An Investigation of the Representation of Social Determinants of Health in the UMLS

Bhanu Pratap Singh Rawat, PhD[1], Heather Keating, PhD[2,3], Raelene Goodwin, BS[2,3], Emily Druhl, MPH[2,3], Hong Yu, PhD[1,2,3]

[1] **College of Information and Computer Science, University of Massachusetts-Amherst**

[2] **Center for Healthcare Organization and Implementation Research, VA Bedford Healthcare System**

[3] **Center of Biomedical and Health Research in Data Sciences, University of Massachusetts-Lowell**

Abstract

Social Determinants of Health (SDOH) are the conditions in which people are born, live, work, and age[1]. Unified Medical Language System (UMLS) incorporates SDOH concepts[2]; but few have evaluated its coverage and quality. With 15,649 expert-annotated SDOH mentions from 3176 randomly selected electronic health record (EHR) notes, we found that 100% SDOH mentions can be mapped to at least one UMLS concept, indicating a good coverage of SDOH. However, we discovered a few challenges for the UMLS's representation of SDOH. Next, we developed a multi-step framework to identify SDOH concepts from UMLS, and a clinical BERT-based classification algorithm to assign each identified SDOH concept to one of the six general categories. Our multi-step framework extracted a total of $198,677$ *SDOH concepts from the UMLS and the SDOH category classification system attained an accuracy of* 91%*. We also built* EASE*: an open-source tool to* ***E****xtr****a****ct* ***S****DOH from* ***E****HRs.*

Introduction

World Health Organization (WHO) identifies social determinants of health (SDOH) as *the conditions in which people are born, grow, live, work, and age, which are shaped by the distribution of money, power and resources*[1]. These include social and environment factors that impact the health of individuals. The Kaiser Family Foundation (KFF) has identified six general categories of SDOH[3] as shown in Table 1.

Table 1: Six general categories of Social Determinants of Health (SDOH).

Economic Stability	Factors affecting the economic stability of the individual such as current employment, income, expenses, debts and/or medical bills.
Physical Environment	Factors related to individuals' housing situation, availability of parks and playgrounds, transportation. These are mainly factors related to the geographical location of the individual.
Education	Factors related to the individuals' literacy, early childhood education, higher education, language proficiency and/or vocational training.
Food	Factors related to hunger, availability of food and access to healthy food options.
Community and Social Context	Factors related to the social support system of the individual along with their community engagements. This category also contains factors related to social discrimination and stress.
Healthcare System	Factors related to health coverage, health provider availability and overall quality of care.

Health systems have long recognized the importance of SDOH[4][5]. More than 80% of US physicians document SDOH in electronic health records[6], and a majority of SDOH ($> 90\%$) are described in electronic health record (EHR) notes[7][8]. Therefore it is important to identify SDOH from EHR notes. Previously, rule-based and machine-learning approaches have been developed[9]; however, most work focuses on identifying only certain types of SDOH (e.g., homelessness[10] or substance abuse[11]). In this study, we developed a framework to identify all six general types of SDOH concepts from UMLS that can be identified in EHRs with the help of existing tools such as MetaMap[12] and cTakes[13].

Unified Medical Language System[14] (UMLS) is a comprehensive biomedical meta-thesaurus that integrates concepts from 156 biomedical vocabularies. Arons et al.[2] examined controlled medical terminologies and extracted $1,095$ SDOH codes from four controlled medical coding vocabularies (LOINC[15], SNOMED CT[16], ICD-10-CM, and CPT). UMLS is widely used for different clinical applications with sophisticated NLP tools, such as MetaMap[12] and

cTakes[13], to map free text to corresponding UMLS concepts[17–19]. If we can identify all SDOH in the UMLS, we could use an existing NLP system (e.g., MetaMap and cTakes) to automatically identify SDOH from free text.

Arons et al.[2] used a keyword-based approach to extract codes for 20 SDOH-related domains such as Access to health care, Stress and Ethnicity. However, due to the comprehensive coverage and expressive flexibility of medical vocabularies, one medical concept can be mapped to multiple concept codes in a vocabulary. Due to this one-to-many mapping, previous research such as Andrews et al.[20] and Quan et al.[21] show that heterogeneity exists when mapping codes from SNOMED-CT and ICD-10 to different free text datasets. Hence, keyword-based search alone is not sufficient to identify all codes relevant to different SDOH concepts in clinical vocabularies. In this study, we developed a multi-step framework to identify SDOH concepts from the UMLS.

For gold standard, we randomly selected 3176 EHR notes from patients prescribed with opioid who received care in the Veteran Health Administration (VHA) system in FY2016. Three domain experts chart-reviewed these 3176 notes, identified SDOH, and mapped them to the UMLS concepts. Our results (details in Results section) show that all 15,649 SDOH identified in these 3176 notes could be mapped to the UMLS concepts. This result has demonstrated a 100% recall for the UMLS to incorporating SDOH. However, we also found some quality issues of the UMLS for SDOH representation.

