be mapped to a structured format for utilization in secondary use applications. Table 1 summarizes the arguments for each event type: *Alcohol*, *Drug*, *Tobacco*, *Employment*, and *Living Status*. There are two categories of arguments: (1) *span-only* arguments (green labels in Figure 1), which include an annotated span (eg, "IV") and argument type (eg, *Method*) and (2) *labeled* arguments (blue labels in Figure 1), which include an annotated span (eg, "Prior"), argument type (eg, *Status Time*), and argument subtype (eg, *past*) that normalize the span to key SDOH concepts. The argument subtype labels associated with the labeled arguments provide discrete features for downstream applications and improve the utility of the extracted information. Additional

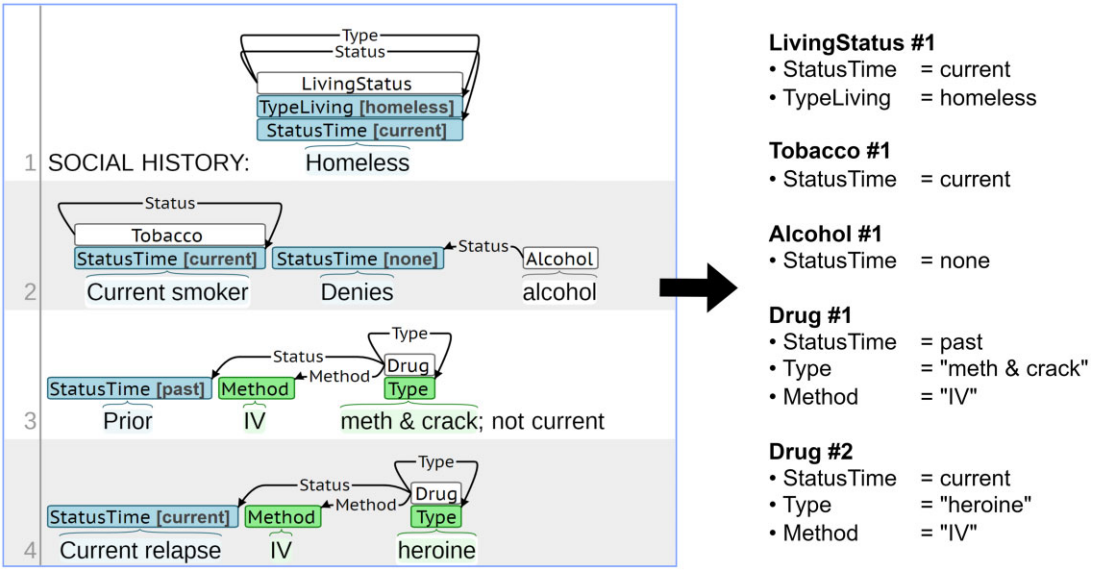


Figure 1. SHAC annotation example (left side) with slot filling interpretation of the annotated events (right side).

Table 1. Annotation guideline summary

Event type	Argument type	Argument subtypes	Span examples
Alcohol, Drug, and Tobacco	Status Time[a]	{none, current, past}	"drinks," "reports"
	Duration	–	"for 10 years"
	History	–	"2 years ago"
	Type	–	"whiskey," "meth"
	Amount	–	"1-2 drinks," "1 pack"
	Frequency	–	"a day," "weekly"
Employment	Status Employ[a]	{employed, unemployed, retired, on disability, student, homemaker}	"working," "retired"
	Duration	–	"for 15 years"
	History	–	"last year"
	Type	–	"construction," "lawyer"
Living status	Status Time[a]	{current, past, future}	"living," "resides"
	Type Living[a]	{alone, with family, with others, homeless}	"with family," "homeless"
	Duration	–	"for 2 years"
	History	–	"until last year"

[a]Indicates a labeled argument. The labeled arguments are required for each event.

