

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

4 Sujoy Roy, PhD<sup>1</sup>, Shane Morrell<sup>2</sup>, Lili Zhao, PhD<sup>3</sup>, and Ramin Homayouni, PhD<sup>1,4\*</sup>

5 <sup>1</sup>Foundational Medical Studies, Oakland University William Beaumont School of Medicine, Oakland University,  
6 Rochester, Michigan, United States of America

7 <sup>2</sup>Quire Inc., Memphis, Tennessee, United States of America

8 <sup>3</sup>Biostatistics, Beaumont Research Institute, Corewell Health, Royal Oak, Michigan, United States of America

9 <sup>4</sup>Population Health & Health Equity Research, Beaumont Research Institute, Corewell Health, Royal Oak,  
10 Michigan, United States of America

11

12 **\*Correspondence:**

13 Ramin Homayouni, PhD

14 Professor, Foundational Medical Studies; and Director, Population Health Informatics

15 Oakland University William Beaumont School of Medicine

16 586 Pioneer Dr, 460 O'Dowd Hall, Rochester, Michigan, 48309-4482, United States of America

17 Ph: (248) 370-2874; Email: rhomayouni@oakland.edu

18 **Keywords:**

19 Social determinants of health, electronic health records, machine learning, natural language processing,  
20 clinical notes

## 21 Abstract

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

26 **Methods:** Latent Semantic Indexing (LSI) was applied to 2,083,180 clinical notes corresponding to 46,146 patients in the  
27 MIMIC-III dataset. Using LSI, patients were ranked based on conceptual relevance to a set of keywords (lexicons) pertaining to  
28 15 different SBDH categories. For Generative Pretrained Transformer (GPT) models, API requests were made with a Python  
29 script to connect to the OpenAI services in Azure, using gpt-3.5-turbo-1106 and gpt-4-1106-preview models. Prediction of  
30 SBDH categories were performed using logistic regression model that included age, gender race and SBDH ICD-9 codes with a  
31 natural cubic spline of 2 degrees of freedom for age.

32 **Results:** LSI retrieved patients according to 15 SBDH domains, with an overall average PPV  $\geq$  83%. Using manually curated  
33 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  
34 (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  
35 GPT-3.5 (55.8%) and GPT-4 (94.2%), compared to LSI (100%). Using common GS subsets for nine different SBDH categories,  
36 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,  
37 0.83-0.94) was not significantly different. After including age, gender, race and ICD-9 in a logistic regression model, the AUC  
38 for prediction of six out of the nine SBDH categories was higher for LSI compared to GPT-4.0.

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

42

## 43 Background

44 There is growing evidence that Social and Behavioral Determinants of Health (SBDH) are associated with  
45 a wide variety of health outcomes and that including SBDH data can improve prediction of health risks.<sup>1,2</sup>

46 While many studies focus on using neighborhood level SBDH indicators, evidence suggests that using  
47 individual-level SBDH significantly improves prediction of outcomes such as medication adherence, risk  
48 of hospitalization, HIV risk, suicide attempts, or the need for social work.<sup>1</sup> In contrast, most studies that  
49 used external neighborhood-level data showed minimal contribution to individual risk prediction.<sup>1</sup> Currently,

50 documentation of individual-level SBDH is sparse and incomplete in the structured fields within the EHR,<sup>3</sup>  
51 but there are increasing efforts to implement screening tools in clinical workflow to document patient-level  
52 SBDH factors.<sup>4</sup> However, screening tools add a significant burden on the healthcare staff at a time when  
53 provider burnout is a major concern.<sup>5</sup>

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

62 Recent work has focused on developing natural language processing (NLP) and machine learning approaches  
63 to extract or infer SBDH from clinical narratives.<sup>8,9</sup> NLP approaches are rule-based and identify SBDH  
64 lexicons (keywords and/or phrases) using keyword matching or regular expressions. Identification of SBDH  
65 lexicons and NLP rules require considerable manual refinement.<sup>10,11</sup> More recently, supervised machine  
66 learning approaches have been explored for identification of SBDH from notes, by combining a variety of  
67 embedding methods, such as bag-of-words, n-grams, word2vec or Bi-directional Encoder Representation from  
68 Transformers (BERT), with supervised classification methods such as support vector machines, random  
69 forests, logistic regression, convolutional neural network and feed-forward neural network methods.<sup>8</sup> More  
70 recent methods that combine transformer-based embeddings learned from large volumes of documents (Large  
71 Language Models, LLM) and deep learning classifiers have demonstrated superior performance in extracting  
72 SBDH from clinical notes.<sup>12-15</sup> However, these models require training large amount of external data sources  
73 and fine-tuning using positive and negative gold standard cases. Thus, these approaches still require a  
74 considerable amount of manual effort for fine-tuning and may not be applicable to SBDH factors with  
75 low prevalence.<sup>9</sup> Recent studies explored augmentation of low prevalence SBDH using simulated synthetic  
76 data and showed that fine-tuned Flan-T5 models outperformed zero-shot Generative Pretrained Transformer  
77 (GPT) models.<sup>16</sup>

78 In this study, using the publicly available MIMIC-III dataset,<sup>17</sup> we analyzed all clinical notes for over 46,000  
79 patients to identify 15 different SBDH categories using a well-known mathematical approach, called Latent

80 Semantic Indexing (LSI). Using a subset of gold standard patient documents, we compared the performance  
81 of LSI with more recent GPT models.

## 82 **Methods**

### 83 **Latent Semantic Indexing**

84 The overview of our approach is shown in Figure 1.

85 For each patient, a patient-document was created by concatenating the individual notes sequentially in  
86 the same order as present in the database. Terms (keywords) were extracted from patient documents  
87 using Text-to-Matrix Generator (TMG) package.<sup>18</sup> Punctuation (excluding hyphens and underscores) and  
88 capitalization were ignored. Additionally, articles and other common, non-distinguishing words were filtered  
89 out using the SMART stop list.<sup>19</sup> A term-by-patient matrix was created where the entries of the matrix  
90 were *tf-idf* weighted frequencies of terms across the patient document collection. Latent semantic indexing, a  
91 well-known factorization (Singular value decomposition) was performed on this matrix, subsequent to which  
92 each term and patient were represented as numeric vectors in reduced dimensions. The similarity between  
93 any two entities was calculated as the cosine between their respective vectors. The details of this process and  
94 various applications have been previously described by our group<sup>20–28</sup> and are documented in Additional file  
95 1.