Then, we developed a 3-step framework to identify SDOH concepts from the UMLS. Each UMLS concept in the UMLS is defined and assigned a CUI (Concept Unique Identifier). Each CUI is also assigned one or multiple semantic types (STY)[1], such as Clinical Drug, Anatomical Structure, Social Behavior and Health Care Activity. Arons et al[2] used a keyword-based search to identify 1095 SDOH related concepts while limiting to four vocabularies (SNOMED, ICD-10, LOINC, and CPT). We expanded the list of the keywords provided by Arons et al.[2] by adding additional SDOH domains such as loss of relationship and substance abuse while adding more keywords to the existing domains and performed a keyword-based search over all 156 vocabularies identifying thousands of concepts. We then performed graph search across UMLS to identify additional CUIs and filter them with the help of selected semantic types resulting in $198,677$ SDOH concepts. In order to categorise the identified SDOH concepts amongst six SDOH categories, we developed a hybrid rule + BERT (Bidirectional Encoder Representations from Transformers) based system to assign each of the $198,677$ concepts to its corresponding SDOH category (refer Table 1). We also built `EASE`: an open-source tool to **E**xtr**a**ct **S**DOH risk factors from **E**HRs. `EASE` can be used by anyone with minimal programming knowledge and can be used to extract SDOH risk factors from clinical notes of patients for better health assessment and medical interventions by medical professionals.

Our three-fold contributions can be summarised as:

1. We extracted $198,677$ SDOH concepts (CUIs) from the UMLS.
2. We developed a hybrid classification system to assign each SDOH concept a SDOH category. The system achieved an accuracy of 91% on a randomly selected test set. The categories provide a higher level information regarding the SDOH risk factors present in the natural text.
3. An open-sourced Python package (`EASE`) would be released for researchers that can extract SDOH risk factors from natural language text using only 2 lines of code. We have created this easy-to-use, readily available tool to push research forward in this direction.

Methodology

We first discuss the method to evaluate the comprehensiveness of the UMLS for representation of SDOH concepts. Next, we discuss a 3-step framework to extract the SDOH concepts (CUIs) from UMLS. Finally, we describe a hybrid approach to assign each SDOH a SDOH category. Our complete approach is illustrated in Figure 1.

EHR Notes Annotation to Assess the Coverage of UMLS for SDOH

In order to evaluate the comprehensiveness of SDOH concepts in UMLS, we conducted an expert chart-review study to examine whether SDOH documented in the clinical notes can be mapped to the UMLS.

[1] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

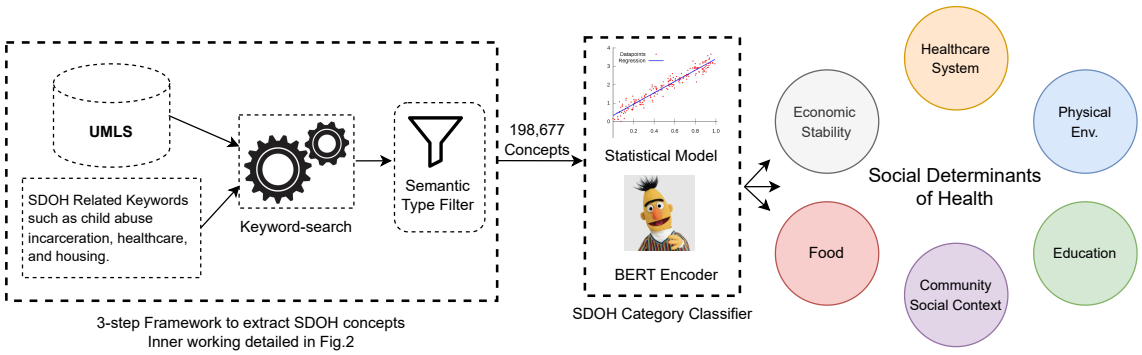

Figure 1: The figure illustrates the framework used for extracting SDOH concepts (CUIs) from UMLS. First we extracted concepts (CUIs) by searching SDOH related keyword in the UMLS graph. Then, we passed all the CUIs through a semantic type filter to avoid the unrelated concepts. Finally, we built a hybrid classification model to assign one of the six SDOH categories[3] to each CUI.