Table 2. Data sources used in the EHR case study

Data type	Name	Total records	Total records with social history	Unique patients
Structured	Flowsheets	83 235	–	7875
	Social history	733 591	–	297 581
	Occupation history	120 733	–	42 115
	Employment status	560 940	–	560 940
	Total	**1 498 499**	–	**618 363**
Free-Text	Progress Notes	3 063 025	283 423	140 820
	ED Notes	147 114	19 120	14 619
	Social History Doc.	127 863	127 863	127 863
	Total	**3 338 002**	**430 406**	**225 089**

Note: "Total Records" indicates the total counts of structured data records and free-text documents. "Total Records with Social History" indicates the number of progress and emergency (ED) notes with social history sections and number of social history entries.

information regarding SHAC, including the distribution of note counts by source and annotation details, is available in the original SHAC paper and the n2c2/UW SDOH Challenge paper.[17,22]

To assess the SDOH coverage in the clinical narrative relative to structured data, we created a clinical dataset from UW from January 1 to December 31, 2021, which we refer to as the *UW Data Set*. The UW Data Set includes structured and narrative text data from the UW Epic EHR from outpatient, emergency, and inpatient settings, including 20 medical specialities. Table 2 summarizes the total records and unique patients. UW Data Set contained more than 3.3 million notes for 225 089 patients. In the case study, we processed 430 406 notes with social history sections with mSpeRT. To validate mSpERT, we created the *UW Validation Set*, which consists of 750 randomly sampled documents with social history sections with equal proportions of progress notes, emergency notes, and social history documents.

Information extraction

Event extraction

The SHAC events can be decomposed into a set of relations, where the head is the trigger, tail is an argument, and relation type is the argument role. To extract SHAC events, we introduce mSpERT, which builds on Eberts and Ulges's SpERT.[29] SpERT jointly extracts entities and relations using BERT[20] with output layers that classify spans and predict span relations. SpERT achieved state-of-the-art performance in multiple extraction tasks.[29] SpERT's span-based architecture allows overlapping span predictions but only allows a single label to be assigned to each span; however, the SHAC annotations frequently assign multiple labels to a single span. To adapt SpERT to SHAC, we developed mSpERT. Figure 2 presents the mSpERT framework, which includes three classification layers: (1) Entity Type, (2) Entity Subtype, and (3) Relation. The input is a sentence, and the output is extracted events. The Entity Type and Relation layers are identical to the original SpERT, and the Entity Subtype layer is incorporated to generate multi-label span predictions.

Input encoding. BERT generates a sequence of word-piece embeddings $(\boldsymbol{h}_{CLS}, \boldsymbol{h}_1, \ldots \boldsymbol{h}_t, \ldots, \boldsymbol{h}_n)$, where $\boldsymbol{h}_{CLS}$ is the sentence representation, $\boldsymbol{h}_t$ is the t^{th} word piece embedding, and n is the sequence length.

Entity Type. The Entity Type classifier labels each span, $s_i = (t, t+1, \ldots t+k)$, where i is the span index and $k+1$ is the span width. Learned span width embeddings, $\boldsymbol{w}$, incorporate a span width prior. The span representation, $g(s_i)$, is generated from the BERT embeddings of s_i and the width embeddings, as:

$$g(s_i) = MaxPool(\boldsymbol{h}_t, \boldsymbol{h}_{t+1}, \ldots \boldsymbol{h}_{t+k}) \circ \boldsymbol{w}_{k+1}, \quad (1)$$

where $\circ$ denotes concatenation. The Entity Type classifier is a linear layer, ϕ_e, operating on $\boldsymbol{x}_{s,i}$, defined as"

