Journal of the American Medical Informatics Association, 2023, **30(8)**, 1363–1366
https://doi.org/10.1093/jamia/ocad121
Advance access publication 16 February 2023
Editorial

Editorial

Advancements in extracting social determinants of health information from narrative text

INTRODUCTION

Social determinants of health (SDoH) are the conditions in which people are born, live, work, and age that affect personal well-being, health outcomes, and life expectancy.[1] SDoH include a range of nonmedical factors, including substance use, quality of domestic life, marital status, employment status, education, race, geography, and other factors that impact health. Understanding patient SDoH can inform patient health care and has the potential to improve health outcomes and reduce health disparities.[2,3] Patient SDoH information is documented in the electronic health record (EHR) and other health-related databases through structured data and free-text (natural language) documents, including patient notes. For many SDoH, the free-text descriptions capture social and behavioral factors with higher prevalence and more detail than is available through structured data. Utilizing free-text SDoH information in large-scale studies, clinical decision-support systems, and other secondary use applications, requires the automatic extraction of key aspects of the SDoH using natural language processing (NLP). NLP-based information extraction maps the unstructured, free-text descriptions of SDoH to structured semantic representations that can be combined with available structured data to create more complete patient profiles.

OVERVIEW

This focus issue highlights studies that advance the extraction and utilization of free-text SDoH information through the: (1) development of state-of-the-art NLP architectures, (2) exploration of unstudied/understudied SDoH, and (3) formulation of patient-centric ethical design and implementation guidelines. The focus issue includes 10 studies selected through rigorous peer review of 24 submissions. Table 1 summarizes the focus issue studies. The studies present resources, such as lexicons,[4] ontologies,[5,6] and annotated datasets[4–8] that support the extraction of free-text SDoH and present methods for utilizing these resources to generate semantic SDoH representations from free-text descriptions.[4–12] Collectively, these papers explore a range of SDoH, including housing stability,[5–12] social support,[5,6] finances,[5,6] food security,[6] transportation access,[6] insurance,[6] marital status,[6] employment,[8–12] alcohol use,[5,8–12] drug use,[4,5,8–12] and tobacco use.[8–12] They investigate the extraction of free-text SDoH from the Veterans Health Administration (VHA),[7] the MIMIC-III dataset,[4,7–13] the National Violent Death Report System (NVDRS),[5] and university hospitals, including University of Florida (UF),[4] University of Washington (UW),[8–12] and University of California San Francisco (UCSF).[6] The studied populations include veterans,[7] suicide victims,[5] aging patients,[4] patients with chronic pain,[6] and others. The presented NLP methods include rule-based[8] and data-driven machine learning (ML)[4–12] approaches. Transformer-based, pretrained large language models (LLM) were used in all extraction studies, and LLM-based approaches consistently outperformed rule-based systems, discrete models (eg, logistic regression), and non-LLM neural networks (eg, convolutional neural networks).[4–12]

The issue includes an overview of the 2022 National NLP Clinical Challenges (n2c2) track 2 on extracting SDoH (n2c2/UW SDoH Challenge) in which participants developed NLP methods for SDoH extraction from EHR narratives,[8] using data from UW and MIMIC-III to explore information extraction, generalizability, and transfer learning. Three papers present methodological solutions that directly responded to the n2c2/UW SDoH Challenge.[9–11] A fourth paper presents an EHR case study at UW, where an extraction model was developed using the n2c2/UW SDoH Challenge data, the extractor was applied to a large clinical data set, and the information gain achieved by combining free-text SDoH information with existing structured data was quantified.[12] The focus issue includes 4 studies that complement the n2c2/UW SDoH Challenge papers by expanding the patient populations, SDoH types, and data explored.[4–7] A guidance paper provides a patient-centric perspective of SDoH extraction, focusing on ethical design and implementation considerations for SDoH extraction systems.[14]