For this, we randomly selected 3,176 social work and rehab notes, as they are most likely to contain SDOH, of patients with a history of opioid-use from the Veteran Health Administration's (VHA) EHR data. Three trained annotators (BS, MPH or PhD in the biomedical sciences), under the supervision of a senior physician, annotated 3,176 notes. Each note was annotated by at least two annotators. Discrepancies were addressed through discussion among the annotators and the senior physician. We built the annotation guideline based on the definitions by the Kaiser Family Foundation[3] (refer Table 1). To facilitate the annotation, we built a keyword-based information retrieval engine so that the annotator could search for a list of UMLS concepts based on a keyword. The annotations also provide us a gold standard set of SDOH concepts ($\mathcal{S}_{\mathcal{G}}$) from UMLS that is used to calculate the coverage accuracy of our SDOH extraction framework at different retrieval steps. Appendix A describes all the the mathematical notations used in this paper.

Identification of SDOH from UMLS

Keyword Search: UMLS (ver. 2020AB) incorporates 156 medical vocabularies, 133 semantic type groups (STYs) and $4,413,092$ concepts (CUIs) representing $13,560,262$ strings (STRs). Each CUI is provided with a brief string definition (STR) which is provided in the `MRCONSO` table of UMLS. To perform a keyword search over these STRs, we expanded the list of keywords provided by Arons et al.[2] by adding more SDOH domains such as loss of relationship and substance abuse while adding more keywords to the existing domains. All the SDOH domains and keywords would be provided along with the tool (`EASE`). The keyword-based search identifies the first set of SDOH concepts from UMLS ($\mathcal{S}_{\mathcal{K}}$).

Graph Search: In medical vocabularies, one medical concept can be mapped to multiple medical codes[20,21]. Hence, just keyword-based search would not retrieve all the relevant SDOH concepts. To extract the SDOH concepts missed by keyword search, we did graph search starting from the concepts of the set $\mathcal{S}_{\mathcal{K}}$ towards their neighboring concepts that are connected through different relationship edges. While performing the graph-search, we only focused on the concepts which were connected to the earlier extracted concepts through the relations: RL (similar or alike), RN (narrow), RB (broad), SIB (sibling) and CHD (child) relationship. This helped in constraining our search to strictly relevant concepts. A single hop (hop=1) graph search would result in finding the concepts ($\mathcal{S}_{\mathcal{GR},hop=1}$) which are directly connected to the concepts extracted via keyword search ($\mathcal{S}_{\mathcal{G}}$). A double hop (hop=2) graph search would result in extracting concepts ($\mathcal{S}_{\mathcal{GR},hop=2}$) which are directly related to concepts extracted during single hop graph search ($\mathcal{S}_{\mathcal{GR},hop=1}$) and so on. At each search stage, we calculated the overlap between the extracted ($\mathcal{S}_{\mathcal{GR},hop=n}$) and gold-standard concepts $\mathcal{S}_{\mathcal{G}}$ and stopped the further search once more than 90% of the concepts from $\mathcal{S}_{\mathcal{G}}$ are covered without extracting a lot of irrelevant concepts. The final set of extracted SDOH concepts ($\mathcal{S}_{\mathcal{F}}$) is a combination of all previously extracted sets using keyword-based search ($\mathcal{S}_{\mathcal{K}}$) and graph search ($\mathcal{S}_{\mathcal{GR},hop=1\ldots n}$)

$$\mathcal{S}_{\mathcal{F}} = \mathcal{S}_{\mathcal{K}} \cup \mathcal{S}_{\mathcal{GR},hop=1\ldots n}$$

