Extracting social determinants of health from electronic health records: development and comparison of rule-based and large language models-based methods

Bo Wang, PhD[a,b,c,d], Dia Kabir, MS[a,b,c], Cheryl R. Clark, MD, ScD[e], Karmel W. Choi, PhD[a,b,c,d*], Jordan W. Smoller, MD, ScD[a,b,c,d,f*]

[a]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, United States

[b]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, United States

[c]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

[d]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, United States

[e]Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States

[f]Center for Suicide Research and Prevention at MGH and Harvard, Massachusetts General Hospital, Boston, MA, United States

* These authors contributed equally to this work and are co-corresponding authors.

ABSTRACT

Objectives

Social determinants of health (SDoH) are critical drivers of health outcomes but are often under-documented in structured electronic health record data. This study aimed to develop and evaluate scalable methods for extracting seven SDoH domain categories and 23 subcategories from unstructured clinical notes using both rule-based and large language model (LLM)-based approaches.

Methods

We constructed a gold-standard SDoH corpus comprising clinical text segments from 171 patients in the Mass General Brigham Research Patient Data Registry. A rule-based system (RBS) was developed and its performance compared with seven OpenAI GPT models (GPT-4o, 4.1, 4.1-mini, o4-mini, GPT-5, GPT-5-mini, and o3) under zero-shot and few-shot settings with multiple prompting strategies. We also implemented ensemble models combining RBS and LLM outputs via late fusion.

Results

The RBS achieved the highest precision for SDoH domain categories (0.97) but substantially lower recall (0.62). GPT-based models outperformed RBS in overall F1 scores, with GPT-5 and GPT-5-mini (few-shot) achieving the best domain-level F1 of 0.88 and o4-mini achieving the highest subcategory F1 of 0.79. The RBS-GPT ensemble improved domain-level performance to 0.89 F1 with balanced precision (0.90) and recall (0.89). Model performance was consistent across demographic groups.

Conclusion

State-of-the-art GPT models with advanced reasoning capabilities, including the recently released “mini” models (e.g., o4-mini and GPT-5-mini), demonstrated robust performance for SDoH extraction without fine-tuning and outperformed rule-based NLP. Integrating rule-based and LLM approaches further enhanced performance. Our results provide a scalable, cost-efficient framework for accurate identification of SDoH from clinical text, supporting downstream population health research and clinical informatics applications.

INTRODUCTION

Social determinants of health (SDoH) are increasingly recognized as critical factors that influence health outcomes and contribute to health disparities. Unmet social needs, such as financial hardship, food insecurity, housing instability and lack of social support, are estimated to account for 30-55% of health outcomes [1]. In recent years, electronic health records (EHRs) have been a crucial resource of real-world data with applications for risk stratification, pharmacoepidemiology, treatment response prediction, and more. However, the value of EHR-based research has been limited by the under-documentation of important patient information such as SDoH within structured EHR data. Instead, this information is more commonly captured in unstructured clinical notes [2–4]. Unlocking the potential of this SDoH information could have far-reaching implications for population health research and inform clinical decision making [2].

Given the value of SDoH information for a wide range of applications, a growing body of work has focused on extracting SDoH from narrative clinical notes using natural language processing (NLP). Historically, these efforts have relied on either rule-based [5–8] or supervised machine learning approaches [3,9,10]. Although rule-based approaches are interpretable and customizable, they often suffer from low sensitivity due to their dependence on fixed, manually engineered rules. Conversely, supervised learning methods require significant amounts of high-quality annotated training data which can be extremely cost- and labor-intensive to generate. Recent advances in large language models (LLMs) present an opportunity to develop highly scalable solutions for identifying SDoH without the need for substantial annotated data [11–13].

While existing studies have explored applying NLP to extract SDoH from clinical notes with varying degrees of success, several gaps remain. First, the scope of SDoH domains addressed in most studies is limited. A 2021 review found that smoking status is among the most commonly studied SDoH-related domains using NLP, followed by substance abuse and homelessness [2]. In contrast, SDoH factors such as financial problems, social support, employment, food security, and particularly health insurance coverage, remain relatively underexplored [13,14]. Second, rule-based approaches accounted for approximately one-quarter of the reviewed studies (22 of 82) [2]. Among studies exploring the use of LLMs, most have employed either open-source models such as LLaMA-2 and FLAN-T5 [11,13,15,16] or earlier version of proprietary LLMs like GPT-3.5 [12,17], while newer models such as GPT-5 and o4-mini, which offer superior reasoning capabilities, including in healthcare contexts [18,19], remain largely unexplored in SDoH extraction. Notably, a recent study by Keloth et al. [13] emphasized the importance of evaluating the latest LLMs for this task. Lastly, while fine-tuning open-source models has gained traction in recent years for SDoH extraction [11,13,15,17], the development and systematic evaluation of different prompting strategies with state-of-the-art LLMs, including fusion with rule-based systems, remains understudied despite its relevance in resource-constrained settings.

This study addresses these gaps by developing and evaluating methods to extract seven social and behavioral determinants of health (SDoH) factors from unstructured clinical notes across diverse note types, including less-explored domains such as social resources, physical activity, food security, and health insurance status. To capture a more nuanced understanding of patients' social contexts, we developed a fine-grained classification system for each SDoH factor. For example, within health insurance status, we included sub-categories such as *adequate insurance*

coverage, *lack of insurance* and *government-assisted insurance* to capture different levels of access to care and coverage. In addition, to ensure robust evaluation of our NLP systems, we added four contextual attributes to each sub-category in our validation dataset, namely *temporality*, *experiencer*, *hypothetical status* and *uncertainty*, allowing for a deeper analysis of each model's capacity for contextual understanding. This granular approach provides more precise social context information for downstream applications such as health risk stratification. Here, we present and compare two complementary NLP approaches: a rule-based system (RBS) and LLM-based methods leveraging state-of-the-art OpenAI Generative Pretrained Transformer (GPT) models with diverse prompting strategies, for SDoH extraction from clinical text. Both approaches require minimal training data and computational resources. For the LLM-based approach, we examined seven GPT models in both zero-shot and few-shot settings, without fine-tuning, and compared their performance against RBS. Furthermore, we investigated various ensemble strategies using late fusion to combine the two approaches and achieve optimal extraction performance. By systematically evaluating prompting strategies, state-of-the-art GPT models, and RBS–GPT ensembles, this study contributes to addressing gaps in SDoH research and provides a scalable framework for future clinical NLP applications.

