Automatic Extraction of Social Determinants of Health from Medical Notes of Chronic Lower Back Pain Patients

Dmytro Lituiev, PhD*1, Benjamin Lacar, PhD*1,2, Sang Pak, PT, DPT3, Peter L Abramowitsch1, Emilia De Marchis, MD, MAS**4, Thomas Peterson, PhD**1,5

1Bakar Computational Health Sciences Institute, University of California San Francisco
2Berkeley Institute for Data Science, University of California, Berkeley
3Department of Physical Therapy and Rehabilitation Science, University of California San Francisco
4Department of Family & Community Medicine, University of California San Francisco
5Department of Orthopaedic Surgery, University of California San Francisco

*Co-first authored
**Co-senior authored

Abstract

Background. Adverse social determinants of health (SDoH), or social risk factors, such as food insecurity and housing instability, are known to contribute to poor health outcomes and inequities. Our ability to study these linkages is limited because SDoH information is more frequently documented in free-text clinical notes than structured data fields. To overcome this challenge, there is a growing push to develop techniques for automated extraction of SDoH. In this study, we explored natural language processing (NLP) and inference (NLI) methods to extract SDoH information from clinical notes of patients with chronic low back pain (cLBP), to enhance future analyses of the associations between SDoH and low back pain outcomes and disparities.

Methods. Clinical notes (n=1,576) for patients with cLBP (n=386) were annotated for seven SDoH domains: housing, food, transportation, finances, insurance coverage, marital and partnership status, and other social support, resulting in 626 notes with at least one annotated entity for 364 patients. We additionally labelled pain scores, depression, and anxiety. We used a two-tier taxonomy with these 10 first-level ontological classes and 68 second-level ontological classes. We developed and validated extraction systems based on both rule-based and machine learning approaches. As a rule-based approach, we iteratively configured a clinical Text Analysis and Knowledge Extraction System (cTAKES) system. We trained two machine learning models (based on convolutional neural network (CNN) and RoBERTa transformer), and a hybrid system combining pattern matching and bag-of-words models. Additionally, we evaluated a RoBERTa based entailment model as an alternative technique of SDoH detection in clinical texts. We used a model previously trained on general domain data without additional training on our dataset.

Results. Four annotators achieved high agreement (average kappa=95%, F1=91.20%). Annotation frequency varied significantly dependent on note type. By tuning cTAKES, we achieved a performance of F1=47.11% for first-level classes. For most classes, the machine learning RoBERTa-based NER model performed better (first-level F1=84.35%) than other models within the internal test dataset. The hybrid system on average performed slightly worse than the RoBERTa NER model (first-level F1=80.27%), matching or outperforming the former in terms of recall. Using an out-of-the-box entailment model, we detected many but not all challenging wordings missed by other models, reaching an average F1 of 76.04%, while
matching and outperforming the tested NER models in several classes. Still, the entailment model may be sensitive to hypothesis wording and may require further fine tuning.

**Conclusion.** This study developed a corpus of annotated clinical notes covering a broad spectrum of SDoH classes. This corpus provides a basis for training machine learning models and serves as a benchmark for predictive models for named entity recognition for SDoH and knowledge extraction from clinical texts.

**Introduction**

Adverse social determinants of health (SDoH), or social risk factors, such as food insecurity and housing instability, are recognized for their deleterious impacts on health outcomes and disparities. There is growing recognition of the role of social risks in chronic low back pain (cLBP), as highlighted in a recent systematic review that found strong associations of cLBP prevalence with educational attainment and socioeconomic status. Outcomes for cLBP, a leading cause of disability worldwide, are known to be worse in patients who are economically and socially disadvantaged. This can be attributed in part to treatment biases, including greater provision of non-evidence-based care, as well as patients’ prior experience of discrimination and beliefs about pain and pain treatment, which may influence engagement with the treatment offered. Much of the current research exploring disparities in cLBP care has been limited to stratifying analyses by socioeconomic status and race, as social risk data are not readily available in electronic health records (EHRs). While some social risk information exists in structured data fields, such as patient demographics and problem lists, these data are known to be under identified by clinical teams, under reported by patients, and under documented in structured fields. When clinical teams screen for and identify social risks, that information is more often documented within free-text fields, or unstructured data. Novel approaches to identifying SDoH in EHRs, such as Natural Language Processing (NLP) and other machine learning techniques, leverage existing health information technology to scan unstructured data fields, resulting in automated extraction.

Several studies have applied NLP methods to obtain SDoH information from clinical notes. Methods have included regular expressions, neural networks, and rule-based algorithmic approaches. Named entity recognition (NER), an NLP task of extracting phrases and their positions from texts, has been most frequently used. SDoH such as housing situation, finances, and social support have been less studied using NLP methods than behavioral determinants of health (BDoH), such as smoking status, and substance and alcohol use. For BDoH, both rule-based and machine learning approaches to NER have been applied. SDoH, most notably housing situation, have most commonly been identified through rule based approaches such as keyword matching. Most other SDoH domains, such as transportation access and finances, have been understudied by either technique. There has been no comprehensive comparison of the strengths and weaknesses of ML and rule-based methods when applied to SDoH on various dataset sizes. We are not aware of any studies that apply NLP for SDoH for patients with cLBP.

