Discovering Social Determinants of Health from Case Reports using Natural Language Processing: Algorithmic Development and Validation

Shaina Raza [1,2*]; Elham Dolatabadi [1,3], Nancy Ondrusek[1,2], Laura Rosella[2], Brian Schwartz[1,2]

[1] Public Health Ontario (PHO), Toronto, ON, Canada.
[2] Dalla Lana School of Public Health, University of Toronto, Toronto, ON,
Canada.
[3] Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, ON, Canada.

Corresponding Author:
Shaina Raza, PhD
Email: shaina.raza@utoronto.ca,

Abstract

Background: Social determinants of health are non-medical factors that influence health outcomes (SDOH). There is a wealth of SDOH information available via electronic health records, clinical reports, and social media, usually in free text format, which poses a significant challenge and necessitates the use of natural language processing (NLP) techniques to extract key information.

Objective: The objective of this research is to advance the automatic extraction of SDOH from clinical texts.

Setting and Data: The case reports of COVID-19 patients from the published literature are curated to create a corpus. A portion of the data is annotated by experts to create gold labels, and active learning is used for corpus re-annotation.

Methods: A named entity recognition (NER) framework is developed and tested to extract SDOH along with a few prominent clinical entities (diseases, treatments, diagnosis) from the free texts.

Results: The proposed NER implementation achieves an accuracy (F1-score) of 92.98% on our test set and generalizes well on benchmark data. A careful analysis of

case examples demonstrates the superiority of the proposed approach in correctly classifying the named entities.
Conclusions: NLP can be used to extract key information, such as SDOH from free texts. A more accurate understanding of SDOH is needed to further improve healthcare outcomes.

Keywords: Natural language processing; data; COVID-19; named entity; BERT.

INTRODUCTION

Background and significance

Social determinants of health (SDOH) refer to the non-medical factors such as birth, education, occupation, and living conditions, that influence the health outcomes of individuals [1]. There is extensive evidence [2–5] that SDOH significantly influences a broad range of health outcomes including mortality rate, elderly care, mental health, and risks for chronic diseases such as asthma, cancer, heart disease, and obesity [6]. Diabetes, depression, hypertension, and suicidal behavior are all outcomes of SDOH [7]. According to some studies, medical care accounts for 10-20% of health factors, while SDOH accounts for 80% to 90% [8]. Thus, it is of utmost importance to address the SDOH to improve the health systems.

SARS-CoV-2 is a virus that infects humans, causing severe upper respiratory problems and coronavirus disease 2019 (COVID-19) disease [9]. Recent statistics suggest that SDOH, such as race, ethnicity, gender, social-economic factors, and related population characteristics are also among the risk factors for COVID-19 [10]. There is evidence [11,12] that those who are homeless have a higher prevalence of COVID-19 disease than those who are housed. As a result, understanding the SDOH in the context of a pandemic is critical for improving population health outcomes.

Despite advancements in technology, the collection of SDOH data remains a challenge for the public health community. Electronic Health Records (EHRs) constitute a significant portion of clinical data [13], but their use in clinical research is limited due to difficulties in automating the process of extracting information from unstructured data. Natural Language Processing (NLP) techniques can

potentially overcome these limitations by facilitating the extraction of relevant information from unstructured data, thereby enabling its use in analysis and decision-making.

Objectives

The objectives of this study are:

- To create a dataset from case reports of COVID-19 patients, prepare a portion of the data for annotations and apply active learning for data re-annotation.
- To propose an SDOH NLP framework that extracts both SDOH and clinical information from case reports data to provide a more holistic view of the patient's condition and how it is affected by social and economic factors.
- To utilize the key findings from the detailed analysis as an estimate for COVID-19 surveillance.

Through this work, we try to address the research question “How to extract SDOH from the case reports to increase its applicability in practice and study?”. To achieve this, the study proposes an NLP framework that extracts both SDOH and clinical information from case reports data. This work adds to previous efforts [14–17] by including both clinical and SDOH factors as named entities for a more comprehensive understanding of patient clinical outcomes, which is a contribution. The proposed framework includes a data processing module that parses the contents of free text and prepares a dataset, an NLP module that extracts SDOH information from the texts, and an evaluation module that assesses the NLP module's accuracy and effectiveness. The experimental results show that the proposed approach outperforms the baseline methods across multiple datasets. A detailed examination of the information extracted reveals the key findings, which can be used as an estimate for COVID-19 surveillance and to inform public health interventions.

MATERIALS AND METHODS

As part of a research project, we developed an NLP framework and validated the performance of its NLP module for SDOH identification. The proposed framework is extensible and applicable to a variety of public health use cases, including surveillance, epidemiology, and policy-making processes.

Data

We created a data cohort derived from COVID-19 patients' electronic case reports. A Case Report [18] is a publication that describes and interprets an individual case of a patient. These case reports adhere to CARE (CAse REports) principles [19], which state that no patient-identifiable information be included in the case reports. We use the LitCOVID [20] API, which allowed us to extract a variety of published material, such as research papers, scoping reviews, and case reports. We filtered only the case reports during the data acquisition phase.

As such, 4000 case reports were collected with the following inclusion criterion:

- Time period between the 1st of January 2022 and the 30th of June 2022.
- Reports written in English.
- Population in age groups of 6-12 years (child), 13-18 years (adolescent), 19-44 years (adult), 45-64 years (middle-aged), 65+ years (elder).
- Peer-reviewed "Case Reports" (grey literature, preprints, and clinical trial registers are excluded). Because we use the LitCOVID API, which only considers published data, this published material constraint is met during the data acquisition process phase.

Proposed SDOH NLP Framework

The SDOH NLP framework, shown in Figure 1, features three modules: (1) data processing, (2) NLP, and (3) evaluation.

Data processing module

The input to the data processing module is the free text case reports that are acquired through LitCOVID [20] API. These case reports are fed into a scientific parser, where the text contents of the case reports are extracted. These contents are saved in a database, with each row representing a case report content and columns containing the case report identifier.

A gold-standard [21] dataset was created by randomly sampling 200 case reports and having them manually annotated by a group of experts. The Spark NLP annotation tool [22] was used in the annotation process, it was shared

among the annotators. Consistency in the annotation process was ensured by using guidelines for annotating named entities in literature sources [23,24] and a simple inter-annotator agreement (IAA) [25] approach was employed to calculate the percentage of annotations that match among all annotators. Both clinical and SDOH entities were extracted as they co-occur in the text or are associated [26]. A full list of named entities chosen for annotation purposes and further annotation guidelines can be found in Appendix 1, Table S1.