METHODS

Data source

The data for this study were obtained from the MGB Research Patient Data Registry (RPDR) [20], a centralized data registry of clinical information from the EHRs across the Mass General Brigham (MGB) health system. The RPDR database includes approximately 7 million patients from 8 hospitals, including two major teaching hospitals, Massachusetts General Hospital and

Brigham and Women's Hospital, encompassing a wide range of patient characteristics including structured data (demographics, diagnoses, medications, procedures, and lab tests) and unstructured narrative notes. We acquired all narrative clinical notes from 95,157 patients enrolled in the MGB Biobank, spanning March 1976 to November 2021. For this study, we used discharge summaries and progress notes, the latter including inpatient, outpatient, and emergency visit notes. The study protocol was approved by the MGB Institutional Review Board, with a waiver of consent for the analysis of electronic health record data.

SDoH categories

We established our SDoH classification schema (Figure 1) through iterative consultations with subject matter experts in psychology, psychiatry, and health disparities research. The final schema encompasses seven domains, i.e., six social and one behavioral (physical activity) determinants of health that are commonly screened in clinical practice [21–23]. To capture granular social contexts relevant to health outcomes, we further defined multiple subcategories within each of the seven domain categories. Given the inherent overlap between certain subcategories (e.g., "patient recently got laid off" could be characterized as both *job loss* and *unemployment*), we developed comprehensive annotation guidelines (Appendix S2) to ensure consistency and reproducibility in SDoH data generation. The schema was also refined to draw more nuanced distinctions. For example, within the *Social Resources* domain, we distinguished subcategories of *living with family/friends* and *living alone* from *good social resources* and *poor social resources* to more precisely distinguish living situation from quality of social support. The initial taxonomy of 27 subcategories was refined to 23 based on data review and feasibility

assessment. Any instance that fell within one of our domain categories but did not match any specific subcategory were assigned to a subcategory labeled "NA" (not applicable).

Gold-standard corpus development

An initial set of 2,000 patients was sampled from the MGB Biobank cohort using stratified sampling based on several key sociodemographic variables: sex, self-reported race, age group (≥65 vs. <65), and health insurance type (public vs. private payer). To create a gold-standard dataset for model development and validation, we conducted a pre-screening phase using predefined keywords from our rule-based system (see next section) to identify sentences in patients' notes likely containing SDoH information. Following the approach of prior work [9,24], for each potential SDoH mention, we extracted text segments with 150 characters of context on either side, providing sufficient information for accurate manual annotation. We randomly sampled 79 unique text segments for annotator training, and a separate set of 226 text segments (10 for each subcategory except for "*Other job insecurity and employment problems*" which had 6 instances) for final model validation.

We developed our annotation workflow using Label Studio (https://labelstud.io/), an open-source data labeling platform (Figure S1). Two annotators (BW: a postdoctoral researcher trained in biomedical informatics; DK: a medical student with experience in neuropsychiatric research) underwent three training sessions and calibration meetings before beginning the final annotation process. Any disagreements between annotators were discussed and resolved in consensus meetings with a clinical psychologist (KWC) serving as the tiebreaker. Based on the finalized

SDoH classification schema (Figure 1) and annotation guidelines, we defined two multi-label annotation tasks for each text segment:

- Task 1: Identify any mention of the 7 SDoH domain categories (bold text in Figure 1).
- Task 2: Identify any mention of the 23 SDoH subcategories (italicized text in Figure 1).

The annotators underwent multiple rounds of training, each beginning with detailed discussion of the annotation guidelines and concluding with inter-annotator agreement (IAA) assessment. After completing three training sessions, they achieved acceptable IAA as measured by Krippendorff's alpha [25] ($\alpha = 0.88$ for domain categories, $\alpha = 0.74$ for subcategories). They then independently annotated the final validation set of 226 text segments, achieving Krippendorff's alpha of 0.95 for domain categories, 0.84 for subcategories alone, and 0.78 for subcategories with contextual attributes (temporality, experiencer, hypothetical status, and uncertainty), respectively, all well above standard thresholds for reliability [26].

Rule-based system

As illustrated in Figure 2, we developed a rule-based system (RBS) to identify mentions of SDoH categories and subcategories from clinical notes. The system was implemented using medspaCy [27], an open-source clinical NLP library that supports flexible integration of rule-based and machine learning-based algorithms. It employs a pipeline of NLP functions spanning sentence and section detection to concept extraction and contextual analysis, to locate, match and disambiguate SDoH mentions, and an iterative process of manual testing, review and refinement based on feedback to optimize model performance. Descriptions for each pipeline component are provided in Appendix S1.3.

Lexicon creation and expansion

RBS requires a lexicon (a dictionary of relevant query terms) and a set of rules for identifying matches in clinical text. We developed our SDoH lexicon through a multi-step process: seed term curation, lexicon expansion, and iterative filtering and refinement (Figure 3). The initial lexicon comprised terms related to each category derived from prior studies and reviews [6,8,22,23,28–32] and SDoH screening tools and surveys (see Appendix S1.1). These seed terms were reviewed by a domain expert (CRC) and systematically expanded into a comprehensive lexicon using: (1) synonym identification and hierarchy mapping (hypernymy-hyponymy) through UMLS Metathesaurus, and (2) semantic similarity search using text embedding models [33–38] to capture related terms and variations. For instance, we encoded the MEDLINE N-Gram Set (1-5 grams) [39,40] using an MPNet-based Sentence-BERT model (*paraphrase-mpnet-base-v2*) [37] (which we make publicly available at our GitHub repository: https://github.com/bwang482/SDoH_Extraction) and performed semantic searches using seed terms as queries against the embedded corpus to identify additional relevant terms for each SDoH category. The iterative lexicon filtering step is detailed in Appendix S1.2.

LLM-based models

We employed a series of OpenAI GPT models through the MGB Azure Enclave, a HIPAA-compliant, secure data platform that provides a dedicated environment for processing protected health information. Using OpenAI's Chat Completions API, we prompted the models to identify SDoH categories and subcategories in clinical text, akin to a hierarchical, multi-label classification task.

While open-source LLMs continue to improve, frontier proprietary models (e.g., GPT-4-class and successors) remain state-of-the-art on many benchmarks, particularly those involving reasoning tasks [19,41,42]. For this study, we selected seven GPT models for SDoH extraction based on their performance in benchmarks including Healthbench [18], as well as their inference cost: GPT-4o, 4.1, 4.1-mini, 5, 5-mini, o3, and o4-mini. We compared their precision and recall against RBS. Additionally, we evaluated their performance (in precision, recall and F1 scores) across different demographic groups and conducted error analysis to understand model limitations and performance variability.

