


---

## Research and Applications

# Identification of social determinants of health using multi-label classification of electronic health record clinical notes

Rachel Stemerman <sup>1</sup>, Jaime Arguello,<sup>2</sup> Jane Brice,<sup>3</sup> Ashok Krishnamurthy,<sup>4</sup> Mary Houston,<sup>3</sup> and Rebecca Kitzmiller<sup>5</sup>

<sup>1</sup>Carolina Health Informatics Program, The University of North Carolina, Chapel Hill, North Carolina, USA, <sup>2</sup>School of Information and Library Sciences, The University of North Carolina, Chapel Hill, North Carolina, USA, <sup>3</sup>Department of Emergency Medicine, The University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA, <sup>4</sup>Department of Computer Science, The University of North Carolina, Chapel Hill, North Carolina, USA and <sup>5</sup>School of Nursing, The University of North Carolina, Chapel Hill, North Carolina, USA

Corresponding Author: Rachel Stemerman, PhD, MPH, NRP, Carolina Health Informatics Program, The University of North Carolina, Chapel Hill, NC, USA; rstemerman@gmail.com

Received 16 September 2020; Revised 16 November 2020; Accepted 20 November 2020

### ABSTRACT

**Objectives:** Social determinants of health (SDH), key contributors to health, are rarely systematically measured and collected in the electronic health record (EHR). We investigate how to leverage clinical notes using novel applications of multi-label learning (MLL) to classify SDH in mental health and substance use disorder patients who frequent the emergency department.

**Methods and Materials:** We labeled a gold-standard corpus of EHR clinical note sentences ( $N=4063$ ) with 6 identified SDH-related domains recommended by the Institute of Medicine for inclusion in the EHR. We then trained 5 classification models: linear-Support Vector Machine, K-Nearest Neighbors, Random Forest, XGBoost, and bidirectional Long Short-Term Memory (BI-LSTM). We adopted 5 common evaluation measures: accuracy, average precision–recall (AP), area under the curve receiver operating characteristic (AUC-ROC), Hamming loss, and log loss to compare the performance of different methods for MLL classification using the F1 score as the primary evaluation metric.

**Results:** Our results suggested that, overall, BI-LSTM outperformed the other classification models in terms of AUC-ROC (93.9), AP (0.76), and Hamming loss (0.12). The AUC-ROC values of MLL models of SDH related domains varied between (0.59–1.0). We found that 44.6% of our study population ( $N=1119$ ) had at least one positive documentation of SDH.

**Discussion and Conclusion:** The proposed approach of training an MLL model on an SDH rich data source can produce a high performing classifier using only unstructured clinical notes. We also provide evidence that model performance is associated with lexical diversity by health professionals and the auto-generation of clinical note sentences to document SDH.

**Key words:** social determinants of health, electronic health records, machine learning, natural language processing

---

### LAY SUMMARY

Health is influenced by many factors, of which one are social determinants of health (SDH) that encompasses economic and social conditions that influence the health of people and communities. Addressing SDH is a primary approach to achieving health equity and a growing body of research highlights the importance of integrating these factors into clinical practice. However, SDH are rarely systematically documented and collected in the electronic health record (EHR). This lack of information is a problem; health professionals can only intervene on SDH if they are aware, document, and communicate these issues throughout the health system. Recent studies have shown promising results using natural language processing to identify patient cohorts and conditions within EHR clinical notes. SDH rarely occur in isolation and often are interconnected. For example, employment is linked to health insurance in the United States resulting in unemployment and lack of insurance. Therefore we are studying the feasibility of identifying and classifying multiple SDH characteristics within a single sentence from a large collection of EHR clinical notes.

## INTRODUCTION

Emergency departments (EDs) are often called the “safety net” of the US healthcare system. Patients with poor mental health, depression, and high ratings of psychological distress have greater odds of being frequent ED users.<sup>1</sup> As a result, this population makes up 1 in 8 ED visits contributing to ED overcrowding, suboptimal quality of care, and represents 1 of the 5 most costly conditions in the United States with expenditures at \$57.5 billion annually.<sup>2,3</sup> Social, psychological, and behavioral factors, or social determinants of health (SDH), key contributors to health, are rarely captured and measured in a systematic way in health care settings.<sup>4–6</sup> For example, housing insecurity is associated with poor health including chronic diseases, substance abuse, and frequent ED visits.<sup>7</sup> Ku et al.<sup>8</sup> found that frequent ED patients expressed a variety of other social needs including the inability to meet essential expenses, having a telephone service disconnected, worrying about running out of food, and inability to afford a balanced meal. These related disparities have a direct link to health and often result from overlapping factors.<sup>9,10</sup> Okin et al. examined interventions to reduce frequent ED visits and found that case management had the most rigorous evidence base and yielded moderate cost savings<sup>11</sup> with the greatest reduction in median per-patient hospital costs was \$7473.<sup>12</sup> The authors attributed most of the cost savings to addressing housing instability and long-term substance abuse care coordination.<sup>12</sup> In order to reduce negative health outcomes associated with SDH, health professionals will require information about their patients’ individual SDH characteristics to better address their needs. For example, while ED physicians may attribute a patient’s frequent visits for depression to poorly managed mental health issues because of a patient’s unwillingness to follow up with specialists (ie, willful noncompliance), frequent visits might instead be caused by lack of transportation or financial constraints. Thus, medical treatment of a disease such as depression, without regard to the SDH, suffers the danger of being ineffective. Just as fluid volume overload cannot be treated without first understanding the physiology of the kidney, heart, lungs, and their interaction, a patient’s mental health and substance use disorders (MHSUDs) treatment will be substandard without understanding associated SDH. In 2014, the National Library of Medicine underscored the importance of capturing these SDH in electronic health records (EHRs) to improve clinical care.<sup>6,13</sup> However, SDH are rarely captured and measured in a systematic way in EHRs<sup>5,6</sup> and, therefore, remain largely unused in care decision-making.