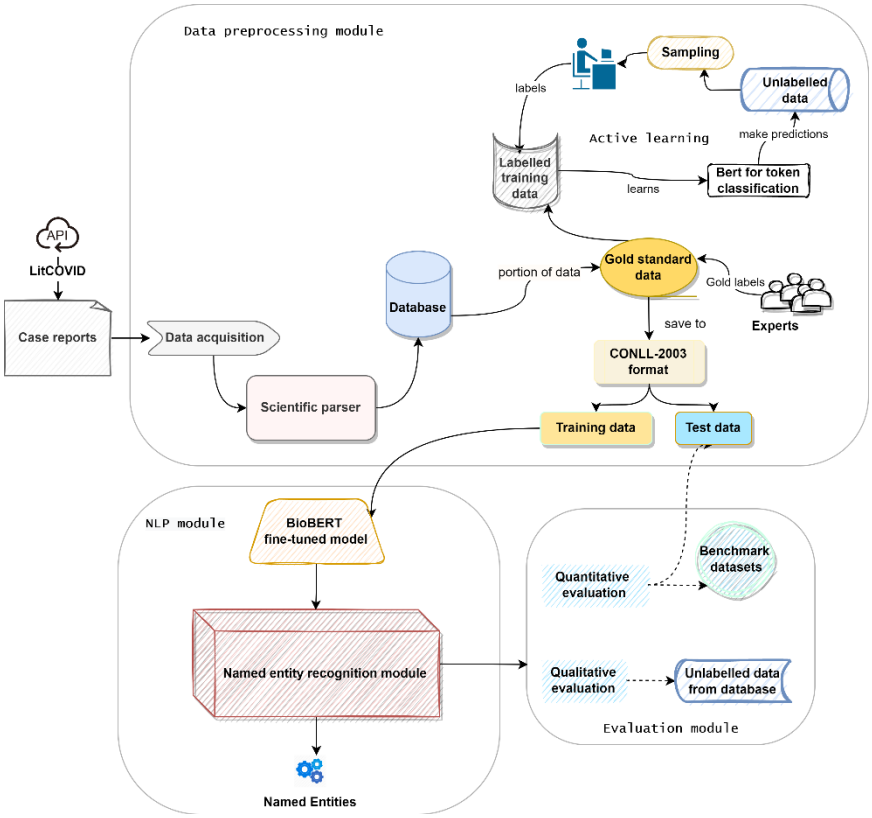


Figure 1: SDOH NLP proposed architecture

In this research project, an active learning technique [27] is employed to address the issue of label shortage. This technique strategically selects samples to be annotated. A BERT[38] model for token classification is trained on the annotated data and applied to the unlabeled data, creating an active learning loop. The loop starts with around 450 sentences from 200 case reports and continues to add batches of 100 samples until reaching approximately 500 samples, with an accuracy of approximately 94.18%. The information is stored in the CONLL-2003 format [28], which

is a commonly used format for NER tasks. By the end of this task, there are approximately 25,000 sentences with 280,000 named entities. The NER model, discussed in the NLP layer, is trained using this annotated data. A portion of the annotated data (30%) is reserved for evaluation purposes.

Natural language processing (NLP) module

The NLP layer of the proposed framework performs the Named Entity Recognition (NER) task as a sequence classification task, where every token in the input sequence is assigned a label. For example, given the sentence "The patient has COVID-19", the output would be "O, I-PERSON, O, I-DISEASE", where "O" indicates a non-entity type and "I-DISEASE" indicates a single-token DISEASE type.

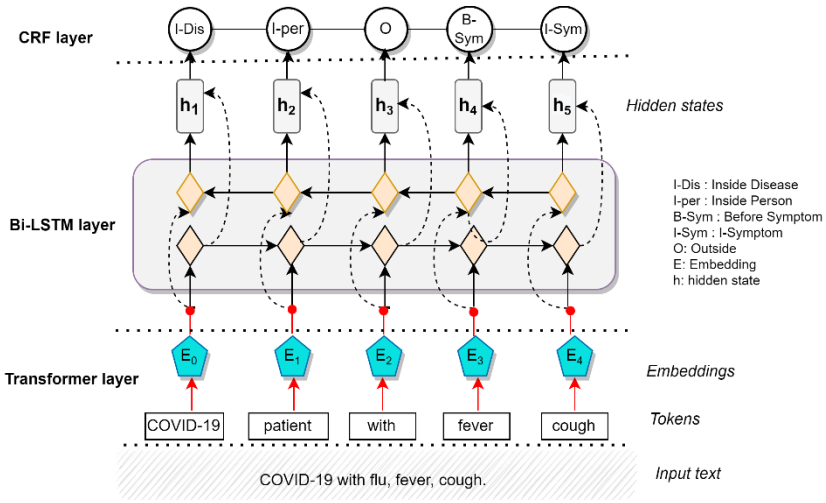


Figure 2: Proposed framework for named entities task.

The proposed NER model is composed of a Transformer layer, a Bidirectional Long Short-Term Memory (BiLSTM), and a Conditional Random Field (CRF) layer [29] as shown in Figure 2. The innovation of the proposed approach is in its subtle blend: the Transformer module converts the input sequence into a comprehensive embedding; the BiLSTM layer learns context features, and the CRF layer captures hidden constraint rules of tags when training data.

The first layer is the Transformer layer, BioBERT [30] is fine-tuned on our annotated dataset and produces contextual representations for each input text data. This fine-tuned model is a task-specific Transformer model that is more robust than task-agnostic pre-trained models and

suffices our purpose. After the Transformer layer is the BiLSTM layer that captures the context features to obtain more comprehensive semantic information from the texts. The BiLSTM network uses the forward and backward LSTM to obtain hidden information in the previous text and the following text. The output sequence of the BiLSTM layer is the input of the CRF layer, which is the last layer in the NLP pipeline. The CRF model captures the dependencies between the output sequence of the BiLSTM layer. The CRF layer captures the dependency relationship between the named tags and constrains them to the final predicted labels [31].

We convert the IOB representation (output of CRF) to a user-friendly format by associating identified named entities with their respective labels. The output of this layer is the named entities (Appendix 1, Table S1) and an example of a case report is shown in Appendix 1, Figure S1.

