BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Large-scale identification of social and behavioral determinants of health from clinical notes: comparison of Latent Semantic Indexing and Generative Pretrained Transformer (GPT) models

Sujoy Roy[1], Shane Morrell[2], Lili Zhao[3] and Ramin Homayouni[1,4*]

Abstract

Background Social and behavioral determinants of health (SBDH) are associated with a variety of health and utilization outcomes, yet these factors are not routinely documented in the structured fields of electronic health records (EHR). The objective of this study was to evaluate different machine learning approaches for detection of SBDH from the unstructured clinical notes in the EHR.

Methods Latent Semantic Indexing (LSI) was applied to 2,083,180 clinical notes corresponding to 46,146 patients in the MIMIC-III dataset. Using LSI, patients were ranked based on conceptual relevance to a set of keywords (lexicons) pertaining to 15 different SBDH categories. For Generative Pretrained Transformer (GPT) models, API requests were made with a Python script to connect to the OpenAI services in Azure, using `gpt-3.5-turbo-1106` and `gpt-4-1106-preview` models. Prediction of SBDH categories were performed using a logistic regression model that included age, gender, race and SBDH ICD-9 codes.

Results LSI retrieved patients according to 15 SBDH domains, with an overall average PPV $\geq$ 83%. Using manually curated gold standard (GS) sets for nine SBDH categories, the macro-F1 score of LSI (0.74) was better than ICD-9 (0.71) and GPT-3.5 (0.54), but lower than GPT-4 (0.80). Due to document size limitations, only a subset of the GS cases could be processed by GPT-3.5 (55.8%) and GPT-4 (94.2%), compared to LSI (100%). Using common GS subsets for nine different SBDH categories, the macro-F1 of ICD-9 combined with either LSI (mean 0.88, 95% CI 0.82-0.93), GPT-3.5 (0.86, 0.82-0.91) or GPT-4 (0.88, 0.83-0.94) was not significantly different. After including age, gender, race and ICD-9 in a logistic regression model, the AUC for prediction of six out of the nine SBDH categories was higher for LSI compared to GPT-4.0.

Conclusions These results demonstrate that the LSI approach performs comparable to more recent large language models, such as GPT-3.5 and GPT-4.0, when using the same set of documents. Importantly, LSI is robust, deterministic, and does not have document-size limitations or cost implications, which make it more amenable to real-world applications in health systems.

*Correspondence:
Ramin Homayouni
rhomayouni@oakland.edu
Full list of author information is available at the end of the article

Keywords Social determinants of health, Electronic health records, Machine learning, Natural language processing, Clinical notes

Background

There is growing evidence that Social and Behavioral Determinants of Health (SBDH), such as housing insecurity, financial insecurity, drug abuse, depression and others, are associated with a wide variety of health outcomes and that including SBDH data can improve prediction of health risks [1, 2]. While many studies focus on using neighborhood level SBDH indicators, evidence suggests that using individual-level SBDH significantly improves prediction of outcomes such as medication adherence, risk of hospitalization, HIV risk, suicide attempts, or the need for social work [1]. In contrast, most studies that used external neighborhood-level data showed minimal contribution to individual risk prediction [1]. Currently, documentation of individual-level SBDH is sparse and incomplete in the structured fields within the EHR [3], but there are increasing efforts to implement screening tools in clinical workflow to document patient-level SBDH factors [4]. However, screening tools add a significant burden on the healthcare staff at a time when provider burnout is a major concern [5].

SBDH topics may arise during informal communications between the patient and healthcare provider, which are often documented in the clinical notes rather than the structured fields in the EHR [5]. As an alternative strategy to screening questionnaires and diagnosis codes, several groups have evaluated SBDH documented in the clinical notes in the EHR. Navathe et al. reported that the highest rates of social characteristics were found in physician notes and that the frequency of six out of the seven social characteristics increased when comparing data from physician notes with billing codes [6]. Similarly, in a larger study, Hatef et al. reported that the prevalence of SBDH in notes was vastly higher compared to billing codes for social isolation (2.59% vs 0.58%), housing issues (2.99% vs 0.19%), and financial strain (0.99% vs 0.06%) [7].

Recent work has focused on developing natural language processing (NLP) and machine learning approaches to extract or infer SBDH from clinical narratives [8, 9]. NLP approaches are rule-based and identify SBDH lexicons (keywords and/or phrases) using keyword matching or regular expressions. Identification of SBDH lexicons and NLP rules require considerable manual refinement [10, 11]. More recently, supervised machine learning approaches have been explored for identification of SBDH from notes, by combining a variety of text transformation methods, such as bag-of-words, n-grams, Word2Vec or Bi-directional Encoder Representation from Transformers (BERT), with supervised classification methods such as support vector machines, random forests, logistic regression, convolutional neural network and feed-forward neural network methods [8]. More recent methods that combine transformer-based embeddings learned from large volumes of documents (Large Language Models, LLM) and deep learning classifiers have demonstrated superior performance in extracting SBDH from clinical notes [12–15]. However, these models require training large amount of external data sources and fine-tuning using positive and negative gold standard cases. Thus, these approaches still require a considerable amount of manual effort for fine-tuning and may not be applicable to SBDH factors with low prevalence [9]. Recent studies explored augmentation of low prevalence SBDH using simulated synthetic data and showed that fine-tuned Flan-T5 models outperformed zero-shot Generative Pretrained Transformer (GPT) models [16]. In another study, the performance of various LLM models were evaluated for extraction of 10 different SDOH event types and arguments from clinical notes for a small corpus of pediatric patients [17].

In this study, using the publicly available MIMIC-III dataset [18], we analyzed all clinical notes for over 46,000 patients to predict 15 different SBDH categories using a well-known mathematical approach, called Latent Semantic Indexing (LSI). Here, we describe the steps in selection of SBDH categories, LSI model development, and the lexicon selection for ranking all patients in the cohort with respect to each SBDH category. The performance of LSI was manually evaluated by chart review. Finally, using a subset of gold standard patients, we compared the performance of LSI with more recent GPT models in predicting SBDH.