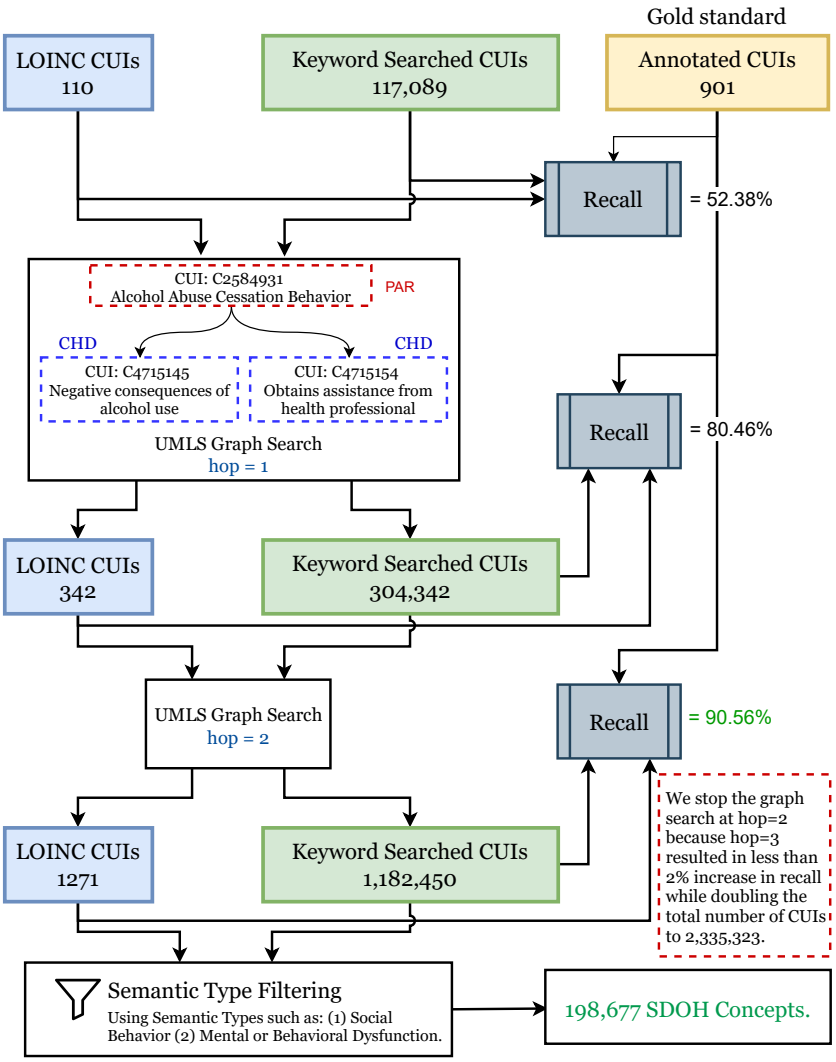

Figure 2: The figure illustrates our multi-step algorithm to extract SDOH related concepts (CUIs) from UMLS. We set our cut-off to hop=2 for our graph search as increasing hop-value resulted in small recall increments while doubling the extracted keywords.

Semantic Type Filtering: In order to filter out the concepts that are unrelated or very broadly related to SDOH risk factors, we decided to use the semantic type information of the CUIs. The semantic type of a concept provides a higher level category for a concept. For ex, concept 'C2584931' defined as 'Alcohol abuse cessation behavior (observable entity)' is assigned the semantic type 'Individual Behavior', and concept 'C0870171' defined as 'Atypical paranoid disorder' has the semantic type 'Mental or Behavioral Dysfunction'. These semantic type groups can provide important information regarding a concept for filtering out the noisy extracted concepts. We manually selected 28 semantic type groups (STY_S) which are relevant to SDOH risk factors. The list of the semantic type groups is provided in Table 2. Each medical concept can be assigned more than one semantic type as well, for ex. 'C0020021' defined as 'Psychiatric hospital' is assigned two semantic types: 'Health Care Related Organization' and 'Manufactured Object'. After filtering, the concepts that are assigned at least one semantic type from the selected group of semantic types (STY_S) would remain whereas the rest of them would be filtered out. After semantic type filtering, our final set of SDOH risk factors $\mathcal{S}_\mathcal{F}$ reduces to $\mathcal{S}_{\mathcal{F}'}$.

Our multi-step framework to extract SDOH concepts from UMLS is illustrated in Figure 2.

Selected Semantic Types			
Age Group	Individual Behavior	Group	Mental or Behavioral Dysfunction
Behavior	Family Group	Group Attribute	Human-caused Phenomenon or Process
Injury or Poisoning	Food	Health Care Activity	Biomedical Occupation or Discipline
Daily or Recreational Activity	Finding	Health Care Related Organization	Environmental Effect of Humans
Educational Activity	Mental Process	Language	Governmental or Regulatory Activity
Social Behavior	Occupation or Discipline	Population Group	Self-help or Relief Organization
Patient or Disabled Group	Occupational Activity	Professional or Occupational Group	Geographic Area

Table 2: The list of selected 28 semantic types.

Classification of SDOH into General Categories

Kaiser Family Foundation (KFF) has grouped SDOH factors into six main categories[3]. Each extracted concept from UMLS can be assigned one of the six categories. We decided to build a hybrid approach to assign a category to each of the concept using their semantic type (STY) group and their string definition (STR) provided in the `MRCONSO` table of UMLS. Some of the semantic type groups can be manually mapped to one of the six SDOH category, for ex, the semantic type 'Educational Activity' can be assigned to the 'Education' SDOH category, 'Health Care Activity' can be assigned to 'Healthcare System', and 'Social Behavior' can be assigned to 'Community and Social Context'. We mapped total 18 semantic type groups ($STY_{\mathcal{C}}$) to one of the six SDOH categories.

Category Classification Models: The rule-based semantic type group to SDOH category mapping helped us in assigning a category to a fraction of concepts ($\mathcal{S}_{\mathcal{C}}$) from $\mathcal{S}_{\mathcal{F}'}$. We used $\mathcal{S}_{\mathcal{C}}$ to train, validate and test our category classification models to categorize the rest of the extracted concepts ($\mathcal{S}_{\mathcal{F}'} - \mathcal{S}_{\mathcal{C}}$). We decided to use the definition (STR) of concepts provided by UMLS. The definition of a concept can be a small descriptive phrase, such as for 'C0687129': 'lack of housing', or a long sentence, such as for 'C2911076': 'Encounter for mental health services for spousal or partner abuse problems'. Hence, we needed a reliable free text encoder that could be used for classification. We decided to use Bidirectional Encoder Representations from Transformers (BERT)[22], a pre-trained language model that encodes contextualised language representations for natural texts. We used the clinically fine-tuned version of BERT (clinicalBERT[23]) to extract the vector representation for the STR of each concept.

