Original Paper

Extracting Social Determinants of Health From Electronic Health Records: Development and Comparison of Rule-Based and Large Language Model Methods

Bo Wang[1,2,3,4], PhD; Dia Kabir[1,2,3], MS; Cheryl Renee Clark[5], MD, SCD; Karmel W Choi[1,2,3,4], PhD; Jordan W Smoller[1,2,3,4], MD, SCD

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

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

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

[4]Broad Institute, Cambridge, MA, United States

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

Corresponding Author:
Jordan W Smoller, MD, SCD
Center for Precision Psychiatry
Massachusetts General Hospital
Richard B. Simches Research Building
185 Cambridge Street, 2nd Floor
Boston, MA, 02114
United States
Phone: 1 6177249447
Email: jsmoller@mgb.org

Abstract

Background: Social determinants of health (SDoH) are critical drivers of health outcomes but are often underdocumented in structured electronic health record (EHR) data. Instead, SDoH are more commonly recorded in unstructured clinical notes, and unlocking this information could have far-reaching implications for advancing population health research and informing clinical decision-making.

Objective: This study develops and systematically evaluates cost-efficient methods for extracting SDoH information from unstructured clinical text using rule-based natural language processing (NLP) and large language model (LLM)–based approaches.

Methods: We constructed a gold-standard annotated corpus comprising clinical text segments from 171 patients in the Mass General Brigham Research Patient Data Registry, covering 7 SDoH domain categories and 23 subcategories. A rule-based system (RBS) was developed and evaluated alongside 7 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 using multiple prompting strategies. We additionally implemented late-fusion ensemble approaches that combined outputs from rule- and LLM-based methods. Performance was assessed using precision, recall, and F_1-score, alongside qualitative error analysis.

Results: The RBS achieved high precision for SDoH domain categories (0.96) but substantially lower recall (0.68). GPT-based models consistently outperformed the RBS in overall recall and F_1-scores. The best domain-level performance was observed for GPT-5 and GPT-5-mini in few-shot settings (F_1-score=0.89), while o4-mini achieved the highest subcategory-level performance (F_1-score=0.88). A late-fusion ensemble integrating RBS and GPT outputs further improved domain-level performance (F_1-score=0.92), with balanced precision (0.93) and recall (0.90), but did not improve subcategory-level performance.

Conclusions: Recent GPT models with advanced reasoning capabilities, including the newly released mini models (eg, o4-mini and GPT-5-mini), demonstrated strong performance for SDoH extraction without task-specific fine-tuning and consistently outperformed the rule-based NLP system. Integrating rule- and LLM-based methods via late fusion further enhanced domain-level extraction performance. Our results demonstrate a cost-efficient framework for the accurate identification of SDoH from clinical text, facilitating downstream population health research and clinical informatics applications.

(JMIR Med Inform 2026;14:e89534) doi: 10.2196/89534

KEYWORDS

social determinants of health; electronic health records; natural language processing; large language model; rule-based system; information extraction; social and behavioral determinants of health

Introduction

Social determinants of health (SDoH) are increasingly recognized as critical factors influencing health outcomes and contributing 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 source of real-world data, with applications in risk stratification, pharmacoepidemiology, treatment response prediction, and more. However, the value of EHR-based research has been limited by the underdocumentation 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].

Consequently, 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 because of their dependence on fixed, manually engineered rules. Conversely, supervised learning methods require significant amounts of high-quality annotated training data, which can be cost- and labor-intensive to generate. This dependence on annotated data is reflected in prior studies such as the 2022 n2c2/UW shared task [11,12], in which the top-performing systems used transformer-based models fine-tuned on annotated corpora. Recent advances in large language models (LLMs) present an opportunity to develop scalable solutions for identifying SDoH without the need for substantial annotated data [13-15].

Despite progress, several important gaps remain. First, the scope of SDoH domains addressed in most studies is limited. A 2021 review by Patra et al [2] found that smoking status is among the most commonly studied SDoH-related domains, followed by substance abuse and homelessness. By contrast, SDoH factors such as financial problems, social support, food security, and health insurance coverage remain relatively underexplored [15,16]. Second, rule-based approaches accounted for approximately one-quarter of the reviewed studies (22 out of 82) [2]. Among studies exploring the use of LLMs, most have used either open-weight models such as LLaMA-2 (Large Language Model Meta AI 2) and FLAN-T5 (Fine-Tuned Language Net—Text-to-Text Transfer Transformer) [11,13,15,17] or earlier versions of proprietary LLMs like GPT-3.5 [12,14], while more advanced LLMs with superior reasoning capabilities, including in health care contexts [18,19], remain largely unexplored for SDoH extraction. Notably, a recent study by Keloth et al [15] emphasized the importance of evaluating the latest LLMs for this task. Lastly, while fine-tuning open-weight models has gained traction [12,13,15,17], the development and systematic evaluation of different prompting strategies with state-of-the-art LLMs, including fusion with rule-based systems (RBSs), remain understudied despite their relevance in resource-constrained settings.

Our study addresses the aforementioned gaps by developing and evaluating methods to identify 7 SDoH domains from clinical text derived from multiple note types. In addition to commonly studied domains, we emphasize less-explored determinants such as social resources and health insurance status, as well as physical activity, a key behavioral determinant of health. A recent study by Lituiev et al [20] used a 2-tier taxonomy with first- and second-level SDoH classes. Similarly, we developed a fine-grained classification system for each domain. For example, within health insurance status, we included subcategories such as adequate insurance coverage, lack of insurance, and government-assisted insurance to capture different levels of access to care and coverage. Inspired by the n2c2/UW shared task [21], we further annotated each subcategory with 4 contextual attributes (temporality, experiencer, hypothetical status, and uncertainty). These attributes served as exclusion criteria for defining positive SDoH cases and supported qualitative analysis of the models' handling of contextual distinctions. This annotation framework was designed to capture richer social context information for downstream applications such as health risk stratification.

We present 2 complementary NLP approaches for SDoH extraction requiring minimal training and computational resources: an RBS and LLM-based method leveraging state-of-the-art OpenAI GPT models available at the time of experimentation. For the LLM-based approach, rule-based prescreening was used to identify candidate text segments from clinical notes, which were then classified by GPT models. We examined 7 GPT models in both zero-shot and few-shot settings without task-specific fine-tuning, evaluating performance at the segment level. Our prompting strategies incorporated examples that varied in annotation difficulty. This work extends prior studies by systematically evaluating advanced reasoning models and comparing their performance with that of the RBS. Finally, we investigated various ensemble strategies using late fusion to combine both approaches and further improve extraction performance. Overall, this study contributes to addressing key gaps in SDoH research and provides a cost-efficient framework for future clinical NLP applications.

Methods

Ethics Considerations

This study was reviewed and approved by the Mass General Brigham (MGB) Institutional Review Board (protocol 2018P002642), with an informed consent waiver for the use of retrospective medical record data without patient interaction. The study procedures were conducted in accordance with the ethical standards of the relevant national and institutional

committees on human experimentation and with the Declaration of Helsinki. Although the clinical notes used in this study were not deidentified, all data access, modeling, and analyses were restricted to authorized researchers and conducted within secure environments behind the MGB firewall. No identifiable patient information is included in the manuscript or its multimedia appendices. Participants did not receive compensation.

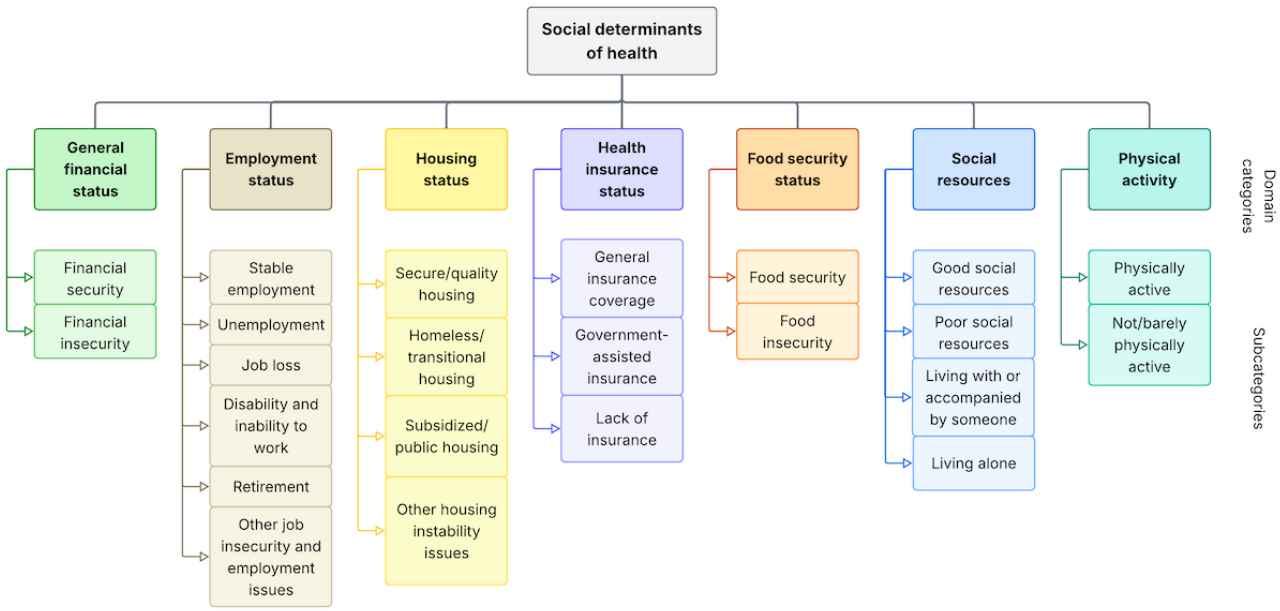
Data Source

Data were obtained from the MGB Research Patient Data Registry (RPDR) [22], a centralized data registry of clinical information from EHRs across the MGB health system. The RPDR database includes approximately 7 million patients from 8 hospitals, including 2 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 laboratory tests) and unstructured narrative notes. We acquired all narrative clinical notes from 95,157 patients enrolled in the MGB Biobank [23], with notes 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.