Methods

Latent Semantic Indexing

The overview of our approach is shown in Fig. 1. Out of a total of 46,520 patients in the MIMIC-III dataset, 46,146 patients had clinical notes. The number of notes associated with these patients ranged from 1 to 1420, with the median being 21 notes. For each patient, a patient-document was created by concatenating the individual notes sequentially in the same order as present in the database. Terms (keywords) were extracted

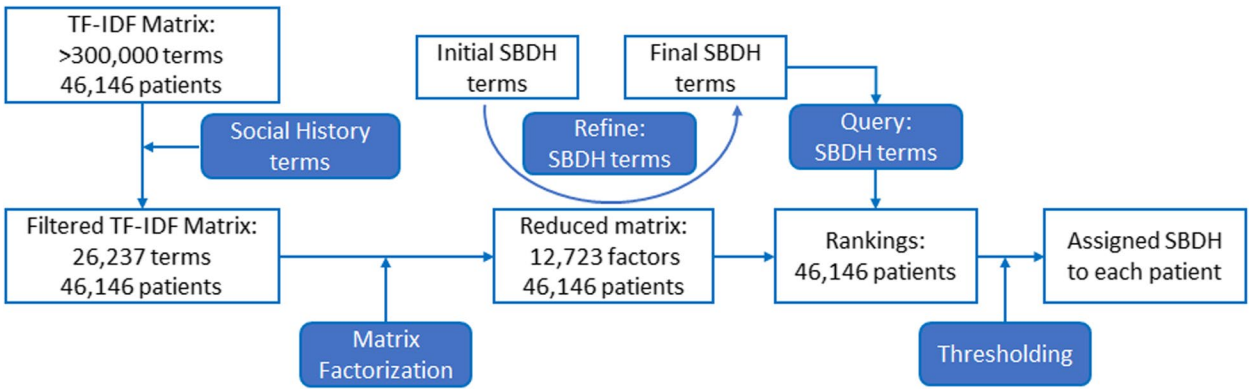


Fig. 1 Workflow diagram of extracting and assigning SBDH factors to each patient in MIMIC-III dataset

from patient-documents using Text-to-Matrix Generator (TMG) package [19]. Punctuation (excluding hyphens and underscores) and capitalization were ignored. Additionally, articles and other common, non-distinguishing words were filtered out using the SMART stop list [20]. After processing, the resulting dictionary included >300,000 terms. To reduce the dictionary size and to focus on terms that are relevant to SBDH, the dictionary was filtered to include only terms that were present in the social history sections of the clinical notes. This resulted in a final dictionary size of 26,237 terms. Each term in the 26,237 terms-by- 46,146 patients matrix was weighted using *tf-idf*, and the matrix was then factorized using Latent Semantic Indexing (Singular Value Decomposition) into three sub-matrices: 26,237 terms-by- 26,237 factors sub-matrix; 46,146 patients-by- 26,237 factors sub-matrix; and 26,237 singular values (scaling factors) diagonal sub-matrix. The optimal number of factors (dimensions) was calculated to be 12,723. Subsequently, each term and patient were represented as numeric row-vectors in reduced (12,723) dimensions.

The relationship between patients and a term can be calculated using the cosine between their vectors. A term query will produce an ordering of all patients based on the cosine associations between their respective vectors. By using lower dimensional sub-matrices, the terms/patients can be grouped together more conceptually, whereas by using higher dimensions, terms/patients can be grouped more literally. The details of this process (and various applications) have been previously described by our group [21–29] and are documented in Additional file 1.

SBDH categories

To develop a comprehensive set of SBDH categories for benchmarking the text-based approaches, we combined Social Determinants of Health (SDoH) categories defined by Torres et al. [30], and chronic behavior categories defined by the Center for Medicaid and Medicare Services (CMS) [31]. The number of patients coded for SDoH ICD-9 codes are shown in Supplementary Figure S1 in Additional file 1. Only five SDoH categories had more than six patients: *V600 housing insecurity* (202), *V1541 physical & sexual abuse* (37), *V620 financial insecurity* (15), *V625 legal circumstances* (13), and *V602 financial circumstances* (6). In addition, we included four behavioral chronic conditions defined by CMS and several other SBDH categories such as *suicidal ideation* and *compliance*, which are represented in ICD-10 but not in ICD-9. Altogether, this study focused on 15 SBDH categories (Table 1), although only nine categories were documented by ICD-9 billing codes in this data set (Supplementary Table S1 in Additional file 1).

Table 1 Number of patients within the cohort who were ICD-9 coded with the following SBDH categories

SBDH Category	Patients (n)
Tobacco use	3005
Alcohol abuse	2988
Opiate abuse	672
Cocaine abuse	545
Housing insecurity	202
Physical & sexual abuse	37
Financial insecurity	15
Legal Circumstances	13
Financial circumstances	6
Compliance	0
Mobility issues	0
Lack of English proficiency	0
Caregiver dependency	0
Suicidal ideation	0
Lack of transportation	0

Lexicon development and patient ranking

To determine the best lexicons to represent various SBDH categories, we manually constructed a set of 134 terms (including variants and plurals) corresponding to the SBDH categories (Supplementary Table S2 in Additional file 1). The lexicons were iteratively refined manually according to the following steps: 1) The pairwise Pearson correlations between terms (treating each term as a vector of term-patient cosine similarities for all patients in the collection) were used to filter out synonyms and closely associated or redundant terms; 2) When applicable, the recall of ICD-9 coded patients at a defined cosine threshold (described below) was used to choose the most representative SBDH category keyword; 3) The precision of the top ranked patients for each keyword query was used to select the best keyword that represented each category. Table 2 lists the categories and their representative keywords and Supplementary Table S1 in Additional file 1 lists the available ICD-9 codes for 9 of the 15 categories.

For each of the 15 SBDH representative keywords, all 46,146 patients were ranked in descending order of the cosine similarity between their vectors. Patients with a cosine value above a cutoff threshold (τ), defined by $\tau > Q3 + (3.0 * IQR)$, were assigned to the respective SBDH category. The IQR (interquartile range) was calculated as Q3 (75^{th} percentile) - Q1 (25^{th} percentile). The patients with a cosine value above τ for each SBDH term query were evaluated manually by chart review to determine the positive predictive value (PPV) of the top 10, median 10 and last 10 ranked patients.

Generative Pretrained Transformers (GPT)

All GPT API requests were made using a Python script which uses the `openai` library to connect to the OpenAI services in Azure, using `gpt-3.5-turbo-1106` and `gpt-4-1106-preview` models. The Azure OpenAI Service is a secure enterprise utility that is fully controlled by Microsoft and does not interact with any services operated by OpenAI (e.g. ChatGPT, or the OpenAI API) [32]. Using this platform mitigated any potential risks to data sharing agreements or to patient privacy. Each API call included two components: 1) A function definition for the SBDH category, and 2) The contents of a patient-document. GPT identifies the presence of the SBDH category in a document based on the name of the function and parameter names, with no other domain-specific information provided to the API. Each SBDH domain had its own function definition in the format of a JSON object (Additional file 1). Below is an example function definition for 'Housing insecurity':

```
            "name": "identify_housing_insecurity",
            "parameters": {
                    "type": "object",
                    "properties": {
                            "housing_insecurity": {
                                    "type": "string",
                                    "enum": ["Yes", "No"]
                            }
                    }
            },
            "required": ["housing_insecurity"]
    }
```