Each concept string text is first passed through a tokenizer to get the tokenized version of phrase or sentence. `[CLS]` is appended at the start of the token sequence and passed through the encoder (clinicalBERT), for example, 'lack of housing' would be tokenized to the sequence: '[CLS]', 'lack', 'of', 'housing'. The contextualised representation, hidden representation of [CLS], from the encoder is considered as the vector representation ($\mathcal{R}_{CUI}$) of the concept string (STR).

$$\text{Concept Definition (STR)} \rightarrow \text{Tokenizer} \rightarrow \text{[CLS] } [t_1] \text{ .... } [t_n] \rightarrow \text{clinicalBERT} \rightarrow \mathcal{R}_{CUI} \quad (1)$$

Here, t_1 .... t_n refer to the tokens of a concept's definition (STR). These vector representations ($\mathcal{R}_{CUI}$) for the concepts are used as input features and the SDOH categories are used as labels for the category classification model.

To train, test, and validate our category classification model, we divided $\mathcal{S}_{\mathcal{C}}$ into training, validation and testing set in the ratio of $70:10:20$. We trained six different classification models on the $\mathcal{R}_{CUI}$ representations of the concepts in the train set: (1) Logistic regressor (2) Ridge classifier (3) Passive Aggressive classifier (4) Perceptron (5) Random Forest classifier and (6) k-Nearest Neighbor classifier. For evaluation, we used both accuracy and macro-averaged F1-score as accuracy shows the overall performance of the model and macro-averaged F1-score shows the average performance across different classes.

Results

The manual annotation process resulted in the gold standard set $\mathcal{S}_{\mathcal{G}}$ of 901 unique SDOH concepts from UMLS. The annotators were able to assign at least one medical concept from UMLS to all the SDOH phrases in the EHRs. This result demonstrated a 100% recall for UMLS to incorporate the SDOH risk factors in the free text. Though our EHR annotation of SDOH concepts revealed substantial challenges for using the UMLS as a resource for representation of the SDOH. The challenges and inconsistencies for incorporating SDOH risk factors in UMLS are listed below.

Non-uniform negations across concepts: Negation is not always encoded within concepts (CUIs). Such as, 'C0425083' refers to 'Loss of job' and can be used to annotate the risk factor 'Economic Stability' at instances where subject lost their current job. Whereas there is no concept code to identify loss of housing which refers to the 'Physical Environment' of the subject and experts have to annotate an extra negation attribute along with the concept code 'C0020056' which refers to housing. There are concept codes that refer to homelessness but they can mainly be used for instances where the subject never had stable housing and cannot be used for cases where the subject recently lost their house.

Overlapping codes for similar concepts: Multiple codes refer to similar concepts such as, 'C0237154' refers to 'homelessness', 'C0019863' refers to 'homeless' or 'Homeless person', and 'C0425241' refers to 'Homeless family'. All three concepts refer to the homelessness situation which makes it difficult for the experts to choose one of the concept code to annotate an instance of homelessness. This also results in inter-annotator disagreements.

Similar concept codes not connected in the graph: At times, UMLS codes that refer to similar concepts are not connected to each other via any relation in the graph. For example, concepts 'C0019863' ('homeless') and 'C0425241' ('homeless family') are both connected to concept code 'C0237154' ('homelessness') via `CHD` relationship but they are not connected to each other. Essentially, they should be connected to each other by atleast `RB` (broad) relationship. Similarly, 'C0021672' ('Insurance') and 'C3845555' ('Private insurance') are also not connected to each other while referring to similar concept. For such cases, our multiple hop graph search becomes even more important as it provides most coverage of relevant concept coder from UMLS for SDOH risk factors.

Inconsistent assignment of semantic type groups to concepts: In UMLS, the concept codes are not assigned consistent semantic type (STY) groups. Such as, concepts 'C0019863' ('homeless') and 'C0425241' ('homeless family') are assigned 'Population Group' as the semantic type whereas 'C0237154' ('homelessness') is assigned 'Finding' even though the former concepts codes are connected to 'C0237154' via `CHD` relationship but did not inherit the parent's semantic type group. Similarly, 'C0021672' ('Insurance') is assigned the semantic type 'Idea or Concept' whereas 'C3845555' ('Private insurance') is assigned 'Governmental or Regulatory Activity' and 'C0481840' ('Insurance medical') is assigned 'Finding'. Similarly, 'C0687758' ('probation') is assigned 'Regulation or Law' whereas 'C4699973' ('I have current charges/trial pending, noncompliance with probation/parole') is assigned 'Finding'. Along with 'Finding', the concept code 'C4699973' should also be assigned the STY 'Regulation or Law'. The inconsistency in semantic type groups makes it harder to group similar concept codes in the same SDOH category.