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 7 domains, that is, 6 social and 1 behavioral (physical activity) determinants of health commonly screened in clinical practice [24-26]. To capture granular social contexts relevant to health outcomes, we further defined multiple subcategories within each domain category. Given the inherent overlap between certain subcategories (eg, "patient recently got laid off" could be characterized as both job loss and unemployment), we developed comprehensive annotation guidelines (see Appendix S2 in Multimedia Appendix 1) to ensure consistency and reproducibility in data generation. The schema was also refined to draw more nuanced distinctions. For example, within social resources, we distinguished the subcategories of living with or accompanied by someone and living alone from good social resources and poor social resources to separate the living situation from the quality of social support. The initial taxonomy of 27 subcategories was refined to 23 based on data review and feasibility assessment. Instances falling within a domain category but not matching any specific subcategory were assigned to "N/A" (not applicable).

Figure 1. Social determinants of health classification schema comprising 7 domains: general financial status (green), employment status (beige), housing status (yellow), health insurance status (purple), food security status (orange), social resources (blue), and physical activity (teal). Domain names are shown in bold, with their corresponding subcategories listed beneath each domain (23 subcategories total).



Gold-Standard Corpus Development

An initial set of 2000 patients was sampled from the MGB Biobank cohort using stratified sampling based on several key sociodemographic variables: sex, self-reported race, age group (≥65 years vs <65 years), and health insurance type (public vs private payer). This cohort was held out from system development, including lexicon refinement of the RBS, to ensure separation from evaluation. To create a gold-standard dataset, we conducted prescreening using predefined patterns from our RBS (see the next section), including keywords, rules, and regular expression-based matching, to identify sentences in patients' notes likely containing SDoH information. Following prior work [9,27], for each potential SDoH mention, we extracted text segments with 150 characters of context on either side, providing sufficient information for annotation. We randomly sampled 79 unique text segments for annotator training and a separate set of 226 text segments (10 per subcategory, except "Other job insecurity and employment issues," which had 6 instances) for primary model validation. These 226 segments were drawn from notes with a median documentation year of 2016 (IQR 2012-2019).

We developed our annotation workflow using Label Studio [28], an open-source data labeling platform (Figure S1 in Multimedia Appendix 1). Two annotators (BW, a postdoctoral researcher trained in biomedical informatics, and DK, a medical student with experience in neuropsychiatric research) underwent 3 training sessions and calibration meetings before final annotation. Disagreements between annotators were discussed and resolved in consensus meetings, with a clinical psychologist (KC) serving as the tiebreaker. Based on the finalized SDoH classification schema (Figure 1) and annotation guidelines, we defined 2 multilabel annotation tasks for each text segment:

- Task 1: Identify mentions of the 7 SDoH domain categories.
- Task 2: Identify mentions of the 23 SDoH subcategories.

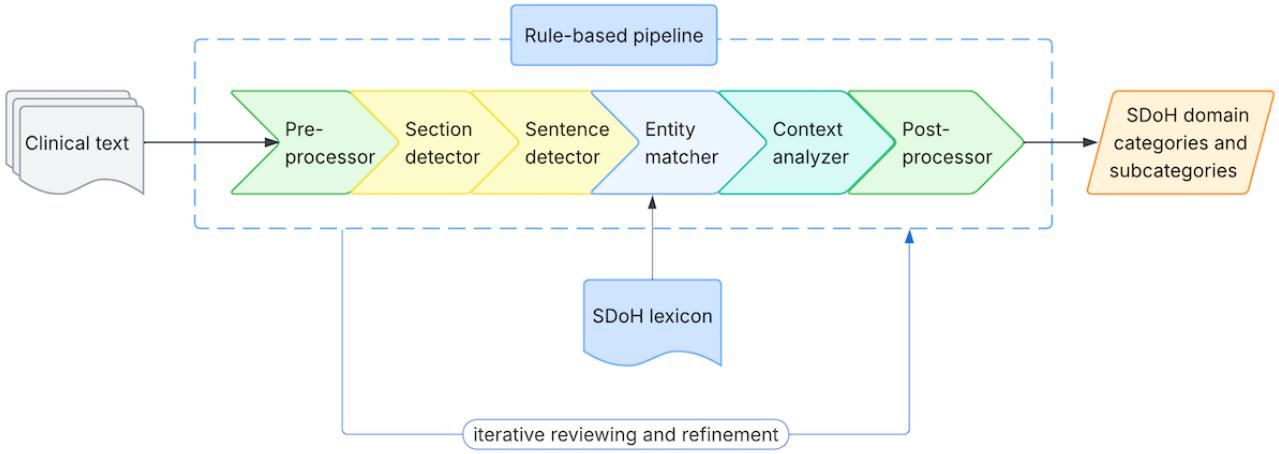
Each training session began with a detailed discussion of the annotation guidelines and concluded with an interannotator agreement assessment. After completing 3 training sessions, annotators achieved acceptable interannotator agreement as measured by Krippendorff α [29] (α=.88 for domain categories and α=.74 for subcategories). They then independently annotated the final 226 text segments, achieving α=.95 for domain categories, .84 for subcategories alone, and .78 for subcategories with contextual attributes (temporality, experiencer, hypothetical status, and uncertainty), all well above standard thresholds for reliability [30].

Rule-Based System

Overview

As illustrated in Figure 2, we developed an RBS to identify SDoH mentions in clinical notes. The system was implemented using medspaCy [31], an open-source clinical NLP library that supports flexible integration of rule-based and machine learning–based algorithms. It uses a pipeline of NLP functions spanning sentence and section detection to concept extraction and contextual analysis to locate, match, and disambiguate SDoH mentions, along with an iterative process of manual testing, review, and refinement based on feedback to optimize model performance. Descriptions of each pipeline component are provided in Appendix S1.3 in Multimedia Appendix 1.