Evaluation module

In this research, a two-fold evaluation protocol has been adopted to determine the reliability and effectiveness of the NER approach. The performance has been evaluated quantitatively by comparing the method against state-of-the-art baseline models on commonly used biomedical datasets, as shown in Appendix 1, Table S2. Additionally, to measure the performance of the model in extracting SDOH entities, it has been evaluated on the test set, which is 30% of the annotated data. The goal of this evaluation scheme is to measure the correctness of different methods in identifying named entities.

In the qualitative evaluation, the effectiveness of the proposed approach to COVID-19 case reports has been evaluated and a detailed analysis has been provided. This evaluation scheme aims to demonstrate the usefulness of the proposed approach in pandemic surveillance and to provide a deeper understanding of the approach applied to case reports data.

To evaluate the performance of the proposed method, a train-test split strategy is used. The train-test split information from the publication of each dataset is utilized, if available. If there is no official train-test split in the original datasets, a 5-Folds stratified cross-validation strategy is employed. Additionally, a test set consisting of 30% of our annotated data, with ground truth labels, is set aside for evaluation.

The benchmark methods for comparison are categorized as Bi-LSTM-based and Transformer-based models. *BiLSTM-based models* include BiLSTM-CRF [31], BiLSTM-CNN-Char[32], BiLSTM-CRF-MTL[33], Doc-Att-BiLSTM-CRF[34], CollaboNet[35]. *Transformer-based NER models* include BLUE-BERT[36], ClinicalBERT [17], BioBERT [30].

To ensure a fair comparison, all the baselines are optimized to their optimal hyper-parameter settings. Both the BiLSTM-based and Transformer-based baseline methods are fine-tuned on the annotated datasets including our test set before evaluation. Fine-tuning these models on the annotated datasets allows them to learn the specific characteristics and features of the dataset and thus perform better on it. The precision, recall, F1-score, and macro-average F1 are used as evaluation metrics, following the practice in previous works [15,32]. This allows for a comprehensive evaluation of the proposed method against the state-of-the-art baselines, providing a clear picture of its performance and potential impact.

Training details: In this study, the experimental environment utilized is Google Colab Pro, which provides access to cloud-based GPUs (K80, P100, or T4) and 32GB RAM. The use of Google Colab allows for efficient and scalable training of the models, while also providing ample storage for the transfer learning process through its integration with Google Drive. The general parameters used for implementing BiLSTM and Transformer-based architectures are provided in Appendix 1: Table S3. The BERT encoder layers are implemented using the PyTorch BERT implementation available from Huggingface.co. This ensures consistency in results across different runs of the experiments.

RESULTS

Quantitative Analysis

Benchmarking against baselines: The proposed method has been extensively tested and proven to be highly effective in different NER tasks. As seen in Table 1, our method consistently outperforms all baseline methods on a wide range of datasets. This performance gain of our model can be attributed to its unique design, which combines the strengths of sequence labeling through a BiLSTM-CRF architecture and the use of powerful contextualized pre-

trained embeddings from the Transformer architecture. This approach has resulted in a significant improvement of 1% to 5% compared to other methods.

Table 1: Test results in NER task (bold means best and italic as second-best) based on macro-averaged F1-scores.

	NCBI[47]	BC5CDR[48]	tmVar[50]	BC4CHEMD[51]	BC2GM[52]	i2b2-clinical[54]	Our test set
BiLSTM-CRF [31]	85.81	86.18	86.28	89.48	81.08	85.66	87.24
BILSTM-CNN-Char [32]	88.19	*88.58*	87.20	90.06	83.29	84.08	89.25
BiLSTM-CRF-MTL [33]	88.85	84.93	83.35	89.42	82.12	83.25	86.78
Doc-Att-BiLSTM-CRF [34],	88.61	87.33	83.31	88.2	81.80	85.18	86.94
CollaboNet [35]	84.08	87.12	81.75	87.12	79.73	85.61	87.13
BLUE-BERT [36]	88.37	87.62	87.24	90.19	82.93	86.09	88.10
ClinicalBERT [17]	87.01	84.19	79.10	80.13	78.13	84.10	84.93
BioBERT [30]	*89.71*	88.37	*89.10*	*91.08*	*88.72*	*86.73*	*92.34*
Our approach	**90.08**	**89.98**	**89.13**	**91.58**	**89.15**	**89.1**	**92.98**

The proposed method has shown a consistent pattern of performance on a variety of datasets including NCBI-disease, tmVar, BC4CHEMD, BC2GM, JNLBPA, BC5CDR, and i2b2-clinical. These results probably indicate that the use of transformer layers in inferring contextualized representations has a positive impact on the performance of NER tasks. On the BC5CDR dataset, our method obtained a macro-averaged F1-score of 89.98%, which is higher than the score of 88.58% achieved by the BiLSTM-CNN-Char method. This demonstrates that while simpler deep neural networks can be useful, contextualized embeddings, as used in our model, greatly enhance the performance of the extraction task. Additionally, on the i2b2 dataset, annotated with clinical named entities, our method and the BioBERT method have been shown to perform well. This suggests that the entity types on which a method is pre-trained, typically perform well, and overall, our model has been observed to be versatile and apply to a variety of entity types, performing best when trained on a dataset with a variety of entity types and also when trained on a different dataset with different entity types.

To evaluate the performance of our proposed Transformer-BiLSTM-CRF model, we conducted a comparison against the best-performing baseline model, BioBERT, using the McNemar test [37]. The experiment was conducted using gold labels for our test set, with both models trained on the training set. We set the significance level to 0.05 and the degree of freedom to 1. The results of the McNemar test (shown in Appendix 1: Table S4) revealed

a statistically significant difference in performance between our proposed model and BioBERT, as indicated by a p-value of less than 0.05. This confirms that our Transformer-BiLSTM-CRF model significantly outperforms the BERT-like model.

Performance analysis on medical and social determinants of health

The fine-grained performance of our proposed NLP approach in extracting 10 SDOH entity classes from our data is shown in Table 2.

Table 2: Performance of our approach in extracting the most occurring (greater than 90%) SDOH terms from our test set.