96 A total of 15 SBDH categories were considered, inspired from Social Determinants of Health (SDoH) cat-  
97 egories defined by Torres et al.,<sup>29</sup> and chronic behavior categories defined by the Center for Medicaid and  
98 Medicare Services (CMS).<sup>30</sup> The representative keyword for each category was finalized after consultation  
99 with a group of care managers. Table 1 lists the categories and their representative keywords while Supple-  
100 mentary Table S1 in Additional file 1 also lists the available ICD-9 codes for 9 of the 15 categories. For each  
101 keyword, patients were ranked in descending order of the cosine similarity between their truncated vectors.  
102 Patients with cosine scores  $> Q3 + (3.0 * IQR)$  were assigned to the respective SBDH category. The IQR  
103 (interquartile range) was calculated as  $Q3$  (75<sup>th</sup> percentile) –  $Q1$  (25<sup>th</sup> percentile).

### 104 **Generative Pretrained Transformers (GPT)**

105 All GPT API requests were made using a Python script which uses the "openai" library to connect to the  
106 OpenAI services in Azure, using gpt-3.5-turbo-1106 and gpt-4-1106-preview models. The Azure OpenAI

107 Service is a secure enterprise utility that is fully controlled by Microsoft and does not interact with any  
108 services operated by OpenAI (e.g. ChatGPT, or the OpenAI API).<sup>31</sup> Using this platform mitigated any  
109 potential risks to data sharing agreements or to patient privacy. Each API call included two components:  
110 1) A function definition for the SBDH category, and 2) The contents of a patient document. GPT identifies  
111 the presence of the SBDH category in a document based on the name of the function and parameter names,  
112 with no other domain-specific information provided to the API. Each SBDH domain had its own function  
113 definition in the format of a JSON object (Additional file 1). Below is an example function definition for  
114 "Housing Insecurity":

```
115  
116     sbdh_function = {  
117         "name": "identify_housing_insecurity",  
118         "parameters": {  
119             "type": "object",  
120             "properties": {  
121                 "housing_insecurity": {  
122                     "type": "string",  
123                     "enum": ["Yes", "No"]  
124                 }  
125             }  
126         },  
127         "required": ["housing_insecurity"]  
128     }  
129
```

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

```
135  
136     response = openai.ChatCompletion.create(  
137         engine = "gpt model name",  
138         messages = [{"role": "user", "content": "Contents of patient  
139             document here..."}],
```

```
144     functions = [sbdh_function],  
145     function_call = {"name": "identify_housing_insecurity"},  
146     temperature = .01  
147 )  
148
```

145 The "temperature" argument controls the determinism of the GPT model, accepting a value between 0  
146 (more deterministic) and 2 (less deterministic). The API call and SBDH function definitions are identical  
147 for GPT-3.5 and GPT-4. All prompts were zero-shot, with no fine-tuning examples provided in the prompt.

## 148 **Analysis and Evaluation**

149 The patient rankings pertaining to each SBDH keyword query was evaluated manually by chart review  
150 to determine the positive predictive value (PPV) of the top 10, median 10 and last 10 ranked. SBDH  
151 classification performance was evaluated using precision, recall and F1 score on manually curated gold  
152 standard (GS) samples. A random sample of up to 20 ICD-9 coded (when applicable) and up to 20 LSI-  
153 predicted cases were balanced with an equal number of non-coded and non LSI-predicted cases for each of  
154 the nine SBDH categories (that had at least six ICD-9 coded patients). This resulted in random samples  
155 ranging from 46 (financial circumstances) to a maximum of 80 (Tobacco use, Alcohol abuse and Opiate  
156 abuse). All cases were manually evaluated by chart review to determine actual positive (P) and negative (N)  
157 cases. Supplementary Table S3 in Additional file 1 includes the summary characteristics of the GS samples  
158 for each SBDH category. The performance of the text-based approaches was evaluated by Precision, Recall  
159 and F1 score.

160 To determine the overall performance of the text-based predictions using either LSI or GPT-4 in addition  
161 to ICD-9 coding, we used a logistic regression model including age, gender, race and ICD-9 for binary  
162 classification of GS patients corresponding to each SBDH category. In all three models, age was fit using  
163 a cubic spline with 2 degrees of freedom. The performance of each model was evaluated by 10-fold cross-  
164 validation and the Area Under the Receiver Operating Curve (AUROC).

## 165 **Results**

166 A number of previous studies have demonstrated that International Disease Classification (ICD) codes cor-  
167 responding to social and behavioral determinants of health are not commonly used in the EHR.<sup>7</sup> Similarly,

168 analysis of the MIMIC-III dataset showed that out of 44 potential Social Determinants of Health (SDoH)  
169 ICD-9 codes,<sup>29</sup> only 17 were used in MIMIC-III and only nine SDoH categories were assigned to three or  
170 more patients (Supplementary Figure S1 in Additional file 1). To develop a comprehensive set of SBDH for  
171 benchmarking the text-based approaches, we included the following SDoH categories in order of frequency:  
172 *Lack of housing* (202), *history of physical abuse* (37), *unemployment* (15), *legal circumstances* (13), *inade-*  
173 *quate material resources* (6). In addition, we included four behavioral chronic conditions defined by CMS<sup>30</sup>  
174 and several other SBDH categories such as suicide ideation and compliance, which are represented in ICD-10  
175 but not in ICD-9. Altogether, this study focused on 15 SBDH categories (Table 1), although only nine  
176 categories were documented by ICD-9 billing codes in this data set (Supplementary Table S1 in Additional  
177 file 1).

Table 1: Performance of LSI predictions of SDBH categories. The terms in parentheses indicate the query word used to rank all patients in the dataset.

SBDH Category (Keyword Query)	ICD Coded N	Predicted N	PPV of LSI Predictions			
			Top 10	Median 10	Bottom 10	Average
Tobacco use disorder (Smokes)	3005	2195	100%	90%	80%	90%
Alcohol abuse (EtOH)	2988	1080	100%	100%	100%	100%
Drug abuse (Opiate)	672	444	100%	60%	50%	70%
Drug abuse (Cocaine)	545	1852	100%	70%	40%	70%
Housing insecurity (Homeless)	202	470	100%	80%	70%	83%
Physical/Sexual abuse (Abused)	37	121	80%	50%	30%	53%
Financial insecurity (Unemployed)	15	809	100%	90%	100%	97%
Legal Circumstances (Legal)	13	1052	80%	50%	20%	50%
Financial circumstances (Financial)	6	402	100%	60%	90%	83%
Compliance (Noncompliant)	0	402	100%	100%	90%	97%
Mobility issues (Walker)	0	3235	90%	100%	90%	93%
Lack of English proficiency (Interpreter)	0	1621	100%	90%	80%	90%
Caregiver dependency (Caretaker)	0	443	100%	90%	60%	83%
Suicidal ideation (Suicide)	0	1090	100%	60%	40%	67%
Lack of transportation (Transportation)	0	452	60%	70%	70%	67%

## 178 Latent Semantic Indexing and Lexicon Development

179 Latent Semantic Indexing (LSI) is a well-known matrix factorization method, which reduces the dimension-  
180 ality of terms and documents in to lower rank matrices.<sup>20-28</sup> By using a lower rank matrix, the terms can  
181 be grouped together more conceptually, whereas by using higher ranks, terms can be grouped more literally.  
182 In addition, patients can be grouped together in more conceptual or literal fashion based on the content in  
183 their clinical notes.