Both machine learning (ML) and traditional rule-based system are being used to extract clinical and SDoH concepts from healthcare narratives. Tried-and-true rule-based systems such as clinical Text Analysis and Knowledge Extraction System (cTAKES), leveraging a Unified Medical Language System (UMLS) ontology developed by National Library of Medicine (NLM), could provide reasonable performance with no training data by leveraging standard ontologies and rule-based pattern matching. Modern machine learning techniques have revolutionized
NLP and natural language understanding, with self-supervised methods such as BERT\textsuperscript{32}, allowing for efficient utilization of large amounts of un-labelled data. For optimal performance in supervised tasks in specialized domains, however, these techniques may require corpora of several (tens of) thousands of documents of manually annotated text\textsuperscript{33}. While still providing important utility in domains where labelled data is scarce\textsuperscript{34} and being more interpretable than ML, rule-based systems often yield to ML methods in performance\textsuperscript{35,36}. In terms of development costs, both rule-based and machine learning based models are resource intensive. For rule-based models, significant labor is required for manual vocabulary tuning. On the other hand, machine learning models require manually annotated data. Still, improvement of machine learning systems is normally less demanding in terms of technical development effort as compared to rule-based systems.

While several applications of natural language inference (NLI) methods, such as Question Answering (QA)\textsuperscript{37–39} and Recognizing Textual Entailment (RTE)\textsuperscript{40–42} have been pursued in the wider medical domain, we were not able to find any SDoH or BDoH studies leveraging these new methods. The fact that this has not been attempted is surprising, as one would expect NLI models to generalize to SDoH and BDoH better than other clinical domains. Knowledge about social and behavioral clues may be more widely available than knowledge about anatomy and physiology, which is necessary for understanding clinical specialties, in general domain texts (i.e., fiction and news) that state-of-the-art NLI models are predominantly trained on. This prompted us to apply NLI models in our study. Out of NLI approaches, we chose to use entailment (RTE) models instead of QA models\textsuperscript{41} given our interest in off-line knowledge extraction (i.e., without a need to present to a clinician for immediate decision making) with a pre-defined ontology in the setting of off-line knowledge extraction (i.e., without a need to present to a clinician for immediate decision making) with a pre-defined ontology. While extractive QA models (returning a piece of original text) may not be able to summarize text into a limited set of categories by design, abstractive QA models suffer from biases of characteristic to all generative models such as ability to learn toxic and biased attitudes\textsuperscript{42,43} as well as exhibit learned attitudes characteristic to mental illness and addiction\textsuperscript{44}.

In this study, our aim was to create a dataset to evaluate and compare NLP tools to extract individual SDoH from free-text clinical notes for patients with cLBP. In this study, we make several contributions to this important topic: 1) we label and evaluate a dataset with social and behavioral determinants of health in cLBP patients, 2) we tune and evaluate a rule-based NER system (cTAKES), 3) we train and evaluate two machine learning NER pipelines, and 4) we evaluate a previously trained common domain natural language inference (NLI) entailment model. We believe this study is a step forward in both applying cutting-edge NLP technology to this important topic and in improving the structured ontology for identifying SDoH and BDoH.

**Methods**

**Study population**
Our study population consisted of cLBP patients, defined by low back pain lasting at least three months\(^4\), from a large, urban academic medical center at the University of California, San Francisco (UCSF) and was approved by institutional review (IRB #19-29016). All patients were from the UCSF Integrated Spine Service (ISS), comprised of an integrated care team of psychiatrists, pain management specialists, primary care physicians, and physical therapists\(^4\). For all patients, we extracted progress notes, history and physical (H&P) notes, emergency department (ED) provider notes, patient instructions, and telephone encounters (TE) from the UCSF clinical data warehouse between March 2017 to April 2020. The cohort demographics is shown in Supplementary Table 1.

**Annotation ontology**

To create a model for extracting named entities related to SDoH from clinical free-text notes, we defined an ontology, annotated the dataset, and tuned and trained algorithms for named entity recognition (Fig 1). SDoH domains of interest included: Housing, Food Transportation, Finances, Insurance Coverage, Marital and Partnership Status and Social Support. These SDoH domains were selected as they are commonly included in social risk screening tools that are used in clinical practice\(^4\)–\(^9\). Additionally, we included mental health factors (anxiety and depression) as well as pain scores, as these are relevant for our ongoing work in characterizing cLBP population\(^2\),\(^5\)–\(^6\). During the study period, none of these SDoH were systematically screened for in our healthcare system. Our ontology builds upon a selection of social risk factors identified earlier by the UCSF Social Interventions Research and Evaluation Network (SIREN)\(^3\) mapped to SNOMED ontology. Initial ontology containing 68 second-level classes within 10 first-level classes was pruned after the label review step, as described in the next section, down to 52 second-level classes. The final ontology is shown in Fig 1B and Supplementary Table 2. Entities within the first-level classes not captured by specific second-level labels due to ambiguity or where low occurrence did not warrant a separate group were designated by a second-level label of “NA” for “not applicable”.