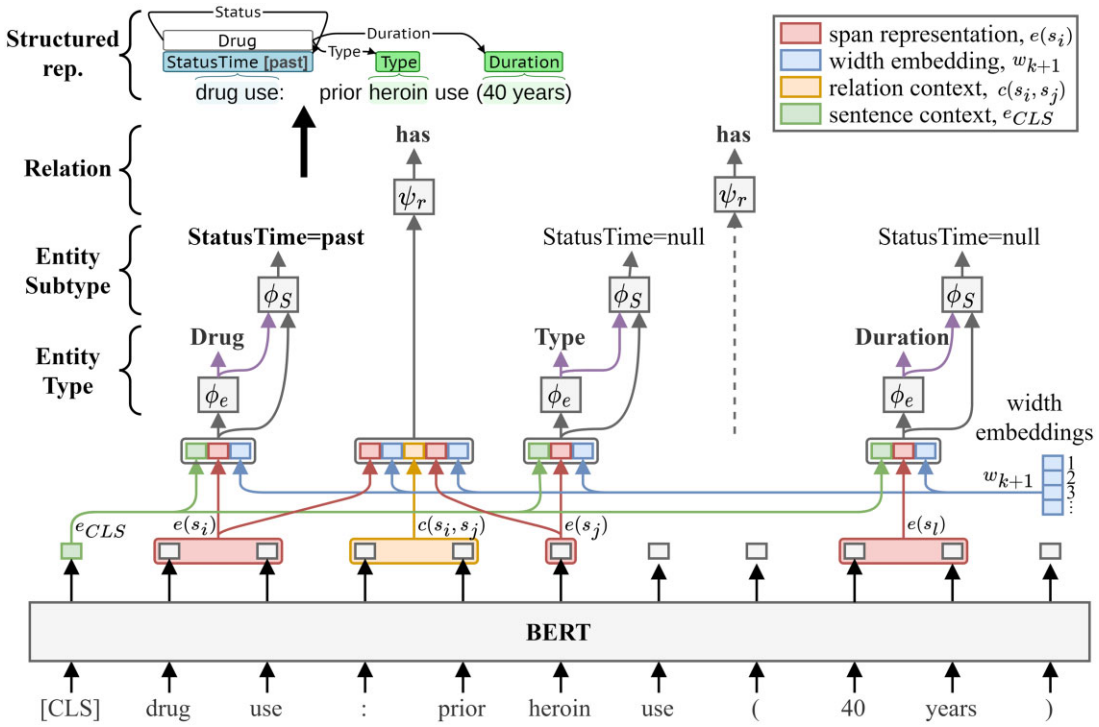


Figure 2. Multi-label Span-based Entity and Relation Transformer (mSpERT) model, which builds on the original SpERT framework.[29]

$$\boldsymbol{x}_{s,i} = g(s_i) \circ \boldsymbol{h}_{CLS}. \quad (2)$$

Entity Subtype. The Entity Subtype classifiers consist of separate linear layers, $\phi_{s,v}$, where v indicates the argument type. The Entity Subtype classifiers operate on the same span representation as the Entity Type classifier and incorporate the Entity Type classifier logits, as:

$$\boldsymbol{x}_{s,v,i} = \boldsymbol{x}_{s,i} \circ \phi_e(\boldsymbol{x}_{s,i}). \quad (3)$$

The Entity Type logits are incorporated to improve the consistency between entity type and subtype predictions.

Relation. The Relation classifier predicts the relationship between a candidate head span, s_i, and a candidate tail span, s_j, with input:

$$\boldsymbol{x}_{r,i,j} = g(s_i) \circ \boldsymbol{c}(s_i, s_j) \circ g(s_j), \quad (4)$$

where $g(s_i)$ and $g(s_j)$ are the head and tail span embeddings and $\boldsymbol{c}(s_i, s_j)$ is the max pooling of the embedding sequence between the head and tail spans. The Relation classifier consists of a linear layer, ϕ_r.

Target Labels. The Entity Type label set, Φ_e, includes the *null* label, event types (*Alcohol*, *Drug*, *Tobacco*, *Employment*, and *Living Situation*), and span-only arguments (*Amount*, *Duration*, *Frequency*, *History*, and *Type*) ($|\Phi_e| = 11$). For all classifiers, *null* is the negative label. There are three Entity Subtype classifiers (*Status Time*, *Status Employ*, and *Type Living*), and the label set for each classifier includes *null* and the applicable subtype labels (eg {*null*, *none*, *current*, *past*} for *Alcohol*). In SHAC, the links between the arguments and triggers can be interpreted as binary connectors (*has* vs *does not have*). Consequently, the Relation label set, Φ_r, is {*null*, *has*}. Only spans predicted to have a non-*null* label by the Entity Type classifier are considered in relation classification.