Although reducing avoidable ED visits is a primary health system goal<sup>14,15</sup>; however, systematically identifying contributors to patients’ frequent ED use is challenging, particularly in the emer-

gency care setting.<sup>11,16</sup> EHR documentation of SDH needs and SDH services delivered are captured in both structured (eg, procedure codes) and unstructured data (free-text).<sup>17,18</sup> Unfortunately, structured data (ie, administrative codes) fail to capture the breadth of SDH characteristics while methods for extracting a patients’ complete SDH history from clinical text are less well developed.<sup>19,20</sup> Hybrid techniques that combine natural language processing (NLP) and machine learning (ML) are the most common biomedical approach to extracting clinical text.<sup>21</sup> Various studies effectively applied NLP approaches, including information extraction techniques, to different types of SDH classification including homelessness,<sup>19,22,23</sup> employment status,<sup>23,24</sup> and exposure to violence.<sup>19,25</sup> These techniques included regular expressions, named-entity recognition (NER), and distributional semantic. Patients with SDH, such as someone who has lost their job (employment insecurity), frequently experience several SDH in relation to the job loss such as the loss of health insurance associated with employment. Furthermore, healthcare domain text data is characterized by long sentences with a large number of technical words and typos/misspellings.<sup>26</sup> New approaches in ML, such as multi-label learning (MLL) may be a viable candidate for modeling the profile of patients affected by several SDH. MLL differs from classical ML by tackling the learning problem from a different perspective. In contrast to traditional classification tasks where each observation belongs to only one mutually exclusive class, in MLL decision areas of labels (ie, classes) overlap. Binary relevance, a traditional approach to solving the multi-label text classification problem, decomposes the problem into multiple independent binary classification tasks (1 for each label).<sup>27,28</sup> A review of MLL algorithms can be found in Min-Ling and Zhi-Hua.<sup>28</sup>

In this article, we investigate novel applications of MLL to classify financial resource strain and poor social support from clinical note data for MHSUD patients who frequent the ED. We assess the feasibility of developing a model to classify SDH using only clinical notes. We then evaluate the performance of 5 approaches to classification: a linear Support Vector Machine (SVM)-baseline, K-Nearest Neighbors (KNN), Random Forest (RF), XGBoost, and bidirectional Long Short-Term Memory (Bi-LSTM). Finally, we develop a multi-label setting (up to 6 labels per instance) and apply the model to single sentences, the most granular level of clinical notes. We rely on clinical notes from a large academic health system to validate our experiments with a gold-standard corpus and highlight the elements in the sentence that explain and support the predicted labels to promote transparency. While research exists for each individual SDH characteristic in our model,<sup>19,23,29</sup> we believe we are the first to tackle multi-labeling in the clinical domain. Our results demonstrate

the feasibility of developing ML models to classify clinical note sentences with multiple SDH labels with XGBoost, SVM, and Bi-LSTM yielding the most promising results.

## MATERIALS AND METHODS

### Setting and sample

Clinical notes were obtained from the clinical data warehouse at the University of North Carolina Health System, a large academic medical center serving much of North Carolina and the surrounding regions. Clinical notes were collected from April 2014 to December 2019, a time period that encompassed the health system's transition to a single EHR. Clinical notes that met the following inclusion criteria were retained: (1) visited University of North Carolina at Chapel Hill Emergency Department (UNC-CH ED) between 2014 and 2019. Patients who had less than 4 ED visits in the year 2017 or 2018 within a rolling 365-day period were excluded, (2) greater than 18 years old in the CDW-H as of 2014, and (3) documented MHSUD "final primary diagnosis" as defined by the International Classification of Diseases and Related Health Problems 10th Revision (ICD-10 CM) code F00-F99 "mental and behavioral disorders". The study was approved by the University of North Carolina's Institutional Review Board.

### Curation of social determinants of health

We created a gold-standard corpus of clinical notes containing information SDH characteristics from MHSUD patients who frequent the ED. Sentences with a high likelihood of SDH characteristics were identified through an SDH dictionary that was developed by training 2 word embedding models (unigram and bigram) using seed terms abstracted from published research studies. These models detected and identified semantically similar terms to characterize financial resource strain and poor social support, yielding 109 terms or phrases ([Supplementary Appendix 1](#)). In this study, we focused on multiple SDH characteristic classification of financial resource strain and poor social support. The selected labels included (1) housing insecurity (homelessness, unstable housing), (2) food insecurity (food stamps, unable to afford food), (3) employment and income insecurity (unemployment, insufficient income), (4) general financial insecurity (lack of transportation, other financial issues), (5) insurance insecurity (uninsured, underinsured), and (6) poor social support (social isolation, lack of social support) as guided by the IOM's "Capturing Social and Behavioral Domains and Measure in Electronic Health Records".<sup>30</sup>

### Gold-standard corpora purposeful sampling

Purposeful sampling is widely used in qualitative research to identify and select information-rich examples related to the target of interest.<sup>31</sup> In contrast, probabilistic or random sampling is used to ensure the generalizability of findings by minimizing the potential for bias in selection. In this study, we developed a data-level hybrid approach to address our imbalanced dataset.

The frequency of redundant text in clinical notes, created by copy and paste or auto-generation, undermines machine learning training and evaluation due to over representation of an SDH single occurrence.<sup>32</sup> To derive an unbiased estimate of likely SDH documentation, we removed auto-generated and copy and paste entries that appeared to duplicate sentences. We removed sentences that were exactly the same as another sentence within an individual patient's clinical record regardless of the time period between the

occurrences of these entries. We then isolated 2 corpora (1 unigram, 1 bigram) with a combined 1 596 166 sentences with likely SDH documentation based on dictionary of terms and phrases developed through an SME driven word embedding expansion approach. We then took a randomized sampling of 150–200 sentences from each SDH class pre-labeled by their associated dictionary term. A randomized sampling of negative sentences (ie, lacking an SDH term) were added to the dataset to adjust for over representation of SDH in the dataset. No duplicates were found between the 2 corpora that were then annotated and combined for model training. This newly formed dataset was used by annotators to complete the annotation process and create a gold-standard corpus.

### Gold-standard annotation guidelines

To produce higher quality SDH analysis and downstream applications, we chose to obtain sentence-level annotations rather than document-level annotations because we wanted to evaluate the feasibility of classifying SDH on a granular-level. For example, we observed mentions of SDH, such as "patient's current stressors include: unemployment, homelessness and recent relapse on illicit substances" and "patient reports that he lost his job in June, lost his girlfriend and then lost his home," that would not be amenable to extraction by document or NER.

Two annotators manually reviewed extracted clinical note sentences to classify documentation using 6 SDH characteristic categories described earlier. A third annotator adjudicated disagreements to determine the final classification. The annotators represented an interdisciplinary group of health professionals that serve study population: a clinical social worker for UNC ED (Author 5), a paramedic and PhD candidate in Health Informatics (Author 1), and a registered nurse and clinical informatician (Author 6). Annotators (Author 1,5) read each clinical note sentence in its entirety to assess the presence of SDH documentation. Any confirmatory mention of SDH associated regardless of status was treated as a positive finding, for example, "patient is currently homeless" or "patient states he has been homeless in the past," resulted in a positive label for housing insecurity. Detailed annotation guidelines are in [Supplementary Appendix 2](#).