**Manual Annotations**

The notes were annotated according to a pre-defined two-level taxonomic ontology by four trained annotators using MAE labelling software\(^5\) for NER task. That is, words or phrases describing a relevant class were highlighted (as shown in Fig 1B and 4). Each note was annotated by at least two annotators (Supplementary Figure 1A). Notes were additionally spot-checked and corrected. Initial annotations occasionally contained overlapping or duplicate labels that were subsequently resolved using Python spacy package. Heuristics for duplicate resolution aimed to both maximize span length for each resulting label and to conform semantics (e.g., “lives at home with her boyfriend” may be labelled with overlapping labels describing Housing, Social Support, and Marital and Partnership Status; the corrected version would contain “lives at home” for Housing, “with her” for Social Support, and “boyfriend” for Marital and Partnership Status). Labels were further manually reviewed and rare second-level categories with low agreement and consistency labels from initial ontology were combined with more general second-level categories after data inspection. E.g., Depressed mood and Symptoms of depression were aggregated under the latter. Thus 68 initial second-level classes were pruned down to 52. After all pre-processing steps, document-level inter-rater agreement was evaluated using span-level F\(_1\) and document-level Cohen’s kappa\(^5\).
Where $p_{\text{obs}}$ is the fraction of observations where experts agreed and $p_{\text{exp}}$ is the expected probability of agreement between the two experts at random. For document-level evaluation, the presence or absence of at least one annotation of a given class in a clinical note was considered a binary event. Presence or absence was compared between each pair of annotators in a set of notes that they had both reviewed. A kappa per annotator pair was calculated by weighting each category with its mean occurrence according to either annotator (as described in Supplementary Methods). The final kappa was calculated as an average weighted by the number of notes annotated by each pair of annotators. We chose $F_1$ as a span-level metric as it is both symmetric (as kappa is) and is less sensitive to number of true negatives, which are abundant and would dominate in NER tasks. In downstream applications, such as training and model evaluation, annotations from each expert were included. For subsequent analysis, the data was split in 6 folds, while stratifying by the number of entities and number of annotators per note and keeping multiple annotations of the same note in the same fold.

The variance of annotation frequencies was analyzed using ANOVA in R software. Pairwise difference between models was assessed using Wilcoxon test, based on $F_1$ scores from each 6-fold cross-validation split.

**cTAKES configuration**

We used three configurations of cTAKES software$^{28}$. cTAKES is a modular framework developed for named entity extraction from medical texts that leverages UMLS vocabulary$^{57}$. We used (1) a basic default out-of-the box configuration, (2) an InfoCommons configuration$^{58}$ developed for general purpose medical texts, and (3) a customized configuration that was tuned to identify pain scores, mental health outcomes (anxiety and depression), and our SDoH domains. Tuning was done using an unlabeled set of notes and the labeling ontology definition. See Supplementary Methods for a detailed description of these three configurations.

**Spacy-based CNN model configuration**

Machine learning for named entity recognition (NER) was performed with Spacy software$^{59}$ (https://spacy.io/) using a transition-based parser leveraging a transition-based parser based on either convolutional neural network (CNN)$^{60}$ or RoBERTa$^{61}$ model, targeting the NER ontology described above. Word embeddings were initialized with the values from the “en_core_web_md” spacy model. The model was trained with a variable batch size of 100 to 1000 examples, drop-out of 10%, learning rate of 0.0002, and for maximum 30 epochs with early stopping. Model was trained and evaluated in nested 6-fold cross-validation, where four parts of data was used as training set, one part as validation set, and one part as test set.

**Hybrid model leveraging pattern matching and classification with Bag-of-Words**

A classification model was applied to text extracted based on keyword matching patterns, which were compiled for each first-level class based on subject matter identification and literature. A
JSON file of the matcher patterns is included in the code repository (https://github.com/BCHSI/social-determinants-of-health-clbp). Text was extracted around the matching pattern (±4 tokens), pre-processed by lemmatizing and removing stop words, then embedded using a term frequency-inverse document frequency (TF-IDF) vectorizer. Extracted text for each of the first-level set of patterns was then classified using a Bag-of-Words (BoW) multinomial logistic regression model into respective second-level subclasses according to manual labels, with unmatched text spans receiving a special exclusion label. Model was trained and evaluated in 6-fold cross-validation as described for CNN model above.

Evaluation of NER models
The results of the automated extraction systems were evaluated using precision, recall, and F1 scores. The metrics were calculated on two levels of the labeling taxonomy (first-level and second-level). For cTAKES model, predicted UMLS terms overlapping ground truth labels were manually annotated as being synonyms, children, or wrong matches of respective first-level and second-level ground truth categories, while accounting for negation, history, and family history modifiers. Synonyms and child concepts were subsequently scored as matches. The scores were calculated in using scikit learn package, with partial overlaps counted as matches. Pairwise comparison between model performance metrics was performed using paired t-test across all cross-validation folds (for CNN and hybrid models) or respective data split folds (for cTAKES). Variation of model performance was analyzed using ANOVA and nested linear model ANOVA.