Training. The classification layers were learned while fine-tuning BERT. The training spans include all the gold spans, S^g, as positive examples and a fixed number of spans with label *null* as negative examples. The training relations include all the gold relations as positive samples, and negative relation examples are created from entity pairs in S^g not connected through a relation. Hyperparameters were tuned using the SHAC training and development sets, and final performance was assessed on the UW partition of the withheld SHAC test set.

Evaluation. We used the n2c2/UW SDOH Challenge evaluation criteria, which interprets event extraction as a slot filling task.[22] In secondary use, there may be multiple semantically similar annotations, and the evaluation uses relaxed criteria that reflect the clinical meaning of the extractions.

Trigger: A trigger is defined by an event type and multi-word span. Trigger equivalence is defined using *any overlap* criteria where triggers are equivalent if: (1) the event types match and (2) the spans overlap by at least one character.

Arguments: Events are aligned based on trigger equivalence, and the arguments of aligned events are compared using different criteria for *span-only* and *labeled* arguments.

Span-only arguments: A span-only argument is defined by an argument type, argument span, and trigger connection. Span-only argument equivalence is defined using *exact match* criteria; span-only arguments are equivalent if: (1) the connected triggers are equivalent, (2) the argument types match, and (3) the spans match exactly.

Labeled arguments: A labeled argument is defined by an argument type, argument subtype, argument span, and trigger connection. Label arguments are defined using a *span agnostic* approach, where labeled arguments are equivalent if: (1) the connected triggers are equivalent, (2) the argument types match, and (3) the argument subtypes match. The argument span is not considered, and the span of the connected trigger is used as a proxy for argument location.

Extraction performance was evaluated using the SHAC gold standard labels. A more detailed description of the scoring criteria

and its justification is available in the n2c2/UW SDOH Challenge paper.[22]

EHR case study

Our EHR case study consisted of two experiments: (1) we validated mSpERT on the UW Data Set using 750 human-annotated documents, called the UW Validation Notes, and (2) we identified 1.4 million SDOH-related structured records for 618 363 patients and compared directly with NLP-derived data from 430 406 documents with social history sections for 225 089 patients written in 2021. The existing structured data did not capture the same granularity of SDOH information as mSpERT, so we mapped the mSpERT output to note-level labels that can be directly compared with structured fields.

Data sources

Structured data. In this study, we used four database tables in the UW Epic EHR: flowsheets, social history, patient employment status, and patient occupation. From the flowsheets table, we leveraged the SDOH-related records identified by Phuong et al[30] to identify employment and housing status. The social history table is primarily composed of Boolean yes/no columns related to alcohol, tobacco, illicit, and recreational drug use. The patient employment status table provides current categorical employment status, such as *Student*, *Full—time*, and *Retired*, while the patient occupation table provides a longitudinal record of free-text occupation titles, such as "Mechanic" or "Therapist." The employment status table does not include timestamps to determine when records were updated, so we limited records to only patients with a completed visit in 2021.

Narrative text data. We used three narrative text sources: (1) progress notes, (2) emergency department (ED) notes, and (3) narrative descriptions of social documentation from an SDOH-related module within our EHR. Progress notes typically document patient clinical status or related health events in outpatient and inpatient settings. ED notes document patient care within an ED setting. Social history documentation is stored in our EHR as longitudinal records with the same text carried forward and edited in subsequent encounters. For simplicity of analysis and to avoid duplicate information, we only analyzed the latest social documentation records for each patient in 2021. Progress and ED notes were pre-processed to extract the social history sections (typically with a header of "SOCIAL HISTORY"). Notes without this section were discarded.

Note classification evaluation

Note-level classification performance was assessed by annotating 750 UW Validation Notes with five multi-class labels (one for each event type): *Alcohol*, *Drug*, and *Tobacco* had labels {*unknown*, *current*, *past*, *none*}; *Employment Status* had labels {*unknown, employed, unemployed, retired, on disability, student, homemaker*}, and *Living Status* had labels {*unknown, alone, with family, with others, homeless*}. The *unknown* label is analogous to the *null* label in mSpERT. Where the patient's status was described multiple times in a document, the most recent value was used. Five medical students annotated our gold standard data set. The initial annotation training round consisted of all annotators labeling the same 15 randomly selected social history sections with the extracted trigger spans from mSpERT pre-labeled. After the initial training round, 750 social history sections were single-annotated. Classification performance was evaluated using accuracy, precision (P), recall (R), and F1. The inter-rater agreement on 15 notes in the initial round of annotation was 0.95 F1.