### Collection of clinical notes

Clinical notes were obtained from the clinical data warehouse at the University of North Carolina (UNC-CDW), North Carolina's largest academic health system. The data files (JSON) received from the UNC-CDW were not exclusively free-text notes, but empty screening tools, blank auto-generated narratives, and meta-data. We initially isolated 2 corpora that were hypothesized to contain documentation of financial resource strain and poor social support. Sentences were derived from a variety of note types such as "Emergency Department progress note," "Psychiatry initial consult," and "social work psychosocial assessment." A word embedding terminology expansion approach was used to identify a subset of notes most likely to contain SDH documentation, therefore, only a small proportion of all notes collected from the EHR system and housed in the UNC-CDW were used in this study A.<sup>33,34</sup> The output of the word embedding expansion approach was used by annotators to complete the annotation process and significantly increased the yield of SDH positive annotations compared to traditional manual annotation of all documents in a corpus.<sup>29,35</sup>

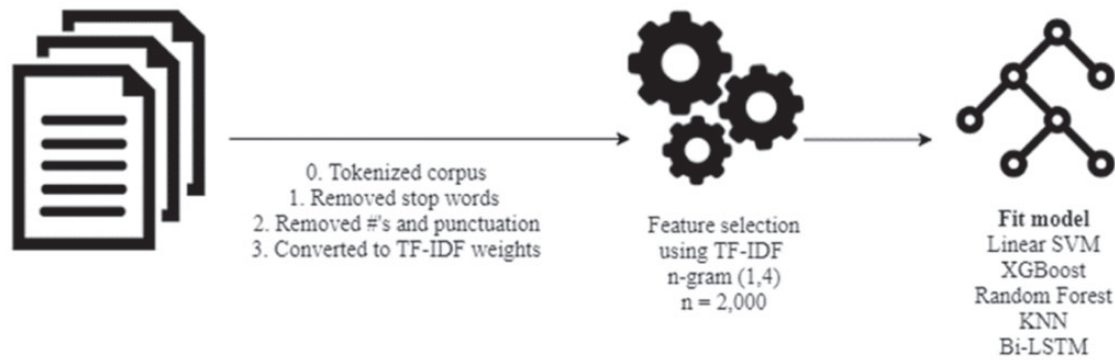


Figure 1. Overview of methods for machine learning.

### Input texts

For clinical notes, we completed the following preprocessing steps: (i) tokenized all input texts using Natural Language Toolkit (NLTK)<sup>36</sup>; (ii) removed all punctuation from each sentence; (iii) removed all non-alphabetical tokens; (iv) converted all letters to lower case; (v) normalized text through stemming and lemmatization that transform words to their root forms; and (vi) removed English stop-words (ie, me, my, myself, etc.).

### Outcome variables

We trained a binary relevance model, an ensemble of single-label binary classifiers, one for each class. Each classifier predicts either the membership or non-membership of one class. The union of all predicted classes then formed the multi-label output.<sup>28</sup> We classified whether a given SDH topic (eg, housing insecurity and/or food insecurity) was either documented or not document in the clinical note sentence.

### Experimental design

We developed and examined the outcome of 5 models: Random Forest (RF), XGBoost, KNN, LSTM, and SVM as our baseline. These models were trained (80%,  $N = 3250$ ) and tested (20%,  $N = 813$ ) on a randomized gold-standard corpus. Inputs into the classification models included a single free-text clinical note sentence. Figure 1 depicts an overview of methods for developing a machine learning classifier to identify SDH in clinical notes.

### Evaluation

Precision, recall, and F1 scores were computed across the SDH models using 5-fold cross-validation. Because the decision to optimize precision or recall depends on the specific clinical application, we considered F1 as the primary evaluation metric.<sup>29</sup> F1 represents the harmonic mean of precision and recall and takes both metrics into account. Since a multi-label neural network lacks a computational library from which to measure each label, we adopted 5 common evaluation measures: accuracy, average precision–recall (AP), area under curve receiver operating characteristic (AUC-ROC), Hamming loss, and log loss to compare the performance of different methods for multi-label SDH classification.<sup>28,37,38</sup> A full mathematical description of all evaluation metrics are found in Figure 2. Additionally, we conducted an error analysis to gain insight into model performance for SDH labels. We reviewed all incorrectly labeled sentences and analyzed false negatives using a classification matrix and attempted to classify each error as an incorrect annotation,

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn}$$

$$F_{\beta} = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}$$

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

$$\text{Average Precision-Recall} = \sum_n (R_n - R_{n-1}) P_n$$

$$\text{AUC-ROC} = \frac{2}{c(c-1)} \sum_{j=1}^c \sum_{k>j}^c p(j \cup k) (\text{AUC}(j|k) + \text{AUC}(k|j))$$

$$L_{\text{Hamming}}(y, \hat{y}) = \frac{1}{n_{\text{labels}}} \sum_{j=0}^{n_{\text{labels}}-1} 1(\hat{y}_j \neq y_j)$$

$$L_{\log}(Y, P) = -\log \Pr(Y|P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k}$$

Figure 2. Evaluation metrics.

unrecognized negation, or confusing auto-generated structure. All source code can be found at [www.github.com/rstem/dissertation](http://www.github.com/rstem/dissertation).

## RESULTS

### Study population

Our gold-standard corpus ( $N = 4063$ , ED visits  $>4$ , over 18 years of age, MHSUD ICD-10 codes) clinical note sentences represented 1119 patients of which 44.6% had at least one positive documentation of SDH. Characteristics of study patients are shown in Table 1. Half ( $N = 548$ , 50.2%) were between the ages of 39–62 and primarily White or Caucasian ( $N = 726$ , 66.9%).