Table 2 Performance of LSI predictions of SDBH categories

SBDH Category (Keyword query)	Predicted N	PPV of LSI Predictions			
		Top 10	Median 10	Bottom 10	Average
Tobacco use (Smokes)	2195	100%	90%	80%	90%
Alcohol abuse (EtOH)	1080	100%	100%	100%	100%
Opiate abuse (Opiate)	444	100%	60%	50%	70%
Cocaine abuse (Cocaine)	1852	100%	70%	40%	70%
Housing insecurity (Homeless)	470	100%	80%	70%	83%
Physical & sexual abuse (Abused)	121	80%	50%	30%	53%
Financial insecurity (Unemployed)	809	100%	90%	100%	97%
Legal circumstances (Legal)	1052	80%	50%	20%	50%
Financial circumstances (Financial)	402	100%	60%	90%	83%
Compliance (Noncompliant)	402	100%	100%	90%	97%
Mobility issues (Walker)	3235	90%	100%	90%	93%
Lack of English proficiency (Interpreter)	1621	100%	90%	80%	90%
Caregiver dependency (Caretaker)	443	100%	90%	60%	83%
Suicidal ideation (Suicide)	1090	100%	60%	40%	67%
Lack of transportation (Transportation)	452	60%	70%	70%	67%

The terms in parentheses indicate the query word used to rank all patients in the dataset

Sending a function ensures that the response from the API will be a predictable, well-formed JSON object with a binary answer of "Yes" or "No" to indicate the presence of the SBDH category in the patient-document. The GPT engine does not actually call the function but instead treats the function like a callback, where the response from GPT includes the "Yes" or "No" value of the function parameter. The Python script calls the API as follows, including the patient-document and the domain function as arguments:

```
response = openai.ChatCompletion.create(
engine = "gpt model name",
messages = [{"role": "user", "content": "Contents of patient-document
    here..."}],
functions = [sbdh_function],
function_call = {"name": "identify_housing_insecurity"},
temperature = .01
)
```

The "temperature" argument controls the determinism of the GPT model, accepting a value between 0 (more deterministic) and 2 (less deterministic). The API call and SBDH function definitions are identical for GPT-3.5 and GPT-4. All prompts were zero-shot, with no fine-tuning examples provided in the prompt. Due to inconsistent responses by GPT-3.5, each prompt was submitted five independent times and the final answer was determined by simple majority. Only one prompt was submitted for GPT-4 because its unresponsiveness was infrequent.

Analysis and evaluation

The classification performance of LSI was compared to ICD-9 coding, GPT-3.5 and GPT-4 using a separate set of 621 gold standard (GS) patient-documents that were randomly chosen from the entire collection of 46,146 patients and then manually labeled. To generate the GS set for each SBDH, a random sample of up to 20 ICD-9 coded (when applicable) and up to 20 LSI predicted patients were balanced with an equal number of non-coded or non-LSI-predicted patients from the rest of the collection. The GS set included only nine of the 15 possible SBDH categories that had at least six ICD-9 coded patients (Table 1). This resulted in random samples ranging from 46 (*financial circumstances*) to a maximum of 80 (*housing insecurity*, *tobacco use*, *alcohol abuse*, *cocaine abuse* and *opiate abuse*). All cases were manually evaluated by chart review to determine actual positive (P) and negative (N) cases for each SBDH category.

During manual chart review, we found that some patients who were ICD-9 coded with specific SBDH did not have any statements in the clinical notes that supported the assignment of the ICD-9 code. We treated these cases as actual positives to represent the real-world situation where diagnosis codes are assigned to patients by healthcare providers based on their professional judgement using other data sources (e.g. labs, imaging, or external questionnaires or self-reported information in the case of SBDH). Supplementary Table S3 in Additional file 1 includes the summary characteristics of the GS samples for each SBDH category. The performance of the text-based approaches (LSI, GPT-3.5, GPT-4) was evaluated by calculating Precision, Recall and F1 score.

To compare the overall performance of the text-based predictions using either LSI or GPT-4 compared to ICD-9 coding alone, we used a logistic regression model to predict each of the nine SBDH categories in the GS subset represented as binary dependent variables (positive or negative). The base regression model included age (numeric), gender (binary), race (categorical) and ICD-9 (binary) as independent variables. The second model included the base model plus LSI-identified cases as an additional binary independent variable, whereas the third model included the base model plus GPT-4-identified cases as the additional binary independent variable. In all three models, age was fitted using a cubic spline with 2 degrees of freedom. The performance of each model was evaluated by 10-fold cross-validation and the Area Under the Receiver Operating Curve (AUROC).

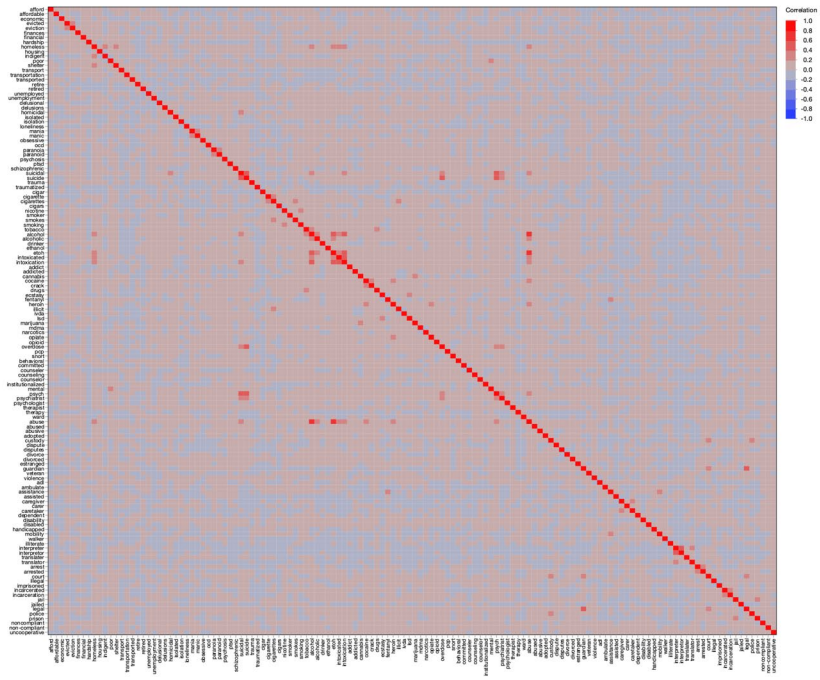
Results

Analysis of the MIMIC-III dataset showed that out of 44 potential Social Determinants of Health (SDoH) ICD-9 codes [30], only 17 were used in MIMIC-III and only five SDoH categories were assigned to six or more patients (Supplementary Figure S1 in Additional file 1).