Structured data and NLP data evaluation

Using the extracted information and existing structured data, we assessed the proportion of patients who had a positive indication for current alcohol, tobacco, and drug use, any description of employment, and current homelessness within the 1-year time period of the UW Data Set. These SDOH were selected because they provide the most direct comparison with existing structured data. In the extracted information and structured data, patients may have multiple descriptions of a given SDOH over time (eg, alcohol use indicated as *current* at one visit but *past* in a subsequent visit). We counted any patient with any positive indication for listed SDOH as positive, regardless of any subsequent changes, given the short time period.

RESULTS

Information extraction

mSpERT was trained on the entire SHAC train set (1316 MIMIC and 1751 UW notes) and evaluated on the UW partition of the SHAC test set (518 notes), as the UW partition is most similar to the UW Data Set. The overall performance in Table 3 is the micro-average across all extracted phenomenon (all event types, triggers, and arguments). The training and test data used to develop the mSpERT SDOH extractor is identical to Subtask C of the n2c2/UW SDOH Challenge.[22] Subtask C included 10 participating teams that used a wide range of extraction approaches including pre-trained transformer-based language models (BERT[20] and T5[21]). The top three teams achieved 0.89, 0.88, and 0.86 overall F1. If compared to shared task systems, the mSpERT performance of 0.86 overall F1 would be similar to the third-place team.

The SHAC event structure most heavily used in the EHR case study includes the triggers and labeled arguments. The triggers

Table 3. Event extraction performance on the UW portion of the SHAC test set

Event type	Argument	# Gold	*P*	R	F1
Substance	Trigger	1310	0.94	0.96	0.95
	Status Time	1310	0.89	0.90	0.89
	Amount	217	0.74	0.76	0.75
	Duration	65	0.77	0.71	0.74
	Frequency	165	0.73	0.74	0.73
	History	103	0.60	0.69	0.64
	Method	102	0.68	0.55	0.61
	Type	319	0.76	0.62	0.68
Employment	Trigger	153	0.94	0.88	0.91
	Status Employ	153	0.90	0.84	0.87
	Duration	5	0.80	0.80	0.80
	History	7	0.80	0.57	0.67
	Type	84	0.80	0.57	0.67
Living Status	Trigger	354	0.88	0.89	0.89
	Status Time	354	0.87	0.87	0.87
	Type Living	354	0.84	0.83	0.83
	Duration	9	0.50	0.33	0.40
	History	2	0.50	0.50	0.50
OVERALL		**5066**	**0.87**	**0.85**	**0.86**

resolve the event's type (*Alcohol*, *Drug*, etc.) and the labeled arguments capture normalized representations of important SDOH. mSpERT achieved high performance in identifying triggers for all events types (0.89–0.95 F1) and resolving the *Status Time*, *Status Employ*, and *Type Living* multi-class labels (0.83–0.89 F1). The span-only argument (eg, *Amount*, *Duration*, etc.) performance varied by argument type and event type.

The *Error Analysis* section of the Supplementary Appendix includes a detailed quantitative and qualitative error analysis, focusing on triggers and labeled arguments. Substance use performance varied by event type (*Alcohol*, *Drug*, and *Tobacco*) and *Status Time* label. Across substance event types, performance was highest for the *Status Time none* label ($\geq 0.94F1$), where descriptions tend to be relatively concise and homogeneous (eg, "Tobacco: denies"). Performance was lower for *current* ($0.80 - 0.91F1$) and *past* ($0.59 - 0.81F1$), which tend to be associated with more heterogeneous descriptions and have higher label confusability. Regarding *Employment*, performance was relatively high for all *Status Employ* labels ($\geq 0.86F1$). The *Type Living* performance was highest for *with family* ($0.91F1$) and *alone* ($0.90F1$) and lower for *homeless* ($0.80F1$) and *with others* ($0.69F1$).