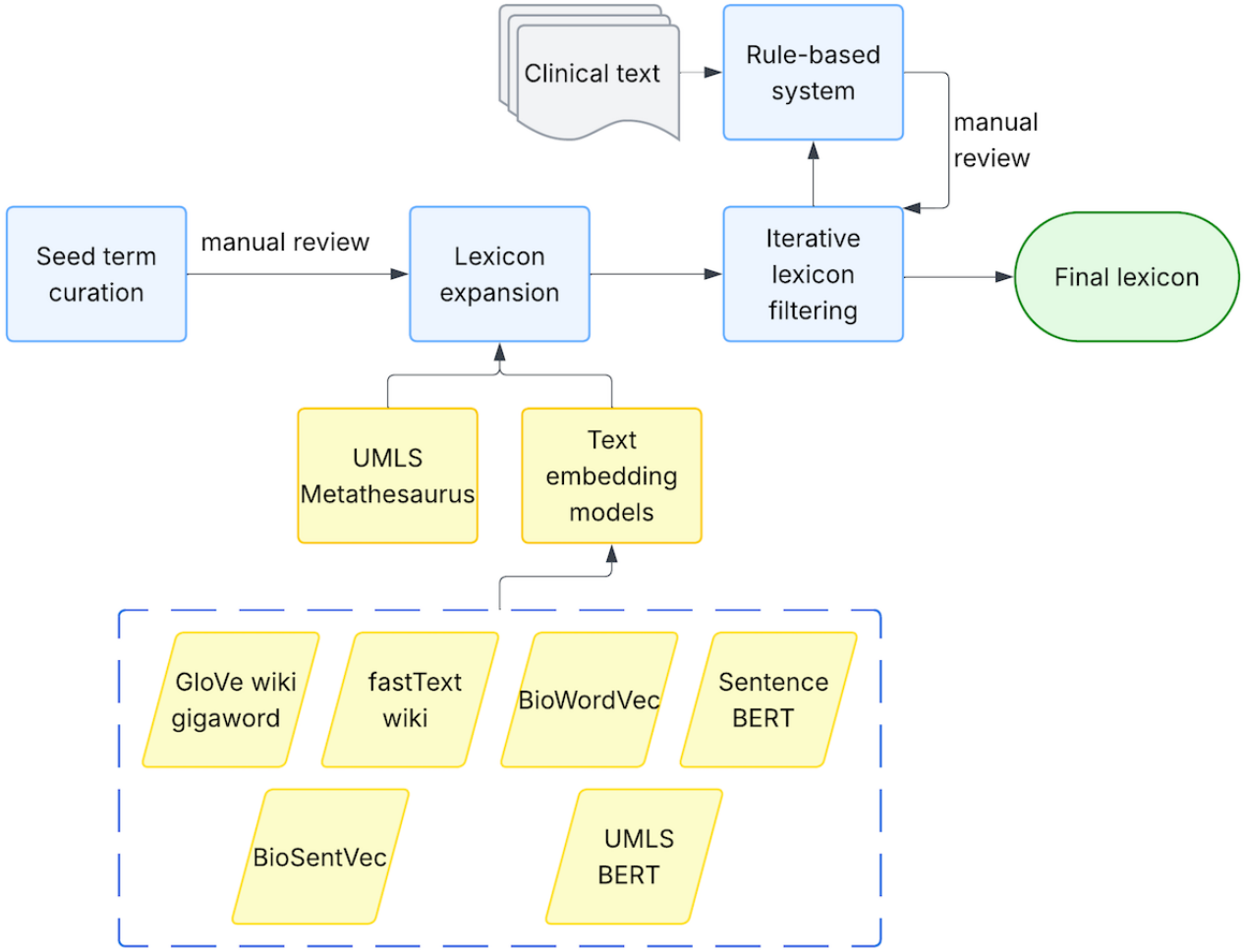
Figure 2. Rule-based system workflow: clinical notes (as the input) are processed through a text pipeline that includes sentence segmentation, lexicon-based entity matching, and context disambiguation, producing sentences labeled with social determinants of health (SDoH) domain categories and subcategories. The rules were developed through an iterative process involving manual review and refinement.



Lexicon Creation and Expansion

RBS requires a lexicon (a dictionary of relevant query terms) and corresponding rules for identifying matches in clinical text. We developed our SDoH lexicon through a multistep process: seed term curation, lexicon expansion, and iterative filtering and refinement (Figure 3). The initial lexicon comprised terms related to each domain derived from prior studies and reviews [6,8,25,26,32-36] and SDoH screening tools and surveys (see Appendix S1.1 in Multimedia Appendix 1). These seed terms were reviewed by a domain expert (CRC) and systematically expanded using (1) synonym identification and hierarchy mapping (hypernymy-hyponymy) through UMLS (Unified Medical Language System) Metathesaurus, and (2) semantic similarity search using text embedding models [37-42]. For instance, we encoded the MEDLINE N-Gram Set (1-5 grams) [43,44] using an MPNet-based Sentence-BERT (Bidirectional Encoder Representations from Transformers) model (paraphrase-mpnet-base-v2) [41] (which we make publicly available [45]) and performed semantic searches using seed terms as queries against the embedded corpus to identify additional relevant SDoH terms. The iterative lexicon filtering step is detailed in Appendix S1.2 in Multimedia Appendix 1.

Figure 3. Lexicon curation process involving seed term selection, expansion using UMLS (Unified Medical Language System) and embedding-based semantic search, and iterative lexicon filtering. BERT: Bidirectional Encoder Representations from Transformers; BioSentVec: Biomedical Sentence Vector; BioVec: Biomedical Word Vector.



LLM-Based Models

Model selection and configuration

We used a series of OpenAI GPT models via the Azure OpenAI Service (Microsoft Foundry), deployed within MGB's Health Insurance Portability and Accountability Act (HIPAA)–compliant Microsoft Azure infrastructure. Using the Chat Completions application programming interface (API), we prompted the models to identify SDoH categories and subcategories in clinical text segments, akin to a hierarchical multilabel classification task.

While open-weight LLMs continue to improve, frontier proprietary models remain state-of-the-art on many benchmarks, particularly those involving reasoning tasks [19,46,47]. For this study, we selected 7 GPT models for SDoH extraction based on their performance on benchmarks, including HealthBench [18], as well as their inference cost: GPT-4o, 4.1, 4.1-mini, 5, 5-mini, o3, and o4-mini (model versions are provided in Appendix S4 in Multimedia Appendix 1). For reasoning models (ie, GPT-5, 5-mini, o3, and o4-mini), the reasoning effort parameter was set to high. We compared their precision and recall with those of the RBS, evaluated performance (precision, recall, and F_1-scores), 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 [48,49], enabling them to follow novel instructions and perform unseen tasks without requiring additional training, a capability known as zero-shot generalization [50,51]. Prior studies have highlighted the importance of systematic prompt engineering [11,15,52], for example, by incorporating annotation guideline instructions. As shown in Figure S2 and Appendix S3 in Multimedia Appendix 1, our zero-shot approach used a 3-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 (eg, doctor's recommendations).

In addition, we evaluated 3 prompting styles, namely, strict, balanced, and liberal, which differ in their threshold for identifying SDoH domains and subcategories (inclusion criteria). The strict style requires explicit evidence, balanced accepts evidence that is strongly implied, and liberal takes the most inclusive approach, accepting not only explicit and strongly implied evidence but also information deemed reasonably likely from the context. We hypothesized that these styles would demonstrate a precision-recall trade-off, with strict achieving higher precision and liberal yielding higher recall.

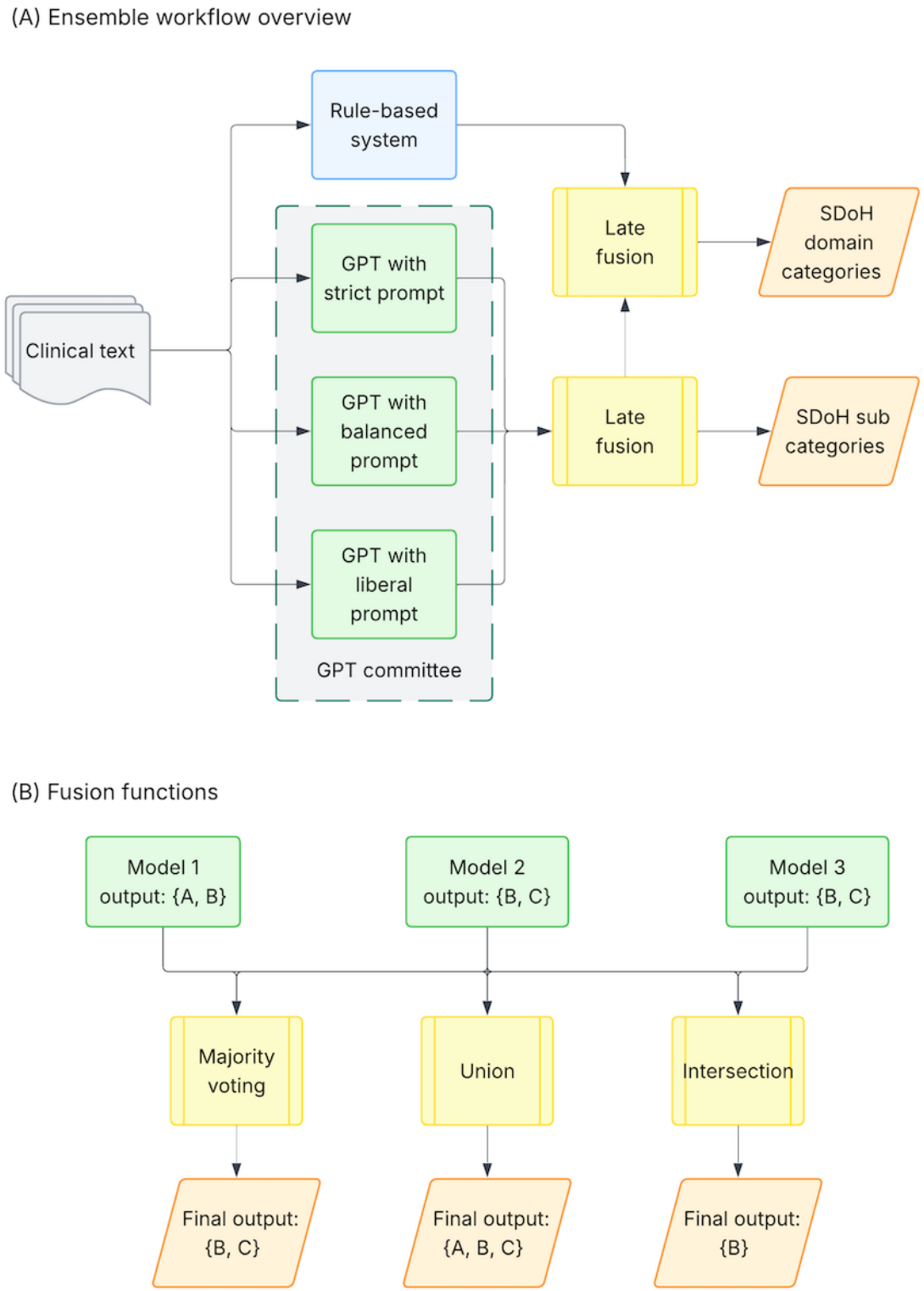
Few-Shot Prompting

Following Consoli et al [14], we tested 6 few-shot prompting strategies using examples from our annotation training: (1) using 5 examples where annotators readily agreed on the correct identification of SDoH information (ie, easy examples); (2) same as (1), but with explanations describing the reasoning behind each annotation; (3) using 5 challenging examples that required annotator adjudication (ie, hard examples); (4) same as (3), but with explanations added; (5) using synthetically generated examples by GPT; and (6) same as (5), but with synthetic explanations. The goal of these experiments was to assess whether advanced GPT models benefit from examples with or without explanations, and whether exposure to difficult, ambiguous, or synthetic cases improves their accuracy and robustness in SDoH extraction. The few-shot prompt structure is illustrated in Figure S3 in Multimedia Appendix 1.