HIGHLIGHTS

The n2c2/UW SDoH Challenge explored the extraction of alcohol, drug, and tobacco use, employment, and living situation information from clinical text. It utilized the Social History Annotated Corpus (SHAC), which consists of deidentified social history sections from UW and MIMIC-III that are annotated using an event-based annotation scheme.[15] In this event scheme, SDoH events in the patient timeline are characterized through triggers and connected arguments that specify status, severity, type, and temporality. The SHAC event extraction task requires identifying trigger and argument spans, resolving the argument roles, linking triggers and

Received: 20 June 2023. **Editorial Decision:** 20 June 2023

Table 1. Description of SDoH focus issue publications

Author	Research area	Study target	Study population	Study data	Contribution	Shared task
Sajdeya et al[4]	Lexicon development and NLP extraction	Preoperative cannabis use status	UF SH surgery patients ≥65 years old (2018–2020)	UF EHR notes (all types) and MIMIC-III notes	Cannabis lexicon, annotated corpus, and LLM extraction	No
Wang et al[5]	Ontology adaptation and NLP extraction	Social, behavioral/lifestyle, and economic factors related to suicide	National suicide victims (2003–2019)	Death investigation narratives from NVDRS	Suicide-specific SDoH-ontology, annotated corpus, and LLM classifier	No
Lituiev et al[6]	Ontology development and NLP extraction	Social support, relationship status, finances, food security, transportation, housing, and insurance	UCSF ISS patients with chronic low back pain (2017–2020)	UCSF EHR notes, patient instructions, and telephone encounters	SDoH ontology, annotated corpus, and LLM classifier	No
Yao et al[7]	NLP extraction	Eviction status	VHA patients with homeless program, social work, or mental health notes (2016–2021)	VHA EHR homeless program, social work, and mental health notes; MIMIC-III	Annotated corpus and prompt-based LLM extraction approach	
Lybarger et al[8]	NLP extraction	Substance use, employment, and living situation	Patients in MIMIC-III (2001–2012) and at UW (2008–2019)	n2c2/UW SDoH Challenge data (notes from MIMIC-III and UW)	Overview of n2c2/UW SDoH Challenge task and results	Yes
Romanowski et al[9]	NLP extraction	Substance use, employment, and living situation	Patients in MIMIC-III (2001–2012) and at UW (2008–2019)	n2c2/UW SDoH Challenge data (notes from MIMIC-III and UW)	LLM seq2seq SDoH event extractor	Yes
Zhao et al[10]	NLP extraction	Substance use, employment, and living situation	Patients in MIMIC-III (2001–2012) and at UW (2008–2019)	n2c2/UW SDoH Challenge data (notes from MIMIC-III and UW)	LLM multistage SDoH event extractor	Yes
Richie et al[11]	NLP extraction	Substance use, employment, and living situation	Patients in MIMIC-III (2001–2012) and at UW (2008–2019)	n2c2/UW SDoH Challenge data (notes from MIMIC-III and UW)	LLM multistage SDoH event extractor	Yes
Lybarger et al[12]	NLP extraction and EHR case study	Substance use, employment, and living situation	UW population, including all medical specialties (2021)	n2c2/UW SDoH Challenge data; UW EHR notes and structured data	LLM extractor, and large-scale EHR case study of narrative SDoH information	No
Hartzler et al[14]	Ethical use of NLP extraction	Ethical considerations for SDoH extraction system design	Marginalized and underrepresented populations emphasized	Perspective article does not use patient data	Ethical guidance for SDoH extraction system design using AI4People framework	No

EHR: electronic health record; LLM: large language models; seq2seq: sequence-to-sequence; UW: University of Washington; VHA: Veterans Health Administration; NVDRS: National Violent Death Reporting System; UF SH: University of Florida Shands Hospital; UCSF ISS: University of California at San Francisco Integrated Spine Service.