Evaluation of LSI-derived SBDH predictions

Figure 2a shows a heatmap of the Pearson correlations between 134 SBDH query terms based on each term's corresponding list of patient cosine values. A magnified view of the heatmap for each SBDH category is provided in Additional file 2. Clustering the term correlations revealed groups of highly synonymous terms deduced from the word usage patterns in the patient-documents. This demonstrates the utility of matrix factorization as an unsupervised machine learning approach which learns conceptually related terms based on the word usage patterns in the clinical notes. For example, factorization revealed that words such as intoxicated/intoxication, crack/cocaine, or manic/mania are synonymously used in the clinical notes (Fig. 2b). In addition, this approach identified short phrases in a rudimentary way, such as legal/guardian (Fig. 2b). Lastly, some of the larger clusters included broader contextual information, such as suicide/overdose/psych/suicidal/psychiatrist (Fig. 2c).

a



b

Cluster #	Within Cluster R^2	Synonymous terms
24	0.723	intoxicated, intoxication
9	0.69	crack, cocain
14	0.667	mania, manic
4	0.623	guardian, legal
19	0.619	cannabis, marijuana
22	0.602	arrested, arrest
27	0.6	incarceration, incarcerated
20	0.595	obsessive, ocd
33	0.59	delusional, delusions
23	0.587	cigar, cigars
36	0.549	smokes, smoker
48	0.523	uncooperative, trauma
43	0.521	behavioral, schizophrenic
45	0.518	psychologist, therapist
37	0.517	noncompliant, isolated
30	0.515	illiterate, institutionalized
40	0.506	hardship, afford
39	0.505	veteran, indigent
46	0.504	isolation, adl

c

Cluster	Number of Members	Total Proportion of Variation Explained	Terms
1	5	0.019	abuse, etoh, alcohol, alcoholic, ethanol
2	5	0.018	suicide, overdose, psych, suicidal, psychiatrist
4	3	0.014	guardian, legal, court
6	5	0.013	interpreter, interpretor, translator, translater, transportation
5	5	0.012	mobility, assistance, narcotics, therapy, retired
24	2	0.011	intoxicated, intoxication
11	3	0.011	cigarettes, illicit, cigarette
12	4	0.011	evicted, eviction, housing, disputes
10	4	0.011	poor, mental, fentanyl, dependent
3	4	0.011	drugs, tobacco, pcp, illegal
9	2	0.01	crack, cocaine
14	2	0.01	mania, manic
8	3	0.01	shelter, homeless, traumatized
7	3	0.01	paranoid, paranoia, psychosis
29	3	0.01	smoking, nicotine, jailed
13	4	0.01	opiate, heroin, opioid, addicted
26	4	0.01	police, custody, violence, dispute

Fig. 2 Relationship between SBDH terms in reduced-dimensional (12,723) vector space model. **a** Heatmap of correlations between terms, where red represent high correlation and blue represents low correlation. **b** List of clusters with the highest intra-cluster correlations, depicting terms that are explicitly or conceptually synonymous as well as terms that share stems. **c** List of terms in clusters that account for 20% of the variability in the entire patient population

A patient was predicted to have a specific SBDH if the cosine value between the query term and the patient was above the cutoff cosine threshold (τ) as defined in the Methods. For all but three SBDH categories (*tobacco use*, *alcohol abuse*, and *opiate abuse*), the number of patients in the collection with an LSI-predicted SBDH was substantially higher than the ICD-9 coded patients (Table 2). To evaluate the classification performance of the LSI-derived SBDH predictions, we determined the PPV by manual evaluation of the top 10, median 10, and bottom 10 patients above τ. In all but four SBDH categories, the PPV of the top 10 ranked patients was 100%. As expected, the PPV decreased with lower rankings, which signifies lower relevancy to the query term. The average PPV for all 15 SBDH categories ranged from 50% (*legal circumstances*) to 100% (*alcohol abuse*), with nine of the SBDH categories having a PPV $\geq$ 83% (Table 2).

Next, we compared the performance of ICD-9 coding to LSI, as well as GPT-3.5 and GPT-4 large language models using different sets of gold standard (GS) patients that were randomly selected for each SBDH category and manually labeled by chart review. Importantly, only LSI was able to process all of the patient-documents. In contrast, due to context window size restrictions, GPT-3.5 processed 55.6% of the gold standard documents and GPT-4 processed 94.2% (Fig. 3). Due to these limitations, the average recall of GPT-3.5 across all of the documents in all nine SBDH categories was low (0.41), compared to LSI (0.70) and GPT-4 (0.77) (Table 3). Overall, the average macro-F1 was highest for GPT-4 (0.8), followed by LSI (0.74), ICD-9 (0.71) and GPT-3.5 (0.54) despite the fact that GPT-4 was unable to process 5.8% of the documents due to context window size limitations.

To be able to directly compare the performance of LSI, GPT-3.5 and GPT-4, the following analyses was performed using a subset of 352 GS patient-documents (out of 621) whose sizes were within the 16K context window limit of GPT-3.5, for each of the nine SBDH categories. Earlier versions of GPT were highly irreproducible such that the same prompt could produce different responses or no response at all. Therefore, for GPT-3.5, the same set of documents were submitted using the same prompt five independent times. GPT-3.5 was unresponsive for 2% (*cocaine abuse*) to 30% (*financial insecurity*) of the patient-documents across the SBDH categories (Table 4). In addition, in all but one SBDH category, GPT-3.5 provided conflicting responses between the five independent prompts. For example, although GPT3.5 provided responses for all 27 patient-documents related to legal circumstances, it provided conflicting responses for

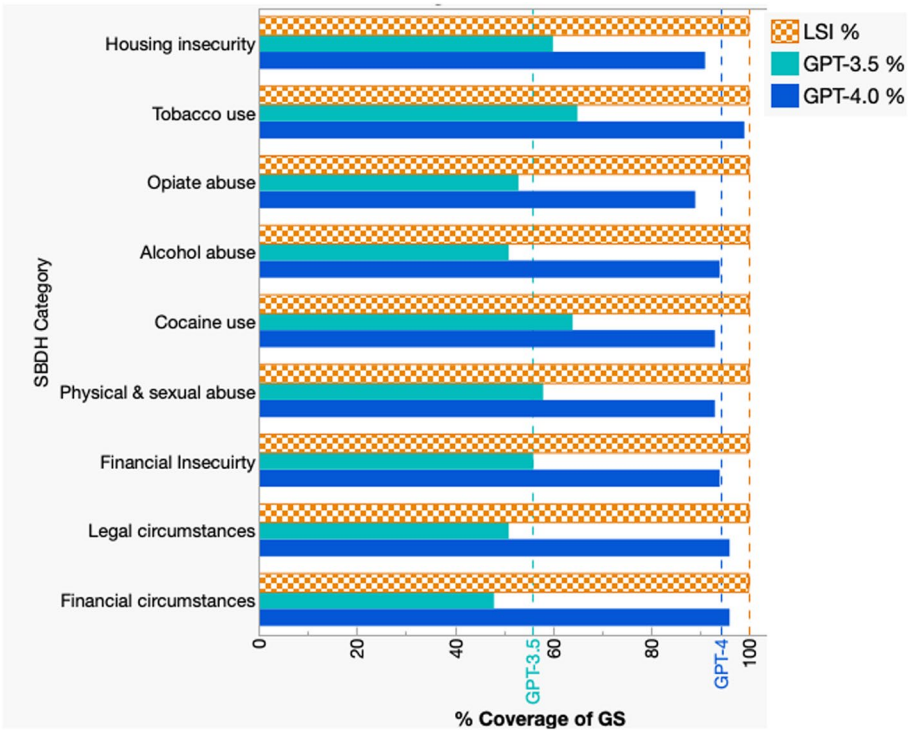


Fig. 3 Proportion of gold standard patient-documents for each SBDH category that yielded results by LSI, GPT-3.5 or GPT-4.0

six (22%) of the patient-documents (Table 4). In contrast, GPT-4 was unresponsive for only two documents (3.8%) in only one SBDH category (*tobacco use*). Averaging across all nine SBDH categories for the subset of GS cases, we found that LSI, GPT-3.5 and GPT-4 performed similarly with respect to precision, recall and F1 when the result of each method was combined with the ICD-9 coded patients (Fig. 4). This demonstrates that the three approaches perform comparably if the document sizes are within the token size limits of GPT models.