RBS-GPT Ensemble

Drawing on prior work in LLM ensemble learning within the broader NLP domain [53-56], we implemented a hybrid ensemble approach that aggregates rule-based and GPT model outputs using late fusion to improve the overall performance of SDoH extraction. Using the annotator training samples as development data, we found that the RBS achieved good precision (0.92) for domain categories but showed poor performance (precision of 0.64) at the subcategory level. Based on these findings, we designed 2 different ensemble strategies (Figure 4): (1) for domain categories, we deployed the best-performing GPT model with 3 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 the RBS; (2) for subcategories, we used only the aggregated GPT committee output to determine the final list of SDoH subcategories. We evaluated several ensemble configurations, testing 3 fusion functions for the GPT committee: (1) majority voting, (2) union, and (3) intersection, and compared them with GPT using strict prompting alone. We also examined 2 fusion functions (union and intersection) for combining RBS and GPT committee outputs (RBS-GPT fusion).

Figure 4. (A) Overview of the RBS-GPT ensemble workflow, illustrating separate fusion pathways for SDoH domain categories and subcategories. For domain categories, outputs from the RBS and multiple GPT prompts (the GPT committee) are combined using late fusion. For subcategories, only outputs from the GPT committee are aggregated. (B) Illustrative example of the fusion strategies, including majority voting, union, and intersection. RBS: rule-based system; SDoH: social determinants of health.



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 annotations, including 7 labeled "none of the above," and 475 subcategory annotations, including 153 labeled "N/A." Table 1 summarizes patient demographic characteristics, showing balanced distributions of gender and self-reported race, with a predominance of older patients and those with public health insurance.

Table 1. Patient demographics from the annotated validation dataset (N=171).

Characteristics and categories	Value, n (%)
Age group (years)	
<25	16 (9.4)
25-44	39 (22.8)
45-64	31 (18.1)
≥65	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 (25.1)
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[a]	
<US $40,000	17 (9.9)
US $40,000–69,000	88 (51.5)
US $70,000–99,000	46 (26.9)
≥US $100,000	17 (9.9)
Unknown	3 (1.8)
Public payer	
Yes	122 (71.3)
No	49 (28.7)
Veteran status	
Yes	4 (2.3)
No	152 (88.9)
Unknown	15 (8.8)

[a]Median household income by ZIP code.

As noted earlier, the 2 annotators achieved strong interannotator agreement (Krippendorff α=.95 for domain categories and .78 for subcategories with attributes). The majority of annotation disagreements for domain categories were attributed to “social resources” (which required a clearer scope definition) and “general financial status” (where 1 annotator [DK] 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 further guidance to correctly parse question-answer pairs within poorly formatted questionnaire text. Only text segments containing current, patient-specific, and nonhypothetical SDoH information were considered SDoH-positive cases.

Model Performance

Comparative Performance of Rule-Based and GPT Models for SDoH Classification

Figure 5 summarizes the macro-averaged performance of the RBS and 5 GPT-based models (additional results in Table S1 in Multimedia Appendix 1); 95% CIs for macro- and micro-averaged F_1-scores are shown in Figures S4 and S5 in Multimedia Appendix 1. While the RBS achieved high precision for domain categories, it demonstrated substantially lower recall than GPT-based models across both classification levels, highlighting the inherent limitation of fixed rules. By contrast, GPT-based models consistently outperformed the RBS in recall and F_1-scores. At the domain-category level, GPT-5-mini (5-shot) and GPT-5 (5-shot) were tied for the highest point-estimate F_1-score (0.89), with precision/recall/F_1-score of 0.91/0.87/0.89 (95% CI for F_1-score 0.86-0.91) and 0.96/0.84/0.89 (95% CI for F_1-score 0.86-0.92), respectively. At the subcategory level, the o4-mini models demonstrated the strongest performance, achieving an F_1-score of 0.87 (95% CI 0.83-0.90) in the zero-shot setting (precision, recall, and F_1-score of 0.85, 0.91, and 0.87, respectively) and 0.88 (95% CI 0.84-0.91) in the 5-shot setting (0.90, 0.87, and 0.88, respectively). As shown in Figure S5 in Multimedia Appendix 1, the 95% CIs overlapped considerably between o4-mini and GPT-5, suggesting that small differences in point estimates should be interpreted cautiously. Overall, few-shot prompting yielded modest gains in precision for subcategories, often accompanied by slight reductions in recall. Table 2 reports micro-averaged metrics and absolute error counts (false positives and false negatives) for subcategories (additional results in Table S2 in Multimedia Appendix 1). Consistent with the macro-averaged results, o4-mini remained the best overall subcategory model, with o4-mini (5-shot) achieving the highest micro-averaged F_1-score (0.88; 95% CI 0.85-0.91; true positive, false positive, and false negative of 283, 37, and 39, respectively). GPT-5-mini (zero-shot) achieved the highest recall (0.93) and the largest number of true positives (298), but at the cost of more false positives (100).

Figure 5. Overall performance comparison between rule-based system (RBS) and GPT-based approaches. Macro-averaged precision, recall, and F_1-score are reported separately for domain categories and subcategories. The corresponding 95% CIs for macro-F_1-score are shown in Figures S4 and S5 in Multimedia Appendix 1. All reasoning models reported here (o4-mini, GPT-5-mini, and GPT-5) were run using the high-reasoning setting. The highest score in each metric column is highlighted with a blue box border. SDoH: social determinants of health.