arguments, and normalizing argument spans to SDoH concepts. The n2c2/UW SDoH Challenge included 3 subtasks focused on extraction, generalizability, and learning transfer. The best performing systems in each subtask used LLM, like Bidirectional Encoder Representations from Transformers (BERT),[16] Robustly Optimized BERT Pretraining Approach (RoBERTa),[17] and Text-To-Text Transfer Transformer (T5).[18] Of the 15 teams that participated in the challenge, 3 teams published NLP system papers through this focus issue.[9–11] Most of the LLM-based systems utilized encoder-only architectures, like BERT and RoBERTa, to map input text into a vector space and then classify the vector representation.[10,11] Richie et al[11] designed a BERT-based pipeline where triggers and arguments were extracted as sequence tagging tasks. In this multistep approach: (1) triggers in the input sample were extracted, and (2) arguments were extracted for each trigger by using the BERT segment ID inputs to focus the argument extraction on the specified trigger location. Romanowski et al[9] utilized an encoder-decoder architecture, T5, to implement a sequence-to-sequence (seq2seq) approach, where the input is the note text and the output is a structured text representation of the SDoH. Romanowski et al's solution included: (1) additional pretraining of T5 on MIMIC-III notes; (2) the introduction of additional negative samples from MIMIC-III that do not include SDoH; (3) the creation of additional training samples by excerpting shorter text snippets spanning annotated events; and (4) the utilization of in-house data with SDoH annotations.

In addition to the n2c2/UW SDoH Challenge overview and submissions, 4 focus issue papers generated resources and methodological solutions for the extraction of: (1) the eviction status of veterans from VHA notes,[7] (2) SDoH-related circumstance surrounding suicide crisis from death investigation narratives from the NVDRS,[5] (3) the cannabis use for older surgery patients,[4] and (4) SDoH related to chronic low back pain.[6] Wang et al[5] developed a suicide-specific SDoH ontology (Suicide-SDoHO), which is a hierarchical ontology of social, behavioral/lifestyle, and economic factors. They implemented multilabel, multiclass classifiers to generate ontology predictions using BERT. Their exploration includes an analysis of crisis trends over time by sex and age groups, as well as the SDoH extraction performance by the US state. Yao et al[7] investigated the extraction of eviction status information from the EHR of the VHA, focusing on characterizing eviction presence (absent, present, pending, etc.) and eviction period (current, history, future, etc.). The authors introduced the Knowledge Injection based on Ripple Effects of Social and Behavioral Determinants of Health (KIRESH) prompt-based BERT approach, where each input sentence is concatenated with extracted SDoH information and natural language statements describing the eviction presence and eviction period. The concatenated SDoH information was identified using an existing extraction model from prior work. The concatenated statements describing eviction presence and eviction period have the target labels masked, and the prediction task is to resolve the masked label (*cloze* prediction task). KIRESH outperforms BERT baselines that only use the input sentence.

These resource and methodological contributions for SDoH extraction from free-text are complemented in this issue with a perspective article, presented by Hartzler et al[14] which used the AI4People[19] framework and prior literature to provide guidance regarding the design and implementation of systems that extract SDoH. The article describes how the use of artificial intelligence for automated SDoH extractions may have unintended consequences related to stigma, privacy, confidentiality, and trust. It presents recommendations for mitigating these unintended consequences and incorporating patient perspectives. Hartzler et al emphasize the inclusive, transparent, and cooperative engagement of patients.

CONCLUSIONS

This focus issue covers a diverse range of research on extracting SDoH, in terms of the studied SDoH types, patient populations, health outcomes of interest, extraction methodologies, and system design guidance. It emphasizes the importance of the SDoH found in narrative text sources, demonstrates effective extraction methods that achieve high performance, and advocates for patient involvement in SDoH extraction system design. The SDoH types encompass a wide array of social, economic, and environmental factors, including some historically understudied factors. The examined narrative text sources include notes from EHRs and a national suicide database. LLM extraction models are the dominant extraction approach in this body of work, both in terms of prevalence and performance.