184 Out of a total of 46,520 patients in the MIMIC-III dataset, 46,146 patients had clinical notes. The number

185 of notes associated with these patients ranged from 1 to 1420, with the median being 21 notes. A patient  
186 document was constructed by concatenating all clinical notes together for each patient, which resulted in a  
187 term dictionary of >300,000 terms. To reduce the dictionary size to terms that are relevant to SBDH, we  
188 filtered the dictionary to include only terms that were extracted from social history sections, resulting in a  
189 final dictionary size of 26,237 terms. Each term in the 26,237 terms-by- 46,146 patients matrix was weighted  
190 using *tf-idf* and then factorized to 12,723 dimensions (see Additional file 1 for details).

191 To determine the best lexicons (terms) to represent various SBDH categories, we manually constructed a set  
192 of 134 keywords (including variants, plurals and common misspellings) corresponding to the SBDH categories  
193 described above (Supplementary Table S2 in Additional file 1). Both the SBDH categories and the lexicons  
194 were iteratively refined as described below based on: 1) The correlations between terms with respect to the  
195 vector of all ranked patients in the MIMIC-III dataset (Figure 2a), 2) the precision of the top ranked patients  
196 for the keyword query, 3) the recall of ICD-9 coded patients.

197 Clustering of the term correlations revealed groups of highly synonymous terms deduced from the word usage  
198 patterns in the patient documents. This demonstrates the utility of matrix factorization as an unsupervised  
199 machine learning approach which learns conceptually related terms based on the word usage patterns in the  
200 clinical notes. For example, factorization revealed that words such as intoxicated/intoxication, crack/cocaine,  
201 or manic/mania are synonymously used in the clinical notes (Figure 2b). In addition, this approach identified  
202 short phrases in a rudimentary way, such as legal/guardian (Figure 2b). Lastly, some of the larger clusters  
203 included broader contextual information, such as suicide/overdose/psych/suicidal/psychiatrist (Figure 2c).

## 204 **Evaluation of LSI-derived SBDH Predictions**

205 All patients in the collection were ranked based on a representative keyword query for each of the 15 SBDH  
206 categories. Application of interquartile outlier detection method determined the cosine threshold for each  
207 query where the patients ranked above the threshold ( $> Q3 + (3.0 * IQR)$ ) are highly associated with the  
208 query and thus predicted to have the specific SBDH. In all but three SBDH categories (Tobacco use, Alcohol  
209 abuse, and Drug abuse - Opiate), the number of patients in the collection with an LSI-predicted SBDH were  
210 substantially higher than the ICD-9 coded patients (Table 1).

211 To evaluate the classification performance of the SBDH predictions, we determined the PPV by manual  
212 evaluation of the top 10, median 10, and bottom 10 patients within the cut-off threshold (Table 1). In all  
213 but four SBDH categories, the PPV of the top 10 ranked patients was 100%. As expected, the PPV decreased



214 with lower rankings. The average PPV for all 15 SBDH categories ranged from 50% (legal circumstances)  
 215 to 100% (alcohol abuse), with nine of the SBDH categories having a PPV  $\geq$  83%.

216 Next, we compared the performance of ICD-9 coding with either LSI, GPT-3.5 or GPT-4 large language  
 217 models using different sets of gold standard (GS) patients that were randomly selected for each SBDH  
 218 category and manually labeled by chart review. The characteristics of the GS sets of patients for each SBDH  
 219 category are provided in Supplementary Table S3 in Additional file 1. Only nine SBDH categories that had  
 220 at least six ICD-9 coded patients were included in this analysis. Importantly, only LSI was able to process  
 221 all of the patient documents. In contrast, due to context window size restrictions, GPT-3.5 processed 55.6%  
 222 of the gold standard documents and GPT-4 processed 94.2% (Figure 3).

223 Earlier versions of GPT were highly irreproducible such that the same prompt could produce different  
 224 responses or no response at all. To evaluate this phenomenon, we compared the responsiveness of GPT-  
 225 3.5 and GPT-4 to the same set of shared documents within the 16K context window limit of GPT-3.5 for  
 226 each of the nine SBDH categories (Table 2). For GPT-3.5, the same set of documents were submitted  
 227 using the same prompt five independent times. GPT-3.5 was unresponsive for 2% (Cocaine use) to 30%  
 228 (unemployed) of the patient documents across the SBDH categories. In addition, in all but one SBDH  
 229 category, GPT-3.5 provided conflicting responses between the five independent prompts. For example,  
 230 although GPT3.5 provided responses for all 27 patient documents related to legal circumstances, it provided  
 231 conflicting responses for six (22%) of the patient documents (Table 2). In contrast, GPT-4 was unresponsive  
 232 for only two documents (3.8%) in only one SBDH category (tobacco use).

Table 2: Unresponsiveness of GPT-3.5 and GPT-4. 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 Category	N	GPT-3.5		GPT-4
		% 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 use	51	0.0%	2.0%	0.0%
Physical & sexual abuse	39	2.6%	5.1%	0.0%
Unemployed	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%

233 As expected, due to the limitations described above, the average recall of GPT-3.5 across all of the documents

234 in all nine SBDH categories was low (0.41), compared to LSI (0.70) and GPT-4 (0.77) (Table 3). Overall,  
235 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)  
236 despite the fact that GPT-4 was unable to process 5.8% of the documents due to context window size  
237 limitations (Figure 3 & Table 3).

238 It is important to note that in some cases, although a patient was assigned an ICD-9 code for a particular  
239 SBDH, supporting documentation in the clinical notes could not be found. In such cases, the ICD-9 coded  
240 individuals were assumed to be actual positives. Therefore, to retrieve all possible SBDH in a given GS  
241 set, the text-based prediction of SBDH was combined with ICD coded individuals across the nine SBDH  
242 categories. On average, all three methods performed similarly with respect to precision, recall and F1 when  
243 combined with ICD-9 (Figure 4).