Entity	Precision	Recall	F1
Disease	0.862	0.746	0.800
Gender	0.832	0.784	0.808
Employment	0.785	0.777	0.781
Relationship_Status	0.789	0.753	0.770
Smoking	0.837	0.709	0.767
Race_Ethnicity	0.917	0.698	0.793
Psychological_Condition	0.820	0.775	0.797
Death_Entity	0.919	0.790	0.850
BMI	0.924	0.834	0.877
Diabetes	0.937	0.841	0.886
Macro-average	0.862	0.771	0.813
Micro-average	0.841	0.767	0.802

Though the performance varies across entities, the model performed best in the categories of diabetes, death, psychological condition, BMI, Diabetes, and gender (F1>80%).

Error analysis

In this sub-section, we provide three running examples to further demonstrate the efficacy of our SDOH extraction approach.

Table 3: Running examples, each example contains a sentence extracted from a case report and ground-truth labels. For each sentence, there are 3 sets of predicted labels generated by the top two best-performing models, BiLSTM-CNN-Char and BioBERT, as well as our proposed NLP model. A NER task presents three major challenges: (i) identifying a long entity without breaking part, (ii) failing to identify an entity type, and (iii) misclassifying a non-entity.

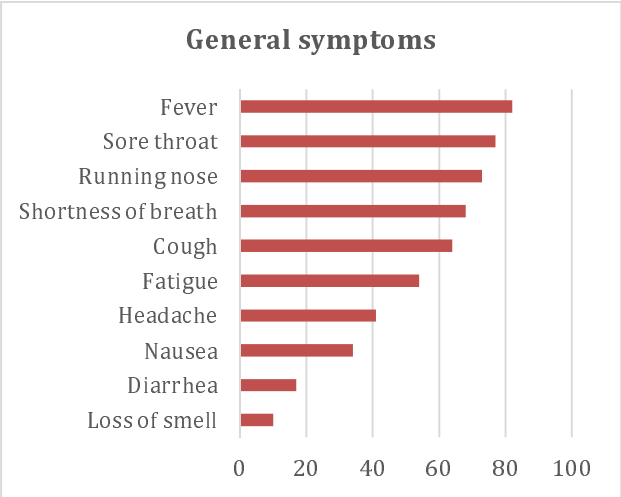
Diseases/Symptoms

Case 1	True label	We report a 23-year-old man who developed, after asymptomatic COVID-19, a complex symptomatology characterized by fatigue, episodic shortness of breath, nocturnal tachycardia, and chest pain.
	BILSTM-CNN-Char	We report a 23-year-old man who developed, after asymptomatic COVID-19, a complex symptomatology characterized by fatigue, episodic shortness of breath, nocturnal tachycardia, and chest pain. *Error:* Missing a few named entities (‘nocturnal tachycardia). Missing a part from long entity chunk (‘episodic shortness of breath’).Identifying a wrong entity chunk (‘symptomatology’) as symptom/disease.
	BioBERT	We report a 23-year-old man who developed, after asymptomatic COVID-19, a complex symptomatology characterized by fatigue, episodic shortness of breath, nocturnal tachycardia, and chest pain. *Error:* Missing a named entity (‘COVID-19’). Breaking a long entity chunk (‘episodic shortness of breath’) into parts and identifying separately. Missing a part from long chunk (‘nocturnal tachycardia).
	Our approach	We report a 23-year-old man who developed, after asymptomatic COVID-19, a complex symptomatology characterized by fatigue, episodic shortness of breath, nocturnal tachycardia, and chest pain.
Social Determinants of Health		
Case 2	True label	'A 13-year-old Caucasian girl was presented to the emergency department of her local hospital in December 2020 during the COVID-19 pandemic.
	BILSTM-CNN-Char	'A 13-year-old Caucasian girl was presented to the emergency department of her local hospital in December 2020 during the COVID-19 pandemic. *Error:* Missing a few named entities (‘girl’, ‘her’, ‘local hospital’). Unable to detect a part of long entity chunk (‘December 2020’).
	BioBERT	'A 13-year-old Caucasian girl was presented to the emergency department of her local hospital in December 2020 during the COVID-19 pandemic. *Error:* Missing a few named entities (‘Caucasian’, ‘emergency department’, ‘local hospital’).
	Our approach	'A 13-year-old Caucasian girl was presented to the emergency department of her local hospital in December 2020 during the COVID-19 pandemic.
Case 3	True label	There were 68% more men than women identified as being unemployed, with more men claiming benefits.
	BILSTM-CNN-Char	There were 68% more men than women identified as being unemployed, with more men claiming benefits. *Error:* Missing a few named entities (‘68%’, ‘unemployed’).
	BioBERT	There were 68% more men than women identified as being unemployed, with more men claiming benefits. *Error:* Missing a few named entities (‘68%’, ‘unemployed’).
	Our approach	There were 68% more men than women identified as being unemployed, with more men claiming benefits. *Error:* Missing a named entity (‘68%’.)

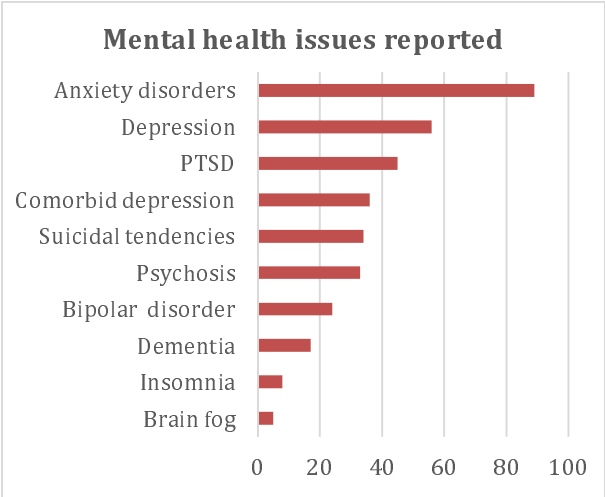
Overall, we see in these examples in Table 3 that BILSTM-CNN-Char and BioBERT missed a few named entities, but BioBERT did not misclassify any. Our model did not misclassify or miss any part from the long entities. However, in one case, our model failed to predict one named entity which may be due to the lack of full context.