Evaluation of an entailment model
A RoBERTA-based entailment model, previously fine-tuned in an adversarial human-in-the-loop setting, was evaluated without additional training on our data. Given two sentences, the entailment model produces probabilities premise and hypothesis for three categories: entailment (premise statement is consistent with hypothesis), contradiction (premise not consistent with hypothesis), or neutral relation (premise indicates neither consistency nor contradiction with hypothesis). To evaluate performance of entailment model, a set of hypothesis sentences was constructed based on NER labels. E.g., “Has access to car” category was phrased as “Patient has access to a car”, “Patient drives a car”, and “Patient drives a vehicle”. Classes with final node labelled as “NA” were phrased as “This describes patient’s (first-level topic)” and were excluded from precision analysis as more specific categories would fall under that premise (See Supplementary Table 3, column “premise”). The word “patient” was replaced with a gender-specific pronoun for quantitative evaluation. Qualitative comparison of using “patient” vs gender specific pronouns indicated that replacement increases accuracy (Supplementary Methods). Each premise sentences (i.e., original text containing a label) was fed together with the each of the hypothesis sentences (See example in Fig 4). For computation expediency, only sentences with at least one NER label were examined. The results were aggregated per premise sentence (i.e., original text) and label combination either (1) by taking an average rate of entailment prediction over alternative hypotheses (“mean”) or (2) by scoring a match if at least one of the hypotheses was deemed to be entailed (“max”). Data was split into 6 folds matching the folds of NER analysis. For classes and fold sets, where positive prediction was missing and thus precision was undefined, F1 was assigned the value of recall. Pain scores, as well as PHQ-9 and GAD-7 scores were excluded from this analysis as they were originally included with the goal of integer value extraction and are not easily amenable for textual entailment task.

Results
Gold Standard Dataset Annotation

Four trained annotators annotated a set of 1,576 clinical notes with an average inter-rater variability of 95.28%, as measured by Cohen’s kappa (per-annotator metrics are shown in Suppl. Table 3) and weighted F1 of 91.20% and macro average F1 of 88.93% (Suppl Table 4). The inter-rater agreement per each label is shown in Suppl. Table 5. Of the 1,576 notes, 39.72% (626) contained at least one entity and was reviewed by at least two annotators (Supplementary Fig 1A). Notes without annotations were discarded from further analysis. ED notes contained on average the most annotations (5.6), followed by H&P (6.2), Progress notes (4.4), TE (1.8), and Patient Instructions (1.5). The number of annotations varied significantly with note type, but not with the annotator on aggregate ($p=5e-14$ and $p=0.3$ in ANOVA). The number of labels differed between note types (Fig 2). For example, Marital and partnership status and Pain Scores were frequently found in H&P and Progress notes, but rare in ED provider notes; Insurance coverage was mainly present in TE and Patient instructions notes. A detailed breakdown of variance p-value per note type and annotator is presented in Suppl Fig 1B. All categories except Pain Scores, Depression, Finances, and Food were significantly associated with the note type ($p<0.05$ in ANOVA). When annotations were present in a note, the association between the length of note and the number of annotations was significant ($p<0.001$, linear regression $R^2 = 0.52$, Supplementary Fig 1C).

Tuning cTAKES tool for rule-based NER

We tuned and assessed cTAKES for the task of NER of targeted SDoH domains, pain scores and anxiety and depression. As the out-of-the-box version of cTAKES has shown limitations, we used a version of cTAKES configured for high throughput NER at our institution (“InfoCommons” configuration), and additionally customized cTAKES to identify SDoH for the purpose of this study (“SDoH” configuration). Tuning resulted in performance improvement across the three configurations of cTAKES (Fig. 3B, Supplementary Table 7), with weighted F1 achieving 38.95%, 37.95%, and 47.11% on first level for the default, InfoCommons, and SDoH versions respectively. Performance at the final (second) level of taxonomy was poor for all configurations (17.48%, 22.23%, and 34.32%, respectively). The default configuration of cTAKES captured Depression ($F_1 = 70.39%$), Anxiety ($F_1 = 74.46%$), and Marital and Partnership Status (first-level $F_1 = 45.15%$) concepts relatively well (see Fig 3 and Suppl Table 6) but performed poorly on other categories (first-level $F_1<25%$). Additional tuning improved the performance in Marital and Partnership Status (first-level $F_1=63.80%$ for SDoH configuration, $p=1e-05$ compared to the default configuration). However, tuning also lead to a minor degradation of performance for the categories of Anxiety (down to first-level $F_1 = 63.98%$), and Depression (first-level $F_1 = 68.58%$). The observed false negatives (resulting in suboptimal recall, 57.20% at the first taxonomy level for SDoH cTAKES) were commonly due to free-text wording that may not use keywords that cTAKES is configured to look for. Often wrong history or negation modifiers were attributed to the terms due to challenges with sentence boundary segmentation. SDoH words and abbreviations are occasionally misinterpreted as abbreviations from other semantic domains (e.g., “bus” and “bf” in Fig. 4). Additional qualitative analysis is provided in Supplementary Results.