### Characteristics of gold-standard corpus

A total of 4063 clinical note sentences associated with 1119 patients treated at a large academic medical system were manually reviewed

for characteristics of SDH; 502 clinical note sentences were associated with housing insecurity, 530 with poor social support, 321 with food insecurity, 686 with employment and/or income insecurity, 437 insurance insecurity, and 428 with general financial insecurity. 19.7% of SDH clinical note sentences in the entire corpus had 2 or more SDH labels documented; however, among positive SDH sentences ( $N=1066$ ) 75.0% of them had 2 or more SDH labels documented (Figure 3). To balance an initially overly positive data-

set, an additional 2252 negative SDH sentences were added to the corpus. Each sentence had an average length of 83.2 words with top N words being patient, discharge, care, and history. All clinical note sentences in the corpus were double annotated by clinical experts, with an overall Kappa statistic of 86.6% agreement (79.1–90.6%). The mean time our abstractors spent reviewing and coding notes was 65 s per clinical note sentence (~58 per hour). Annotators disagreed about 255 sentences. Disagreement occurred among sentences with the greatest length, an average of 145.8 words as compared to the corpus average of 83.2. Figure 3 shows the Pearson correlation coefficient between SDH labels with highest between general financial insecurity and poor social support (0.29).

**Table 1.** Study population characteristics<sup>a</sup>

Characteristic	All (%)	SDH positive (%)
N patients	1119	1066
Age	51.4 (±15.9)	51.2 (±16)
18–29	106 (9.5%)	104 (9.8%)
30–39	180 (16.1%)	173 (16.2%)
40–49	222 (16.1%)	215 (20.2%)
50–59	263 (19.8%)	251 (23.5%)
60–69	203 (23.5%)	186 (17.4%)
70–79	95 (8.5%)	88 (8.3%)
>80	50 (4.5%)	49 (4.6%)
Sex		
Male	573 (51.2%)	542 (50.8%)
Female	546 (48.8%)	524 (49.2%)
Race		
American Indian or Alaska Native	6 (0.0%)	6 (0.0%)
Asian	4 (0.0%)	4 (0.0%)
Black or African American	304 (27.2%)	285 (26.7%)
Native Hawaiian or other Pacific Islander	1 (0.0%)	1 (0.0%)
Other race	29 (0.0%)	29 (0.0%)
Patient refused	1 (0.0%)	1 (0.0%)
Unknown	24 (0.0%)	24 (0.0%)
White or Caucasian	750 (67.0%)	716 (67.2%)

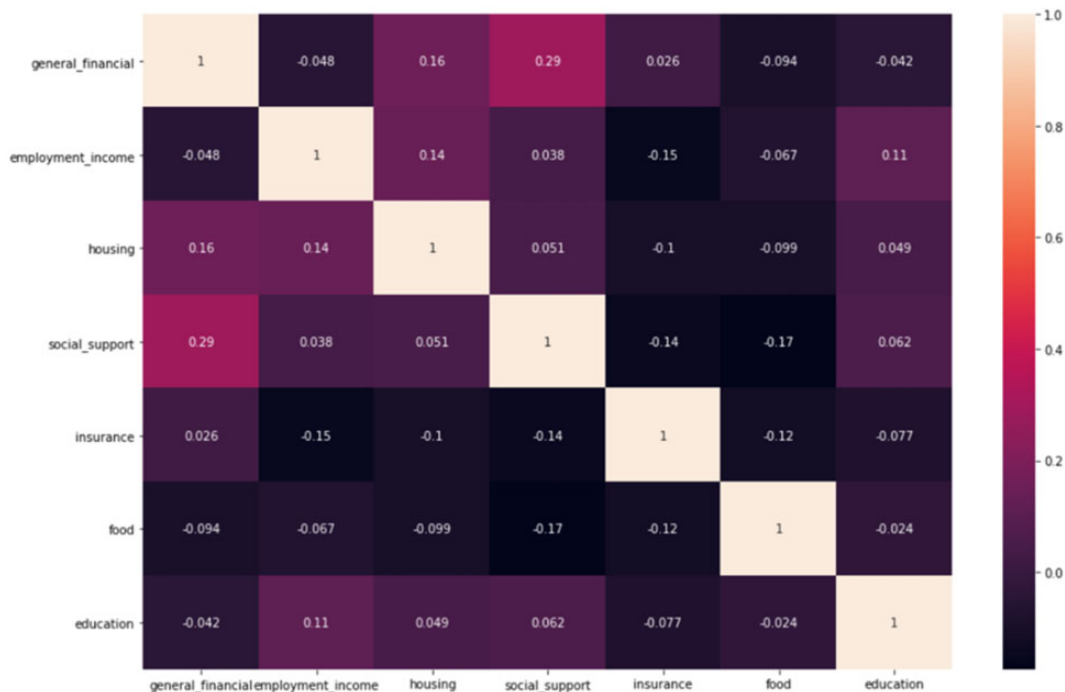
<sup>a</sup>Percentage based on non-missing data.

**Features used for SDH label classification**

Text features used by the classifiers included explicit indicators of SDH, as well as co-occurring determinants. For example, top features for a sentence classified as poor social support and employment/income insecurity included “limited social support,” “alcohol use disorder,” “cocaine use disorder,” and “financial concerns.” Meanwhile, other SDH labels had text features that were more closely associated with SDH risk factors. For example, top features for a sentence classified as food insecure included “food stamp,” “afford,” and “money”.

**Classifier performance**

Classification results inferring the presence of topic-specific SDH documentation are presented in Table 3 and ranged from F1 = 0.82 for XGBoost to F1 = 0.45 for KNN (F1 micro-averaged across all labels). XGBoost had the highest average precision micro-averaged across all labels (0.85), while the highest average recall micro-averaged across all labels was SVM (0.88). XGBoost had the lowest Hamming loss (4.13) with all other algorithms having nearly double the loss (Table 2). When comparing the precision-recall micro-averaged across all labels, Bi-LSTM (0.76) (Figure 4) out-performed all



**Figure 3.** Correlation matrix of SDH labels.

other classifiers, produced the lowest Hamming and log loss (0.12, 0.17), and the highest average ROC (93.5). The XGBoost poor social support model was the best performing SDH classifier (F1 = 0.89; Table 3), while the RF insurance insecurity model was the lowest performing model (F1 = 0.17). Because the prediction task relied on imbalanced data, we report area under ROC curve (AUC). We observed that the Bi-LSTM outperformed all other models (Figure 4) with the exception of the XGBoost classifying employment and/or income insecurity (Table 4). Figure 5 shows the AUC-ROC curve of the best model on each outcome. We achieve relatively high AUC's with the SVM, XGBoost, and Bi-LSTM.

### Error analysis

Among the 118 incorrect sentences classified by the SVM-baseline, 52 were false negatives and 20 were attributed to confusion between general financial insecurity and employment/income insecurity. Many false negatives were within the insurance insecurity label due to "Medicaid" being the only included text feature. However, many abbreviations for Medicaid were found in false negative insurance

insecurity sentences. For example, "mcd" and "mcaid" were common abbreviations found in sentences.