EHR case study

Extractor validation

Table 4 presents mSpERT validation results for the note-level extraction performance on the UW Validation Notes. Precision, recall, and F1 were calculated by considering *unknown* as the negative label, and accuracy was calculated using direct comparisons of all class labels. Comparing Tables 3 and 4 suggests some reduction in performance associated with mapping the events extracted by mSpERT to document-level labels. However, the note-level performance is relatively high across event types (0.77–0.86 F1).

Comparison of extracted and structured information

Table 5 compares the extracted and structured SDOH information, including the proportions of unique patients with *current* substance use (*Alcohol*, *Drug*, and *Tobacco*), any *Employment* information (*employed*, *unemployed*, etc.), and *Living Status* of *homeless*. These selections are most directly comparable between the structured data and extracted SDOH. *Tobacco*, *Drug*, and *Living Status* showed the most significant gains in the number of patients for whom extracted SDOH revealed risk factors not captured by structured data; 32% of homeless patients, 19% of current tobacco users, and 10% of current drug users only have these SDOH captured in the clinical notes without corresponding structured information. Employment showed the lowest relative gain with 11% of patient employment found in both sources and 1% found only by NLP. The *Note Distribution* section of the Supplementary Appendix presents the distribution of extracted event types by note type and provider speciality. The *Alcohol and Drug usage types* section of the Supplementary Appendix presents normalized past and present substance counts.

Table 4. Note-level performance of mSpERT on the UW validation notes

Category	Acc.	*P*	R	F1
Alcohol	0.93	0.88	0.84	0.86
Tobacco	0.92	0.87	0.85	0.86
Drug	0.94	0.87	0.87	0.87
Employment	0.86	0.82	0.72	0.77
Living Status	0.87	0.77	0.80	0.79
OVERALL	0.90	0.84	0.81	0.83

DISCUSSION

Our SDOH extraction approach provides a promising way to identify patients' SDOH and social needs. We demonstrate high performance, especially in identifying SDOH events (triggers) and determining status and type labels. The SDOH from prior work that is most comparable to our EHR case study is tobacco use. 19% of potential smokers identified were found only by NLP, and Navathe et al similarly identified 15% of patients, though only among cardiovascular disease patients and as compared to ICD-9 codes.[6] Yu et al's study of cancer cohorts similarly found 18% of smoking information using only NLP.[13] Our findings differed from Wang et al's, who found 52% of a small cohort with smoking habits using NLP, but all of whom also had corresponding structured data.[15] This may be due to differing institutional practices or other confounding factors.

Unlike previous studies which extracted SDOH using text classification or NER, our detailed SDOH representation may better aid clinicians in identifying SDOH documented in notes by determining chronicity, duration, frequency, and type. This event-based approach can automatically generate detailed summaries of patient SDOH risk factors, reducing clinician chart review time. While EHRs offer structured fields to document social needs, the consistency of this information collection depends on competing priorities.[31] Given the expanding quantity of EHR data, it is increasingly important for clinicians to efficiently identify key information that informs patient care, including SDOH and social needs. NLP serves to bridge the gap between unstructured clinical narratives and structured data by augmenting existing structured data and identifying otherwise unknown social needs. Our study explores the entire patient population at a health system in an urban setting. Our findings may be most generalizable to other urban hospital systems; however, we leave this examination to future work.

Developing SDOH extraction capabilities is timely, given new guidelines released by the Center for Medicare & Medicaid Services that will request screening rates for SDOH and social needs in 2023 and require reporting in 2024.[32] Our investigation indicates important SDOH information can be extracted from the clinical narrative with high performance to augment structured data.

Limitations and future work

While we extracted SDOH from a large clinical data set spanning the UW medical system, our investigation only used progress notes, emergency notes, and social history text, which are a subset of documentation and likely do not represent all documented SDOH. Our EHR case study was limited to data from one year, and the performance of mSpERT for other time periods or institutions is not well understood. The prevalence and patterns of SDOH descriptions in narrative text may vary over time and by institution.