Machine learning based NER prediction with CNN and RoBERTa

Next, we trained and evaluated performance of a CNN NER model implemented using spacy software. CNN NER and RoBERTa NER models achieved an average weighted F1 of 82.65% and 84.38% respectively on first-level and 69.39% and 72.24% on second-level (see Fig 3, Supplementary Tables 6 and 7). On more granular level, RoBERTa model outperformed CNN in most categories, reaching significant level only in category of Depression (second level, $p=0.043$) and approaching significant level for Social Support ($p=0.08$ and $p=0.06$ for first and
second level respectively). Similar to cTAKES, some of the false positives in machine learning results were due to misinterpretation of context. For example, the word “depressed” referencing an anatomical finding, was identified as a mental health symptom. Lack of context sensitivity was a widespread cause of mislabeling of housing related concepts. For example, “plan to discharge home” does not always reflect a patient’s housing status, and “works as a home CNA” does not refer to patient housing status. In some cases, negation markers such as “denies” were not accounted for. Misspellings were also missed, such as “derpession” for “depression.”

Hybrid text extraction and classification system
Next, we applied a hybrid approach, which extracts relevant text segments by pattern matching and classifies with a logistic regression BoW model. The hybrid model displayed a high recall (86.52% at first level), outperforming other methods (Fig 3, Supplementary Table 6, 7, and 8), with an average first-level F₁= 80.27%. Several cases of failure may be attributed to fundamental limitations of the underlying BoW model, such as cases that refer to people other than the patient, e.g., “her father lives with a partner” was interpreted as “Lives with” category pertaining to the patient.

Comparison of NER model performance
The machine learning and hybrid models significantly outperformed the rule-based system in terms of F₁ on the first level (p<1e-5) and the second level (p<0.03) of the taxonomy for all classes except Finances and Food (Fig 3, Suppl. Table 6 and 8.) While the NER-RoBERTa model achieved the highest first-level F₁=84.38% among all models (p<0.04), it scored on par (72.24%) with the hybrid model (68.91%) at the second level F₁ (p=0.13). The hybrid model had a consistently higher recall for all first-level categories but was outperformed by the RoBERTa based NER in terms of first-level F₁ in Marital status (p=0.03) and both first and second level in Social isolation (p<0.001). All NER models performed poorly for the classes of Food and second level domains with low F₁ agreement (35%--60%, Suppl. Table 5), performed poorly in all methods (Suppl. Fig 2), including classes of Depression: NA, Marital or Partnership status: NA, Social Support: NA, Housing: Housing unsuited to needs, Marginally housed. To evaluate contribution of various factors to the model performance, we conducted a multifactor ANOVA of F₁ metric. Both the method and the ontology label are significantly associated with the performance, but not data split fold (p<2e-31 Suppl. Table 9). Additionally, per-class performance was correlated across methods (p<2e-5, Suppl. Table 10). Inter-rater agreement F₁ explains 6.76% of between-label variation in F₁ model performance (p=2e-20 in one-way ANOVA).

Evaluation of transformer-based models for entailment
We hypothesized that general domain NLI models may be directly applied to SDoH and BDoH domains due to semantic overlap between SDoH and general domain texts. Thus, we evaluated performance of an entailment model without fine-tuning, whereby a subset of premise sentences that contained labelled named entities was taken from clinical notes and was passed together with a set of hypothesis sentences compiled to reflect semantics of the respective named entities. The aggregated metrics are shown in Fig. 3 and examples of prediction are shown in Fig 4, while granular metrics are presented in Supplementary Fig 2. Max-aggregation (relying on an assumption that at least one of the hypotheses per class must holds) performed better than mean-aggregation (one assuming all hypotheses must hold), F₁=76.04% vs F₁=30.65% at the first level respectively. The model scored best on categories of Marital Status,
Depression, Anxiety, and Social Support (Fᵢ>70%). The entailment model outperformed the NER models evaluated above in the categories of Transportation, Finances, and Food (Fig. 3). Qualitative analysis reveals sensitivity of the results to hypothesis wording, which in turn may be sensitive to the way the patient is referred to for proper reference resolution. Additionally, entailment models performed poorly on long lists such as lists of family history conditions and lists of diagnosis due to both sentence segmentation issues and difficulties with incomplete sentences. Further qualitative details are provided in Supplementary Results.

Discussion

In this study, we developed a corpus of annotated clinical notes covering a broad spectrum of SDoH domains, together with anxiety, depression, and pain scores for cLBP patients. This corpus provides a basis for training machine learning models and serves as a benchmark for both machine learning and other types of predictive models for NER for SDoH. H&P notes were the richest source of SDoH, while patient instructions contained little relevant information.

By evaluating various NER pipelines, we identified strengths and weaknesses of both rule-based (cTAKES) and machine learning based approaches for identifying SDoH domains. For most of our evaluated categories, the machine learning methods outperformed cTAKES, except for Finances and Food security, domains for which very little training data was available. This may be expected as cTAKES is rule-based and thus does not require training data and relies instead on the richness of synonym vocabularies in NLM Metathesaurus. Some of cTAKES's performance limitations appeared to be due to a failure to account for non-affirmative mentions in questionnaires or missing rare keywords. Inclusion of these keywords may improve cTAKES recall. Additionally, poor cTAKES performance on the second-level taxonomy may be attributed to difficulty detecting nuanced differences, such as distinguishing between symptoms (e.g., anxiety) vs disorders (e.g., generalized anxiety disorder), which depends heavily on the context. Polysemous abbreviations present a challenge even for human intelligibility in general medical practice64, and are even harder for multi-domain rule-based tools such as cTAKES that have no means of factoring in domain-specific contextual knowledge. Additional tuning of the cTAKES model based on the wordings present in annotations obtained in this study may overcome some of these challenges. Due to study resource constraints, we were limited in our ability to continue tuning cTAKES, but our work provides an understanding of the current limitations and shortcomings that may be improved upon.