244 Lastly, to evaluate the overall predictive performance of LSI with GPT-4 when combined with ICD-9 coding,  
245 we compared the prediction AUC of three different logistic regression models: 1) base model including  
246 gender, age, race and SBDH ICD-9 codes, 2) base model plus LSI identified SBDH, 3) base model plus  
247 GPT-4 identified SBDH (Figure 5). Using only ICD-9 coding (base model), the AUCs for the nine SBDH  
248 categories ranged between 0.69 (*housing insecurity* and *financial circumstances*) to 0.85 (*history of physical*  
249 *and sexual abuse*). In all nine categories, inclusion of LSI or GPT-4 improved the AUCs compared to  
250 ICD-9. Interestingly, LSI outperformed GPT-4 in six of the nine SBDH categories (*housing insecurity*,  
251 *unemployment*, *opiate abuse*, *alcohol abuse*, *legal circumstances*, and *financial circumstances*).

Table 3: Retrieval performance of each method alone using a set of sampled Gold Standard cases. The bold text indicate the highest precision, recall, or F1 for each SBDH category (row).

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

## 252 Discussion

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

258 One major advantage of LSI is its ability to process all of the notes for a given patient without the imposed  
259 context window token size limitations of GPT. As pointed out in Figure 3, only 55.6% and 94.2% of the  
260 GS cases could be processed by GPT-3.5 and GPT-4, respectively. At the time of our analysis, the input  
261 context window size limits for GPT-3.5 and GPT-4 were 16K and 128K tokens, respectively. However,  
262 other LLMs may have larger context windows. Even with the context window limits, it is possible to  
263 process larger documents by ‘chunking’, a method where a large document is split into smaller overlapping  
264 documents that are smaller than the token limits. In our analysis, we did not attempt to process all of the  
265 GS documents, instead we directly compared the performance of LSI with GPT-3.5 and GPT-4 using the  
266 same set of documents (Table 3 and Figures 4 & 5). Another reason for limiting the analysis to a subset of  
267 GS documents was cost. At the time of the analysis, the cost for GPT-3.5 and GPT-4 using the Microsoft  
268 Azure OpenAI<sup>31</sup> services per query was USD \$0.001 and \$0.01 per 1K input tokens, respectively. Thus,  
269 it would have been more costly to chunk the larger GS documents. Another way to reduce the number of  
270 GPT queries would have been to perform multi-class labeling. In our analysis, we performed single class  
271 labeling, where each document was processed individually to identify a single SBDH category. Although this  
272 approach would be useful, it may require considerable fine-tuning and may not be feasible for identifying all  
273 15 SBDH categories at once.

274 Another major advantage of LSI is that it does not require external training on a large dataset and fine-  
275 tuning for domain specific applications. For this study, the LSI model was built using all of the clinical  
276 notes for all of the > 46,000 patients at once. In contrast, GPT and other LLM require extensive training  
277 using large amounts of external data sources. For example, GPT 3.5 was trained on 175 billion parameters  
278 using training data up to September 2021. Although the models perform well for general text analysis,  
279 they may not perform well on specialized clinical tasks. For example, Lybarger et al. developed an event  
280 based deep-learning extractor for SBDH that determines chronicity, duration, frequency and type of event.<sup>12</sup>  
281 However, their models apply only to a subset of SBDH categories, including employment, living status, as

282 well as alcohol, tobacco and drug use. They point out that training these models required significant manual  
283 effort by human experts to develop both positive and negative gold standard datasets for fine-tuning.<sup>12</sup> In  
284 addition, since these methods require large amounts of training data for fine-tuning, they can have limited  
285 usefulness for SBDH categories that are rare (low prevalence).

286 Another major advantage of LSI is that, unlike GPT, it is deterministic (reproducible) and 100% responsive  
287 to all queries. For a given factorization rank, LSI produces the same exact ranking of the documents based  
288 on the same query. On the other hand, we showed (Table 2) that GPT-3.5 produces conflicting responses to  
289 the same prompt on the same set of documents. Moreover, we demonstrated that both GPT-3.5 and GPT-4  
290 may not respond, a phenomenon commonly referred to as ‘laziness’. Although the GPT-4 model has been  
291 improved to reduce laziness, we found that it can be unresponsive as the document size reaches its maximum  
292 context window size limits.

293 Our findings indicate that using clinical notes to identify SBDH should not replace efforts in health systems  
294 to screen for SBDH, rather provide a complementary approach to enhance estimates of the SBDH burden  
295 (prevalence) in large populations. During chart review for developing the GS sets, we found a few ICD-9  
296 coded individuals who had no supporting documentation for the codes. For example, some patients had  
297 few encounters with the health system and had no social history notes, yet were coded for homelessness or  
298 alcohol abuse. As reported by others, this observation illustrates the importance of combining the information  
299 provided by ICD-9 codes and other structured data (e.g., questionnaires) with unstructured data in the EHR  
300 to obtain a more representative assessment of the SBDH prevalence in a population.<sup>7,10–12,32</sup> On the other  
301 hand, implementing SDoH screening tools across a large health system is impractical and potentially biased.  
302 Studies have shown that SDoH screening forms are primarily implemented in outpatient and primary care  
303 settings. However, it is thought that socioeconomically disadvantaged individuals are less likely to go to  
304 primary care, instead use the emergency department (ED) for their healthcare needs.<sup>33</sup> Moreover, a recent  
305 study demonstrated that only 3.7% of the patients in a large health care system in South Carolina had  
306 answered all 11 questions on the SDoH screening forms.<sup>34</sup> Therefore, for better assessment of SBDH burden  
307 in a population, information must be aggregated from a variety of sources in the EHR, including the clinical  
308 notes.

309 It is worth highlighting that the costs associated with OpenAI services make it currently unrealistic to  
310 implement in health systems to assess SBDH burden in large populations of patients. To address this issue,  
311 future research will focus on using LSI to narrow large populations of patients into smaller groups that

312 are conceptually predicted to have SBDH and then process those documents using GPT to contextualize  
313 and validate the LSI predictions. Factorization provides value beyond keyword searching alone because it  
314 contextualizes keywords as vectors in reduced ranked space, thereby grouping words that are frequently used  
315 together in the context of SBDH keywords. This approach provides a general advantage by automatically  
316 grouping synonyms, misspellings, and conceptually related terms that are often used together in narratives  
317 (Figure 2). For example, a homeless individual is often unemployed and has drug/alcohol abuse problems.  
318 Also, factorization is able to infer that ‘shelter’ and ‘homelessness’ are synonymously used in the narratives.  
319 By lowering the rank of the factorized matrix, one can identify a subset of patients who are conceptually  
320 related to the SBDH, achieving higher recall than precision. By subsequently processing these patient  
321 documents with GPT-4, the specific evidence in support of the SBDH can be readily deduced while keeping  
322 the overall processing cost low.