## DISCUSSION

This study aimed to develop an MLL model for identifying the SDH characteristics financial resource strain and poor social support using only clinical notes. Our findings suggest that an MLL approach trained on an SDH rich corpus can produce a high performing model. We also provide evidence that model performance is associated with lexical diversity by health professionals and the auto-generation of clinical note sentences to document SDH.

Based on our results, we recommend the neural network model, Bi-LSTM, because it performed well across all evaluation metrics. However, if the classification task requires transparency, gradient decision tree algorithms such as XGBoost, performed well across traditional evaluation metrics precision, recall, and micro-averaged F1 across all SDH labels. Our model outperformed (F1; 0.89–0.43) a similar study by Feller et al.<sup>29</sup> who used a multi-class gradient boosting tree to classify SDH sexual risk factors with F1 ranging from 79.2 for LGBT status to 27.3 for intravenous drug abuse. Our results are most likely due to our significantly larger training and testing dataset. Our model also outperformed a similar MLL algorithm classification task by Zufferey et al.<sup>37</sup> whose top-performing algorithm (SVM) had a Hamming loss of 16.94 and AP of 0.72, as compared to our model (0.12, 0.76). The disparity in results may be attributed to our use of only 6 labels as opposed to Zufferey et al. 15 labels. On the other hand, the poor performance of our KNN algorithm suggests that in multi-label medical domains the correlation between features may be an important characteristic to take into consideration. The successful results of the binary relevance SVM (F1 = 0.74) approach assumes independence among SDH characteristics suggests the features are not as correlated as previously assumed. It is difficult to give a complete explanation about these

**Table 2.** Performance of models using 5-fold cross-validation<sup>a</sup>

Metric	SVM	XGBoost	KNN	RF	Bi-LSTM
Hamming loss	7.18	4.13	9.51	8.79	0.12
Accuracy	70.92	81.38	64.46	62.62	93.3
Log loss	2.71	3.98	4.76	5.16	0.17
Average precision–recall	0.58	0.69	0.31	0.34	0.76
Average ROC	90.5	88.2	65.6	65.8	93.9
Micro-average precision	0.64	0.85	0.70	0.81	NA
Micro-average recall	0.88	0.78	0.33	0.33	NA
Micro-average F1	0.74	0.82	0.45	0.46	NA

<sup>a</sup>Averages are across all labels.

**Table 3.** Performance of models inferring SDH labels using 5-fold cross-validation

Algorithm	Label	Precision	Recall	F1	Support
SVM	General financial insecurity	0.51	0.91	0.66	69
	Employment/income insecurity	0.66	0.81	0.73	98
	Housing insecurity	0.62	0.89	0.73	83
	Poor social support	0.7	0.97	0.82	89
	Insurance insecurity	0.62	0.84	0.71	73
	Food insecurity	0.92	0.82	0.87	44
Xgboost	General financial insecurity	0.84	0.71	0.77	69
	Employment/income insecurity	0.85	0.76	0.8	98
	Housing insecurity	0.88	0.69	0.77	83
	Poor social support	0.85	0.93	0.89	89
	Insurance insecurity	0.81	0.75	0.78	73
	Food insecurity	0.93	0.86	0.89	44
KNN	General financial insecurity	0.66	0.33	0.44	69
	Employment/income insecurity	0.76	0.3	0.43	98
	Housing insecurity	0.53	0.12	0.2	83
	Poor social support	0.68	0.46	0.55	89
	Insurance insecurity	0.56	0.3	0.39	73
	Food insecurity	1	0.59	0.74	44
RF	General financial insecurity	0.78	0.42	0.55	69
	Employment/income insecurity	0.78	0.29	0.42	98
	Housing insecurity	0.83	0.23	0.36	83
	Poor social support	0.8	0.57	0.67	89
	Insurance insecurity	0.7	0.1	0.17	73
	Food insecurity	1	0.34	0.51	44