Extraction of SDOH from UMLS

The keyword-based search extracted $117,089$ unique SDOH concepts across UMLS. We added another 110 concepts from LOINC vocabulary that have already been identified and categorised into six SDOH categories[2]. This resulted in $117,199$ concepts in the set $\mathcal{S}_{\mathcal{K}}$. To evaluate the performance of our keyword-based search, we calculated the overlap between the extracted concepts $(117,199)$ via keyword search and the gold standard concepts (901) which were annotated by experts over $3,176$ EHRs. The overlap match came out to be 52.38% which is quite low. This is in-line with our previous suggestion that keyword search alone is not enough to extract the relevant SDOH concepts from UMLS.

Through graph search ($hop = 2$), we achieved 90.56% overlap between the sets $\mathcal{S}_{\mathcal{G}}$ and $\mathcal{S}_{\mathcal{F}}$. Further graph search of $hop = 3$ increased the extracted concepts by 200% while only increasing the overlap by 2% suggesting that most of the extracted concepts at $hop = 3$ are either unrelated or very broadly connected to SDOH risk factors. After graph search, $\mathcal{S}_{\mathcal{F}}$ consists of $1,182,450$ SDOH concepts. After filtering the set $\mathcal{S}_{\mathcal{F}}$ using $STY_{\mathcal{S}}$ we got the final set $\mathcal{S}_{\mathcal{F}'}$ consisting of $198,677$ SDOH concepts. We also performed manual reviews over two mutually exclusive sets of randomly selected 500 concepts and observed that at least 75% of them are strictly SDOH concepts. The rest of them are broadly related to the 'Healthcare' SDOH category and belong to one of the three semantic type groups: 'Injury or Poisoning', 'Health Care Activity', and 'Finding'.

SDOH Category Classification

From the final set $\mathcal{S}_{\mathcal{F}'}$, a total of $65,726$ ($\mathcal{S}_{\mathcal{C}}$) were assigned one of the 18 semantic type groups ($STY_{\mathcal{C}}$) that had a mapping to one of the SDOH category. We then assigned SDOH categories to the $\mathcal{S}_{\mathcal{C}}$ set according to their semantic

[2]Listing of all six SDOH categories in LOINC: https://loinc.org/LG41762-2/

Table 3: Frequency count of CUIs for each SDOH category along with examples.

Category	# Concepts (mapping)	# Concepts (classification)	Example Concepts
Community and Social Context	4,496	7,802	C0557128: Lives with Friend C0580097: Poor Social Circumstances
Economic Stability	6,590	1,157	C2371583: Seeking employment C0337629: Economic problem
Physical Environment	2,206	2,023	C0518530: Drives own car C2136015: Lives in Assisted Living Facility
Healthcare System	44,492	119,258	C4551672: Medicare coverage C0012605: Disability Evaluation
Education	2,232	238	C3841898: Completed a GED C0814966: Student Failing Academically
Food	5,808	2,473	C0681144: Food Stamps C0522060: Inadequate Diet

type groups, their distribution is shown in Table 3. 'Healthcare System' category has the most number of concepts because 6 of the semantic group types (STY), out of 18, were assigned to 'Healthcare System'. The semantic type group to SDOH category mapping is provided in Appendix B.

We used the $\mathcal{S}_C$ concept set to build our classification model. We divided this set into training, validation and testing set in the ratio of $70 : 10 : 20 = 46,008 : 6,572 : 13,146$. The results for all the classification models are provided in Table 4. The Logistic Regression model performs the best with the highest macro-averaged F1-score of 0.83 and accuracy of 0.91. The k-Nearest Neighbor model also achieved the same accuracy but lower F1-score than the Logistic Regression model. We used the Logistic Regression model to assign SDOH category to each of the $132,951$ remaining concepts. The distribution of SDOH categories for this set of $132,951$ SDOH concepts is provided in Table 3.

Table 4: Macro-averaged F1-score and Accuracy for all the classification models. The highest F1-score and accuracy are highlighted in the table.

Model	F1-score	Accuracy
Logistic Regressor	**0.83**	**0.91**
Ridge Classifier	0.78	0.89
Passive Aggressive Classifier	0.74	0.89
Perceptron Classifier	0.77	0.85
Random Forest Classifier	0.72	0.87
k-Nearest Neighbor Classifier	0.82	**0.91**

EASE 1.0

With the help of $198,677$ SDOH concepts identified in the UMLS, we developed `EASE`: an open-source tool to **E**xtr**a**ct **S**DOH from **E**HRs. `EASE` integrates the MetaMap output with the framework we developed to automatically identify SDOH and their SDOH categories. It is written in `Python` programming language as it is widely used amongst researchers. In this section, we will first discuss how to use `EASE` and how to modify the tool according to the researcher's requirement.