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

335 Feller et al. were among the first groups to apply NLP methods to infer SBDH from clinical notes. After  
336 feature selection, they included 2-4,000 individual words as independent variables in various machine learning  
337 classifiers to identify sexual history, sexual orientation, alcohol use, substance use and housing status. They  
338 found that combining clinical notes and structured data enabled reasonably accurate inference of these SBDH  
339 categories.<sup>35,36</sup> Bejan et al., using a vector embedding approach to expand SDoH lexicons, demonstrated  
340 better performance of identification of homelessness and adverse childhood experiences (ACEs) from clinical  
341 notes.<sup>37</sup> Our process, which combines the bag-of-words approach with factorization for embedding, allows

342 an automated method to identify a broad set of SBDH categories.

343 The LSI approach has several limitations. First, it is a bag-of-words embedding technique, which does not  
344 account for word context (phrases) and negated terms. In addition, the performance of our approach was  
345 affected by the presence of forms and templated text in the clinical notes, such as “Family Information” or  
346 social history forms, where there are many negations and repeated text. The performance of our approach  
347 would improve if certain note types, forms and templates were removed during pre-processing. Lastly, our  
348 approach does not provide temporal relations and event-types. As stated above, many of these limitations  
349 would be addressed by combining the advantages of LSI (e.g., robustness, determinism, and no cost) with  
350 the advantages of LLM (i.e., contextualization, removal of negation, and multi-label classification).

## 351 **Conclusions**

352 In this study, we demonstrated that using an unsupervised machine learning factorization approach on  
353 clinical notes is a robust way to enhance SBDH identification from the EHR. This work is significant because  
354 it provides an automated way to extract SBDH for patients in a health system without the additional burden  
355 of implementing standardized surveys in clinical workflows. By providing better estimates of SBDH burden  
356 in populations, this work sets the stage for developing patient level health risk and utilization prediction  
357 models that incorporate SBDH factors in addition to standard clinical and structured data from the EHR.

## 358 **Declarations**

### 359 **Ethics approval and consent to participate**

360 Not applicable.

### 361 **Consent for publication**

362 Not applicable.

### 363 **Availability of data and materials**

364 The MIMIC-III dataset is available publicly through [physionet.org](https://physionet.org).

365 **Competing interests**

366 RH & SM hold equity in Quire Inc.

367 **Funding**

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

370 **Authors' contributions**

371 SR designed and implemented the methods, generated data, interpreted results and contributed to writing  
372 of the manuscript. SM generated data and performed analysis. LZ analyzed data, interpreted results and  
373 contributed to writing of the manuscript. RH designed the study, interpreted results, performed chart reviews  
374 and wrote the manuscript.

375 **Acknowledgements**

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

379 **Figure titles and legends**



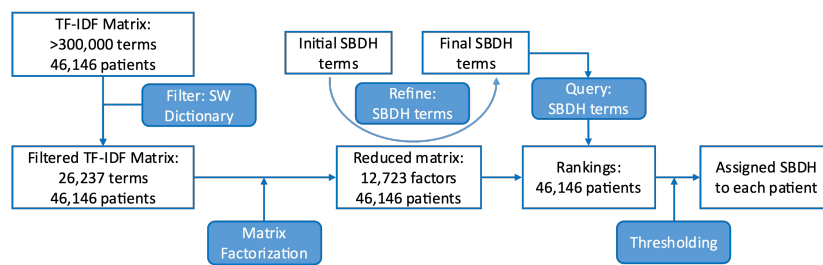
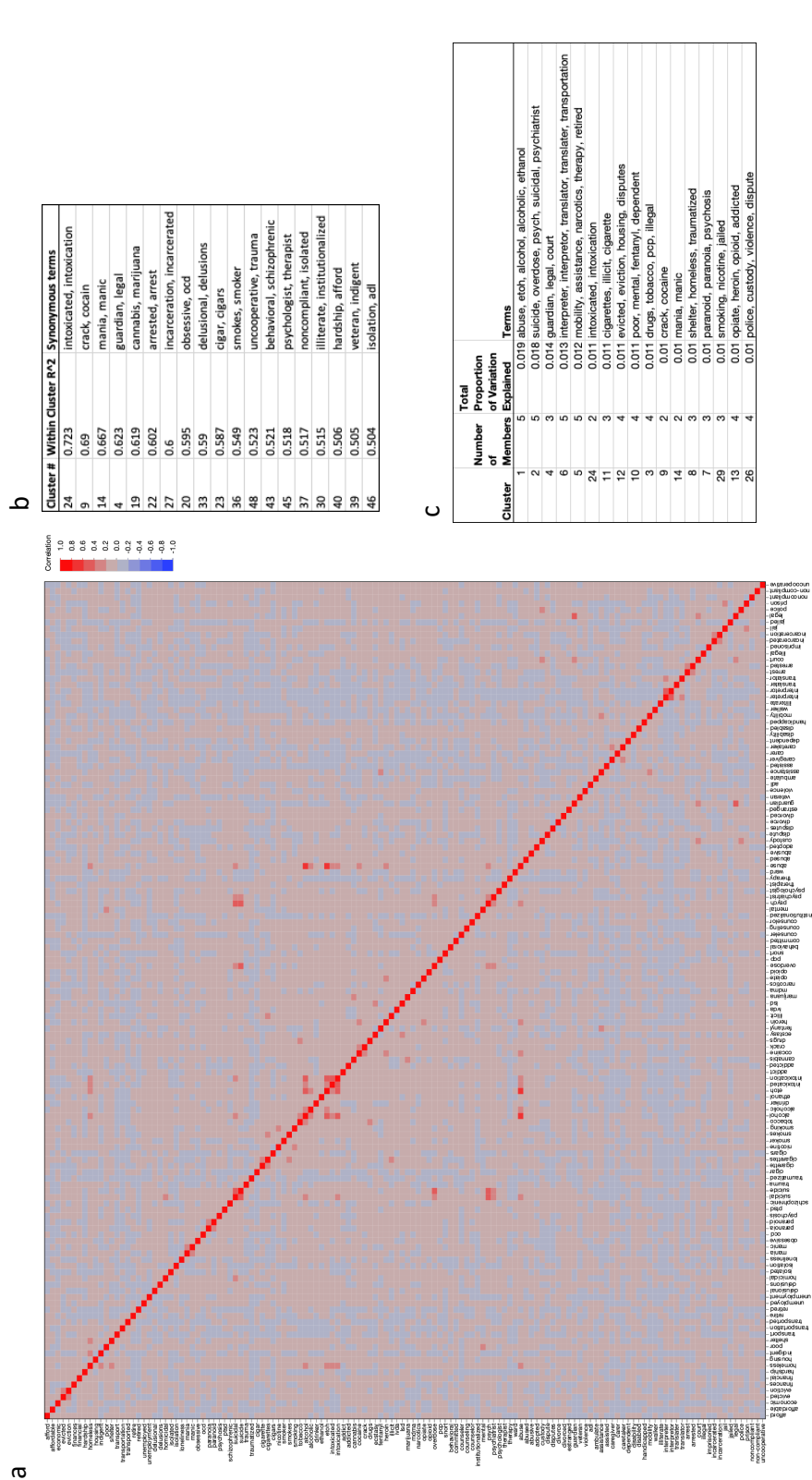


Figure 1: Workflow diagram of extracting and assigning SBDH factors to each patient in MIMIC-III dataset.