There are several challenges and opportunities for the continued advancement of SDoH extraction related to privacy restrictions, data availability, studied SDoH types, text data sources, and innovations in ML and NLP. Privacy concerns and the associated data restrictions remain a fundamental challenge for SDoH extraction and clinical NLP more broadly. The n2c2/UW SDoH Challenge provides a multi-institution SDoH resource to the research community; however, the challenge data only spans a subset of SDoH of interest and may differ in format and content from other hospital systems. More publicly available annotated data sets are needed to facilitate inter-institution collaboration. Even with the contributions of this focus issue, there remain many unstudied and understudied SDoH types, and there are additional sources of text data with important SDoH information, beyond EHR data, for example social media data. Recent developments for generative LLM with billions or trillions of trainable parameters are fundamentally changing NLP research, including clinical NLP. In the context of SDoH extraction, recent advancements in LLM may improve performance, enable previously intractable tasks, and reduce annotated data requirements.

FUNDING

This work was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under Award Numbers R13LM013127 (to OU) and the NIH NLM Biomedical and Health Informatics Training Program at the University of Washington, Grant Number T15LM007442 (to KL). 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

All authors managed the peer-review of the papers in this Focus Issue and made selection decisions. All authors conceptualized the editorial and KL, OU, and MY wrote the editorial. All authors approved the final draft.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The n2c2/UW SDoH Challenge data, SHAC, will be made available through the University of Washington. The availability of other data sets used in this focus issue varies by study.

Kevin Lybarger[1,*], Oliver J. Bear Don't Walk IV[2], Meliha Yetisgen[2], and Özlem Uzuner[1]

[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

*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

REFERENCES

1. Centers for Disease Control and Prevention. Social Determinants of Health [Internet]. 2021. https://www.cdc.gov/socialdeterminants/index.htm. Accessed May 1, 2023.
2. 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 (2): 139–64.
3. Blizinsky KD, Bonham VL. Leveraging the learning health care model to improve equity in the age of genomic medicine. *Learn Health Sys* 2018; 2 (1): e10046.
4. Sajdeya R, Mardini MT, Tighe PJ, *et al.* Developing and validating a natural language processing algorithm to extract preoperative cannabis use status documentation from unstructured narrative clinical notes. *J Am Med Inform Assoc* 2023; 30(8): 1418–28.
5. Wang S, Dang Y, Sun Z, *et al.* An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives. *J Am Med Inform Assoc* 2023; 30(8): 1408–17
6. Lituiev DS, Lacar B, Pak S, Abramowitsch PL, De Marchis EH, Peterson TA. Automatic extraction of social determinants of health from medical notes of chronic lower back pain patients. *J Am Med Inform Assoc* 2023; 30(8): 1438–47.
7. Yao Z, Tsai J, Liu W, *et al.* Automated identification of eviction status from electronic health record notes. *J Am Med Inform Assoc* 2023; 30(8): 1429–37.
8. 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.
9. Romanowski B, Ben Abacha A, Fan Y. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *J Am Med Inform Assoc* 2023; 30(8): 1448–55.
10. Zhao X, Rios A. A marker-based neural network system for extracting social determinants of health. *J Am Med Inform Assoc* 2023; 30(8): 1398–407.
11. Richie R, Ruiz VM, Han S, Shi L, Tsui FR. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. *J Am Med Inform Assoc* 2023; 30(8): 1379–88.
12. Lybarger K, Dobbins NJ, Long R, *et al.* Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J Am Med Inform Assoc* 2023; 30(8): 1389–97.
13. Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 160035.
14. Hartzler AL, Xie SJ, Wedgeworth P, *et al.* Integrating patient voices into the automatic extraction of social determinants of health from clinical records: ethical considerations and recommendations. *J Am Med Inform Assoc* 2023; 30(8): 1456–62.
15. 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.
16. Devlin J, Chang MW, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: *N Am Chapter Assoc Comput Linguist*; 2019: 4171–86; Minneapolis, MN. doi: 10.18653/v1/N19-1423.
17. Zhuang L, Wayne L, Ya S, *et al.* A robustly optimized BERT pre-training approach with post-training. In: *Proceedings of Chinese National Conference on Computational Linguistics*; 2021: 1218–27; Huhhot, China.
18. 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.
19. Floridi L, Cowls J, Beltrametti M, *et al.* AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach (Dordr)* 2018; 28 (4): 689–707.