The `Python` package developed can be easily installed using the command: `pip install .` inside the repository of `EASE`[3]. Figure 3 shows a code snippet where the tool is being used to extract the SDOH concepts from a given sentence of a clinical note. The tool is able to extract important SDOH information related to the patient which could impact patient's mental or physical health. This extracted information can help the clinical professionals in early medical interventions and modify the existing treatment plan and supports.

[3]https://github.com/bsinghpratap/ease

Code

```
import ease

text = 'The patient is widowed and lives alone in [**Location (un) **].'
tool = ease.sdoh_extractor(config='/home/usr/configs/config_sdoh.json')
extracted_cuis = tool.get_sdoh_cuis(text)

for each_cui_tuple in extracted_cuis:
        print('Extracted Phrase:', each_cui_tuple['phrase'])
        print('CUI: ', each_cui_tuple['CUI'])
        print('Semantic Type: ', each_cui_tuple['sem_type'])
        print('Start, End: ', each_cui_tuple['st_ind'], each_cui_tuple['end_ind'])
        print('----------')
```

Output

```
Extracted Phrase: widowed
CUI: C1510465
Semantic Type: Family Group (famg)
Start, End: 15, 22
----------
Extracted Phrase: alone
CUI: C0679994
Semantic Type: Group Attribute (grpa)
Start, End: 33, 38
----------
Extracted Phrase: Location
CUI: C0450429
Semantic Type: Spatial Concept (spco)
Start, End: 45, 53
----------
```

Figure 3: *EASE* only needs two lines of code (yellow and red pointer) to extract the SDOH risk factors. In the example above, we can see that the tool was able to extract SDOH-related information regarding the patient. The tool was able to extract the token 'widowed' which tells us about the current relationship status of the patient. The next extracted token is 'alone' which tells about the social isolation of the patient and can potentially cause mental and physical toll on patient's health. The next extracted token is 'Location' which is a de-identified value but if the tool is being used as an internal tool then it would provide the actual geographical location of the patient that is closely related to 'Neighborhood and the Physical Environment' SDOH.

Configuring EASE

The tool can be easily configured to extract specific set of information such as concepts related only to 'Mental or Behavioral Dysfunction' of the patient or extract all the concepts from a sentence or a paragraph. As shown in Figure 3, the model takes as input a `json` file that helps in configuring the tool according to the researchers' requirements. The tool can also be used for several other generic purposes such as extracting the text definition of a CUI or checking if there exists a relationship between two CUIs or list all the CUIs which have a relationship with a specific CUI.

Conclusion

In this work, we show that UMLS has a good representation of SDOH risk factors. We also developed a 3-step framework to extract $198,677$ SDOH concepts from UMLS. We also developed a hybrid classification technique using a clinically finetuned BERT-based encoder to identify a SDOH category for all the extracted SDOH concepts. Our SDOH category classification system achieved 91% accuracy and 83% f1-score. Furthermore, we developed `EASE` an open-source tool to extract SDOH risk factors from free-text and assign them a related UMLS concept along with a SDOH category. `EASE` can be used by anyone with minimal programming experience and can help in building smarter clinical decision support systems by extracting and providing SDOH information about patients from their EHRs.

Acknowledgement

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number 5R01MH125027. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Commission on Social Determinants of Health et al. *Closing the gap in a generation: health equity through action on the social determinants of health: final report of the commission on social determinants of health*. World Health Organization, 2008.

2. Abigail Arons, Sarah DeSilvey, Caroline Fichtenberg, and Laura Gottlieb. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA open*, 2(1):81–88, 2019.

3. Samantha Artiga and Elizabeth Hinton. Beyond health care: the role of social determinants in promoting health and health equity. *Health*, 20(10):1–13, 2019.

4. Oliver J Bear Feller, Daniel J and, Jason Zucker, Michael T Yin, Peter Gordon, Noémie Elhadad, et al. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Applied clinical informatics*, 11(01):172–181, 2020.

5. John R Blosnich, Ann Elizabeth Montgomery, Melissa E Dichter, Adam J Gordon, Dio Kavalieratos, Laura Taylor, Bryan Ketterer, and Robert M Bossarte. Social determinants and military veterans' suicide ideation and attempt: a cross-sectional analysis of electronic health record data. *Journal of general internal medicine*, pages 1–9, 2019.

6. Michael Wang, Matthew S Pantell, Laura M Gottlieb, and Julia Adler-Milstein. Documentation and review of social determinants of health data in the ehr: measures and associated insights. *Journal of the American Medical Informatics Association*, 28(12):2608–2616, 2021.