This study is limited by extractor performance and target SDOH. Although performance was relatively high for most SDOH information, extraction errors negatively impact the case study, and certain SDOH will be disproportionately affected. For example, substance abstinence was extracted with higher performance than current or past substance use. The SHAC annotations and this study capture substance use, employment, and living status information;

Table 5. Proportions of unique patients with current SDOH found only in structured data, only by NLP extraction, or in both

	All patients ($N = 618$K)				Patients with social history text ($N = 225$K)			
		SDOH source				SDOH source		
SDOH	# Patients with SDOH info	Struct.	NLP	Both	# Patients with SDOH info	Struct.	NLP	Both
Alcohol (current)	148 221	87%	5%	8%	106 658	82%	7%	11%
Tobacco (current)	34 871	69%	19%	12%	25 675	56%	26%	17%
Drug (current)	42 309	86%	10%	5%	31 222	81%	13%	6%
Employment (any status)	575 278	88%	1%	11%	200 926	66%	4%	30%
Living Status (homeless)	11 567	64%	32%	4%	9390	55%	39%	5%

Note: For purposes of comparison, in the right-most three columns, we further limit structured data to only patients who had social history narrative text as well.

however, there are many important SDOH that are not addressed, such as living environment, access to care, and food security.

Future studies are needed to understand how extracted SDOH should be incorporated into social needs screening. Topics of interest could include methods for integration into the medical record and reducing the need for manual data entry,[33] impact of false positives on stigmatization,[34] and influence on patient access to healthcare or social services.

CONCLUSIONS

SDOH are increasingly recognized for their impact on patient well-being and public health. The clinical narrative contains rich descriptions of SDOH, and the automatic extraction of SDOH from these narratives can enable large-scale use of the information they contain. We introduce a multi-label version of the entity and relation extraction SpERT architecture, mSpERT, which can extract overlapping spans (entities) and assign multiple labels to spans. mSpERT achieves high performance on the UW partition of the SHAC test set at 0.86 F1 overall. mSpERT achieves especially high performance for event (trigger) identification (0.89–0.95 F1) and the status and type arguments (0.83–0.89 F1) that characterize the most salient aspects of the SDOH.

In an EHR case study, we processed 430 406 free-text descriptions of SDOH using mSpERT and automatically compared the extracted structured semantic representations of SDOH to existing structured EHR data. Based on our analysis, combining the narrative SDOH information with the existing structured data yields a more comprehensive patient representation that can be used to guide patient care, assess health risks, and identify social needs.

FUNDING

This work was supported in part by the National Institutes of Health (NIH)—National Cancer Institute (Grant No. R21CA258242-01S1), NIH—National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at the University of Washington (Grant No. T15LM007442), NIH—National Center for Advancing Translational Sciences (NCATS) (Institute of Translational Health Sciences, Grant No. UL1 TR002319). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

KL and NJD are co-lead authors for the manuscript and contributed equally. All authors contributed to the study design. KL, NJD, and RL developed and implemented algorithms and analyzed the data. KL and NJD drafted the initial manuscript. All authors contributed to the interpretation of the data, manuscript revisions, and intellectual value to the manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

This work was done in collaboration with the UW Medicine Analytics Department.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The SHAC data set used in this work will be made available through the University of Washington (see https://github.com/Lybarger/sdoh_extraction for details).

REFERENCES

1. Friedman NL, Banegas MP. Toward addressing social determinants of health: a health care system strategy. *Perm J* 2018; 22 (4S): 18-095. doi:10.7812/TPP/18-095.
2. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. *Milbank Q* 2019; 97 (2): 407–19.
3. Ma J, Bao YP, Wang RJ, *et al*. Effects of medication-assisted treatment on mortality among opioids users: a systematic review and meta-analysis. *Mol Psychiatry* 2019; 24 (12): 1868–83.
4. Nijhawan AE, Metsch LR, Zhang S, *et al*. Clinical and sociobehavioral prediction model of 30-day hospital readmissions among people with HIV and substance use disorder: beyond electronic health record data. *J Acquir Immune Defic Syndr* 2019; 80 (3): 330–41.
5. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *J Am Med Inform Assoc* 2020; 27 (11): 1764–73.
6. Navathe AS, Zhong F, Lei VJ, *et al*. Hospital readmission and social risk factors identified from physician notes. *Health Serv Res* 2018; 53 (2): 1110–36.