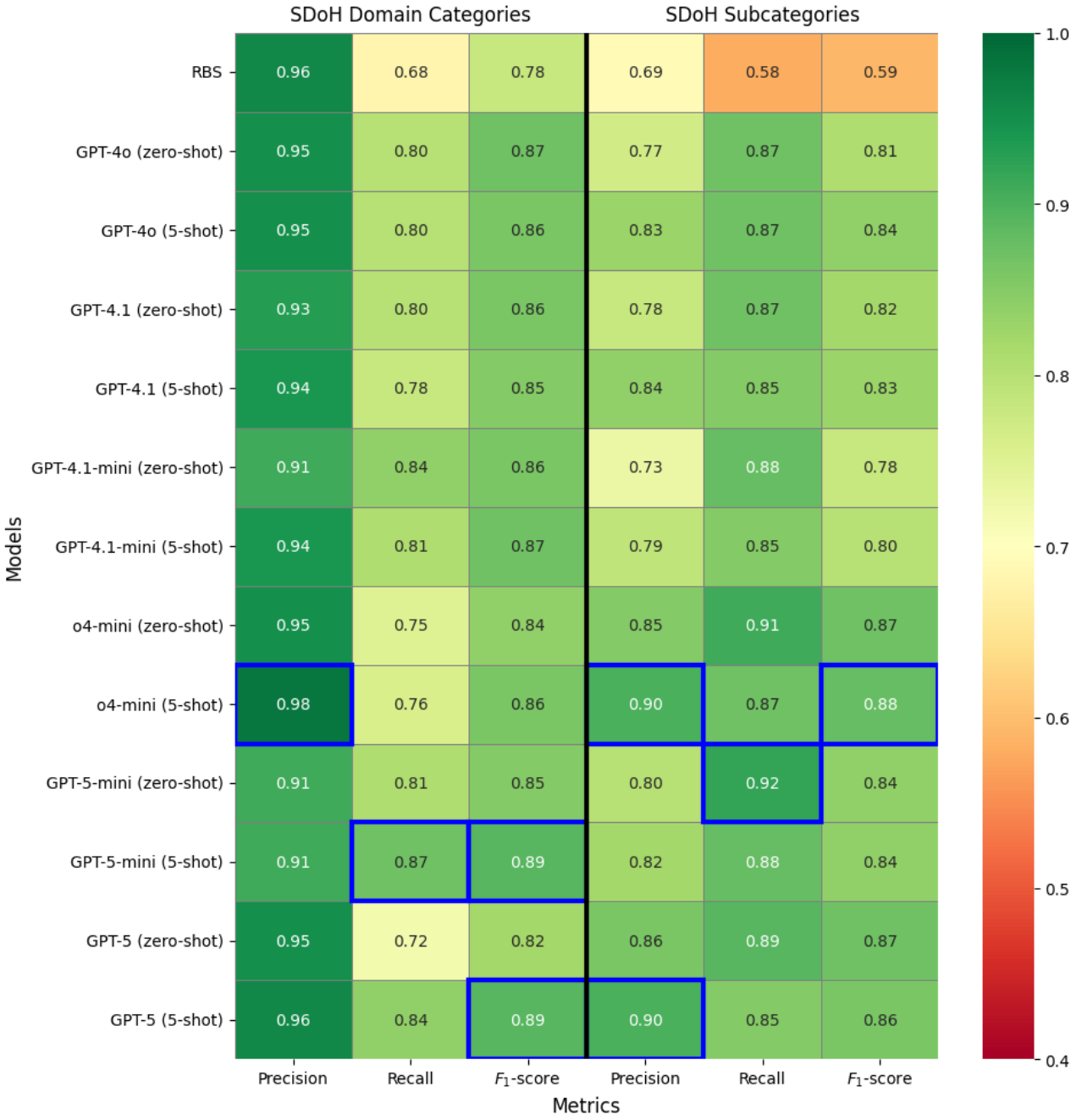


Table 2. Micro-averaged performance metrics and absolute error counts for subcategories[a,b].

Models	Precision	Recall	F_1-score	True positive, n	False positive, n	False negative, n
Rule-based system	0.67	0.49	0.57	159	79	163
GPT-4o (zero-shot)	0.77	0.85	0.81	274	83	48
GPT-4o (5-shot)	0.82	0.84	0.83	272	58	50
GPT-4.1 (zero-shot)	0.79	0.85	0.82	275	75	47
GPT-4.1 (5-shot)	0.84	0.83	0.83	266	50	56
o4-mini (zero-shot)	0.84	0.90	0.87	289	55	33
o4-mini (5-shot)	0.88	0.88	0.88	283	37	39
GPT-5-mini (zero-shot)	0.75	0.93	0.83	298	100	24
GPT-5-mini (5-shot)	0.79	0.90	0.84	289	78	33
GPT-5 (zero-shot)	0.84	0.88	0.86	282	54	40
GPT-5 (5-shot)	0.88	0.84	0.86	272	37	50

[a]Absolute counts denote true positives, false positives, and false negatives relative to gold-standard annotations.

[b]True positive + false negative sums to 322 rather than the 475 total annotations because 153 annotations labeled "NA" do not contribute to true positive or false negative for any positive subcategory label.

Appendix S5 in Multimedia Appendix 1 benchmarks the inference time (client wall-clock) and cost across o4-mini, GPT-5-mini, and GPT-5 under a high-reasoning setting, using a single-worker sequential deployment. While both mini models performed competitively with GPT-5, they demonstrated substantially lower inference latency and estimated per-segment cost. For example, o4-mini required approximately 7.0 seconds per segment in zero-shot and 8.9 seconds in 5-shot prompting, compared with 33.0 and 34.6 seconds, respectively, for GPT-5. API-reported reasoning tokens accounted for most completion tokens across models (94%-98%; see Appendix S5 in Multimedia Appendix 1), reflecting the computational overhead of chain-of-thought processing. Based on cost efficiency and competitive performance at both levels, as demonstrated in the per-domain and per-subcategory results (Table S3 in Multimedia Appendix 1), we selected o4-mini (5-shot) as our primary model for further experiments involving ensemble modeling.

Prompting Comparison

Among the strict, balanced, and liberal prompting styles, we observed the anticipated precision-recall trade-off (Table S4 in Multimedia Appendix 1): strict prompting achieved the highest precision and F_1-scores for subcategories in both zero-shot and 5-shot settings, while liberal prompting had the best recall, and balanced prompting yielded intermediate results.

Performance varied modestly across example types (easy, hard, or synthetic), with F_1-scores ranging from 0.81 to 0.86 for domain categories and 0.87 to 0.88 for subcategories (Table S5 in Multimedia Appendix 1). Hard examples requiring annotator adjudication yielded better performance for subcategories, particularly in precision (0.90 vs 0.86 for easy examples and 0.88 for synthetic examples). Adding explanations to the prompt improved precision for subcategories in the hard-example model (from 0.88 to 0.90) but had little impact on other models or metrics.

Ensemble Performance

As described in the "Methods" section, we used different ensemble strategies by classification level: 2-step fusion for domain categories and GPT-committee ensembling for subcategories. As shown in Table 3, majority voting on GPT-committee outputs followed by union ($\cup$) with RBS outputs achieved the highest macro-averaged F_1-score (0.92; 95% CI 0.89-0.94) and the most balanced performance for domain categories, with precision above 0.93 and recall above 0.90. However, the overlapping CIs suggest that performance differences among top fusion strategies should be interpreted cautiously. Of the 402 domain-category predictions produced (of which 21 were "none of the above"), 32 (8.0%) were contributed by the RBS alone, demonstrating its added value; 140 (34.8%) came from the GPT committee, indicating its ability to capture patterns missed by the RBS; and the remaining 230 (57.2%) were identified by both systems. For subcategories, the strict-only configuration achieved the best performance (precision of 0.90, recall of 0.87, F_1-score=0.88; 95% CI 0.84-0.91), consistent with our objective of maximizing F_1-score with balanced precision and recall. The intersection configuration followed closely, achieving an F_1-score of 0.87, with higher precision (0.92) but lower recall (0.84). Micro-averaged results and absolute error counts are reported in Table S15 in Multimedia Appendix 1 and showed a similar overall pattern.

Table 3. Performance of ensemble models across fusion strategies[a,b].

SDoH[c] label level and GPT-committee	RBS[d]-GPT fusion	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
Domain categories				
Majority voting	∪	0.93 (0.90-0.96)	0.90 (0.87-0.93)	*0.92* (0.89-0.94)
Majority voting	∩	*0.99* (0.98-1.00)	0.60 (0.55-0.66)	0.74 (0.70-0.78)
Union	∪	0.89 (0.85-0.92)	*0.93* (0.90-0.96)	0.91 (0.88-0.93)
Union	∩	*0.99* (0.97-1.00)	0.63 (0.58-0.68)	0.76 (0.72-0.80)
Intersection	∪	0.95 (0.92-0.98)	0.85 (0.81-0.89)	0.90 (0.87-0.92)
Intersection	∩	*0.99* (0.98-1.00)	0.55 (0.49-0.60)	0.70 (0.65-0.74)
Strict-only	∪	0.95 (0.92-0.98)	0.88 (0.84-0.91)	0.91 (0.88-0.93)
Strict-only	∩	*0.99* (0.98-1.00)	0.56 (0.51-0.62)	0.71 (0.67-0.76)
Subcategories				
Majority voting	N/A[e]	0.85 (0.82-0.89)	0.91 (0.87-0.94)	0.87 (0.84-0.90)
Union	N/A	0.73 (0.69-0.78)	*0.95* (0.92-0.97)	0.82 (0.78-0.84)
Intersection	N/A	*0.92* (0.89-0.95)	0.84 (0.79-0.88)	0.87 (0.83-0.90)
Strict-only	N/A	0.90 (0.87-0.94)	0.87 (0.83-0.91)	*0.88* (0.84-0.91)

[a]Macro-averaged precision, recall, and F_1-scores are reported for domain categories and subcategories. Majority voting, union, and intersection denote label assignment by most, any, or all GPT committee members, respectively. Strict-only uses the strict-prompt GPT output without aggregation. For domain categories, ∪ denotes assignment by either RBS or the GPT committee, and ∩ denotes assignment by both. 95% CIs were computed using 2000 patient-level cluster bootstrap resamples.

[b]The highest scores for each metric (column) are italicized.

[c]SDoH: social determinants of health.

[d]RBS: rule-based system.

[e]For subcategories, RBS-GPT fusion was not applied; N/A (not applicable) indicates GPT-committee ensembling only.

Comparison With International Classification of Diseases (ICD) Codes

Both the RBS and GPT models identified substantially more SDoH information than documentation based on ICD (International Classification of Diseases) codes (V and Z codes). Of the 226 annotated samples, our RBS identified SDoH information in 225, and the 5-shot GPT models identified SDoH in 194 (o4-mini), 219 (GPT-5-mini), and 206 (GPT-5) cases. In comparison, relevant ICD codes were present in far fewer visits across all 3 time windows examined: using broad Z and V codes (see Table S12 in Multimedia Appendix 1), matches were found in 84 visits within 7 days of the corresponding clinical documentation, 128 visits within a 6-month lookback window, and 133 visits within a 12-month lookback window. Using SDoH category–specific Z and V codes (see Table S13 in Multimedia Appendix 1), matches were found in only 11, 12, and 14 visits for the same time windows, respectively. These findings highlight the value of extracting SDoH from clinical text, as structured ICD coding substantially undercaptures SDoH information even when extended lookback periods are used.