Zero-shot prompting

State-of-the-art LLMs are instruction-tuned on diverse tasks [43,44], enabling them to follow novel instructions and perform unseen tasks without requiring additional training, a capability known as zero-shot generalization [45,46]. Prior studies have highlighted the importance of systematic prompt engineering [13,16,47], for example by incorporating annotation guideline instructions. As shown in Appendix S3, our zero-shot approach employed a three-component system prompt: (1) role-playing instructions to contextualize the task, (2) definitions of SDoH domain categories and subcategories, and (3) step-by-step instructions for extracting SDoH information and structuring the output. The prompt instructed LLMs to follow a hierarchical approach: identify SDoH domain categories first, then their corresponding subcategories, returning "None" if no relevant information was present. We emphasized that only current, patient-specific, and confirmed SDoH information should be extracted, excluding historical, family-member, or hypothetical mentions (e.g., doctor's recommendations).

In addition, we evaluated three prompting styles, namely strict, balanced, and liberal, which differ in their threshold for identifying SDoH categories and subcategories (inclusion criteria). The *strict style* requires explicit evidence, the *balanced style* accepts evidence that is strongly implied, and the *liberal* version takes the most inclusive approach, accepting not only explicit and strongly implied evidence but also information deemed reasonably likely given the clinical text. We hypothesized that these styles would demonstrate a precision-recall trade-off, with the strict style achieving higher precision and the liberal style yielding higher recall.

Few-shot prompting

Following [12], we tested six few-shot prompting strategies using examples from our annotation training: (1) using five examples where annotators readily agreed on the correct identification of SDoH information (i.e., easy examples); (2) same as (1), but with explanations describing the reasoning behind each annotation; (3) using five challenging examples that required annotator adjudication (i.e., hard examples); (4) same as (3), but with explanations added; (5) using synthetically generated examples by GPT; (6) same as (5), but with synthetic explanations. The goal of these experiments was to assess whether state-of-the-art GPT models benefit from examples with or without explanations, and whether exposure to difficult or ambiguous cases, or to synthetic examples, improves their accuracy and robustness in extracting SDoH information from clinical notes.

RBS-GPT ensemble

Drawing on prior work in LLM ensemble learning within the broader NLP domain [48–51], we implemented a hybrid ensemble approach that aggregates the outputs of rule-based and GPT

models using late fusion to improve the overall performance of SDoH extraction. Using the annotator training samples as development data, we found RBS achieved good precision (0.92) in identifying SDoH domain categories but showed poor performance (precision = 0.64) at the subcategory level. Based on these findings, we designed two different ensemble strategies for domain categories and subcategories (Figure 4): (a) for domain categories, we deployed the best performing GPT model with three different prompting styles (strict, balanced, and liberal) to form a small "GPT-committee". The committee's outputs were first aggregated and then combined with those from RBS; (b) for subcategories, we used only the aggregated output of the GPT-committee to determine the final list of SDoH subcategories. We evaluated several ensemble configurations for aggregating outputs from the GPT-committee, testing three fusion functions: (1) majority voting, (2) union, and (3) intersection, and compared them with results from GPT using strict prompting alone. We also examined two fusion functions (union and intersection) for combining the outputs from the RBS and GPT-committee (RBS-GPT fusion).

RESULTS

Gold-standard data annotation

The final annotated validation dataset consisted of 226 text segments from 171 patients, with a total of 410 SDoH domain-category and 475 subcategory mentions. Table 1 summarizes the demographic characteristics of these patients. Notably, our annotated cohort showed balanced distributions of gender and self-reported race, along with a predominance of older patients and those with public health insurance.

Table 1. Patient demographics from the annotated validation dataset. *: Median household income by ZIP code.

Characteristic	**Category**	**N (%)**
Age group	< 25 years	16 (9.4)
	25–44 years	39 (22.8)
	45–64 years	31 (18.1)
	≥ 65 years	85 (49.7)
Gender	Female	86 (50.3)
	Male	85 (49.7)
Self-reported race	Asian	38 (22.2)
	Black/African American	46 (26.9)
	White	43 (25.2)
	Other	44 (25.7)
Ethnicity	Hispanic	17 (9.9)
	Non-Hispanic	154 (90.1)
Marital status	Single	84 (49.1)
	Married	62 (36.3)
	Partner	1 (0.6)
	Divorced	11 (6.4)
	Separated	2 (1.2)
	Widowed	5 (2.9)
	Other/unknown	6 (3.5)
Income level*	< $40 k	17 (10.1)
	$40 k–69 k	88 (52.4)

	$70 k–99 k	46 (27.4)
	≥ $100 k	17 (10.1)
Public payer	Yes	122 (71.4)
	No	49 (28.7)
Veteran status	Yes	2 (2.3)
	No	152 (88.9)
	Unknown	15 (8.8)

As noted earlier, the two annotators achieved strong inter-annotator agreement (Krippendorff's α: 0.95 for categories, 0.84 for subcategories, and 0.78 for subcategories with attributes). The majority (64%) of annotation disagreements for SDoH domain categories were attributed to "Social resources" (which required clearer scope definition) and "General financial status" (where one annotator overlooked questionnaire-derived mentions). Similarly, most subcategory-level disagreements stemmed from poorly formatted in-text questionnaires and inconsistencies in identifying contextual attributes of SDoH mentions. For instance, references to public housing applications prompted discussion about whether they should be labeled as "Hypothetical" under "Subsidized/public housing"; additionally, annotators needed additional guidance to correctly parse question-answer pairs within poorly formatted questionnaire text.

Model performance

Table 2 shows the performance of the rule-based system (RBS) and five GPT-based models (with additional model results reported in Table S1). While RBS achieved the highest precision for SDoH domain categories, it demonstrated substantially lower recall compared to GPT-based

models across both domain categories and subcategories, highlighting the inherent limitation of fixed rules. In contrast, GPT-based models consistently outperformed RBS in recall and F1 scores across both classification levels. At the domain-category level, GPT-5-mini (5-shot) and GPT-5 (5-shot) achieved the highest F1 scores of 0.88, with precision/recall/F1 of 0.91/0.86/0.88 and 0.93/0.83/0.88, respectively. At the subcategory level, the o4-mini models demonstrated the best performance, achieving F1 scores of 0.79 in both zero-shot (precision/recall/F1: 0.80/0.77/0.79) and 5-shot settings (0.81/0.76/0.79). Overall, few-shot prompting yielded modest gains in precision for subcategories, often at the expense of slightly reduced recall. Based on its cost-efficiency and competitive performance relative to larger models at both levels, as demonstrated in the per-category and per-subcategory results (Table S2 and S3), we selected o4-mini (5-shot) as our primary model for further experiments involving ensemble modeling and integration with the rule-based approach.

Table 2. Overall performance comparison between RBS and GPT-based approaches. The o4-mini, GPT-5-mini and GPT-5 models have a "reasoning effort" parameter (low, medium or high) that controls the depth of the model's "chain-of-thought" reasoning processing. The reasoning models reported in this table were run with the high-reasoning setting. Precision, recall and F1 scores are presented for both SDoH domain categories and subcategories. The highest scores for each metric (column) are in bold and underlined.