Lastly, to evaluate the overall predictive performance of LSI and GPT-4, we compared the prediction AUC of three different logistic regression models on the aforementioned subset of 352 GS cases. The base regression model included gender, age, race and SBDH ICD-9 codes as independent variables. The second model included the base variables plus LSI identified SBDH. The third model included the base model plus GPT-4 identified SBDH (Fig. 5). Using only ICD-9 coding (base model), the AUCs for the nine SBDH categories ranged between 0.69 (*housing insecurity* and *financial circumstances*) to 0.85 (*physical & sexual abuse*). In all nine categories, inclusion of LSI or GPT-4 improved the AUCs compared to ICD-9. Importantly, LSI outperformed GPT-4 in six of the nine

Table 3 Retrieval performance of each method alone using a set of sampled Gold Standard cases

		Precision				Recall				F1			
SBDH Category	**Sampled N (P)**	**ICD-9**	**LSI**	**GPT-3.5**	**GPT-4**	**ICD-9**	**LSI**	**GPT-3.5**	**GPT-4**	**ICD-9**	**LSI**	**GPT-3.5**	**GPT-4**
Housing insecurity	80 (53)	0.85	**0.95**	0.78	0.92	0.64	**0.72**	0.47	0.62	0.73	**0.82**	0.59	0.74
Tobacco use	80 (56)	**0.95**	0.93	0.89	0.88	0.68	0.66	0.43	**0.93**	0.79	0.77	0.58	**0.90**
Opiate abuse	80 (36)	**0.75**	0.63	**0.75**	0.67	**0.83**	0.69	0.42	**0.83**	**0.79**	0.66	0.54	0.74
Alcohol abuse	80 (52)	0.85	**0.95**	0.84	0.82	0.65	0.73	0.40	**0.90**	0.74	0.83	0.55	**0.86**
Cocaine abuse	80 (43)	0.78	0.80	0.90	**0.95**	0.72	0.74	0.42	**0.81**	0.75	0.77	0.57	**0.88**
Physical & sexual abuse	67 (37)	0.96	0.67	0.88	**1.00**	0.70	0.49	0.38	**0.73**	0.81	0.56	0.53	**0.84**
Unemployed	54 (36)	**1.00**	**1.00**	0.85	0.91	0.42	0.81	0.31	**0.89**	0.59	0.89	0.45	**0.90**
Legal circumstances	53 (26)	**1.00**	0.72	0.67	0.78	0.50	**0.69**	0.38	**0.69**	0.67	0.71	0.49	**0.73**
Financial circumstances	46 (18)	**1.00**	0.61	**1.00**	0.75	0.33	**0.78**	0.44	0.50	0.50	**0.68**	0.62	0.60

The bold text indicate the highest precision, recall, or F1 for each SBDH category (row)

Table 4 Unresponsiveness of GPT-3.5 and GPT-4

		GPT-3.5		GPT-4
SBDH Category	**N**	**% Disagreement**	**% No Response**	**% No Response**
Housing insecurity	48	6.3%	0.0%	0.0%
Tobacco use	52	3.8%	15.4%	3.8%
Opiate abuse	42	7.1%	0.0%	0.0%
Alcohol abuse	41	2.4%	0.0%	0.0%
Cocaine abuse	51	0.0%	2.0%	0.0%
Physical & sexual abuse	39	2.6%	5.1%	0.0%
Financial insecurity	30	6.7%	30.0%	0.0%
Legal circumstances	27	22.2%	0.0%	0.0%
Financial circumstances	22	13.6%	4.5%	0.0%

On a set of shared patient-documents (N), GPT-3.5 was prompted five independent times, whereas GPT-4 was prompted only once. The % of documents where GPT-3.5 or GPT-4 did not provide a response is indicated for each SBDH category. The % disagreement corresponds to the number of documents where GPT-3.5 provided conflicting binary responses

SBDH categories (*housing insecurity*, *financial insecurity*, *opiate abuse*, *alcohol abuse*, *legal circumstances*, and *financial circumstances*).

Discussion

In this study, we demonstrated the utility of LSI as a robust unsupervised approach for comprehensively processing all clinical notes in the EHR to identify SBDH and to supplement the SBDH documented by ICD-9 diagnosis codes. Importantly, we show that although LSI is a bag-of-words approach, it performed similarly and sometimes better than GPT models. This work highlights several advantages for using LSI in real-world healthcare applications.

One major advantage of LSI is its ability to process all of the notes for a given patient without the imposed context

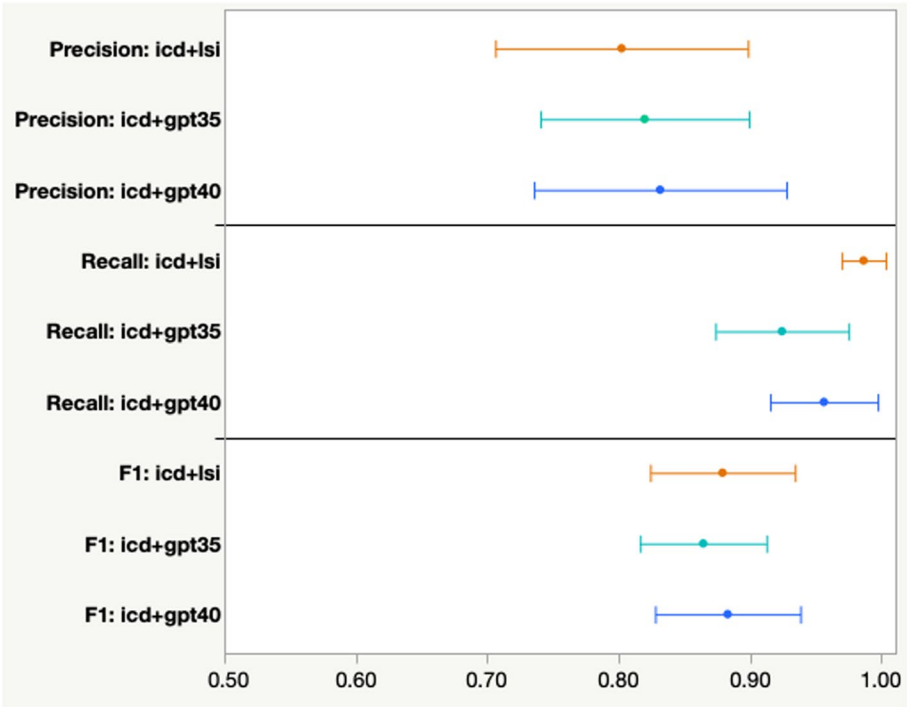


Fig. 4 Retrieval performance of LSI, GPT-3.5 or GPT-4 when combined with ICD-9 coding. Precision (upper panel), recall (middle panel) and F1 (lower panel) of ICD-9 combined with either LSI (orange lines), GPT-3.5 (cyan lines) and GPT-4 (blue lines). Values represent the mean (filled circle) and 95% confidence intervals (error bars) across the nine SBDH gold standard sets