**Figure 2: Relationship between SBDH terms in reduced-rank (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.

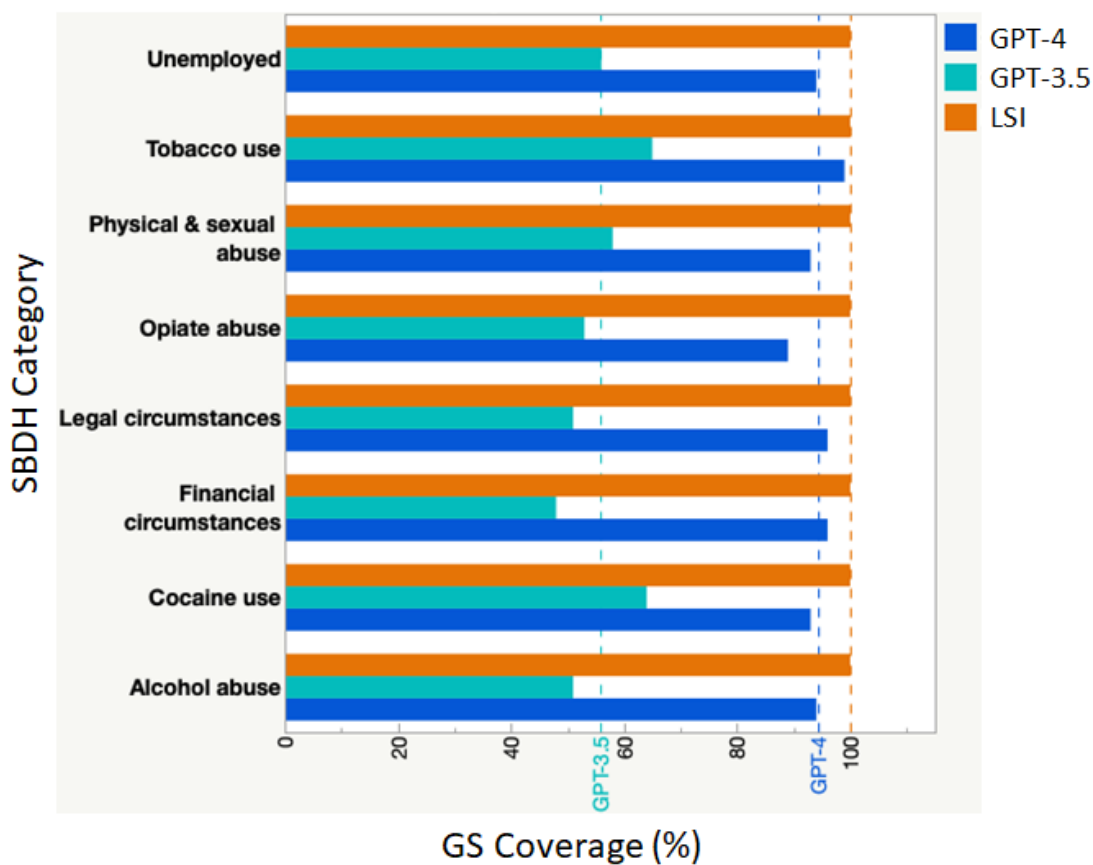


Figure 3: Proportion of gold standard patient documents for each SBDH category that yielded results by LSI, GPT-3.5 or GPT-4.0.

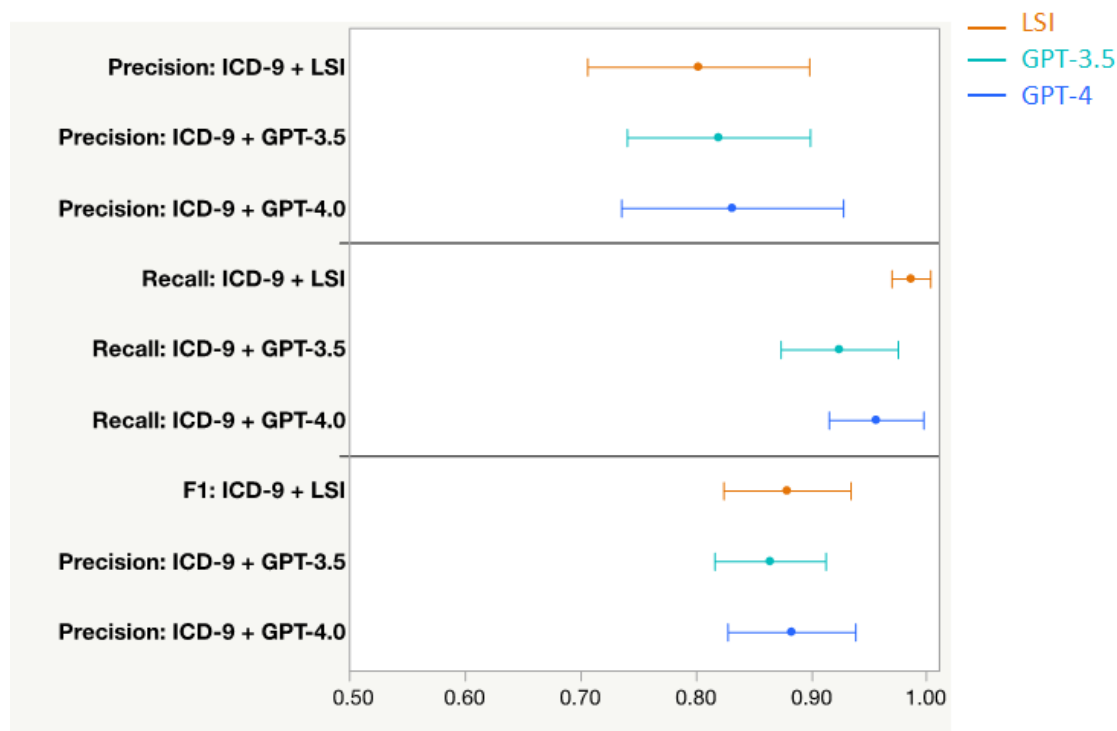


Figure 4: **Retrieval performance of LSI, GPT-3.5 or GPT-4 when combined with ICD-9 coding.** Precision (upper panel), recall (middle panel) and F1 (lower panel) of ICD-9 combined with either LSI (orange lines), GPT-3.5 (cyan lines) and GPT-4 (blue lines). Values represent the mean (filled circle) and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