Comparing all models, we noticed that the detection performance of different SDoH domains was correlated across methods, which suggests that model performance is limited by both fundamental variability of phrasing in free text notes, as well as training data size. Additional analysis showed that inter-rater variability explains this correlation to a small degree. In our dataset, the least documented categories were insurance status, transportation access, food security, and finances, which may be due to a low occurrence of these risk factors in our population; lack of screening for social risk factors; and/or under-documentation of these risks, even when they are known to the health care team. Annotations were also limited by the quality of notes, which could hinder our ability to categorize potentially relevant phrases. Most of these rare categories were detected with low performance. On the other hand, categories of higher frequency such as depression, anxiety, pain scores, as well as marital and partnership status, were the categories detected most accurately by all methods. Machine learning and hybrid
methods outperformed rule-based method (cTAKES) for most classes except Finances and Insurance status which were detected best by the hybrid model. These classes exemplify data scarce situations where rule-based systems may confer advantage\textsuperscript{34,65}, in pure form or as a component of pipeline, as they do not rely on training data. Thus, on average in terms of $F_1$, CNN model performed best, yielding up to the hybrid and rule-based models for a few data-poor classes.

Evaluation of an out-of-the-box entailment RoBERTa\textsuperscript{66} model trained on general domain texts, yielded promising results. Without fine-tuning, this model performed comparably to machine learning models trained on our corpus in most categories. Direct comparison of $F_1$ metric examined in our study (76\%) to the accuracy reported in the original study by Nie and colleagues\textsuperscript{66} (50\%-93\% depending on test set) is not possible, as it would require additional manual annotation to account for cases of contradiction relation. However, both qualitative and quantitative analysis suggests that the model generalizes reasonably well to SDoH domains. As performance of the out-of-the-box entailment model exceeded the characteristics of widely accepted rule-based pipeline (cTAKES), with recall of the former nearing 90\% without fine-tuning, we believe that entailment models may be readily deployed in user-facing tools for clinical text exploration and retrieval, such as EMERSE\textsuperscript{66}, thus allowing researchers to query data based on custom criteria. However, sensitivity of the model to the hypothesis wording observed in this study needs to be taken into consideration when creating fine-tuning data, directly applying to new data, or querying with different hypotheses. Additional fine-tuning on medical- and SDoH-related corpora may further increase performance and should be further explored.

Our study should be interpreted with consideration of its limitations. First, the generalizability of the models validated here may be affected by the fact that clinical notes came from an urban academic medical center, where fewer patients may experience—or fewer providers may be aware to inquire and document—social risk factors, versus the Veterans Affairs or safety-net hospitals, where much of the current SDoH NLP literature is derived. Our study may therefore be better generalizable to other urban academic medical centers than to other settings. This may be a limitation while applying models developed in this study to different kind of institutions, even when they are located in the same geographic area. Related, our study focused on notes for patients with chronic LBP and not other conditions. We included, however, a diverse set of note types from different settings that may improve the generalizability to patients with other conditions. Further validation of the presented models across different medical centers and for patients with different medical conditions is needed to understand their wider applicability. Second, as previously noted, some SDoH domains had particularly low counts of annotations, such as food security and finances. We erred on the side of not making assumptions when annotating text, as notes were not always written clearly due to grammatical errors and/or the shorthand format of notes, which led many annotations to fall under the “NA” (other) category. The strength and capacity of our models overall are limited by the quality of notes that were annotated. Third, dictionaries constructed during tuning of rule-based and hybrid model were composed based on the complete dataset (before annotation), thus, data leakage may not be completely excluded. Lastly, as noted above, our study’s resources limited our ability to continue to fine tune cTAKES and further refine our models overall. Differences in performance between the models may therefore exaggerate the limitations of cTAKES for identifying SDoH as compared to other methods. This study, however, outlines how cTAKES, as well as the other models, can be improved upon.

**Conclusion**
This study is an important step toward understanding the differences between, and strengths and limitations of, a diverse set of NLP methods for detecting SDoH domains within clinical notes. To our knowledge, this is the first study to compare a broad panel of ten domains of SDoH, BDoH, and pain scores with 52 granular sub-domains. This study lays the foundation for better detecting social risk factors in chronic low back pain patients, which can advance our understanding of how social risks impact low back pain treatment access, utilization, and outcomes. Our findings and open-source methods can also be applied to other settings and patient populations, to fuel the growing momentum to apply machine learning techniques to the detection of SDoH in EHRs.