7. Hatef E, Rouhizadeh M, Tia I, *et al.* Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform* 2019; 7 (3): e13802.
8. Patra BG, Sharma MM, Vekaria V, *et al.* Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc* 2021; 28 (12): 2716–27.
9. Uzuner Ö, Goldstein I, Luo Y, *et al.* Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008; 15 (1): 14–24.
10. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open* 2021; 4 (3): ooaa069.
11. 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 (2): e0192360.
12. Feller DJ, Zucker J, Srikishan B, *et al.* Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. *AMIA Annu Symp Proc* 2018; 2018: 422–9.
13. 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. *AMIA Annu Symp Proc*. 2021; 2021: 1225–33.
14. 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.
15. Wang Y, Chen ES, Pakhomov S, *et al.* Investigating longitudinal tobacco use information from social history and clinical notes in the electronic health record. *AMIA Annu Symp Proc* 2016; 2016: 1209–18.
16. Yetisgen M, Vanderwende L. Automatic identification of substance abuse from social history in clinical text. *Artif Intell Med* 2017; 10259: 171–81. doi: 10.1007/978-3-319-59758-4_18.
17. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform* 2021; 113: 103631.
18. Lowery B, D'Acunto S, Crowe RP, *et al.* Using natural language processing to examine social determinants of health in prehospital pediatric encounters and associations with EMS transport decisions. *Prehosp Emerg Care* 2023; 27 (2): 246–51.
19. Reeves RM, Christensen L, Brown JR, *et al.* Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *J Biomed Inform* 2021; 120: 103851.
20. Devlin J, Chang MW, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: proceedings of NAACL-HLT 2019; June 2–7, 2019; Minneapolis, MN: Association for Computational Linguistics; 2019: 4171–86.
21. Raffel C, Shazeer N, Roberts A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020; 21 (140): 1–67.
22. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. *J Am Med Inform Assoc* 2023; 30(8): 1367–78.
23. Zhou L, Plasek JM, Mahoney LM, *et al.* Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. *AMIA Annu Symp Proc* 2011; 2011: 1639–48.
24. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak* 2020; 20 (1): 1–11.
25. Liu Y, Ott M, Goyal N, *et al.* RoBERTa: a robustly optimized bert pretraining approach. *arXiv preprint:190711692*. 2019.
26. 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.
27. 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): 1–10.
28. Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3: 160035.
29. Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training. In: European conference on artificial intelligence; August 29–September 8, 2020; IOS Press; 2020: 2006–13.
30. Phuong J, Zampino E, Dobbins N, *et al.* Extracting patient-level social determinants of health into the OMOP common data model. *AMIA Annu Symp Proc* 2021; 2021: 989.
31. Berg K, Doktorchik C, Quan H, *et al.* Automating data collection methods in electronic health record systems: a Social Determinant of Health (SDOH) viewpoint. *Health Systems* 2022: 1–9.
32. Centers for Medicare & Medicaid Services. FY 2023 Hospital Inpatient Prospective Payment System (IPPS) and Long-Term Care Hospital Prospective Payment System (LTCH PPS) Final Rule—CMS-1771-F. https://www.cms.gov/newsroom/fact-sheets/fy-2023-hospital-inpatient-prospective-payment-system-ipps-and-long-term-care-hospital-prospective. Accessed October 21 2022.
33. Bakken S. Can informatics innovation help mitigate clinician burnout? *J Am Med Inform Assoc* 2019; 26 (2): 93-4. doi: 10.1093/jamia/ocy186.
34. Hartzler AL, Xie SJ, Wedgeworth P, *et al.*; SDoH Community Champion Advisory Board. Integrating patient voices into the extraction of social determinants of health from clinical notes: ethical considerations and recommendations. *J Am Med Inform Assoc* 2023; 30(8): 1456–62.