Models	SDoH domain categories			SDoH subcategories		
	Precision	Recall	F1	Precision	Recall	F1
RBS	**<u>0.97</u>**	0.62	0.75	0.67	0.41	0.51
GPT-4o (zero-shot)	0.92	0.81	0.86	0.75	0.74	0.74
GPT-4o (5-shot)	0.91	0.81	0.85	0.78	0.74	0.76

GPT-4.1 (zero-shot)	0.92	0.81	0.86	0.77	0.73	0.75
GPT-4.1 (5-shot)	0.91	0.77	0.84	0.78	0.72	0.75
GPT-4.1-mini (zero-shot)	0.89	0.84	0.86	0.71	0.73	0.72
GPT-4.1-mini (5-shot)	0.90	0.82	0.86	0.73	0.71	0.72
o4-mini (zero-shot)	0.89	0.76	0.82	0.80	0.77	**0.79**
o4-mini (5-shot)	0.89	0.74	0.81	**0.81**	0.76	**0.79**
GPT-5-mini (zero-shot)	0.89	0.82	0.85	0.72	**0.78**	0.75
GPT-5-mini (5-shot)	0.91	**0.86**	**0.88**	0.74	0.76	0.75
GPT-5 (zero-shot)	0.88	0.75	0.81	0.78	0.76	0.77
GPT-5 (5-shot)	0.93	0.83	**0.88**	0.80	0.74	0.77

Prompting comparison

Among the strict, balanced, and liberal prompting styles, we observed the anticipated precision-recall trade-off (Table S4): strict prompting achieved the highest precision and F1 scores for extracting subcategory concepts in both zero-shot and 5-shot settings, while liberal prompting had the best recall, and balanced prompting yielded intermediate results.

Performance varied only marginally across different example types (easy, hard, or synthetic), with F1 scores ranging from 0.81-0.83 for domain categories and 0.76-0.79 for subcategories, as shown in Table S5. Hard examples that required annotator adjudication yielded better performance for SDoH subcategories, particularly in precision (0.81 vs. 0.76 for easy examples and 0.78 for synthetic examples). Adding explanations to the prompt improved precision for subcategories in the hard-example model (from 0.79 to 0.81) but had little impact on other models or metrics.

Ensemble performance

As described in Methods, we implemented different ensemble strategies at each classification level: a two-step fusion for domain categories (GPT-committee ensemble followed by late fusion with RBS) and a GPT-committee ensemble for subcategories. As shown in Table S6 (Supplementary Material), the ensemble configuration that applied *majority voting* to the GPT-committee outputs followed by a *union* (∪) with the RBS outputs, achieved the highest F1 score and the most balanced performance for domain categories, with precision above 0.90 and recall above 0.89. Among the 402 extracted domain-category labels in total, 33 (8.2%) were contributed by RBS alone, demonstrating its added value; 140 (34.8%) came from GPT-committee, indicating their ability to capture patterns missed by RBS; and the remaining 229 (57.0%) were identified by both systems. For subcategories, the *Strict-only* configuration had the optimal performance (Table S6; precision = 0.81, recall = 0.76, F1 = 0.79), consistent with our objective of maximizing the F1 score with balanced precision and recall. The *Intersection* configuration followed closely, also achieving an F1 score of 0.79, with higher precision (0.87) but lower recall (0.72).

Performance across demographic groups

When stratified by gender, self-reported race, and ethnicity, we found no statistically significant performance differences across demographic groups for either RBS or o4-mini in identifying SDoH domain categories and subcategories (Table 3, Tables S7-S11).

Table 3. F1 scores of o4-mini (5-shot) across SDoH categories by demographic group. M: Male; F: Female; W: Self-reported White; NW: Self-report non-White; H: Hispanic; NH: Non-Hispanic. The

scores reported below are not normally distributed, as confirmed by the Shapiro-Wilk tests. No statistically significant performance differences were observed (Mann–Whitney U test, $p < 0.05$).

SDoH	Overall	M	F	W	NW	H	NH
Social resources	*0.83*	0.86	0.80	0.80	0.84	0.95	0.81
Physical activity	*0.89*	0.76	1.0	0.94	0.86	1.0	0.88
General financial status	*0.72*	0.76	0.69	0.76	0.71	0.86	0.71
Employment status	*0.90*	0.91	0.89	0.88	0.90	0.86	0.90
Housing status	0.87	0.90	0.84	0.88	0.86	1.0	0.85
Food security status	*0.71*	0.73	0.70	0.83	0.67	0.75	0.71
Insurance status	*0.85*	0.80	0.89	0.83	0.85	0.80	0.85
Mann-Whitney U test		*P-value = 1.0*		*P-value = 0.798*		*P-value = 0.200*	

Comparison to ICD codes

As expected, SDoH information was identified more frequently by both rule-based and LLM-based NLP methods than documentation based on ICD codes (V and Z codes). Of the 226 annotated text segments used for validation, our RBS identified SDoH information in 225, and the 5-shot GPT models identified SDoH information in 191 (o4-mini), 201 (GPT-4o), and 204 (GPT-4.1) segments. For comparison, relevant ICD codes appeared in only 84 visits using Z and V codes (see Table S12) or 11 visits using selected Z and V codes specific to each SDoH category (see Table S13) within ±7 days of the corresponding clinical documentation, highlighting the value of extracting SDoH from clinical text.

DISCUSSION

This study aimed to extract SDoH factors across seven domain categories, including less studied domains such as social resources and health insurance status, and 23 corresponding subcategories from discharge summaries and progress notes, including ambulatory visit notes, within a large health system comprising more than eight hospitals. We evaluated the performance of seven GPT models using various prompt engineering approaches and compared their results with those of an expert-designed, iteratively-optimized rule-based system (RBS). The newer GPT models with improved reasoning capabilities, such as GPT-5 and o4-mini, achieved superior overall performance across both SDoH domain categories and subcategories. While RBS demonstrated the highest precision for domain categories, it exhibited consistently low recall due to the inherent rigidity of manually engineered rules. For example, GPT models correctly identified "patient needs WIC" (WIC: Women, Infants, & Children Nutrition Program) as an indicator of "Food insecurity" by recognizing WIC's role as a food assistance program, while appropriately excluding gastrointestinal-related eating difficulties as medical rather than socioeconomic issues. In contrast, the RBS failed to capture the WIC reference due to its reliance on predefined lexicons and misclassified gastrointestinal-related eating problems as food insecurity, lacking the semantic understanding to distinguish medical from social determinants. Our findings differ from Patra et al. [15], who reported superior RBS performance over LLM-based approach for classifying social support and social isolation, primarily due to the bias introduced when their RBS closely mirrored their gold-standard annotation rule book, leading to overfitting and inflated performance. Comparing efficiency in model development, we found that building GPT-based approaches in zero-shot or few-shot settings required substantially less development time and cost than constructing the RBS, which involved an iterative process of

rule creation and refinement. Among the GPT models, our findings suggest that newer "mini" models, such as GPT-5-mini and o4-mini, which retain reasoning capabilities, perform competitively with OpenAI's flagship GPT-5 model while offering significantly reduced per-token costs, making them suitable for cost-sensitive applications. In contrast, earlier small models such as GPT-4o-mini and GPT-4.1-mini, though even more cost-efficient, do not perform on-par with their respective full models.