**Acknowledgements**

This work was supported by Back Pain Consortium (BACPAC) grant, UCSF Core Center for Patient-centric Mechanistic Phenotyping in Chronic Low Back Pain (UCSF REACH), UCSF Social Interventions Research and Evaluation Network (SIREN), the Innovate for Health program, including the UC Berkeley Institute for Data Science, the UCSF Bakar Computational Health Sciences Institute and Johnson & Johnson (to BL). We would like to acknowledge Madhumita Sushil and Hunter Mills for their advice and comments on this manuscript, and Sara Jan, Matthew Lee, and Rafael Lee for their work as annotators.
Figure Legends

Figure 1. Study design. A. Workflow of the study. B. Annotation ontology. Clinical notes were annotated such that text relevant to the seven studied social risk factors (solid border) or three clinical factors (dashed border) were marked. Two levels of labels were used, such that the second level was a subcategory of the first. Level 2 labels for each level 1 annotation are shown in descending order of frequency. Level 2 annotations that comprised less than 1% of the group's annotations are not shown. Text that can be classified to the first level but not the second due to ambiguity or low frequency is designated as "NA". Examples of selected text are shown within the hypothetical clinical note.

Figure 2. Exploratory data analysis. A. Histogram of number of entities in different note types. B. Number of entities per note type and first-level annotated domain.

Figure 3. Comparison of model performance. A. Comparison of F1 performance in four best performing models per model class. Second levels metrics are aggregated using weighted average over first level domains. B. Comparison of F1, precision, and recall in all studied models. Metrics are aggregated using weighted average.

Figure 4. Examples of predictions from four best models per model class. Left: NER models. Right: RoBERTA entailment model. Probabilities of three possible relations are shown as shaded horizontal bars and numerically together with a final relation prediction.

References


49. Social Needs Screening Tool Comparison Table | SIREN. https://sirenetwork.ucsf.edu/tools-resources/resources/screening-tools-comparison.


57. Unified Medical Language System (UMLS).


64. Neil M Davis. Medical abbreviations with multiple meanings: A prescription for disaster. 29.


Extracting Social Determinants of Health from Medical Notes of Chronic Lower Back Pain Patients

Figures
Figure 1

A

386 patients 1576 notes

Expert annotation

e1 e2
e3 e4

Curated cohort of cLBP patients

Two-level taxonomy categories: 10 level-1 68 level-2

364 patients 626 notes with annotations

6 fold split

Quantitative Analysis

NER models

Fig. 2

Supplementary Table 1

Entailment

Fig. 3

S. Tables

CNN

Hybrid: matching + BoW

NER models

Entailment

Fig. 4

RoBERTA

Clinical Note

Social History
- Marital status: Married
- Lives part-time in SF with his son
- Emotional and financial stressors have been high the past few months, gluten-free diet, vegetables, protein. About 1 month ago, he missed a step when leaving a bus and fell....
- Pain coincided with increased family stress (more family moved into his one bedroom apartment). Case manager calling for the following reason:
- BP monitor RX was sent to pharmacy - Not covered under pt's insurance
- Relevant medical comorbidities: Depression, Hypercholesterolemia
- Has a past medical history of Ankle sprain, Anxiety, and high cholesterol. Pain is 8/10 on VAS.

Marital and Partnership Status
- Single
- Partner
- Engaged
- Married
- Partner relationship problem
- Divorced
- Separated
- Widowed
- NA

Financials
- Unable to afford medication
- Financial problem
- Financially secure
- NA

Food
- Able to obtain food
- Fruit and vegetable intake
- Provision of food
- NA

Transportation
- Access to a car
- Access to public transport vehicle
- Transportation problems
- NA

Insurance
- Inadequate healthcare resources
- NA

Housing
- Homeless
- Marginally housed
- Housing unsuited to needs
- Subsidized housing
- Lives in facility
- Stably housed

Social Determinants

Social Support
- Has social support
- Lives with
- Lives alone
- At risk for loneliness
- Personal relationship breakdown
- NA

Pain Scores
- Pain intensity rating scale, current
- Pain intensity rating scale, worst
- NA

Anxiety
- Signs and Symptoms of Anxiety
- Generalized Anxiety Disorder
- History of Anxiety State
- Family history of Anxiety
- GAD-7
- NA

Depression
- Symptoms of Depression
- Major Depressive Disorder
- History of Depression
- Family History of Depression
- PHQ-9
- NA
Figure 2. Analysis of Annotations. Note types are associated with different proportions of social domains. The total number of socially relevant spans are shown above each note type bar. History and progress notes, ED provider notes, and progress notes are more likely to have information on marital or partnership status than telephone encounter and patient instruction notes.
Fig 3. Comparison of model performance
Fig 4. Qualitative analysis