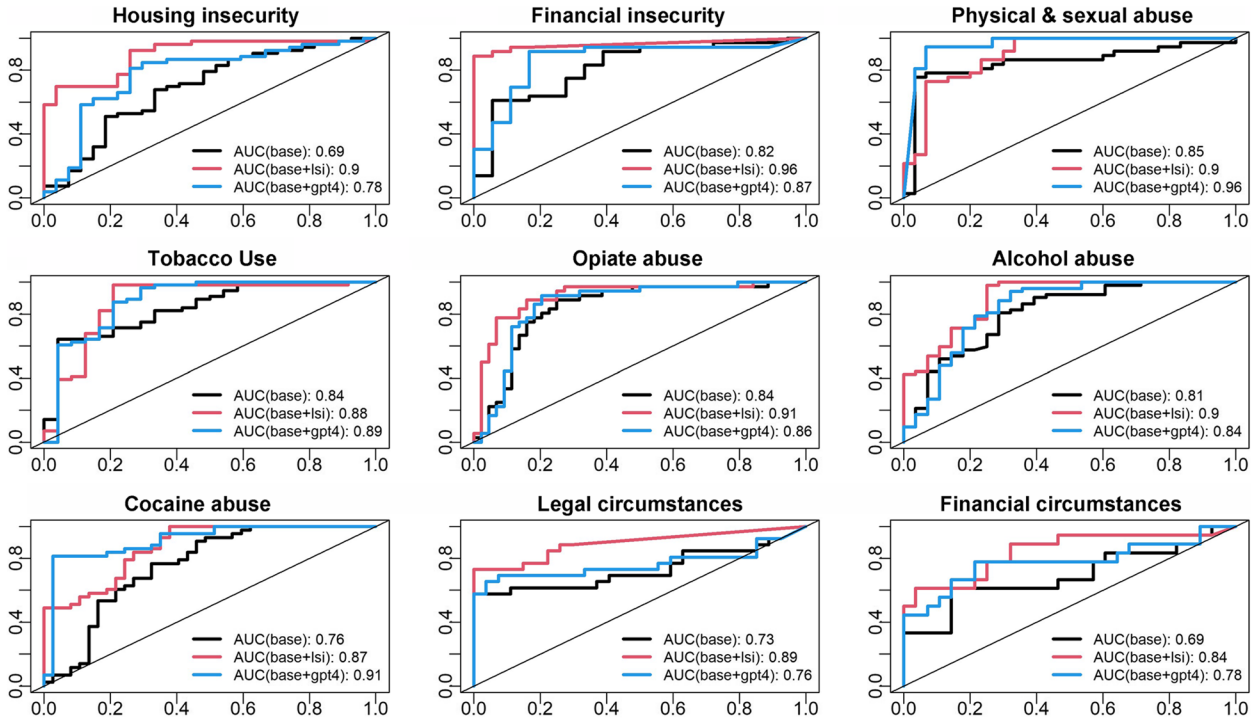


Fig. 5 Comparison of classification performance of ICD-9 and/or text-predicted SBDH categories using multivariable analysis. The AUC is shown for three different models: 1) Base model including age, gender and ICD-9 codes (black lines), 2) Base model plus LSI identified SBDH (red lines), and 3) Base model plus GPT-4 identified SBDH (blue lines)

window token size limitations of GPT. As pointed out in Fig. 3, only 55.6% and 94.2% of the GS cases could be processed by GPT-3.5 and GPT-4, respectively. At the time of our analysis, the input context window size limits for GPT-3.5 and GPT-4 were 16K and 128K tokens, respectively. However, other LLMs may have larger context windows. Even with the context window limits, it is possible to process larger documents by 'chunking', a method where a large document is split into smaller overlapping documents that are smaller than the token limits. In our analysis, we did not attempt to process all of the GS documents, instead we directly compared the performance of LSI with GPT-3.5 and GPT-4 using the same set of documents (Table 4 and Figs. 4 and 5). Another reason for limiting the analysis to a subset of GS documents was cost. At the time of the analysis, the cost for GPT-3.5 and GPT-4 using the Microsoft Azure OpenAI [32] services per query was USD \$0.001 and \$0.01 per 1K input tokens, respectively. Thus, it would have been more costly to chunk the larger GS documents. Another way to reduce the number of GPT queries would have been to perform multi-class labeling. In our analysis, we performed single class labeling, where each document was processed individually to identify a single SBDH category at a time. Although multi-class labeling would be useful, it may require considerable fine-tuning and may not be feasible for identifying all 15 SBDH categories at once.

Another major advantage of LSI is that it does not require external training on a large dataset and fine-tuning for domain specific applications. For this study, the LSI model was built using all of the clinical notes for all of the $> 46,000$ patients at once. In contrast, GPT and other LLM require extensive training using large amounts of external data sources. For example, GPT 3.5 was trained on 175 billion parameters using training data up to September 2021. Although the models perform well for general text analysis, they may not perform well on specialized clinical tasks. For example, Lybarger et al. developed an event based deep-learning extractor for SBDH that determines chronicity, duration, frequency and type of event [12]. However, their models apply only to a subset of SBDH categories, including employment, living status, as well as alcohol, tobacco and drug use. They point out that training these models required significant manual effort by human experts to develop both positive and negative gold standard datasets for fine-tuning [12]. In addition, since these methods require large amounts of training data for fine-tuning, they can have limited usefulness for SBDH categories that are rare (low prevalence).

Yet another major advantage of LSI is that, unlike GPT, it is deterministic (reproducible) and 100% responsive to all queries. For a given number of factors (post-factorization), LSI produces the same exact ranking of the patients based on the same query. On the other hand, we showed (Table 3) that GPT-3.5 produces conflicting responses to the same prompt on the same set of documents. Moreover, we demonstrated that both GPT-3.5 and GPT-4 may not respond, a phenomenon commonly referred to as 'laziness'. Although the GPT-4 model has been improved to reduce laziness, we found that it can be unresponsive as the document size reaches its maximum context window size limits.