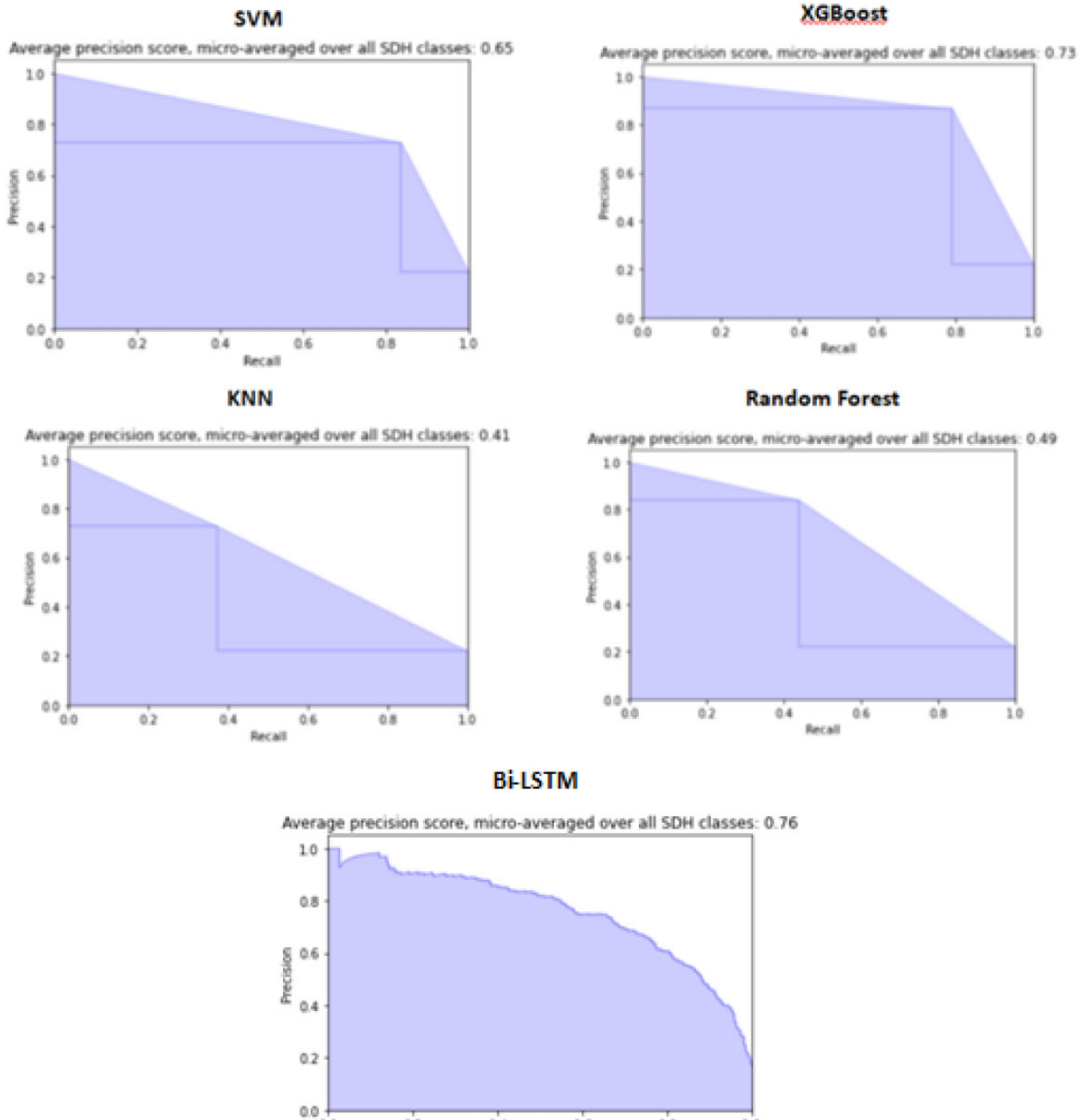


Figure 4. Average precision–recall score, micro-averaged over all SDH labels.

Table 4. Model performance (AUC) by SDH Label

SDH Label	SVM	XGBoost	KNN	Random Forest	Bi-LSTM
General financial insecurity	0.83	0.83	0.6	0.66	0.89
Employment or income insecurity	0.85	0.88	0.66	0.66	0.87
Housing insecurity	0.84	0.88	0.59	0.64	0.94
Poor social support	0.92	0.89	0.74	0.87	0.99
Insurance insecurity	0.88	0.85	0.63	0.63	0.95
Food insecurity	0.93	0.95	0.78	0.8	1

results; however, it may be that the feature extraction process cannot optimally model the correlation between the features in a manner that an MLL approach can exploit. To confirm this, we suggest further studies use algorithms and processing methods that deepen the analysis of SDH interdependence.

Our study is innovative in the following aspects. First, we developed a classifier to identify SDH on sentence-level data. The sentence-level scope can reduce the ambiguity of SDH characteristics and increase the agreement in the classification task as opposed to document level where SDH can still be buried within large quantities of text. Additionally, sentence level data allows for more granular results and thus a better understanding of the SDH documentation

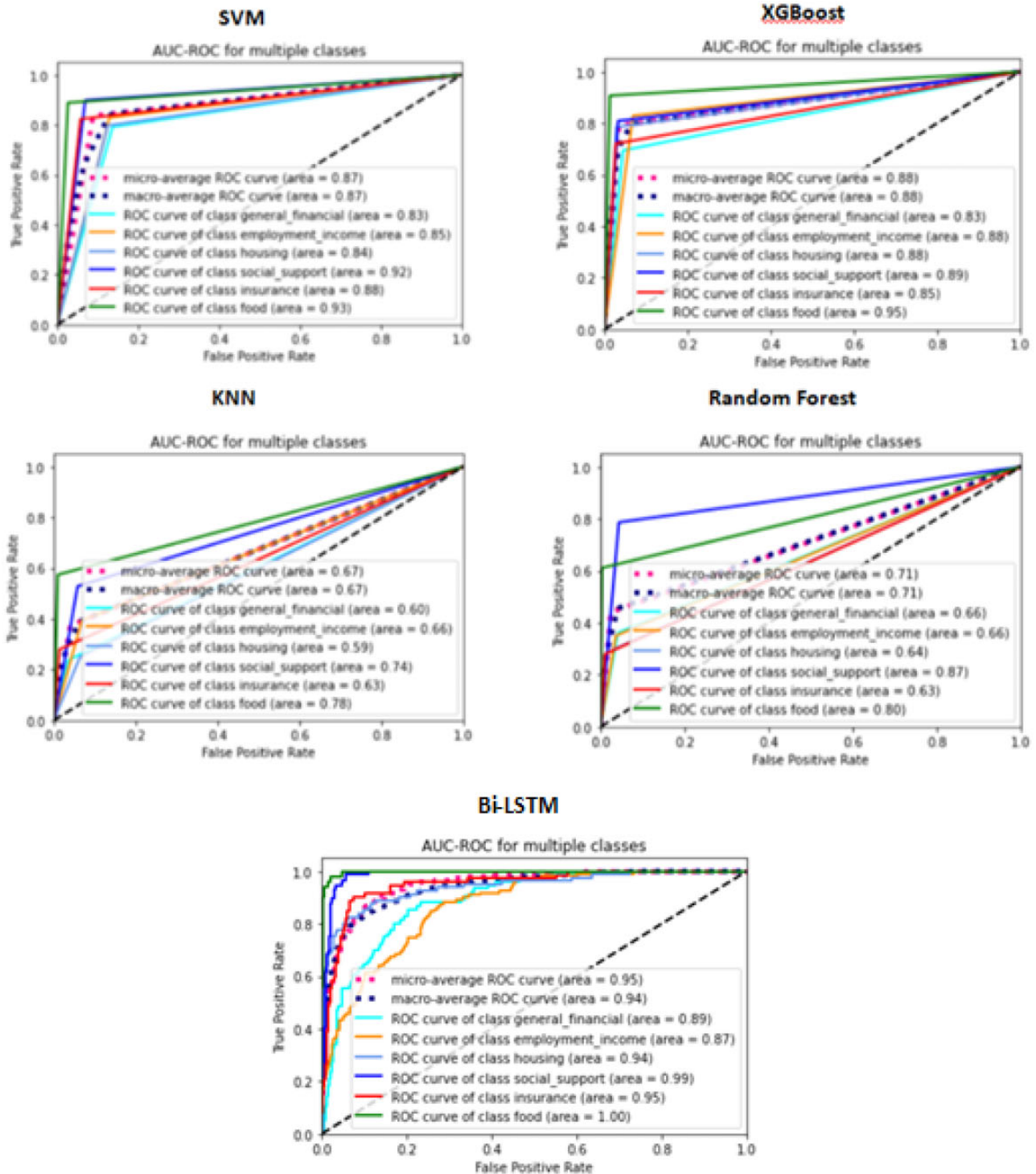


Figure 5. AUC-ROC for SDH labels.