Our prompt engineering evaluation highlighted the sensitivity of model performance to the wording of prompts, showing a precision-recall trade-off can be tuned by adopting stricter or more liberal prompt styles. For few-shot prompting, when comparing the addition of different types of examples, we found that including cases where annotators encountered ambiguity and required clarification or adjudication to reach agreement, along with explanations, yielded optimal performance for SDoH subcategories, suggesting the value of exposing LLMs to examples that humans themselves found challenging and ambiguous. For example, to improve GPT models' ability to parse questionnaire text, we provided examples demonstrating how to identify question-answer pairs and when to recognize missing answers, in which case no subcategory should be assigned. Taken together, our prompt engineering findings suggest that adopting a systematic approach of prompt design and incorporating challenging examples should be considered as best practices for LLM-based extraction method development.

Lastly, we developed and evaluated various model ensemble strategies using late fusion. We found combining the RBS with a committee of GPT models employing different prompting styles (strict, balanced, and liberal) improved extraction performance for domain categories

compared with using GPT models alone. For example, the RBS-GPT hybrid system correctly recognized "involvement in spiritual community" as a mention of "Social resource", whereas the GPT models alone failed to do so. For subcategories, the GPT-committee demonstrated that precision or recall could be selectively improved depending on the fusion function applied, although it did not result in an overall performance gain. These findings highlight the complementary strengths of rule-based and LLM-based systems. However, the lack of improvement at the subcategory level indicates the need for more advanced and adaptive ensemble strategies, such as mixture-of-experts (MoE) approaches where different models ("experts") handle different inputs, a direction we leave for the future investigation.

There are several limitations of the presented study. First, both our RBS and GPT-based models were validated in a single healthcare system, and performance in other systems may vary depending on variation in patient characteristics and documentation practices. Nevertheless, we note that the MGB system comprises more than eight hospitals and multiple community health centers with heterogeneous catchment areas and clinical practices, supporting some degree of generalizability. In addition, our gold-standard validation data were sampled to enrich several key sociodemographic variables, helping to mitigate potential biases related to patient characteristic variations. Second, our primary goal was to develop a cost-efficient and scalable SDoH extraction pipeline for resource-constrained settings, which motivated our focus on rule-based and LLM-prompting based approaches that can be deployed with minimal annotation effort and computational resources. We did not explore the use of synthetic data generation for training supervised models or LLM fine-tuning, which, while potentially reducing annotation cost, still require computation for model training. Future work could examine how these

approaches compare with those evaluated here. Third, although recent GPT models are considered state-of-the-art for many tasks, particularly those requiring reasoning capabilities, we did not conduct a systematic comparison with open-source LLMs. Such a comprehensive evaluation would provide a more complete assessment of model performance for SDoH extraction. Finally, since we applied rule-based filtering during data sampling, the recall and F1 scores reported here are based on an enriched validation set and therefore do not fully reflect performance in the general population of clinical notes. Unlike precision, which remains directly interpretable, recall and F1 should be understood as relative measures within the enriched evaluation setting rather than absolute indicators of real-world performance.

The ability to accurately and efficiently extract SDoH factors from clinical notes has important implications for healthcare systems, as such information is often under-documented in structured EHR data. Our approach provides a scalable method for SDoH extraction, demonstrating consistently high precision and recall across both domain categories and subcategories. By pre-screening clinical notes with a rule-based NLP filter to identify relevant text segments before applying LLM-based extraction, we improved efficiency and substantially reduced computational costs compared with processing entire patient notes directly with LLMs. The approach presented in this study has the potential to advance important population health research and ultimately inform clinical outcomes and evidence-based policy intervention [6,52]. Incorporating SDoH information into clinical prediction models is expected to improve prediction performance, as recent studies suggest [53,54], while also supporting audits of algorithmic biases beyond traditional demographic variables [55]. Finally, as interest grows in

implementing such NLP systems in clinical operations, it will be critical to anticipate and mitigate potential unintended consequences of automated SDoH extraction for patients [56].

Future work will be needed to evaluate the robustness and generalizability of our approach to other healthcare systems, especially those with very different clinical settings, patient demographics and socioeconomic environments. Furthermore, social determinants are highly context dependent. Existing classification models often categorize a patient's social circumstance into predefined labels without capturing context, which limits their ability to provide a comprehensive and contextually relevant understanding of the patient's social needs. Lybarger et al. [57] employed an event-based annotation scheme characterizing each SDoH mention with multiple arguments (e.g., *employment status, duration, history, and type* for "employment") for the task of extracting "substance use", "living status" and "employment" from the social history sections. We plan to explore methods for incorporating richer contextual information that further refines each SDoH extraction, moving beyond the contextual attributes used in our rule-based approach and related work [5,13,57]. In particular, we will investigate leveraging text summarization techniques to generate structured representations that capture a comprehensive view of a patient's SDoH status from each clinical encounter. Lastly, it will also be important to validate the utility of the extracted SDoH information, along with community-level socioeconomic factors [52], in supporting downstream health research applications such as suicide prediction [9,54,58], as well as in conducting model audits, as highlighted in recent studies [55,59].

CONCLUSION

This study presented and compared two NLP approaches, an RBS and GPT-based models, for extracting SDoH information from clinical notes. Recent GPT models with advanced reasoning capabilities achieved superior performance in identifying seven SDoH domain categories and 23 subcategories with high precision and recall, requiring no additional training or fine-tuning. In addition, we developed and validated late-fusion model ensemble combining both approaches to optimize extraction performance. By making our code and prompts to the scientific community, we provide a scalable and cost-efficient solution for accurate SDoH extraction, with significant potential for advancing important downstream health research applications.

FUNDING

This work was supported in part by funding from the Brain and Behavior Research Foundation Young Investigator Award (BW) and NIMH grants P50MH129699, R01MH118233, R01MH137218 (JWS) and K08MH127413 (KWC).

CONFLICT OF INTEREST

JWS reported grants from Biogen, Inc and serving as a scientific advisory board member with options from Sensorium Therapeutics, Inc outside the submitted work. KWC is a paid consultant and member of the Mind Advisory Committee of Sword Health. No other disclosures were reported.

DATA AVAILABILITY

Protected Health Information restrictions apply to the availability of the clinical data here, which were used under IRB approval for use only in the current study. As a result, this dataset is not publicly available. The code of our rule-based system and LLM prompts is available at https://github.com/bwang482/SDoH_Extraction.

FIGURE LEGENDS

Figure 1: Hierarchical schema of SDoH categories and subcategories.

Figure 2: Overview of the rule-based system (RBS) workflow.

Figure 3: Lexicon curation process for RBS.

Figure 4: RBS-GPT ensemble workflow with separate fusion paths for SDoH domain categories and subcategories.

Figure 1. SDoH classification schema, consisting of "General financial status" (cyan), "Employment status" (beige), "Housing status" (yellow), "Health insurance status" (pink), "Food security status" (orange), "Social resources" (blue), and "Physical activity" (purple). The seven domain categories are shown in bold text, and the 23 subcategories are shown in italic text.

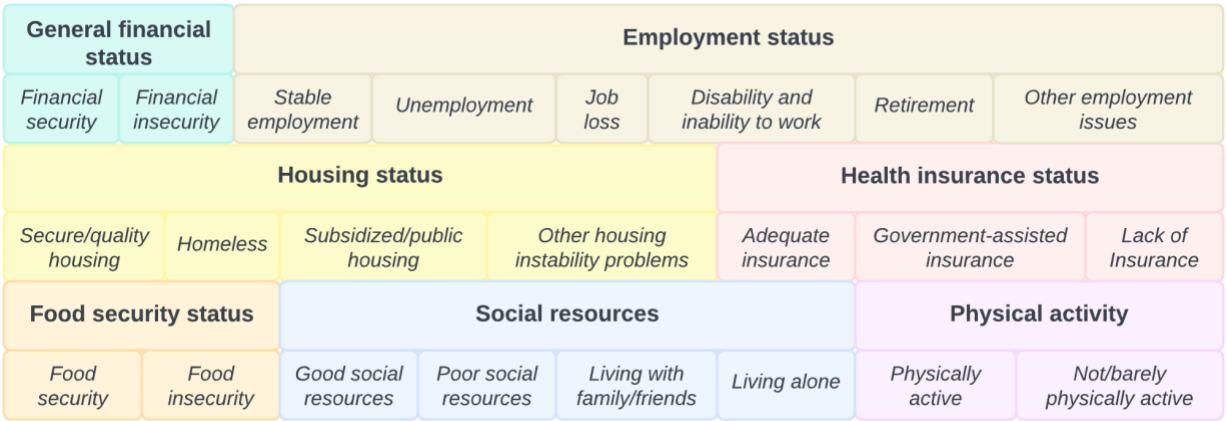


Figure 2. RBS workflow, which takes clinical notes as input, processes them through its text processing pipeline, including sentence segmentation, lexicon-based entity matching, and context disambiguation, and outputs sentences identified with SDoH domains and subcategories. The pipeline was developed iteratively involving manual reviewing and refinement of the rules.

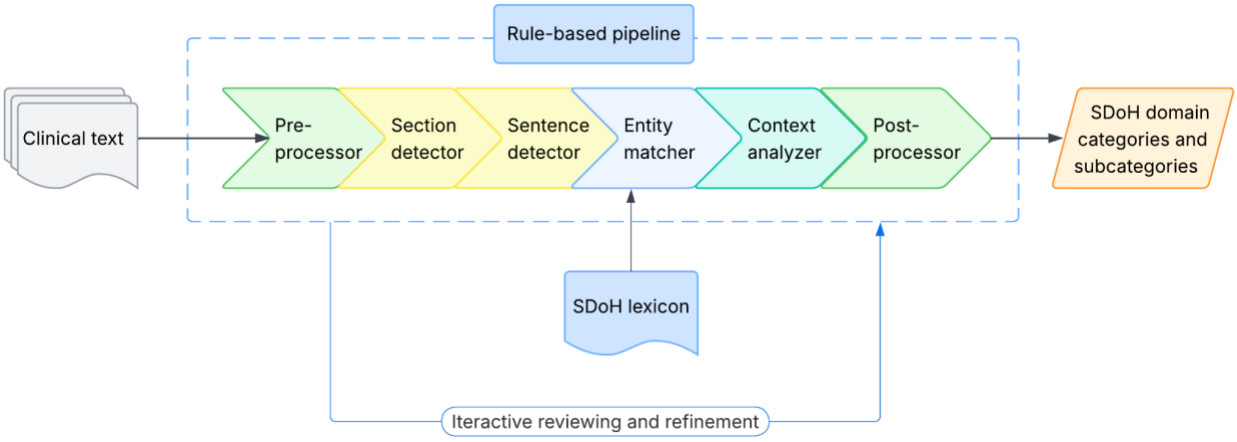


Figure 3. Lexicon curation process, involving seed term curation, lexicon expansion using UMLS and embedding-based semantic search, and an iterative process of lexicon filtering.

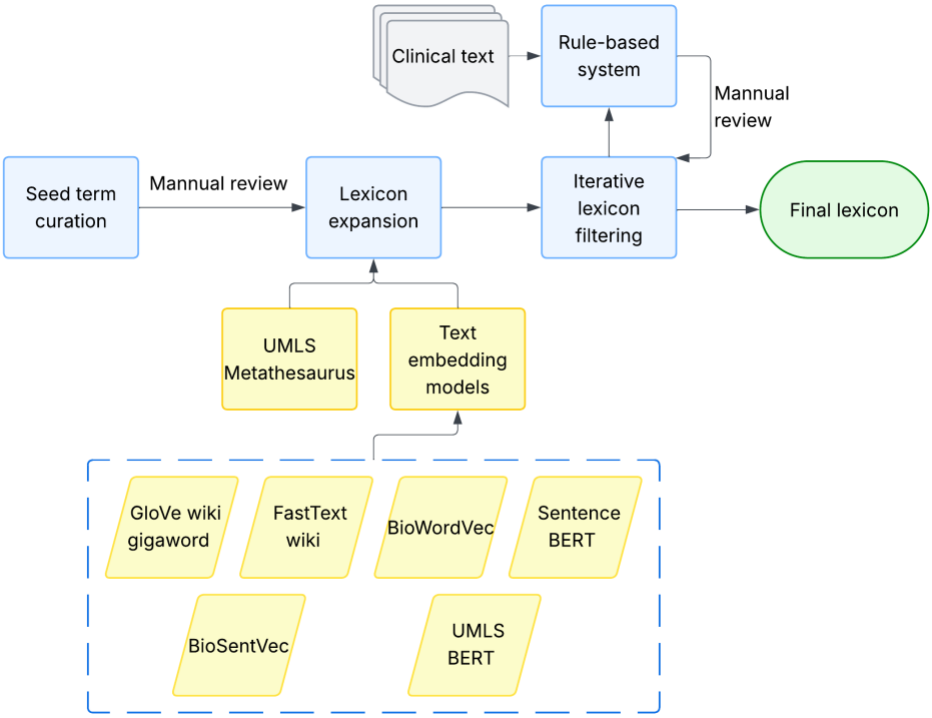
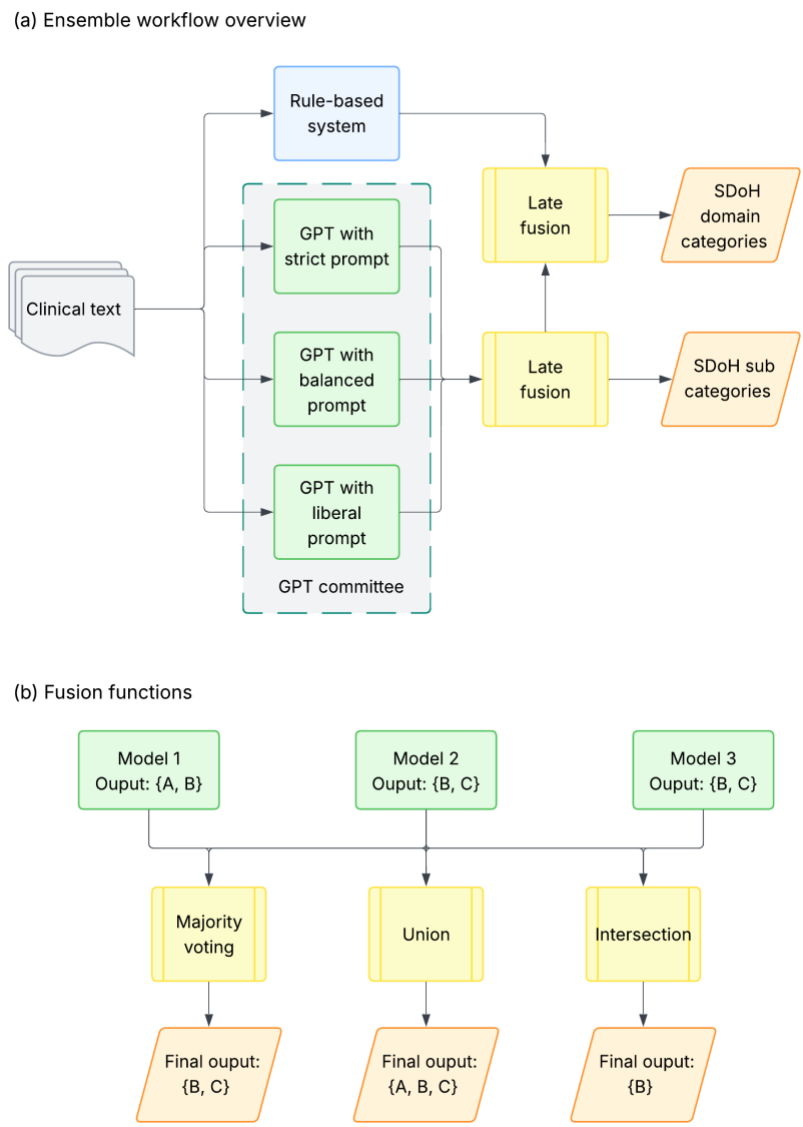


Figure 4. (a) Overview of the RBS–GPT ensemble workflow, showing separate fusion paths for SDoH domain categories and subcategories. For domain categories, outputs from both RBS and GPT with different prompting styles (the GPT-committee) are aggregated via late fusion. For subcategories, only the GPT-committee outputs are aggregated. (b) Illustration of the fusion functions, namely *majority voting*, *union* and *intersection*.



REFERENCES

1 WHO. Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health (accessed 21 Feb2024).

2 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. doi:10.1093/jamia/ocab170

3 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. doi:10.1016/j.jbi.2021.103984

4 Mehta S, Lyles CR, Rubinsky AD, *et al.* Social determinants of health documentation in structured and unstructured clinical data of patients with diabetes: comparative analysis. *JMIR Med Inform* 2023;**11**:e46159. doi:10.2196/46159

5 Chapman AB, Jones A, Kelley AT, *et al.* ReHouSED: A novel measurement of Veteran housing stability using natural language processing. *J Biomed Inform* 2021;**122**:103903. doi:10.1016/j.jbi.2021.103903

6 Bejan CA, Angiolillo J, Conway D, *et al.* Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc* 2018;**25**:61–71. doi:10.1093/jamia/ocx059

7 Conway M, Keyhani S, Christensen L, *et al.* Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *J Biomed Semantics* 2019;**10**:6. doi:10.1186/s13326-019-0198-0

8 Zhu VJ, Lenert LA, Bunnell BE, *et al.* Automatically identifying social isolation from clinical narratives for patients with prostate Cancer. *BMC Med Inform Decis Mak* 2019;**19**:43. doi:10.1186/s12911-019-0795-y

9 Mitra A, Pradhan R, Melamed RD, *et al.* Associations Between Natural Language Processing-Enriched Social Determinants of Health and Suicide Death Among US Veterans. *JAMA Netw Open* 2023;**6**:e233079. doi:10.1001/jamanetworkopen.2023.3079

10 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**:1448–55. doi:10.1093/jamia/ocad071

11 Guevara M, Chen S, Thomas S, *et al.* Large language models to identify social determinants of health in electronic health records. *npj Digital Med* 2024;**7**:6. doi:10.1038/s41746-023-00970-0

12 Consoli B, Wang H, Wu X, *et al.* SDoH-GPT: using large language models to extract social determinants of health. *J Am Med Inform Assoc* Published Online First: 10 June 2025. doi:10.1093/jamia/ocaf094

13 Keloth VK, Selek S, Chen Q, *et al.* Social determinants of health extraction from clinical notes across institutions using large language models. *npj Digital Med* 2025;**8**:287. doi:10.1038/s41746-025-01645-8

14 Yu Z, Peng C, Yang X, *et al.* Identifying social determinants of health from clinical narratives: A study of performance, documentation ratio, and potential bias. *J Biomed Inform* 2024;**153**:104642. doi:10.1016/j.jbi.2024.104642

15 Patra BG, Lepow LA, Kasi Reddy Jagadeesh Kumar P, *et al.* Extracting social support and social isolation information from clinical psychiatry notes: comparing a rule-based natural language processing system and a large language model. *J Am Med Inform Assoc* Published Online First: 18 October 2024. doi:10.1093/jamia/ocae260

16 Gu B, Shao V, Liao Z, *et al.* Scalable information extraction from free text electronic health records using large language models. *BMC Med Res Methodol* 2025;**25**:23. doi:10.1186/s12874-025-02470-z

17 Gabriel RA, Litake O, Simpson S, *et al.* On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proc Natl Acad Sci USA* 2024;**121**:e2320716121. doi:10.1073/pnas.2320716121

18 Arora RK, Wei J, Hicks RS, *et al.* HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv* Published Online First: 2025. doi:10.48550/arxiv.2505.08775

19 Moëll B, Farestam F, Beskow J. Swedish Medical LLM Benchmark: development and evaluation of a framework for assessing large language models in the Swedish medical domain. *Front Artif Intell* 2025;**8**:1557920. doi:10.3389/frai.2025.1557920

20 Nalichowski R, Keogh D, Chueh HC, *et al.* Calculating the benefits of a Research Patient Data Repository. *AMIA Annu Symp Proc* 2006;:1044.

21 Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains in electronic health records: phase 1*. Washington (DC): : National Academies Press (US) 2014. doi:10.17226/18709

22 Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains and measures in*

electronic health records: phase 2. Washington (DC): : National Academies Press (US) 2015. doi:10.17226/18951

23 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**:e13802. doi:10.2196/13802

24 Botelle R, Bhavsar V, Kadra-Scalzo G, *et al.* Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. *BMJ Open* 2022;**12**:e052911. doi:10.1136/bmjopen-2021-052911

25 Krippendorff K. *Content analysis: an introduction to its methodology*. 2455 Teller Road, Thousand Oaks California 91320 : : SAGE Publications, Inc. 2019. doi:10.4135/9781071878781

26 Marzi G, Balzano M, Marchiori D. K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX* 2024;**12**:102545. doi:10.1016/j.mex.2023.102545

27 Eyre H, Chapman AB, Peterson KS, *et al.* Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc* 2021;**2021**:438–47. doi:10.48550/arxiv.2106.07799

28 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* 2021;**4**:ooaa069. doi:10.1093/jamiaopen/ooaa069

29 Kim H, Kim J, Shenvi E, *et al.* Developing a semantic model to describe physical activity data. *Stud Health Technol Inform* 2016;**225**:447–51.

30 Hollister BM, Restrepo NA, Farber-Eger E, *et al.* Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac Symp Biocomput* 2017;**22**:230–41. doi:10.1142/9789813207813_0023

31 Winden TJ, Chen ES, Monsen KA, *et al.* Evaluation of flowsheet documentation in the electronic health record for residence, living situation, and living conditions. *AMIA Jt Summits Transl Sci Proc* 2018;**2017**:236–45.

32 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.

33 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*

Processing (EMNLP). Stroudsburg, PA, USA: : Association for Computational Linguistics 2014. 1532–43. doi:10.3115/v1/D14-1162

34 Bojanowski P, Grave E, Joulin A, *et al.* Enriching Word Vectors with Subword Information. *Trans Assoc Comput Linguist* 2017;**5**:135–46. doi:10.1162/tacl_a_00051

35 Zhang Y, Chen Q, Yang Z, *et al.* BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019;**6**:52. doi:10.1038/s41597-019-0055-0

36 Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE 2019. 1–5. doi:10.1109/ICHI.2019.8904728

37 Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2019. 3973–83. doi:10.18653/v1/D19-1410

38 Michalopoulos G, Wang Y, Kaka H, *et al.* Umlsbert: clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2021. 1744–53. doi:10.18653/v1/2021.naacl-main.139

39 Lu CJ, Tormey D, McCreedy L, *et al.* Generating the MEDLINE N-Gram Set. American Medical Informatics Association 2015. 1569.

40 Lu CJ, Tormey D, McCreedy L, *et al.* Generating a Distilled N-Gram Set - Effective Lexical Multiword Building in the SPECIALIST Lexicon. In: *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications 2017. 77–87. doi:10.5220/0006142000770087

41 Wu S, Koo M, Blum L, *et al.* Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. *NEJM AI* Published Online First: 17 January 2024. doi:10.1056/AIdbp2300092

42 Wang, Shansong, Hu, *et al.* Capabilities of GPT-5 on Multimodal Medical Reasoning. *arXiv* 2025.

43 Chung HW, Hou L, Longpre S, *et al.* Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* 2024.

44 Ouyang L, Wu J, Jiang X, *et al.* Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 2022.

45 Wei J, Bosma M, Zhao V, *et al.* Finetuned Language Models Are Zero-Shot Learners. *International Conference on Learning Representations* 2022.

46 Sanh V, Webson A, Raffel C, *et al.* Multitask Prompted Training Enables Zero-Shot Task Generalization. *International Conference on Learning Representations* 2022.

47 Hu Y, Chen Q, Du J, *et al.* Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc* 2024;**31**:1812–20. doi:10.1093/jamia/ocad259

48 Li J, Zhang Q, Yu Y, *et al.* More Agents Is All You Need. *Transactions on Machine Learning Research* 2024.

49 Lu, Jinliang, Pang, *et al.* Merge, Ensemble, and Cooperate! A Survey on Collaborative Strategies in the Era of Large Language Models. *arXiv* 2024.

50 Chen, Zhijun, Li, *et al.* Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. *arXiv* 2025.

51 Yang H, Li M, Zhou H, *et al.* Large language model synergy for ensemble learning in medical question answering: design and evaluation study. *J Med Internet Res* 2025;**27**:e70080. doi:10.2196/70080

52 Brignone E, LeJeune K, Mihalko AE, *et al.* Self-Reported Social Determinants of Health and Area-Level Social Vulnerability. *JAMA Netw Open* 2024;**7**:e2412109. doi:10.1001/jamanetworkopen.2024.12109

53 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**:1764–73. doi:10.1093/jamia/ocaa143

54 Mitra A, Chen K, Liu W, *et al.* Post-discharge suicide prediction among US veterans using natural language processing-enriched social and behavioral determinants of health. *npj Mental Health Res* 2025;**4**:8. doi:10.1038/s44184-025-00120-2

55 Yang MY, Kwak GH, Pollard T, *et al.* Evaluating the impact of social determinants on health prediction in the intensive care unit. In: Rossi F, Das S, Davis J, *et al.*, eds. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: : ACM 2023. 333–50. doi:10.1145/3600211.3604719

56 Hartzler AL, Xie SJ, Wedgeworth P, *et al.* 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**:1456–62. doi:10.1093/jamia/ocad043

57 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**:1367–78. doi:10.1093/jamia/ocad012

58 Na PJ, Shin J, Kwak HR, *et al.* Social Determinants of Health and Suicide-Related Outcomes: A Review of Meta-Analyses. *JAMA Psychiatry* 2025;**82**:337–46. doi:10.1001/jamapsychiatry.2024.4241

59 Papini S, Hsin H, Kipnis P, *et al.* Validation of a multivariable model to predict suicide attempt in a mental health intake sample. *JAMA Psychiatry* 2024;**81**:700–7. doi:10.1001/jamapsychiatry.2024.0189