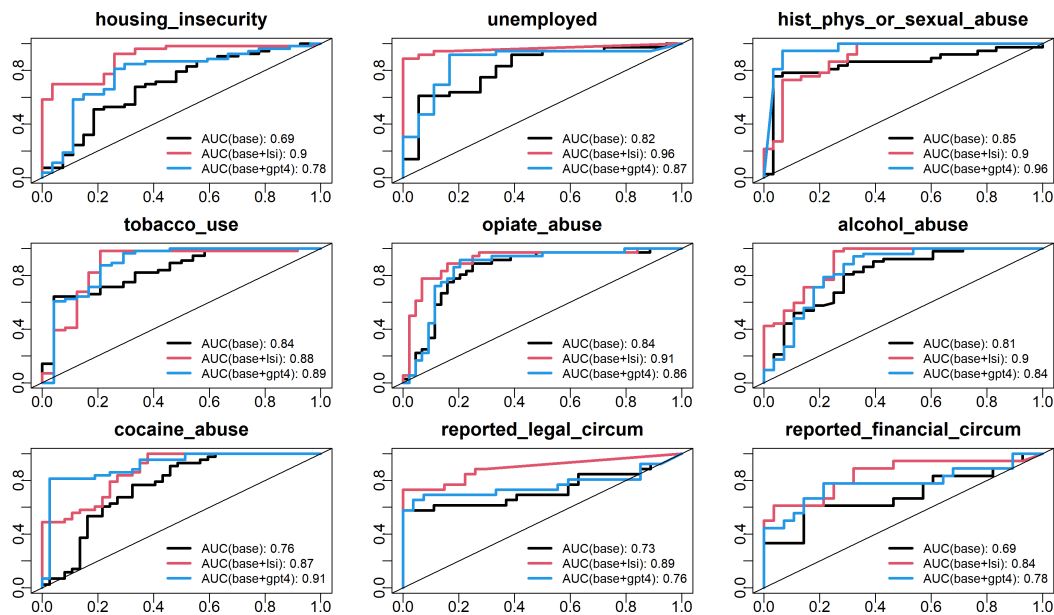


Figure 5: Comparison of classification performance of ICD-9 and/or text-predicted SBDH categories using multivariable analysis. The AUC is shown for three different models: 1) Base model including age, gender and ICD-9 codes (black lines), 2) Base model plus LSI identified SBDH (red lines), and 3) Base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

## 380 REFERENCES

- 381 [1] Chen, M., Tan, X., and Padman, R. Social determinants of health in electronic health records and  
382 their impact on analysis and risk prediction: A systematic review. *Journal of the American Medical*  
383 *Informatics Association*, 27(11):1764–1773, November 2020. ISSN 1527974X. doi: 10.1093/jamia/  
384 ocaa143.
- 385 [2] Tan, M., Hatef, E., Taghipour, D., et al. Including social and behavioral determinants in predictive  
386 models: Trends, challenges, and opportunities. *JMIR Medical Informatics*, 8(9), September 2020. ISSN  
387 22919694. doi: 10.2196/18084.
- 388 [3] Guo, Y., Chen, Z., Xu, K., et al. International classification of diseases, tenth revision, clinical modifica-  
389 tion social determinants of health codes are poorly used in electronic health records. *Medicine (United*  
390 *States)*, 99(52), December 2020. ISSN 15365964. doi: 10.1097/MD.00000000000023818.
- 391 [4] Andermann, A. Screening for social determinants of health in clinical care: Moving from the margins to  
392 the mainstream. *Public Health Reviews*, 39(1), 2018. ISSN 21076952. doi: 10.1186/s40985-018-0094-7.
- 393 [5] Alpert, J., Kim, H., McDonnell, C., et al. Barriers and facilitators of obtaining social determinants of  
394 health of patients with cancer through the electronic health record using natural language processing  
395 technology: Qualitative feasibility study with stakeholder interviews. *JMIR formative research*, 6(12),  
396 December 2022. ISSN 2561-326X. doi: 10.2196/43059.
- 397 [6] Navathe, A.S., Zhong, F., Lei, V.J., et al. Hospital readmission and social risk factors identified from  
398 physician notes. *Health Services Research*, 53(2):1110–1136, April 2018. ISSN 14756773. doi: 10.1111/  
399 1475-6773.12670.
- 400 [7] Hatef, E., Rouhizadeh, M., Tia, I., et al. Assessing the availability of data on social and behavioral  
401 determinants in structured and unstructured electronic health records: A retrospective analysis of a  
402 multilevel health care system. *Journal of Medical Internet Research*, 21(8), 2019. ISSN 14388871. doi:  
403 10.2196/13802.
- 404 [8] Patra, B.G., Sharma, M.M., Vekaria, V., et al. Extracting social determinants of health from electronic  
405 health records using natural language processing: a systematic review. *Journal of the American Medical*  
406 *Informatics Association*, 28(12):2716–2727, December 2021. ISSN 1527-974X. doi: 10.1093/JAMIA/  
407 OCAB170.