within clinical notes. Second, our model performance shows the feasibility of classifying SDH using only clinical notes without structured features of the EHR. However, Feller et al.<sup>29</sup> found that the combination of clinical notes and structured data yielded better performance than either data source alone when inferring SDH sexual risk factors. Future studies should explore whether structured data elements such as demographics and medical codes could enhance

performance. Finally, we used an MLL approach to reduce information loss<sup>28,39</sup> and take advantage of the cardinality and label dependency.<sup>39,40</sup>

We observed a positive correlation between model performance and the prevalence of each specific SDH. This demonstrates the necessity of building gold-standard corpora of adequate size, especially for infrequently documented SDH such as food insecurity. However,



the poor social support label had higher precision and recall than other labels with similar prevalence, likely reflecting the limited lexical diversity used to express this SDH. For example, poor social support was often referenced as “limited social supports” or “lacking social support.” The results of our error analysis suggest several areas for improvement in automated SDH classification. The labels general financial insecurity and employment/income insecurity had the highest false negatives ( $N = 13$ ). We believe this is due to the high lexical diversity associated with these labels, suggesting that clinicians lack a standardized way of expressing those SDH. Potentially, further subdomains describing these SDH could be used to enhance future classification tasks. Our findings also suggest benefit from standardized approaches to collecting SDH data in EHRs.<sup>30,41,42</sup>

We found very little systematic documentation of patients’ SDH data in the EHRs clinical notes including a large amount of incomplete or empty SDH screening surveys. Experts recommend limiting SDH screening to a subset of patients and enabling EHR-based SDH data tools to target this subset to avoid overwhelming or burdening health professionals.<sup>13,42</sup> The SDH classification models we developed may be best applied as a tool to identify patients requiring standardized screening for SDH such as food and housing insecurity. Our work shows many implications for SDH collection in the EHR. Future research from this project may lead to an SDH screening alert within the EHR to increase adoption rates. In this task recall may be prioritized to limit alert fatigue and assure that all those who need a screening, receive one. Gold et al.<sup>13</sup> found that health professionals did not want to collect SDH data themselves, preferring to transfer the responsibility to another team member. With SDH data collected via multiple routes and certain SDH data are already collected regularly by specific health professionals (eg, social workers), future research should explore a need for an EHR-based summary that contains all of a patient’s SDH data. When adequately leveraged, electronic platforms improve integration between medical and social service delivery. EHRs could provide opportunities to improve the evidence by improving data accessibility and standardization, linking SDH interventions with health outcomes, and supporting the examination of individual and population-level data.

### Limitations

First, our SDH classifier was trained using data from a single institution limiting its generalizability, although our data comprise input from geographically distributed, rural and urban, academic and non-academic EDs. Future work should focus on using corpus developed from multiple institutions or publicly available sources. Second, our overall modest results may have resulted from data quality issues in the documentation of SDH and/or inaccurate annotation. Third, most approach this problem as a NER task but because we approached the problem as a sentence labeling task, our experimental design does not allow for direct comparisons to previous work. Future work may explore the proficiency of SDH identification as an NER task. Fourth, our model performance may have been improved by considering negation or by correcting misspelling in text; we did not consider negation due to the fact that not all SDH studied would have benefited from this addition (ie, “the patient denied homelessness despite living in a tent in the woods”). Fifth, our patient population was comprised of those with MHSUD who frequented the ED creating an overly positive dataset of SDH; these records likely differ from the general population of a health system, potentially compromising the generalizability of the classification models. Furthermore, our data collection time period encompassed

the health system’s transition to a single EHR, and may have impacted the quality and consistency of SDH documentation. Sixth, we did not use a “holdout” dataset that was never used in model training since we did not have the requisite volume of data to create training, validation, and test sets and thus the observed model performance may be inflated. If possible, future studies should use a holdout set to estimate unbiased model performance. Seventh, we did not explore other problem transformation approaches to MLL such as classifier chains or label powerset. Eighth, to balance our overly positive dataset we added negative sentences that were reviewed by one annotator leading to the possibility of hidden SDH among the negative. Ninth, there is limited ability to understanding neural networks as they use a hidden layer for pattern recognition in feature selection and thus full explanation of the Bi-LSTM results are not possible at this time. For full transparency into feature selection and decision tree decisions future work could explore transforming this problem into a multi-class classification task despite the information loss risks.

## CONCLUSION

We investigated 5 common ML models for the task of multi-label classification of SDH using only clinical notes. Unlike previous work, we evaluated our models sentence-level data that contained multiple instances of SDH documentation, labels thus making sure our models could be used for real-world SDH clinical decision support tasks. Our MLL approach is a first concrete step toward SDH phenotyping across EHRs as SDH characteristics cross multiple domains and future studies may approach SDH phenotyping as an extreme MLL task. The study findings suggest that SDH prevalence and the lexical diversity used to express a given SDH characteristic have an impact on the performance of classification algorithms. Future studies should explore the standardization of SDH collection and computational methods that can effectively learn models of diverse rare features.

## FUNDING

This study was funded by the following sources: National Library of Medicine—T15 LM012500: “Training in Biomedical Informatics at the University of North Carolina—Chapel Hill.” The project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002489. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## AUTHOR CONTRIBUTIONS

RS designed and performed the experiments, derived the models and analysed the data. RK, AK, and JA assisted with experimental design and evaluation of results. RS wrote the manuscript with support from RK, JA, AK, JB, and LM. JB assisted with domain expertise, JA and AK with technical expertise, and RK supervised the project.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## ACKNOWLEDGMENTS

The authors would like to thank Mary Houston, MSW for their help in preparing the gold-standard corpus.

## CONFLICT OF INTEREST

None declared.

## DATA AVAILABILITY

All source code can be found at [www.github.com/rstem/dissertation](http://www.github.com/rstem/dissertation).

## PROTECTION OF HUMAN AND ANIMAL SUBJECTS

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed by the Institutional Review Board at the University of North Carolina—Chapel Hill.