A key takeaway from our study is that clinical notes provide a valuable source of SBDH information. However, relying solely on clinical notes is not adequate in real-world settings. We show that by combining ICD-9 codes along with SBDH detected from clinical notes allows for better prediction of patient-level SBDH needs than either method alone. During chart review for developing the GS sets, we found some ICD-9 coded individuals who had no supporting documentation in the clinical notes. For example, some patients had few encounters with the health system and had no social history notes, yet were coded for homelessness or alcohol abuse. This is typical in real-world settings where diagnosis codes are assigned by healthcare providers who may use a variety of other sources (e.g. labs, imaging, etc.) to support the diagnosis, but not document them explicitly in the notes. Similarly, SBDH codes may be assigned to a patient based on answers to screening questionnaires. Our findings are consistent with other studies showing the importance of combining the information provided by ICD-9 codes and other structured data (e.g., questionnaires) with unstructured data in the EHR to obtain a more representative assessment of the SBDH prevalence in a population [7, 10–12, 33]. Implementing SDoH questionnaires across a large health system is impractical. Studies have shown that SDoH screening forms are primarily implemented in inpatient and primary care settings. However, it is thought that socioeconomically disadvantaged individuals are less likely to go to primary care, instead use the emergency department (ED) for their healthcare needs [34]. Moreover, a recent study demonstrated that only 3.7% of the patients in a large health care system in South Carolina had answered all 11 questions on the SDoH screening forms [35]. Therefore, for better assessment of SBDH burden in a population, information must be aggregated from a variety of sources in the EHR, including the clinical notes.

It is worth highlighting that the costs associated with OpenAI services make it currently unrealistic to implement in health systems to assess SBDH burden in large populations of patients. To address this issue, future research will focus on using LSI to narrow large populations of patients into smaller groups that are conceptually predicted to have SBDH and then process those documents using GPT to contextualize and validate the LSI predictions. Factorization provides value beyond keyword searching alone because it contextualizes keywords as vectors in reduced-dimensional space, thereby grouping words that are frequently used together in the context of SBDH keywords. This approach provides a general advantage by automatically grouping synonyms, misspellings, and conceptually related terms that are often used together in narratives (Fig. 2). For example, a homeless individual is often unemployed and has drug/alcohol abuse problems. Also, factorization is able to infer that 'shelter' and 'homelessness' are synonymously used in the narratives. By reducing the (number of) factors of the factorized matrix, one can identify a subset of patients who are conceptually related to the SBDH, achieving higher recall than precision. By subsequently processing these patient-documents with GPT-4, the specific evidence in support of the SBDH can be readily deduced while keeping the overall processing cost low.

While LSI was highly sensitive (high PPV) for most SBDH categories, its performance was limited for a few SBDH categories such as legal circumstances. We found that legal circumstances covered a broad range of areas ranging from power of attorney, guardianship issues, hospital liability to encounters with law enforcement for illegal activities. More refinement would be necessary to evaluate the performance of our approach on specific areas pertaining to specific legal circumstances. For example, guardianship issues for clinical decision making could be better identified with a 'guardian' query rather than a general term such as 'legal'. In three cases (alcohol abuse, tobacco use, and opiate abuse), our approach identified fewer cases than ICD coded individuals. This may be due to the fact that drug, alcohol and tobacco use are routinely captured within structured fields in current clinical practice. However, other SBDH categories are not routinely captured. One approach to increase the number of cases identified by our approach would be to relax the thresholding parameter or to combine multiple lexicons representing alcohol abuse in an additive way.

Feller et al. were among the first groups to apply NLP methods to infer SBDH from clinical notes [36]. After feature selection, they included 2-4,000 individual words as independent variables in various machine learning classifiers to identify sexual history, sexual orientation, alcohol use, substance use and housing status. They found that combining clinical notes and structured data enabled reasonably accurate inference of these SBDH categories [37]. Bejan et al., using a vector embedding

approach to expand SDoH lexicons, demonstrated better performance of identification of homelessness and adverse childhood experiences (ACEs) from clinical notes [38]. Our process, which combines the bag-of-words approach with factorization, allows an automated method to identify a broad set of SBDH categories.

This study has several limitations. First, LSI is a bag-of-words technique, which does not account for word context (phrases) and negated terms. Second, the performance of LSI was affected by the presence of forms and templated text in the clinical notes, such as 'Family information' or social history forms, where there are many negations and repeated text. The performance would improve if certain note types, forms and templates were removed during pre-processing. Third, our approach does not provide temporal relations and event-types. Lastly, the performance of the GPT models could be further improved by fine-tuning or providing examples in the prompt, which were not explored in this study. In future work, many of these limitations could be addressed by combining the advantages of LSI (e.g., robustness, determinism, and no cost) with the advantages of LLM (i.e., contextualization, removal of negation, and multi-label classification).

Conclusions

In this study, we demonstrated that using an unsupervised machine learning factorization approach on clinical notes is a robust way to enhance SBDH identification from the EHR. In addition, the results demonstrate the importance of combining SBDH data from both structured and unstructured fields in the EHR to more comprehensively estimate the prevalence of SBDH in patient populations. By providing better estimates of SBDH burden in populations, this work sets the stage for developing patient-level health risk and utilization prediction models that incorporate SBDH factors in addition to standard clinical and structured data from the EHR.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02705-x.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

The authors are grateful to Oakland University for providing the high-performance computing resources and to MIT Laboratory for Computational Physiology for providing the MIMIC-III dataset. We thank Kevin Heinrich (Quire Inc.) and Brad Silver (Quire Inc.) for helpful discussions.

Authors' contributions

S.R. designed and implemented the methods, generated data, interpreted results and contributed to writing of the manuscript. S.M. generated data and performed analysis. L.Z. analyzed data, interpreted results and contributed to writing of the manuscript. R.H. designed the study, interpreted results, performed chart reviews and wrote the manuscript. All authors reviewed the manuscript.

Funding

This work was supported by the funding from Oakland University William Beaumont School of Medicine and the Beaumont Research Institute.

Availability of data and materials

The MIMIC-III dataset is available publicly through https://physionet.org.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

R.H. and S.M. hold equity in Quire Inc.

Author details

[1]Foundational Medical Studies, Population Health Informatics, Oakland University William Beaumont School of Medicine, Oakland University, 586 Pioneer Dr, 460 O'Dowd Hall, Rochester, MI 48309-4482, USA. [2]Quire Inc., Memphis, Tennessee, USA. [3]Biostatistics, Beaumont Research Institute, Corewell Health, Royal Oak, Michigan, USA. [4]Population Health & Health Equity Research, Beaumont Research Institute, Corewell Health, Royal Oak, Michigan, USA.