- 408 [9] Lybarger, K., Bear, O.J., Yetisgen, M., et al. Advancements in extracting social determinants of health  
409 information from narrative text. *Journal of the American Medical Informatics Association*, 30(8):1363–  
410 1366, July 2023. ISSN 1527974X. doi: 10.1093/JAMIA/OCAD121.
- 411 [10] Allen, K.S., Hood, D.R., Cummins, J., et al. Natural language processing-driven state machines to  
412 extract social factors from unstructured clinical documentation. *JAMIA open*, 6(2), July 2023. ISSN  
413 2574-2531. doi: 10.1093/JAMIAOPEN/OOAD024.
- 414 [11] Mehta, S., Lyles, C., Rubinsky, A., et al. Social determinants of health documentation in structured and  
415 unstructured clinical data of patients with diabetes: Comparative analysis. *JMIR medical informatics*,  
416 11, January 2023. ISSN 2291-9694. doi: 10.2196/46159.
- 417 [12] Lybarger, K., Dobbins, N.J., Long, R., et al. Leveraging natural language processing to augment  
418 structured social determinants of health data in the electronic health record. *Journal of the American  
419 Medical Informatics Association*, 30(8):1389–1397, July 2023. ISSN 1527-974X. doi: 10.1093/JAMIA/  
420 OCAD073.
- 421 [13] Lybarger, K., Ostendorf, M., and Yetisgen, M. Annotating social determinants of health using ac-  
422 tive learning, and characterizing determinants using neural event extraction. *Journal of Biomedical  
423 Informatics*, 113(April 2020):103631, 2021. ISSN 15320464. doi: 10.1016/j.jbi.2020.103631.
- 424 [14] Yu, Z., Yang, X., Dang, C., et al. A study of social and behavioral determinants of health in lung  
425 cancer patients using transformers-based natural language processing models. *AMIA Annual Symposium  
426 Proceedings*, 2021:1225, 2021. ISSN 1942597X.
- 427 [15] Yu, Z., Yang, X., Guo, Y., et al. Assessing the documentation of social determinants of health for lung  
428 cancer patients in clinical narratives. *Frontiers in public health*, 10, March 2022. ISSN 2296-2565. doi:  
429 10.3389/FPUBH.2022.778463.
- 430 [16] Guevara, M., Chen, S., Thomas, S., et al. Large language models to identify social determinants of  
431 health in electronic health records. *NPJ digital medicine*, 7(1):6, 2024.
- 432 [17] Johnson, A.E., Pollard, T.J., Shen, L., et al. MIMIC-III, a freely accessible critical care database. *Scientific  
433 data*, 3(1):1–9, 2016.
- 434 [18] Zeimpekis, D. and Gallopoulos, E. Tmg: A matlab toolbox for generating term-document matrices  
435 from text collections. In *Grouping multidimensional data*, pages 187–210. Springer, 2006.

- 436 [19] Salton, G. The smart document retrieval project. In *Proceedings of the 14th annual international ACM*  
437 *SIGIR conference on Research and development in information retrieval*, pages 356–358. ACM, 1991.
- 438 [20] Homayouni, R., Heinrich, K., Wei, L., et al. Gene clustering by latent semantic indexing of medline  
439 abstracts. *Bioinformatics*, 21(1):104–115, 2005.
- 440 [21] Heinrich, K.E., Berry, M.W., Homayouni, R., et al. Gene tree labeling using nonnegative matrix  
441 factorization on biomedical literature. *Computational intelligence and neuroscience*, 2008.
- 442 [22] Roy, S., Heinrich, K., Phan, V., et al. Latent semantic indexing of pubmed abstracts for identification of  
443 transcription factor candidates from microarray derived gene sets. In *BMC bioinformatics*, volume 12,  
444 pages 1–13. Springer, 2011.
- 445 [23] Roy, S., Homayouni, R., Berry, M.W., et al. Nonnegative tensor factorization of biomedical literature  
446 for analysis of genomic data. In *Data Mining for Service*, pages 97–110. Springer, 2014.
- 447 [24] Roy, S., Curry, B.C., Madahian, B., et al. Prioritization, clustering and functional annotation of  
448 micrnas using latent semantic indexing of medline abstracts. In *BMC bioinformatics*, volume 17,  
449 pages 131–142. BioMed Central, 2016.
- 450 [25] Roy, S., Yun, D., Madahian, B., et al. Navigating the functional landscape of transcription factors  
451 via non-negative tensor factorization analysis of medline abstracts. *Frontiers in Bioengineering and*  
452 *Biotechnology*, 5:48, 2017.
- 453 [26] Roy, S. and Berry, M.W. Mining multimodal big data: Tensor methods and applications. In *Handbook*  
454 *of Research on Big Data Storage and Visualization Techniques*, pages 674–702. IGI Global, 2018.
- 455 [27] Roy, S., Zaman, K.I., Williams, R.W., et al. Evaluation of sirtuin-3 probe quality and co-expressed  
456 genes using literature cohesion. *BMC bioinformatics*, 20:31–43, 2019.
- 457 [28] Akbilgic, O., Homayouni, R., Heinrich, K., et al. Unstructured text in emr improves prediction of death  
458 after surgery in children. *Informatics*, 6(1), 2019. ISSN 22279709. doi: 10.3390/informatics6010004.
- 459 [29] Torres, J.M., Lawlor, J., Colvin, J.D., et al. Icd social codes: An underutilized resource for tracking  
460 social needs. *Medical Care*, 55(9):810–816, 2017. ISSN 15371948. doi: 10.1097/MLR.0000000000000764.
- 461 [30] CMS. Chronic conditions data warehouse. URL <https://www2.ccwdata.org/web/guest/home/>.



- 462 [31] Microsoft. Microsoft azure openai. URL [https://learn.microsoft.com/en-us/legal/cognitive-](https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy)  
463 [services/openai/data-privacy](https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy).
- 464 [32] Harle, C.A., Wu, W., and Vest, J.R. Accuracy of electronic health record food insecurity, housing  
465 instability, and financial strain screening in adult primary care. *JAMA*, 329(5):423–424, February 2023.  
466 ISSN 1538-3598. doi: 10.1001/JAMA.2022.23631.
- 467 [33] Capp, R., Camp-Binford, M., Sobolewski, S., et al. Do adult medicaid enrollees prefer going to their  
468 primary care provider’s clinic rather than emergency department (ed) for low acuity conditions? *Medical*  
469 *care*, 53(6):530, 2015.
- 470 [34] Rudisill, A.C., Eicken, M.G., Gupta, D., et al. Patient and care team perspectives on social de-  
471 terminants of health screening in primary care: A qualitative study. *JAMA Network Open*, 6(11):  
472 e2345444–e2345444, 2023.
- 473 [35] Feller, D.J., Bear, O.J., Zucker, J., et al. Detecting social and behavioral determinants of health with  
474 structured and free-text clinical data. *Applied Clinical Informatics*, 11(1):172–181, 2020. ISSN 18690327.  
475 doi: 10.1055/s-0040-1702214.
- 476 [36] Feller, D.J., Zucker, J., Yin, M.T., et al. Using clinical notes and natural language processing for  
477 automated hiv risk assessment. *Journal of acquired immune deficiency syndromes (1999)*, 77(2):160–166,  
478 2018. ISSN 19447884. doi: 10.1097/QAI.0000000000001580.
- 479 [37] Bejan, C.A., Angiolillo, J., Conway, D., et al. Mining 100 million notes to find homelessness and adverse  
480 childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health  
481 records. *Journal of the American Medical Informatics Association*, 25(1):61–71, 2018. ISSN 1527974X.  
482 doi: 10.1093/jamia/ocx059.