7. David Dorr, Cosmin A Bejan, Christie Pizzimenti, Sumeet Singh, Matt Storer, and Ana Quinones. Identifying patients with significant problems related to social determinants of health with natural language processing. *Studies in health technology and informatics*, 264:1456–1457, 2019.

8. Avijit Mitra, Hiba Ahsan, Wenjun Li, Weisong Liu, Robert D Kerns, Jack Tsai, William Becker, David A Smelson, Hong Yu, et al. Risk factors associated with nonfatal opioid overdose leading to intensive care unit admission: A cross-sectional study. *JMIR medical informatics*, 9(11):e32851, 2021.

9. Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631, 2021.

10. Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among us veterans. In *AMIA Annual Symposium Proceedings*, volume 2013, page 537. American Medical Informatics Association, 2013.

11. Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. Automated extraction of substance use information from clinical texts. In *AMIA Annual Symposium Proceedings*, volume 2015, page 2121. American Medical Informatics Association, 2015.

12. Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

13. Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

14. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

15. Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4):624–633, 2003.

16. Ronald Cornet and Nicolette de Keizer. Forty years of snomed: a literature review. In *BMC medical informatics and decision making*, volume 8, pages 1–6. BioMed Central, 2008.

17. Matthias Becker and Britta Böckmann. Extraction of umls® concepts using apache ctakes™ for german language. In *Health Informatics Meets Ehealth*, pages 71–76. IOS Press, 2016.

18. Ruth Reátegui and Sylvie Ratté. Comparison of metamap and ctakes for entity extraction in clinical notes. *BMC medical informatics and decision making*, 18(3):13–19, 2018.

19. Alejandro Rodríguez-González, Roberto Costumero, Marcos Martinez-Romero, Mark D Wilkinson, and Ernestina Menasalvas-Ruiz. Extracting diagnostic knowledge from medline plus: a comparison between metamap and ctakes approaches. *Current Bioinformatics*, 13(6):573–582, 2018.

20. James E Andrews, Timothy B Patrick, Rachel L Richesson, Hana Brown, and Jeffrey P Krischer. Comparing heterogeneous snomed ct coding of clinical research concepts by examining normalized expressions. *Journal of biomedical informatics*, 41(6):1062–1069, 2008.

21. Hude Quan, Bing Li, L Duncan Saunders, Gerry A Parsons, Carolyn I Nilsson, Arif Alibhai, William A Ghali, and IMECCHI investigators. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health services research*, 43(4):1424–1441, 2008.

22. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

23. Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

A Mathematical Notations

Notation	Definition
$\mathcal{S}_{\mathcal{G}}$	Set containing all the gold standard SDOH concepts from the UMLS that were extracted by the annotators (901).
$\mathcal{S}_{\mathcal{K}}$	Set containing all the SDOH concepts extracted from UMLS using keyword search (117,199). This set also includes the LOINC concepts that were already identified as SDOH.
$\mathcal{S}_{\mathcal{GR},hop=n}$	Set containing all the SDOH concepts extracted using graph search via n^{th} hop. After $hop = 2$, this set consisted of 1,182,450 UMLS concepts.
$\mathcal{S}_{\mathcal{F}}$	Set that combines the SDOH concepts of $\mathcal{S}_{\mathcal{K}}$, $\mathcal{S}_{\mathcal{GR},hop=1}$, and $\mathcal{S}_{\mathcal{GR},hop=2}$.
$STY_{\mathcal{S}}$	Semantic type groups that are selected for filtering the UMLS concepts to get relevant SDOH concepts (28).
$\mathcal{S}_{\mathcal{F}'}$	Final set after semantic type group filtering using $STY_{\mathcal{S}}$.
$STY_{\mathcal{C}}$	Selected semantic type groups that could be directly mapped to a SDOH category (18).
$\mathcal{S}_{\mathcal{C}}$	Set of SDOH concepts that could be mapped to a SDOH category using $STY_{\mathcal{C}}$. (65,276)

B Semantic Group Type to SDOH category mapping

Semantic Type Group (STY)	SDOH Category
Biomedical Occupation or Discipline	Economic Stability
Daily or Recreational Activity	Social and Community
Diagnostic Procedure	Healthcare
Educational Activity	Education
Family Group	Social and Community
Food	Food
Geographic Area	Physical Environment
Group	Social and Community
Health Care Activity	Healthcare
Health Care Related Organization	Healthcare
Mental Process	Healthcare
Mental or Behavioral Dysfunction	Healthcare
Occupation or Discipline	Economic Stability
Occupational Activity	Economic Stability
Patient or Disabled Group	Healthcare
Population Group	Social and Community
Professional or Occupational Group	Economic Stability
Social Behavior	Social and Community