Received: 21 March 2024 Accepted: 30 September 2024
Published online: 10 October 2024

References

1. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. J Am Med Inform Assoc. 2020;27(11):1764–73. https://doi.org/10.1093/jamia/ocaa143.
2. Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, et al. Including social and behavioral determinants in predictive models: Trends, challenges, and opportunities. JMIR Med Inform. 2020;8(9). https://doi.org/10.2196/18084.
3. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. Medicine (United States). 2020;99(52). https://doi.org/10.1097/MD.0000000000023818.
4. Andermann A. Screening for social determinants of health in clinical care: Moving from the margins to the mainstream. Public Health Rev. 2018;39(1). https://doi.org/10.1186/s40985-018-0094-7.
5. Alpert J, Kim H, McDonnell C, Guo Y, George TJ, Bian J, et al. Barriers and Facilitators of Obtaining Social Determinants of Health of Patients With Cancer Through the Electronic Health Record Using Natural Language Processing Technology: Qualitative Feasibility Study With Stakeholder Interviews. JMIR Formative Res. 2022;6(12). https://doi.org/10.2196/43059.
6. Navathe AS, Zhong F, Lei VJ, Chang FY, Sordo M, Topaz M, et al. Hospital Readmission and Social Risk Factors Identified from Physician Notes. Health Serv Res. 2018;53(2):1110–36. https://doi.org/10.1111/1475-6773.12670.
7. 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. J Med Internet Res. 2019;21(8). https://doi.org/10.2196/13802.
8. 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. 2021;28(12):2716–27. https://doi.org/10.1093/JAMIA/OCAB170.
9. Lybarger K, Bear OJ, Yetisgen M, Uzuner O. Advancements in extracting social determinants of health information from narrative text. J Am Med Inform Assoc. 2023;30(8):1363–6. https://doi.org/10.1093/JAMIA/OCAD121.
10. Allen KS, Hood DR, Cummins J, Kasturi S, Mendonca EA, Vest JR. Natural language processing-driven state machines to extract social factors from unstructured clinical documentation. JAMIA Open. 2023;6(2). https://doi.org/10.1093/JAMIAOPEN/OOAD024.
11. Mehta S, Lyles C, Rubinsky A, Kemper K, 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. 2023;11. https://doi.org/10.2196/46159.
12. Lybarger K, Dobbins NJ, Long R, Singh A, Wedgeworth P, Uzuner O, et al. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. J Am Med Inform Assoc. 2023;30(8):1389–97. https://doi.org/10.1093/JAMIA/OCAD073.
13. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. J Biomed Inform. 2020;2021(113):103631. https://doi.org/10.1016/j.jbi.2020.103631.
14. Yu Z, Yang X, Dang C, Wu S, Adekkanattu P, Pathak J, et al. A Study of Social and Behavioral Determinants of Health in Lung Cancer Patients Using Transformers-based Natural Language Processing Models. AMIA Ann Symp Proc. 2021;2021:1225.
15. Yu Z, Yang X, Guo Y, Bian J, Wu Y. Assessing the Documentation of Social Determinants of Health for Lung Cancer Patients in Clinical Narratives. Front Public Health. 2022;10. https://doi.org/10.3389/FPUBH.2022.778463.
16. 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. 2024;7(1):6.
17. Fu Y, Ramachandran GK, Dobbins NJ, Park N, Leu M, Rosenberg AR, et al. Extracting social determinants of health from pediatric patient notes using large language models: novel corpus and methods. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Prague: International Committee for Computational Linguistics (ICCL); 2024. p. 7045–56.
18. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):1–9.
19. Zeimpekis D, Gallopoulos E. TMG: A MATLAB toolbox for generating term-document matrices from text collections. In: Grouping multidimensional data. Berlin: Springer; 2006. p. 187–210.
20. Salton G. The Smart document retrieval project. In: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. Hoboken: Prentice Hall; 1991. p. 356–8.
21. Homayouni R, Heinrich K, Wei L, Berry MW. Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics. 2005;21(1):104–15.
22. Heinrich KE, Berry MW, Homayouni R, et al. Gene tree labeling using nonnegative matrix factorization on biomedical literature. Comput Intell Neurosci. 2008;2008(1):276535.
23. Roy S, Heinrich K, Phan V, Berry MW, Homayouni R. Latent semantic indexing of PubMed abstracts for identification of transcription factor candidates from microarray derived gene sets. In: BMC bioinformatics. vol. 12. Heidelberg: Springer; 2011. pp. 1–13.
24. Roy S, Homayouni R, Berry MW, Puretskiy AA. Nonnegative tensor factorization of biomedical literature for analysis of genomic data. In: Data Mining for Service. Berlin: Springer; 2014. pp. 97–110.
25. Roy S, Curry BC, Madahian B, Homayouni R. Prioritization, clustering and functional annotation of MicroRNAs using latent semantic indexing of MEDLINE abstracts. In: BMC bioinformatics. vol. 17. BioMed Central; 2016. pp. 131–42.
26. Roy S, Yun D, Madahian B, Berry MW, Deng LY, Goldowitz D, et al. Navigating the functional landscape of transcription factors via non-negative tensor factorization analysis of MeDline abstracts. Front Bioeng Biotechnol. 2017;5:48.
27. Roy S, Berry MW. Mining multimodal big data: tensor methods and applications. In: handbook of research on big data storage and visualization techniques. Hershey, Pennsylvania USA: IGI Global; 2018. p. 674–702.
28. Roy S, Zaman KI, Williams RW, Homayouni R. Evaluation of Sirtuin-3 probe quality and co-expressed genes using literature cohesion. BMC Bioinformatics. 2019;20:31–43.
29. Akbilgic O, Homayouni R, Heinrich K, Langham MR, Davis RL. Unstructured text in EMR improves prediction of death after surgery in children. Informatics. 2019;6(1). https://doi.org/10.3390/informatics6010004.
30. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD Social Codes: An underutilized resource for tracking social needs. Med Care. 2017;55(9):810–6. https://doi.org/10.1097/MLR.0000000000000764.
31. CMS. Chronic Conditions Data Warehouse. https://www2.ccwdata.org/web/guest/home/. Accessed 30 Oct 2023.
32. Microsoft. Microsoft Azure OpenAI. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy. Accessed 10 Jan 2024.
33. Harle CA, Wu W, Vest JR. Accuracy of Electronic Health Record Food Insecurity, Housing Instability, and Financial Strain Screening in Adult Primary Care. JAMA. 2023;329(5):423–4. https://doi.org/10.1001/JAMA.2022.23631.
34. Capp R, Camp-Binford M, Sobolewski S, Bulmer S, Kelley L. Do adult Medicaid enrollees prefer going to their primary care provider's clinic rather than emergency department (ED) for low acuity conditions? Med Care. 2015;53(6):530.
35. Rudisill AC, Eicken MG, Gupta D, Macauda M, Self S, Kennedy AB, et al. Patient and Care Team Perspectives on Social Determinants of Health Screening in Primary Care: A Qualitative Study. JAMA Netw Open. 2023;6(11):e2345444–e2345444.
36. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. J Acquir Immune Defic Syndr (1999). 2018;77(2):160–6. https://doi.org/10.1097/QAI.0000000000001580.
37. Feller DJ, Bear OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. Appl Clin Inform. 2020;11(1):172–81. https://doi.org/10.1055/s-0040-1702214.
38. 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. 2018;25(1):61–71. https://doi.org/10.1093/jamia/ocx059.

Publisher's Note