## REFERENCES

- Smith JL, De Nadai AS, Storch EA, *et al.* Correlates of length of stay and boarding in Florida emergency departments for patients with psychiatric diagnoses. *Psychiatr Serv* 2016; 67 (11): 1169–74.
- Data on Behavioral Health in the United States. <http://www.apa.org/help-center/data-behavioral-health.aspx> (accessed 2 June 2018).
- Brennan JJ, Chan TC, Hsia RY, *et al.* Emergency department utilization among frequent users with psychiatric visits. *Acad Emerg Med* 2014; 21 (9): 1015–22.
- Phelan JC, Link BG, Anderson N, editor. Fundamental social causes of disease and mortality. In: *Encyclopedia of Health and Behavior*. 2455 Teller Road. Thousand Oaks, CA: SAGE Publications, Inc.; 2004. doi:10.4135/9781412952576.n102.95-102.
- Matthews KA, Adler NE, Forrest CB, *et al.* Collecting psychosocial “vital signs” in electronic health records: Why now? What are they? What’s new for psychology? *Am Psychol* 2016; 71 (6): 497–504.
- Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/18709.
- Chang G, Weiss AP, Orav EJ, *et al.* Predictors of frequent emergency department use among patients with psychiatric illness. *Gen Hosp Psychiatry* 2014; 36 (6): 716–20.
- Ku BS, Fields JM, Santana A, *et al.* The urban homeless: super-users of the emergency department. *Popul Health Manag* 2014; 17 (6): 366–71.
- Foegle WH. Actual causes of death in the United States—Reply. *JAMA* 1994; 271 (9): 660.
- Singh GK, Siahpush M, Kogan MD. Neighborhood socioeconomic conditions, built environments, and childhood obesity. *Health Aff (Millwood)* 2010; 29 (3): 503–12.
- Soril LJJ, Leggett LE, Lorenzetti DL, *et al.* Reducing frequent visits to the emergency department: a systematic review of interventions. *PLoS One* 2015; 10 (4): e0123660.
- Okin RL, Boccellari A, Azocar F, *et al.* The effects of clinical case management on hospital service use among ED frequent users. *Am J Emerg Med* 2000; 18 (5): 603–8.
- Gold R, Cottrell E, Bunce A, *et al.* Developing electronic health record (EHR) strategies related to health center patients’ social determinants of health. *J Am Board Fam Med* 2017; 30 (4): 428–47.
- Bodenmann P, Velonaki V-S, Griffin JL, *et al.* Case management may reduce emergency department frequent use in a universal health coverage system: a randomized controlled trial. *J Gen Intern Med* 2017; 32 (5): 508–15.
- Ondler C, Hegde GG, Carlson JN. Resource utilization and health care charges associated with the most frequent ED users. *Am J Emerg Med* 2014; 32 (10): 1215–9.
- Tsai M-H, Xirasagar S, Carroll S, *et al.* Reducing high-users’ visits to the emergency department by a primary care intervention for the uninsured: a retrospective study. *Inquiry* 2018; 55: 4695801876391.
- Hatef E, Rouhizadeh M, Tia I, *et al.* Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform* 2019; 7 (3): e13802.
- Vest JR, Grannis SJ, Haut DP, *et al.* Using structured and unstructured data to identify patients’ need for services that address the social determinants of health. *Int J Med Inform* 2017; 107: 101–6.
- Bejan CA, Angiolillo J, Conway D, *et al.* Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc* 2018; 25 (1): 61–71.
- Bettencourt-Silva JH, Mulligan N, Sbodio M, *et al.* Discovering new social determinants of health concepts from unstructured data: framework and evaluation. *Stud Health Technol Inform* 2020; 270: 173–7.
- Mishra R, Bian J, Fiszman M, *et al.* Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform* 2014; 52: 457–67.
- Gundlapalli AV, Carter ME, Palmer M, *et al.* Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. *AMIA Annu Symp Proc* 2013; 2013: 537–46.
- Hollister BM, Restrepo NA, Farber-Eger E, *et al.* Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac Symp Biocomput* 2017; 22: 230–41.
- Lindemann EA, Chen ES, Rajamani S, *et al.* Assessing the representation of occupation information in free-text clinical documents across multiple sources. *Stud Health Technol Inform* 2017; 245: 486–90.
- Blosnich JR, Marsiglio MC, Dichter ME, *et al.* Impact of social determinants of health on medical conditions among transgender Veterans. *Am J Prev Med* 2017; 52 (4): 491–8.
- Baumel T, Nassour-Kassis J, Elhadad M, Elhadad N. Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment. *CoRR* 2017. <http://arxiv.org/abs/1709.09587>
- Du J, Chen Q, Peng Y, *et al.* ML-Net: multi-label classification of biomedical texts with deep neural networks. *J Am Med Inform Assoc* 2019; 26 (11): 1279–85.
- Zhang M-L, Zhou Z-H. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng* 2014; 26 (8): 1819–37.
- Feller DJ, Bear Don’t Walk Iv OJ, Zucker J, *et al.* Detecting social and behavioral determinants of health with structured and free-text clinical data. *Appl Clin Inform* 2020; 11 (01): 172–81.
- Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington (DC): National Academies Press (US); 2015. doi:10.17226/18951.
- Palinkas LA, Horwitz SM, Green CA, *et al.* Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Health* 2015; 42 (5): 533–44.
- Cohen R, Aviram I, Elhadad M, *et al.* Redundancy-aware topic modeling for patient record notes. *PLoS One* 2014; 9 (2): e87555.
- Fan Y, Pakhomov S, McEwan R, *et al.* Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open* 2019; 2 (2): 246–53.
- Wang Y, Liu S, Afzal N, *et al.* A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform* 2018; 87: 12–20.

35. Lingren T, Deleger L, Molnar K, *et al.* Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J Am Med Inform Assoc* 2014; 21 (3): 406–13.
36. Natural Language Toolkit—NLTK 3.5 Documentation. <https://www.nltk.org/> (accessed 12 May 2020).
37. Zufferey D, Hofer T, Hennebert J, *et al.* Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Comput Biol Med* 2015; 65: 34–43.
38. Madjarov G, Kocev D, Gjorgjevikj D, *et al.* An extensive experimental comparison of methods for multi-label learning. *Pattern Recognit* 2012; 45 (9): 3084–104.
39. Luaces O, Díez J, Barranquero J, *et al.* Binary relevance efficacy for multi-label classification. *Prog Artif Intell* 2012; 1 (4): 303–13.
40. Bromuri S, Zufferey D, Hennebert J, *et al.* Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. *J Biomed Inform* 2014; 51: 165–75.
41. Monsen KA, Rudenick JM, Kapinos N, *et al.* Documentation of social determinants in electronic health records with and without standardized terminologies: a comparative study. *Proc Singapore Healthc* 2019; 28 (1): 39–47.
42. Gottlieb LM, Tirozzi KJ, Manchanda R, *et al.* Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med* 2015; 48 (2): 215–8.