Error Analysis

To better understand model limitations beyond quantitative performance metrics, we conducted a qualitative review of false-positive and false-negative predictions from the o4-mini (5-shot) model and the RBS at both the domain and subcategory levels. Owing to its lack of semantic understanding, the RBS failed to capture many SDoH mentions that fell outside its predefined keyword rules. For example, it was unable to recognize that “patient owns multiple homes” indicates secure housing. Its finite vocabulary coverage also meant that domain-specific terms such as “Section 8” (a government-funded rent subsidy program) went unrecognized. The majority of the RBS’s false positives stemmed from its failure to correctly recognize negations, temporality, hypotheticals, and experiencer attribution, which were annotated as contextual attributes in our validation data, as well as its inability to perform semantic disambiguation. Table S14 in Multimedia Appendix 1 summarizes the primary error types for both approaches.

In contrast to the RBS, o4-mini demonstrated its ability to interpret semantic meaning and generalize beyond keyword matching. However, it exhibited a different set of errors related to reasoning calibration, evidence evaluation, and instruction adherence. These errors can be classified into 4 primary categories (Textbox 1).

Textbox 1. Error classification.

1. Missing implicit evidence

Under the strict prompting style adopted for subsequent analyses, o4-mini at times failed to infer social determinants of health (SDoH) subcategories that were implied but not explicitly stated.

2. Overinterpretation of insufficient evidence

Conversely, the model also exhibited over-reasoning, assigning labels when the evidence was directionally relevant but fell short of what the annotation guidelines require.

3. Prompt instruction nonadherence

In several cases, the model failed to follow explicit SDoH definitions provided in the prompt.

4. Temporality misclassification

The model occasionally correctly identified an SDoH mention but failed to determine whether it reflected the patient's current or historical status.

The first 2 error types reflect a gap between the model and human annotators regarding what constitutes sufficient evidence for SDoH classification, suggesting that the model lacks a calibrated threshold aligned with the annotation framework. Future efforts should focus on closing this gap through refined SDoH definitions and boundary-clarifying examples.

Discussion

Principal Findings

This study examined the extraction of SDoH information across 7 domain categories, including less-studied domains such as social resources and health insurance status, and 23 corresponding subcategories from discharge summaries and progress notes in a large health system. We evaluated 7 GPT models under multiple prompting strategies and compared their performance with an expert-designed, iteratively optimized RBS. Recently released GPT models with improved reasoning capabilities, such as GPT-5 and o4-mini, achieved the best overall performance at both the domain category and subcategory levels. While the RBS demonstrated high precision for domain categories, it exhibited consistently low recall because of the inherent rigidity of manually engineered rules. For example, GPT models correctly identified "patient needs WIC" (where WIC refers to the 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. By contrast, the RBS failed to capture the WIC reference because it relied 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 [17], who reported superior RBS performance over an LLM-based approach for classifying social support and social isolation, primarily because their RBS closely mirrored the gold-standard annotation rule book, leading to overfitting.

Focusing on our GPT models, the domain-level macro-F_1-scores of 0.82-0.89 are broadly consistent with recent evaluations of LLM-based SDoH extraction [11,13,15]. For example, Keloth et al [15] reported macro-F_1-scores ranging from 0.53 to 0.84 across 4 institutions using instruction fine-tuned LLaMA models, while Guevara et al [13] achieved a macro-F_1-score of 0.71 for sentence-level SDoH classification using fine-tuned Flan-T5 models. At the subcategory level, our GPT models achieved macro-F_1-scores of 0.78-0.88 across 23 subcategories. By contrast, Keloth et al [15] reported lower level-2 macro-F_1-scores (0.45-0.59), although their task additionally required temporality determination for some categories. Earlier-generation LLMs have shown more limited in-context learning performance, with GPT-4 one-shot prompting achieving only 0.65 micro-averaged F_1-score on the n2c2/UW SHAC event extraction task [21,57], and few-shot–prompted LLaMA-2 underperforming fine-tuned models [15]. While these comparisons should be interpreted with caution because of differences in SDoH taxonomies, annotation schemas, note types, evaluation metrics, and the enriched nature of our validation set, the overall pattern suggests that recent reasoning-capable GPT models, when applied to prescreened clinical text in prompted settings, can achieve performance comparable to that reported for fine-tuned approaches and represent a meaningful advance over earlier LLM prompting-based models. Additional evaluation on 500 unfiltered text segments, drawn from notes with a median documentation year of 2014 (IQR 2010-2018), showed low false-positive rates and provided preliminary evidence that the model can identify SDoH concepts in unfiltered clinical text, although the small number of SDoH-positive cases (n=42) precludes definitive conclusions about broader generalizability (see Appendix S7 in Multimedia Appendix 1).

Comparing efficiency in model development, GPT-based approaches in zero-shot or few-shot settings required substantially less development time and cost than constructing the RBS, which involved iterative rule creation and refinement. Among the GPT models, newer "mini" models, such as GPT-5-mini and o4-mini, which retain reasoning capabilities, performed competitively with OpenAI's flagship GPT-5 model while demonstrating substantially lower inference costs and latency. In our benchmarking (Appendix S5 in Multimedia Appendix 1), o4-mini (5-shot) and GPT-5-mini (5-shot) incurred estimated costs of US $0.008 and US $0.005 per text segment, respectively, compared with US $0.029 for GPT-5 (5-shot), representing 72%-82% cost reductions. By contrast, earlier small models such as GPT-4o-mini and GPT-4.1-mini, although offering even lower per-token pricing, did not perform on par

with their respective full models. At the observed single-worker throughput for o4-mini (5-shot), processing 100,000 and 1,000,000 segments would require approximately 10 and 102 days, respectively. In practice, cloud providers offer asynchronous batch processing with higher rate-limit pools and a 50% cost discount, which could substantially reduce wall-clock time relative to sequential single-worker deployment. However, realized throughput remains subject to API quotas, particularly for reasoning models such as o4-mini, whose hidden reasoning tokens increase token consumption. Alternatively, LLM-generated annotations on a representative subset could be used to train a local classifier, reducing API dependence for large-scale deployment.

Our prompt engineering evaluation highlighted the sensitivity of model performance to prompt wording, showing a precision-recall trade-off that can be tuned through stricter or more liberal prompt styles. For few-shot prompting, when comparing the addition of different types of examples, we found that including cases in which annotators required clarification or adjudication, 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, these findings suggest that adopting a systematic approach to prompt design and incorporating challenging examples should be considered best practices for LLM-based extraction methods.

Lastly, we evaluated various model ensemble strategies using late fusion. Combining the RBS with a committee of GPT models employing different prompting styles improved domain-category extraction compared with using GPT models alone. For example, the RBS-GPT hybrid correctly recognized "involved in church 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, though without an overall performance gain. These findings suggest the complementary strengths of rule- and LLM-based systems, although the contribution of the RBS to the ensemble for domain category extraction may be inflated by the keyword-enriched validation set. The lack of improvement at the subcategory level also indicates the need for more advanced and adaptive ensemble strategies, such as mixture-of-experts approaches in which different models (experts) handle different inputs, a direction for future investigation.

Limitations and Future Directions

This study has several limitations. First, both our RBS and GPT-based models were validated in a single health care system, and performance may vary elsewhere depending on documentation practices. This concern may be particularly relevant for the RBS, whose lexicon was filtered and refined through manual review of MGB notes and may therefore be less generalizable to other health systems. Additionally, because the validation set was sampled from the same source corpus, the reported RBS performance may represent an optimistic upper bound in this setting. At the same time, MGB comprises more than 8 hospitals and multiple community health centers with heterogeneous catchment areas and clinical practices, supporting some degree of generalizability. Second, our primary goal was to develop a cost-efficient 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 requires computation for model training. Future work could examine how these approaches compare. 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-weight LLMs. Such an evaluation would provide a more complete performance assessment. Fourth, we annotated text segments rather than entire clinical notes, as manual review suggested that segments contained sufficient information for SDoH extraction. However, this approach may miss SDoH mentions that depend on long-distance context within full notes. Fifth, although exploratory analyses stratified by gender, self-reported race, and ethnicity did not detect statistically significant performance differences across demographic subgroups for o4-mini or the RBS (see Appendix S6 and Tables S6-S11 in Multimedia Appendix 1), small subgroup sizes substantially limited statistical power. These results should not be interpreted as evidence that the models are free of demographic bias, and a larger, more demographically balanced dataset is needed for robust subgroup evaluation. Finally, because we applied rule-based filtering during data sampling, the recall and F_1-scores reported from our primary validation set are based on enriched text and therefore do not fully reflect performance in the general population of clinical notes. Although an additional evaluation on 500 unfiltered text segments (Appendix S7 in Multimedia Appendix 1) suggested that model performance generalizes beyond the enriched setting, the limited number of SDoH-positive cases in unfiltered text prevents definitive conclusions about generalizability and warrants further validation on larger, independently sampled datasets.