Qualitative analysis to see model effectiveness

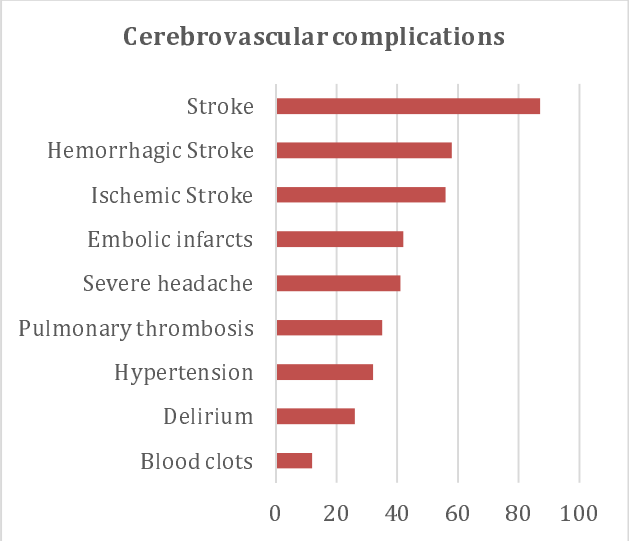
In this section, we see the effectiveness of the proposed method in inferring the named entities from case report data. The prevalence of common COVID-19 symptoms and conditions reported in the case reports is depicted in Figure 3.



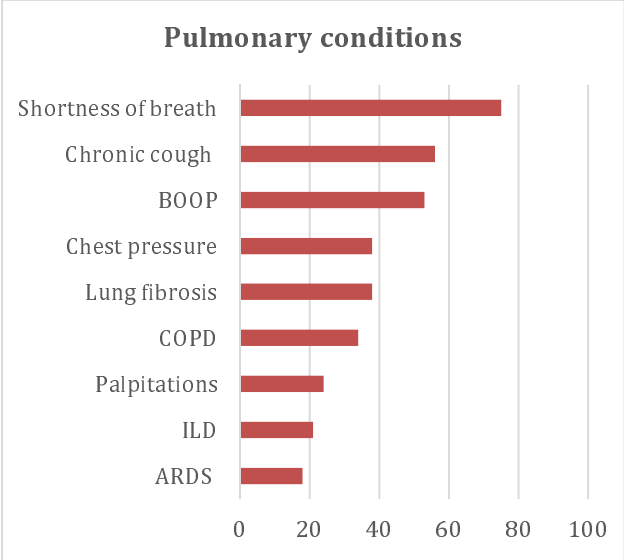
(a) General symptoms of COVID-19 in patients.



(b) Mental health issues reported.



(c) Cerebrovascular complications reported.

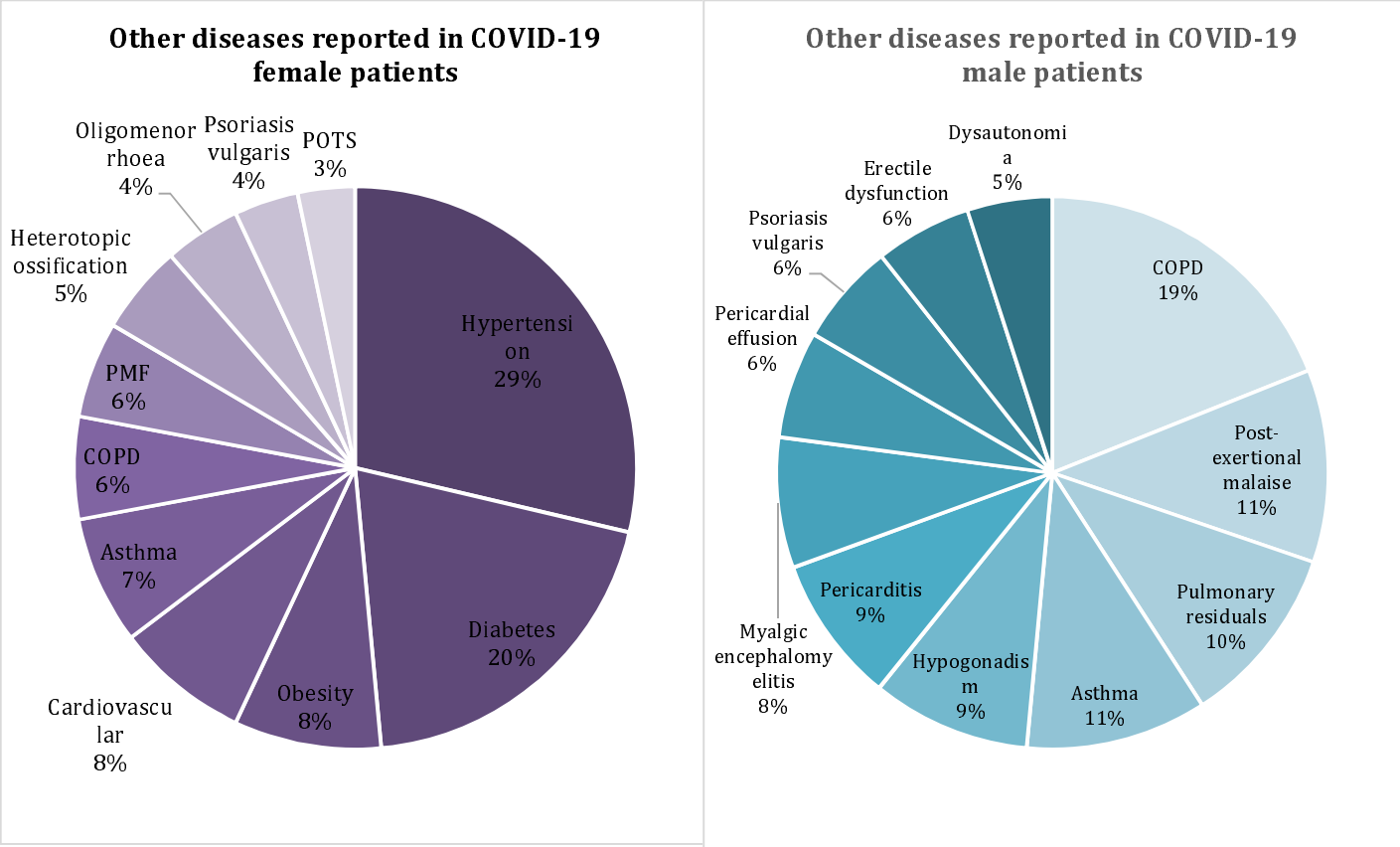


(d) Pulmonary conditions reported.

Figure 3: Symptom prevalence estimates (occurring more than 70%) in COVID-19 patients, divided into general symptoms, mental health, and cerebrovascular and pulmonary (based on common complications reported). Each bar represents the percentage of respondents who experienced that symptom. Some acronyms used are post-traumatic stress disorder (PTSD), distress syndrome (ARDS), bronchiolitis

obliterans organizing pneumonia (BOOP), chronic obstructive pulmonary disease (COPD), and interstitial lung disease (ILD).

Next, we present the prevalence of common disease disorders in COVID-19 patients for both female and male groups in Figure 4. This result allows for a more detailed analysis of how gender may affect the likelihood of certain disease disorders in COVID-19 patients. It is worth noting that these are generalizations and individual cases may vary. Some studies suggest that male COVID-19 patients may have a higher risk of severe illness or hospitalization [38], but more research is needed to confirm this.

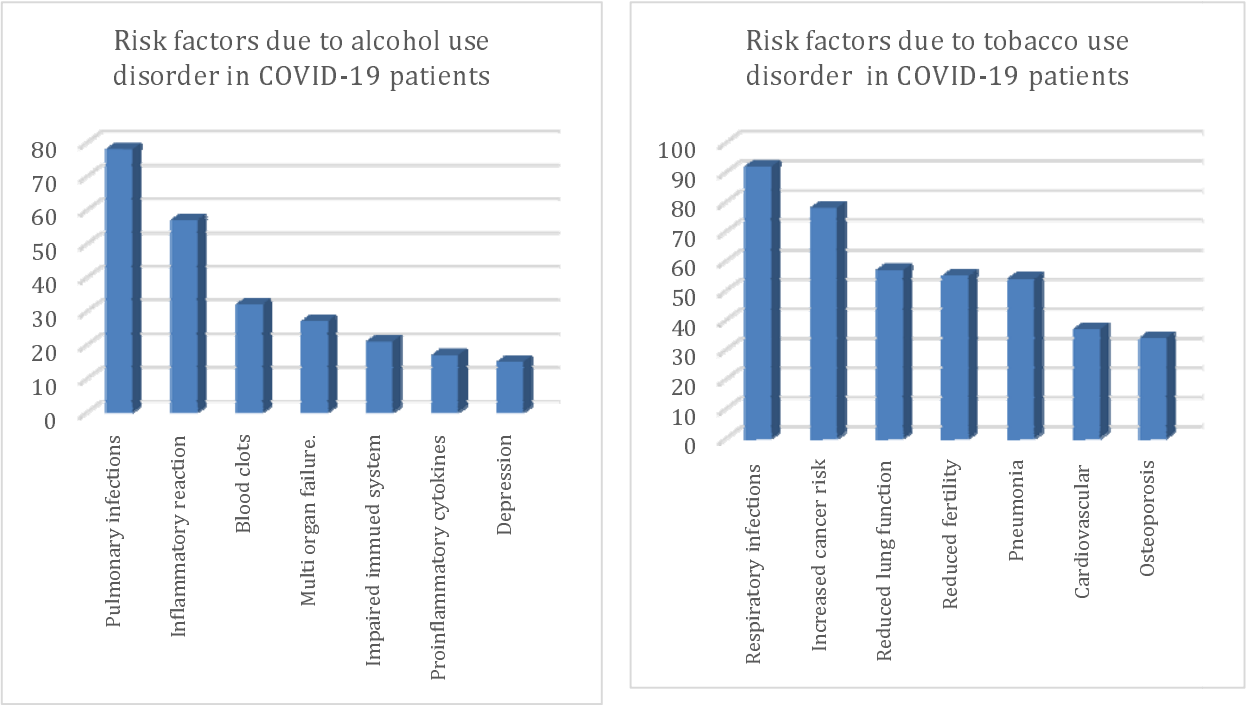


(a) Diseases reported in COVID-19 female patients (b) Diseases reported in COVID-19 male patients

Figure 4: Common diseases in COVID-19 patients to gender.

Next, we present the prevalence of disease disorders in COVID-19 patients split across two substance (alcohol, tobacco) disorders in Figure 5. This allows for a more detailed analysis of how these disorders may affect the likelihood of certain disease disorders in COVID-19 patients. Smoking and heavy alcohol consumption have been shown to increase the risk of pulmonary, cardiovascular,

cerebrovascular, cancer and related diseases in COVID-19 patients, as depicted in Figure 5 results. Studies [39] have also shown that smokers and heavy alcohol consumers are more likely to be hospitalized and require intensive care for COVID-19 scompared to non-smokers and moderate alcohol consumers, and may have a harder time recovering from the virus. This highlights the importance of considering these factors in the management and treatment of COVID-19 patients.



(a) Risk factors due to alcohol overuse

(b) Risk factors due to tobacco overuse

Figure 5: Frequency occurrences of COVID-19 symptoms in patients with substance use disorder. The bars in the charts represent the prevalence of disease disorders among COVID-19 patients who have also been reported with substance use disorders.

Figure 6 depicts the percentage of COVID-19 patients across different race-based categories. We take the nominal race groups [40] and report the results where the race group is reported to account for more than 5% of the population. Overall, we see that in our study COVID-19 has the greatest impact on the Black community. This finding, however, is based on a subset of available case reports from a specific time period, so it may not be an accurate representation of whole population during the COVID-19 outbreak. We also

show the hospitalization, ICU and morality rate in COVID-19 patients with different age groups in Appendix 1, Figure S2.

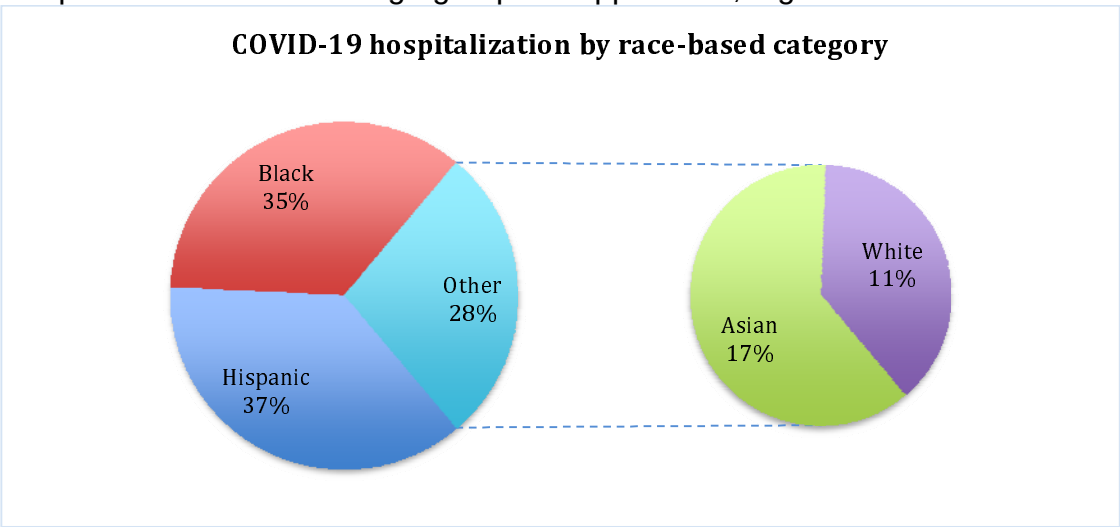


Figure 6: COVID-19 hospitalization by race and ethnicity.

DISCUSSION

Previous works: Previous works have extracted SDOH information from clinical data using different methods such as regular expressions, dictionaries, rule-based methods like cTAKES [41,42], and deep neural networks like CNNs, LSTMs [43], and Transformer-based methods [42]. Language model-based representations have been found to perform well, especially with large training sets. However, even simpler neural network representations like BiLSTM can perform well with enough training data, being only slightly worse than Transformer-based models. It is important to use a model type with sufficient expressive power when using simpler neural network representations.

Practical impact for enhancing clinical knowledge and patient care: This research aimed to investigate the potential of using NLP models on case report datasets to improve clinical knowledge and patient care. Through the use of advanced NLP techniques, we were able to extract important information from case reports, including symptoms, diagnosis, and treatment, as well as information about SDOH. This information can be used to inform targeted interventions and deliver personalized and evidence-based care. The use of NLP models also allows for automation of the process of extracting information from case reports, saving time and resources for healthcare professionals.

Such a framework can be integrated into clinical decision support systems to improve the quality of care.

Limitations: The study acknowledges its limitations and provides opportunities for future research. One limitation is that the dataset may not be representative of all SDOH impacting patients with COVID-19 and publication bias favoring positive outcomes [44,45] may have an impact on the types and frequencies of SDOH observed. To address this, future research could aim to implement EHRs and clinical notes that are updated in real-time. Additionally, data privacy concerns will need to be addressed by masking named entities associated with patients' personal information, such as names, locations, and identifiers.

A case report may not always describe a patient's current condition. For instance, "the patient has a family history of hysteria" can be classified as a psychiatric condition of the patient, even though it is not the patient's current condition. To address this, the annotation scheme could be extended and the rules or semantics defined more clearly, allowing for the retraining of language models. However, this would be a labor-intensive process.

There are several ways to further extend this research. One way is to use an extensive active learning approach [46] to improve model performance. Another direction is to use prompt-based learning, which utilizes the strengths of pre-trained foundation models [47], to improve overall effectiveness. Additionally, experimenting with different model architectures [48] and performing detailed significance tests may also add value to this work. By addressing these limits, we believe that this work will lead to an effective and general-purpose model.

CONCLUSIONS

This study demonstrates that NLP-based methods can be used to identify clinical entities and SDOH from texts. The proposed framework uses a combination of neural networks. To assess the model ability to extract different named entities, a detailed analysis of the SDOH is performed. The proposed methods outperform the state-of-the-art methods by about 1-5% on the macro-averaged F1 score for the NER task and showed effectiveness in determining clinical outcomes. The current study paves the way for future

research and addresses the health disparities that appear in systematic healthcare systems.

CONFLICTS OF INTEREST

None declared.

FUNDING STATEMENT

The authors have received no specific funding for this work.

AUTHOR CONTRIBUTIONS

SR, ED, NO, BS, LR conceived the study design. SR and BS participated in the literature search. SR and BS prepared the search query for the data collection. SR performed the data curation, data pre-processing, and dataset preparation. SR built the framework and the models, and ED validated the framework. SR created the tables, plotted the graphics, and ED and NO interpreted the study findings. SR drafted the initial manuscript. BS and NO validated the results and evaluated the findings and revised the draft. All authors critically reviewed and substantively revised the manuscript. All authors have approved the final version of the manuscript for publication.

SUPPLEMENTARY MATERIAL

Supplementary material is attached.

DATA AVAILABILITY STATEMENT

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1 Who WHO. Social determinants of health. *World Heal Organ WHO* Published Online First: May 2019.https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1

2 Flaskerud JH, DeLilly CR, Flaskerud JH. Social determinants of health status. *Issues Ment Health Nurs* 2012;**33**:494–7.

3 Chang CD. Social determinants of health and health disparities among immigrants and their children. *Curr Probl Pediatr Adolesc Health Care* 2019;**49**:23–30.

4 Donkin A, Goldblatt P, Allen J, *et al.* Global action on the social determinants of health. *BMJ Glob Heal* 2018;**3**:e000603.

5 Braveman P, Gottlieb L. The social determinants of health: It's time to consider the causes of the causes. *Public Health Rep* 2014;**129**:19–31. doi:10.1177/00333549141291s206

6 Brakefield WS, Ammar N, Olusanya OA, *et al.* An urban population health observatory system to support COVID-19 pandemic preparedness, response, and management: Design and development study. *JMIR Public Heal Surveill* 2021;**7**:1–16.

doi:10.2196/28269
7 Whitman A, De Lew N, Chappel A, *et al.* Addressing Social Determinants of Health: Examples of Successful Evidence-Based Strategies and Current Federal Efforts. *Off Heal Policy* 2022;:1–30.
8 Williams DR, Costa M V., Odunlami AO, *et al.* Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities. *J Public Health Manag Pract* 2008;**14 Suppl**:S8--S17. doi:10.1097/01.PHH.0000338382.36695.42
9 Intiso D, Marco Centra A, Giordano A, *et al.* Critical illness polyneuropathy and functional outcome in subjects with COVID-19: Report on four patients and a scoping review of the literature. *J Rehabil Med* Published Online First: 2021. doi:10.2340/jrm.v53.1139
10 Singu S, Acharya A, Challagundla K, *et al.* Impact of social determinants of health on the emerging COVID-19 pandemic in the United States. *Front public Heal* 2020;:406.
11 Baggett TP, Keyes H, Sporn N, *et al.* Prevalence of SARS-CoV-2 infection in residents of a large homeless shelter in Boston. *Jama* 2020;**323**:2191–2.
12 Kiran T, Craig-Neil A, Das P, *et al.* Factors associated with SARS-CoV-2 positivity in 20 homeless shelters in Toronto, Canada, from April to July 2020: a repeated cross-sectional study. *C open* 2021;**9**:E302–8. doi:10.9778/cmajo.20200253
13 Tarcar AK, Tiwari A, Rao D, *et al.* Healthcare NER models using language model pretraining. *CEUR Workshop Proc* 2020;**2551**:12–8. doi:10.1145/3336191.3371879
14 Wang Y, Sohn S, Liu S, *et al.* A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019;**19**:1–13. doi:10.1186/s12911-018-0723-6
15 Lee J, Yoon W, Kim S, *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40. doi:10.1093/bioinformatics/btz682
16 Doğan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J Biomed Inform* 2014;**47**:1–10. doi:https://doi.org/10.1016/j.jbi.2013.12.006
17 Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv Prepr arXiv190405342* Published Online First: 2019.http://arxiv.org/abs/1904.05342
18 Rison RA, Shepphird JK, Kidd MR. How to choose the best journal for your case report. J. Med. Case Rep. 2017;**11**:1–9.
19 IMI. CARE case report guidelines. CARE Case Rep. Guidel. 2019.https://www.care-statement.org/
20 Chen Q, Allot A, Lu Z. LitCovid: An open database of COVID-19 literature. *Nucleic Acids Res* 2021;**49**:D1534–40. doi:10.1093/nar/gkaa952
21 Cardoso JR, Pereira LM, Iversen MD, *et al.* What is gold standard and what is ground truth? *Dental Press J Orthod* 2014;**19**:27–30.
22 Spark NLP 4.0.2 ScalaDoc - com.johnsnowlabs.nlp.annotators.parser.dep.DependencyParser

Model. 2022.https://nlp.johnsnowlabs.com/api/com/johnsnowlabs/nlp/annotators/parser/dep/DependencyParserModel.html

23 Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A, *et al.* A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Med Inform Decis Mak* 2021;**21**:1–19. doi:10.1186/s12911-021-01395-z

24 Nothman J, Ringland N, Radford W, *et al.* Learning multilingual named entity recognition from Wikipedia. *Artif Intell* 2013;**194**:151–75.

25 Artstein R. Inter-annotator agreement. In: *Handbook of linguistic annotation*. Springer 2017. 297–313.

26 Han S, Zhang RF, Shi L, *et al.* Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J Biomed Inform* 2022;**127**:103984. doi:10.1016/j.jbi.2021.103984

27 Chen Y, Lasko TA, Mei Q, *et al.* A study of active learning methods for named entity recognition in clinical text. *J Biomed Inform* 2015;**58**:11–8. doi:10.1016/j.jbi.2015.09.010

28 Tjong Kim Sang EF, de Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proc 7th Conf Nat Lang Learn CoNLL 2003 HLT-NAACL 2003* 2003;:142–7.

29 Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. 2015. doi:10.48550/ARXIV.1508.01991

30 Lee J, Yoon W, Kim S, *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40. doi:10.1093/bioinformatics/btz682

31 Lample G, Ballesteros M, Subramanian S, *et al.* Neural architectures for named entity recognition. *arXiv Prepr arXiv160301360* 2016.

32 Chiu JPC, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans Assoc Comput Linguist* 2016;**4**:357–70. doi:10.1162/tacl_a_00104

33 Wang X, Zhang Y, Ren X, *et al.* Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* 2019;**35**:1745–52.

34 Xu K, Yang Z, Kang P, *et al.* Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Comput Biol Med* 2019;**108**:122–32.

35 Yoon W, So CH, Lee J, *et al.* Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* 2019;**20**:55–65.

36 Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv Prepr arXiv190605474* 2019.

37 Gillick L, Cox SJ. Some statistical issues in the comparison of speech recognition algorithms. In: *International Conference on Acoustics, Speech, and Signal Processing,*. 1989. 532–5.

38 Abate BB, Kassie AM, Kassaw MW, *et al.* Sex difference in coronavirus disease (COVID-19): a systematic review and meta-analysis. *BMJ Open* 2020;**10**:e040129.

39 Hamer M, Kivimäki M, Gale CR, *et al.* Lifestyle risk factors, inflammatory mechanisms, and COVID-19 hospitalization: A community-based cohort study of 387,109 adults in UK. *Brain Behav Immun* 2020;**87**:184–7.
40 Abdi S, Bennett-AbuAyyash C, MacDonald L, *et al.* Provincial implementation supports for socio-demographic data collection during COVID-19 in Ontario's public health system. *Can J Public Heal* 2021;**112**:853–61. doi:10.17269/s41997-021-00551-2
41 Chen ES, Carter EW, Sarkar IN, *et al.* Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. In: *AMIA Annual Symposium Proceedings*. 2014. 366.
42 Lituiev D, Lacar B, Pak S, *et al.* Automatic Extraction of Social Determinants of Health from Medical Notes of Chronic Lower Back Pain Patients. *medRxiv* 2022;:2022.03.04.22271541.http://medrxiv.org/content/early/2022/03/08/2022.03.04.22271541.abstract
43 Quan W, Chen Z, Gao J, *et al.* Comparative study of CNN and LSTM based attention neural networks for aspect-level opinion mining. In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018. 2141–50.
44 Easterbrook PJ, Gopalan R, Berlin JA, *et al.* Publication bias in clinical research. *Lancet* 1991;**337**:867–72.
45 Nissen T, Wynn R. The clinical case report: a review of its merits and limitations. *BMC Res Notes* 2014;**7**:1–7.
46 Ménard PA, Mougeot A. Turning silver into gold: error-focused corpus reannotation with active learning. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019. 758–67.
47 Bommasani R, Hudson DA, Adeli E, *et al.* On the Opportunities and Risks of Foundation Models. 2021;:1–214.http://arxiv.org/abs/2108.07258
48 Davidson EM, Poon MTC, Casey A, *et al.* The reporting quality of natural language processing studies: systematic review of studies of radiology reports. BMC Med. Imaging. 2021;**21**. doi:10.1186/s12880-021-00671-8