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>entailment</th>
<th>neutral</th>
<th>contradiction</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>She broke up with the partner</td>
<td>0.11%</td>
<td>9.96%</td>
<td>99.85%</td>
<td>contradiction</td>
</tr>
<tr>
<td>She came with a family member</td>
<td>1.80%</td>
<td>85.84%</td>
<td>13.34%</td>
<td>neutral</td>
</tr>
<tr>
<td>She experienced a personal relationship breakdown</td>
<td>0.12%</td>
<td>90.62%</td>
<td>9.27%</td>
<td>neutral</td>
</tr>
<tr>
<td>She experiences relationship problems</td>
<td>0.22%</td>
<td>99.22%</td>
<td>0.56%</td>
<td>neutral</td>
</tr>
<tr>
<td>She has a partner</td>
<td>94.80%</td>
<td>4.90%</td>
<td>0.30%</td>
<td>entailment</td>
</tr>
<tr>
<td>She has a significant other</td>
<td>96.88%</td>
<td>3.01%</td>
<td>0.15%</td>
<td>entailment</td>
</tr>
<tr>
<td>She is engaged to be married</td>
<td>0.29%</td>
<td>98.68%</td>
<td>1.04%</td>
<td>neutral</td>
</tr>
<tr>
<td>She is in a relationship</td>
<td>96.02%</td>
<td>3.93%</td>
<td>0.05%</td>
<td>entailment</td>
</tr>
<tr>
<td>She keeps in touch with friends or relatives</td>
<td>1.75%</td>
<td>96.53%</td>
<td>1.72%</td>
<td>neutral</td>
</tr>
<tr>
<td>She lives alone</td>
<td>0.02%</td>
<td>0.09%</td>
<td>99.89%</td>
<td>contradiction</td>
</tr>
<tr>
<td>She lives with someone</td>
<td>83.31%</td>
<td>16.68%</td>
<td>0.12%</td>
<td>entailment</td>
</tr>
<tr>
<td>This describes her marital status</td>
<td>96.03%</td>
<td>3.80%</td>
<td>1.17%</td>
<td>entailment</td>
</tr>
<tr>
<td>This describes her social circumstances</td>
<td>31.38%</td>
<td>38.06%</td>
<td>30.56%</td>
<td>entailment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>entailment</th>
<th>neutral</th>
<th>contradiction</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>She drives a car</td>
<td>0.43%</td>
<td>33.03%</td>
<td>66.55%</td>
<td>contradiction</td>
</tr>
<tr>
<td>She drives a vehicle</td>
<td>12.33%</td>
<td>27.29%</td>
<td>60.39%</td>
<td>contradiction</td>
</tr>
<tr>
<td>She experiences logistic problems</td>
<td>95.65%</td>
<td>4.22%</td>
<td>0.13%</td>
<td>entailment</td>
</tr>
<tr>
<td>She experiences transportation problems</td>
<td>94.85%</td>
<td>4.90%</td>
<td>0.25%</td>
<td>entailment</td>
</tr>
<tr>
<td>She has access to a car</td>
<td>1.64%</td>
<td>54.20%</td>
<td>44.16%</td>
<td>neutral</td>
</tr>
<tr>
<td>She has access to public transportation</td>
<td>98.85%</td>
<td>10.10%</td>
<td>1.05%</td>
<td>entailment</td>
</tr>
<tr>
<td>She has issues getting around</td>
<td>95.15%</td>
<td>34.06%</td>
<td>0.76%</td>
<td>entailment</td>
</tr>
<tr>
<td>She rides public transportation</td>
<td>95.90%</td>
<td>3.86%</td>
<td>0.25%</td>
<td>entailment</td>
</tr>
<tr>
<td>This describes her access to transportation</td>
<td>81.75%</td>
<td>7.44%</td>
<td>2.81%</td>
<td>entailment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>entailment</th>
<th>neutral</th>
<th>contradiction</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient appears depressed</td>
<td>99.76%</td>
<td>0.21%</td>
<td>0.03%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient complains of low mood</td>
<td>99.76%</td>
<td>0.20%</td>
<td>0.03%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient displays symptoms of depression</td>
<td>99.80%</td>
<td>0.17%</td>
<td>0.02%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient feels down</td>
<td>99.85%</td>
<td>0.13%</td>
<td>0.01%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient felt depressed</td>
<td>99.90%</td>
<td>0.08%</td>
<td>0.01%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient had depression earlier</td>
<td>0.04%</td>
<td>99.95%</td>
<td>0.01%</td>
<td>neutral</td>
</tr>
<tr>
<td>Patient is sad</td>
<td>98.85%</td>
<td>2.13%</td>
<td>0.02%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient used to be depressed</td>
<td>0.39%</td>
<td>87.79%</td>
<td>1.80%</td>
<td>neutral</td>
</tr>
<tr>
<td>Patient was diagnosed with MDD</td>
<td>0.03%</td>
<td>99.95%</td>
<td>0.04%</td>
<td>neutral</td>
</tr>
<tr>
<td>Patient was diagnosed with depression</td>
<td>70.21%</td>
<td>29.54%</td>
<td>0.25%</td>
<td>entailment</td>
</tr>
<tr>
<td>Patient was diagnosed with dysthymia</td>
<td>0.03%</td>
<td>99.95%</td>
<td>0.04%</td>
<td>neutral</td>
</tr>
<tr>
<td>Patient was diagnosed with major depressive disorder</td>
<td>0.02%</td>
<td>99.95%</td>
<td>0.03%</td>
<td>neutral</td>
</tr>
<tr>
<td>Patient’s family members had depression</td>
<td>0.02%</td>
<td>99.95%</td>
<td>0.01%</td>
<td>neutral</td>
</tr>
<tr>
<td>Positive for depression</td>
<td>0.21%</td>
<td>26.03%</td>
<td>73.73%</td>
<td>contradiction</td>
</tr>
</tbody>
</table>
Supplementary Figures
Supplementary Fig 1. Annotations
A. Number of annotators per document
B. ANOVA of frequency of annotations note per annotator and note type
C. Scatter plot of number of entities vs note length
Supplementary Figure 2. Granular Level-2 metrics