The ability to accurately and efficiently extract SDoH factors from clinical notes has important implications for health care systems, as such information is often underdocumented in structured EHR data. Our approach provides a cost-efficient method for SDoH extraction, demonstrating consistently high precision and recall across both domain categories and subcategories. By prescreening 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 [6,58]. Incorporating SDoH information into clinical prediction models is expected to

improve prediction performance, as recent studies suggest [59,60], while also supporting audits of algorithmic biases beyond traditional demographic variables [61]. 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 [62].

Future work should assess the robustness and generalizability of our approach to other health care systems, especially those with very different clinical settings, patient demographics, and socioeconomic environments. Furthermore, social determinants are highly context-dependent. Existing models often categorize a patient's social circumstances into predefined labels without capturing context, limiting their ability to provide a comprehensive and contextually relevant understanding of the patient's social needs. Lybarger et al [21] used an event-based annotation scheme characterizing each SDoH mention with multiple arguments (eg, employment status, duration, history, and type for "employment") from the social history sections. We plan to explore methods for incorporating richer contextual information, moving beyond the contextual attributes used in our rule-based approach and related work [5,15,21]. We will investigate text summarization methods to generate structured representations that encapsulate a comprehensive profile of a patient's SDoH status for each clinical encounter. Lastly, it will also be important to validate the utility of the extracted SDoH information, along with community-level socioeconomic factors [58], in supporting downstream health research applications, including suicide prediction [9,60,63] and model auditing [61,64].

Conclusions

This study presented and compared 2 NLP approaches, an RBS and a GPT-based model, for extracting SDoH information from clinical text segments retrieved from clinical notes. Recent GPT models with advanced reasoning capabilities achieved superior performance in identifying 7 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 ensembles combining both approaches to optimize extraction performance. By making our code and prompts available to the scientific community, we provide a cost-efficient solution for accurate SDoH extraction, with the potential to advance important downstream health research applications.

Acknowledgments

We declare the use of the generative artificial intelligence tool ChatGPT (OpenAI) for language editing and formatting of the manuscript.

Data Availability

Protected Health Information restrictions apply to the availability of the clinical data used here, which were accessed under IRB approval for use only in this study. As a result, this dataset is not publicly available. The code for our rule-based system and LLM prompts is available online [45].

Funding

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

Authors' Contributions

Conceptualization: BW, KWC, JWS
Data curation: BW, DK
Formal analysis: BW
Funding acquisition: BW, JWS
Investigation: BW, KWC, JWS
Methodology: BW, KWC, JWS, DK, CRC
Project administration: BW
Resources: KWC, JWS
Supervision: KWC, JWS
Validation: BW, DK
Visualization: BW
Writing – original draft: BW
Writing – review & editing: BW, DK, CRC, KWC, JWS

Conflicts of Interest

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

Multimedia Appendix 1

Supplementary materials.
[DOCX File , 7266 KB-Multimedia Appendix 1]

References

1. Social determinants of health. World Health Organization (WHO). URL: https://www.who.int/health-topics/social-determinants-of-health [accessed 2024-02-21]
2. Patra BG, Sharma MM, Vekaria V, Adekkanattu P, Patterson OV, Glicksberg B, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc. Nov 25, 2021;28(12):2716-2727. [FREE Full text] [doi: 10.1093/jamia/ocab170] [Medline: 34613399]
3. Han S, Zhang RF, Shi L, Richie R, Liu H, Tseng A, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. J Biomed Inform. Mar 2022;127:103984. [FREE Full text] [doi: 10.1016/j.jbi.2021.103984] [Medline: 35007754]
4. Mehta S, Lyles CR, Rubinsky AD, Kemper KE, Auerbach J, Sarkar U, et al. Social determinants of health documentation in structured and unstructured clinical data of patients with diabetes: comparative analysis. JMIR Med Inform. Aug 22, 2023;11:e46159. [FREE Full text] [doi: 10.2196/46159] [Medline: 37621203]
5. Chapman AB, Jones A, Kelley AT, Jones B, Gawron L, Montgomery AE, et al. ReHouSED: a novel measurement of Veteran housing stability using natural language processing. J Biomed Inform. Oct 2021;122:103903. [FREE Full text] [doi: 10.1016/j.jbi.2021.103903] [Medline: 34474188]
6. Bejan CA, Angiolillo J, Conway D, Nash R, Shirey-Rice JK, Lipworth L, 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. Jan 01, 2018;25(1):61-71. [FREE Full text] [doi: 10.1093/jamia/ocx059] [Medline: 29016793]
7. Conway M, Keyhani S, Christensen L, South BR, Vali M, Walter LC, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. J Biomed Semantics. Apr 11, 2019;10(1):6. [FREE Full text] [doi: 10.1186/s13326-019-0198-0] [Medline: 30975223]
8. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate cancer. BMC Med Inform Decis Mak. Mar 14, 2019;19(1):43. [FREE Full text] [doi: 10.1186/s12911-019-0795-y] [Medline: 30871518]
9. Mitra A, Pradhan R, Melamed RD, Chen K, Hoaglin DC, Tucker KL, et al. Associations between natural language processing-enriched social determinants of health and suicide death among US Veterans. JAMA Netw Open. Mar 01, 2023;6(3):e233079. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.3079] [Medline: 36920391]
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. Jul 19, 2023;30(8):1448-1455. [FREE Full text] [doi: 10.1093/jamia/ocad071] [Medline: 37100768]
11. Gu B, Shao V, Liao Z, Carducci V, Brufau SR, Yang J, et al. Scalable information extraction from free text electronic health records using large language models. BMC Med Res Methodol. Jan 28, 2025;25(1):23. [FREE Full text] [doi: 10.1186/s12874-025-02470-z] [Medline: 39871166]
12. Gabriel RA, Litake O, Simpson S, Burton BN, Waterman RS, Macias AA. On the development and validation of large language model-based classifiers for identifying social determinants of health. Proc Natl Acad Sci U S A. Sep 24, 2024;121(39):e2320716121. [FREE Full text] [doi: 10.1073/pnas.2320716121] [Medline: 39284061]
13. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med. Jan 11, 2024;7(1):6. [FREE Full text] [doi: 10.1038/s41746-023-00970-0] [Medline: 38200151]
14. Consoli B, Wang H, Wu X, Wang S, Zhao X, Wang Y, et al. SDoH-GPT: using large language models to extract social determinants of health. J Am Med Inform Assoc. Jan 01, 2026;33(1):67-78. [doi: 10.1093/jamia/ocaf094] [Medline: 40493530]
15. Keloth VK, Selek S, Chen Q, Gilman C, Fu S, Dang Y, et al. Social determinants of health extraction from clinical notes across institutions using large language models. NPJ Digit Med. May 17, 2025;8(1):287. [FREE Full text] [doi: 10.1038/s41746-025-01645-8] [Medline: 40379919]
16. Yu Z, Peng C, Yang X, Dang C, Adekkanattu P, Gopal Patra B, et al. Identifying social determinants of health from clinical narratives: a study of performance, documentation ratio, and potential bias. J Biomed Inform. May 01, 2024;153:104642. [FREE Full text] [doi: 10.1016/j.jbi.2024.104642] [Medline: 38621641]
17. Patra BG, Lepow LA, Kasi Reddy Jagadeesh Kumar P, Vekaria V, Sharma MM, Adekkanattu 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. 2025;32(1):218-226. [doi: 10.1093/jamia/ocae260] [Medline: 39423850]

18. Arora R, Wei J, Hicks R, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al. HealthBench: evaluating large language models towards improved human health. arXiv. Preprint posted online May 13, 2025. [FREE Full text] [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. Jul 10, 2025;8:1557920. [FREE Full text] [doi: 10.3389/frai.2025.1557920] [Medline: 40718621]
20. 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. Jul 19, 2023;30(8):1438-1447. [FREE Full text] [doi: 10.1093/jamia/ocad054] [Medline: 37080559]
21. Lybarger K, Yetisgen M, Uzuner. The 2022 n2c2/UW shared task on extracting social determinants of health. J Am Med Inform Assoc. Jul 19, 2023;30(8):1367-1378. [FREE Full text] [doi: 10.1093/jamia/ocad012] [Medline: 36795066]
22. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a Research Patient Data Repository. AMIA Annu Symp Proc. 2006;2006:1044. [FREE Full text] [Medline: 17238663]
23. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. Building the Partners HealthCare Biobank at Partners Personalized Medicine: informed consent, return of research results, recruitment lessons and operational considerations. J Pers Med. Jan 14, 2016;6(1):2. [FREE Full text] [doi: 10.3390/jpm6010002] [Medline: 26784234]
24. Institute of Medicine. Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. Washington, DC. National Academies Press; 2014.
25. Institute of Medicine. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC. National Academies Press; 2014.
26. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, 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. Aug 02, 2019;7(3):e13802. [FREE Full text] [doi: 10.2196/13802] [Medline: 31376277]
27. Botelle R, Bhavsar V, Kadra-Scalzo G, Mascio A, Williams MV, Roberts A, 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. Feb 16, 2022;12(2):e052911. [FREE Full text] [doi: 10.1136/bmjopen-2021-052911] [Medline: 35172999]
28. Label Studio. URL: https://labelstud.io/ [accessed 2026-04-27]
29. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA. SAGE Publications; 2019.
30. 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. Jun 2024;12:102545. [doi: 10.1016/j.mex.2023.102545]
31. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438-447. [FREE Full text] [Medline: 35308962]
32. 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. Jul 2021;4(3):ooaa069. [FREE Full text] [doi: 10.1093/jamiaopen/ooaa069] [Medline: 34514351]
33. Kim H, Kim J, Shenvi E, Quach J, Sutjiadi B, Richardson A, et al. Developing a semantic model to describe physical activity data. Stud Health Technol Inform. 2016;225:447-451. [Medline: 27332240]
34. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. Pacific Symposium on Biocomputing. 2017;22:230-241. [FREE Full text] [doi: 10.1142/9789813207813_0023] [Medline: 27896978]
35. Winden TJ, Chen ES, Monsen KA, Wang Y, Melton GB. Evaluation of flowsheet documentation in the electronic health record for residence, living situation, and living conditions. AMIA Jt Summits Transl Sci Proc. 2018;2017:236-245. [FREE Full text] [Medline: 29888079]
36. Feller DJ, Zucker J, Don't Walk OB, Srikishan B, Martinez R, Evans H, 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;2018:422-429. [FREE Full text] [Medline: 30815082]
37. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. New York, NY. Association for Computational Linguistics (ACL); 2014. Presented at: The 2014 Conference on Empirical Methods in Natural Language Processing; October 25-29, 2014:1532-1543; Doha, Qatar. [doi: 10.3115/v1/D14-1162]
38. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. Dec 2017;5:135-146. [doi: 10.1162/tacl_a_00051]
39. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data. May 10, 2019;6(1):52. [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
40. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. New York, NY. IEEE; 2019. Presented at: 2019 IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China. [doi: 10.1109/ichi.2019.8904728]

41. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. New York, NY. Association for Computational Linguistics; 2019. Presented at: The 2019 Conference on Empirical Methods in Natural Language Processing; November 3-7, 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1410]
42. Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. UmlsBERT: clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. New York, NY. Association for Computational Linguistics; 2021. Presented at: The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6-11, 2021:1744-1753; Online. [doi: 10.18653/v1/2021.naacl-main.139]
43. Lu C, Tormey D, McCreedy L, Browne A. Generating the MEDLINE N-gram set. 2015. Presented at: AMIA Annual Symposium; November 14-18, 2015:1569; San Francisco, CA.
44. Lu C, Tormey D, McCreedy L, Browne A. Generating a distilled N-gram set - effective lexical multiword building in the SPECIALIST lexicon. 2017. Presented at: The 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017); February 21-23, 2017:77-87; Porto, Portugal. [doi: 10.5220/0006142000770087]
45. GitHub. URL: https://github.com/bwang482/SDoH_Extraction [accessed 2026-04-27]
46. Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. NEJM AI. Jan 17, 2024;1(2):AIdbp2300092. [doi: 10.1056/aidbp2300092]
47. Wang S, Hu M, Li Q, Safari M, Yang X. Capabilities of GPT-5 on multimodal medical reasoning. arXiv. Preprint posted online August 11, 2025. [FREE Full text] [doi: 10.48550/arXiv.2508.08224]
48. Chung H, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. Journal of Machine Learning Research. Jan 01, 2024:3381-3433. [FREE Full text]
49. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing System. 2022. Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; December 9, 2022:27730-27744; New Orleans, LA. [doi: 10.52202/068431-2011]
50. Wei J, Bosma M, Zhao V, Guu K, Yu A, Lester B, et al. Finetuned language models are zero-shot learners. 2022. Presented at: International Conference on Learning Representations; April 25, 2022; Online.
51. Sanh V, Webson A, Raffel C, Bach S, Sutawika L, Alyafeai Z, et al. Multitask prompted training enables zero-shot task generalization. 2022. Presented at: International Conference on Learning Representations; April 25, 2022; Online. [doi: 10.18653/v1/2022.acl-demo.9]
52. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc. Jan 27, 2024:1812-1820. [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
53. Li J, Zhang Q, Yu Y, Fu Q, Ye D. More agents is all you need. Transactions on Machine Learning Research. Oct 2024:1-18. [FREE Full text]
54. Lu J, Pang Z, Xiao M, Zhu Y, Xia R, Zhang J. Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models. arXiv. Preprint posted online July 8, 2024. [FREE Full text] [doi: 10.48550/arXiv.2407.06089]
55. Chen Z, Lu X, Li J, Chen P, Li Z, Sun K, et al. Harnessing multiple large language models: a survey on LLM ensemble. arXiv. Preprint posted online February 25, 2025. [FREE Full text] [doi: 10.48550/arXiv.2502.18036]
56. Yang H, Li M, Zhou H, Xiao Y, Fang Q, Zhou S, et al. Large language model synergy for ensemble learning in medical question answering: design and evaluation study. J Med Internet Res. Jul 14, 2025;27:e70080. [FREE Full text] [doi: 10.2196/70080] [Medline: 40658884]
57. Ramachandran G, Fu Y, Han B, Lybarger K, Dobbins N, Uzuner O, et al. Prompt-based extraction of social determinants of health using few-shot learning. New York, NY. Association for Computational Linguistics; 2023. Presented at: The 5th Clinical Natural Language Processing Workshop; July 14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.clinicalnlp-1.41]
58. Brignone E, LeJeune K, Mihalko AE, Shannon AL, Sinoway LI. Self-reported social determinants of health and area-level social vulnerability. JAMA Netw Open. May 01, 2024;7(5):e2412109. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.12109] [Medline: 38767915]
59. 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. Nov 01, 2020;27(11):1764-1773. [FREE Full text] [doi: 10.1093/jamia/ocaa143] [Medline: 33202021]
60. Mitra A, Chen K, Liu W, Kessler RC, Yu H. Post-discharge suicide prediction among US veterans using natural language processing-enriched social and behavioral determinants of health. Npj Ment Health Res. Feb 22, 2025;4(1):8. [FREE Full text] [doi: 10.1038/s44184-025-00120-2] [Medline: 39987238]
61. Yang M, Kwak G, Pollard T, Celi L, Ghassemi M. Evaluating the impact of social determinants on health prediction in the intensive care unit. New York, NY. Association for Computing Machinery; 2023. Presented at: The 2023 AAAI/ACM Conference on AI, Ethics, and Society; August 8-10, 2023:333-350; Montréal, Canada. [doi: 10.1145/3600211.3604719]
62. Hartzler AL, Xie SJ, Wedgeworth P, Spice C, Lybarger K, Wood BR, 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. Jul 19, 2023;30(8):1456-1462. [FREE Full text] [doi: 10.1093/jamia/ocad043] [Medline: 36944091]
63. Na PJ, Shin J, Kwak HR, Lee J, Jester DJ, Bandara P, et al. Social determinants of health and suicide-related outcomes: a review of meta-analyses. JAMA Psychiatry. Jan 02, 2025;82(4):337-346. [doi: 10.1001/jamapsychiatry.2024.4241] [Medline: 39745761]
64. Papini S, Hsin H, Kipnis P, Liu VX, Lu Y, Girard K, et al. Validation of a multivariable model to predict suicide attempt in a mental health intake sample. JAMA Psychiatry. Jul 01, 2024;81(7):700-707. [doi: 10.1001/jamapsychiatry.2024.0189] [Medline: 38536187]

Abbreviations

API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
EHR: electronic health record
FLAN-T5: Fine-Tuned Language Net—Text-to-Text Transfer Transformer
HIPAA: Health Insurance Portability and Accountability Act
ICD: International Classification of Diseases
LLaMA: Large Language Model Meta AI
LLM: large language model
MGB: Mass General Brigham
NLP: natural language processing
RBS: rule-based system
RPDR: Research Patient Data Registry
SDoH: social determinants of health
UMLS: Unified Medical Language System
WIC: Women, Infants, & Children Nutrition Program

Edited by A Coristine; submitted 15.Dec.2025; peer-reviewed by T Peterson, Á García-Barragán; comments to author 08.Jan.2026; accepted 14.Apr.2026; published 19.May.2026

Please cite as:
Wang B, Kabir D, Clark CR, Choi KW, Smoller JW
Extracting Social Determinants of Health From Electronic Health Records: Development and Comparison of Rule-Based and Large Language Model Methods
JMIR Med Inform 2026;14:e89534
URL: https://medinform.jmir.org/2026/1/e89534
doi: 10.2196/89534
PMID: