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@oahpp.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 texts 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. The proposed model consists of three deep neural networks – A Transformer-based model, a BiLSTM model and a CRF module.
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 motivation

Social determinants of health (SDOH) refer to the non-medical factors such as birth, education, occupation, living conditions, that influence the health outcomes of individuals[1]. There is extensive evidence [2–5] that SDOH significantly influence 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]. Some studies suggest that SDOH accounts for 35-50% of health outcomes[1],[7]. Furthermore, medical care accounts for 10-20% of health factors, while SDOH account for 80% to 90% of health factors [8]. Therefore, it is important to address the SDOH to improve health and delivery of care.

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]. For example, there is evidence[11,12] that those who are homeless had a greater frequency of COVID-19 disease than people who are housed. Therefore, it is important to understand the SDOH in the context of pandemic to improve population health outcomes.

Despite significant advances in technology, a large amount of SDOH data is not systematically collected[13], making it difficult for the public health community to address SDOH. For example, Electronic Health Records (EHRs), which make up about 70% of clinical data[14], are rarely used in clinical research due to limitations in automated data extraction processes. This research is motivated by the fact that natural language processing (NLP) can accelerate the extraction of information from unstructured clinical data, which can be used for analysis and decision-making. 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?"

Contributions

We summarize our contributions as:

1. A dataset is created from the case reports of COVID-19 patients. A portion of the data is prepared for annotations, and the application of active learning is explored for data re-annotation.
2. A SDOH NLP framework is proposed to extract SDOH from clinical texts. The 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.

3. A two-phase evaluation protocol is used, with the first phase evaluating the proposed framework's performance for accuracy and the second phase assessing the method's performance for effectiveness.

Many experiments are carried out to answer the research question stated above. When compared to other methods, the proposed method achieves an F1-score of 92.98% on our test set and outperforms other methods by about 1-5% on benchmark datasets. A detailed analysis is performed based on the identification of SDOH to demonstrate the key findings, which can be used as an estimate for COVID-19 surveillance.

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. Our 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[15] is a publication that describes and interprets an individual case of a patient. These case reports adhere to CARE (CAse REports) principles [16], which state that no patient-identifiable information be included in the case reports. We use the LitCOVID[17] 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[17] 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 group of experts manually annotated a subset of the data (200 case reports) which was sampled randomly to create our gold-standard dataset[18]. Clinical and SDOH entities co-occur in the text (or are associated) [19,20]. As such, we extracted both types of entities in this study (the full list of named entities chosen for this study for annotation purposes are given in Appendix 1, Table S1).

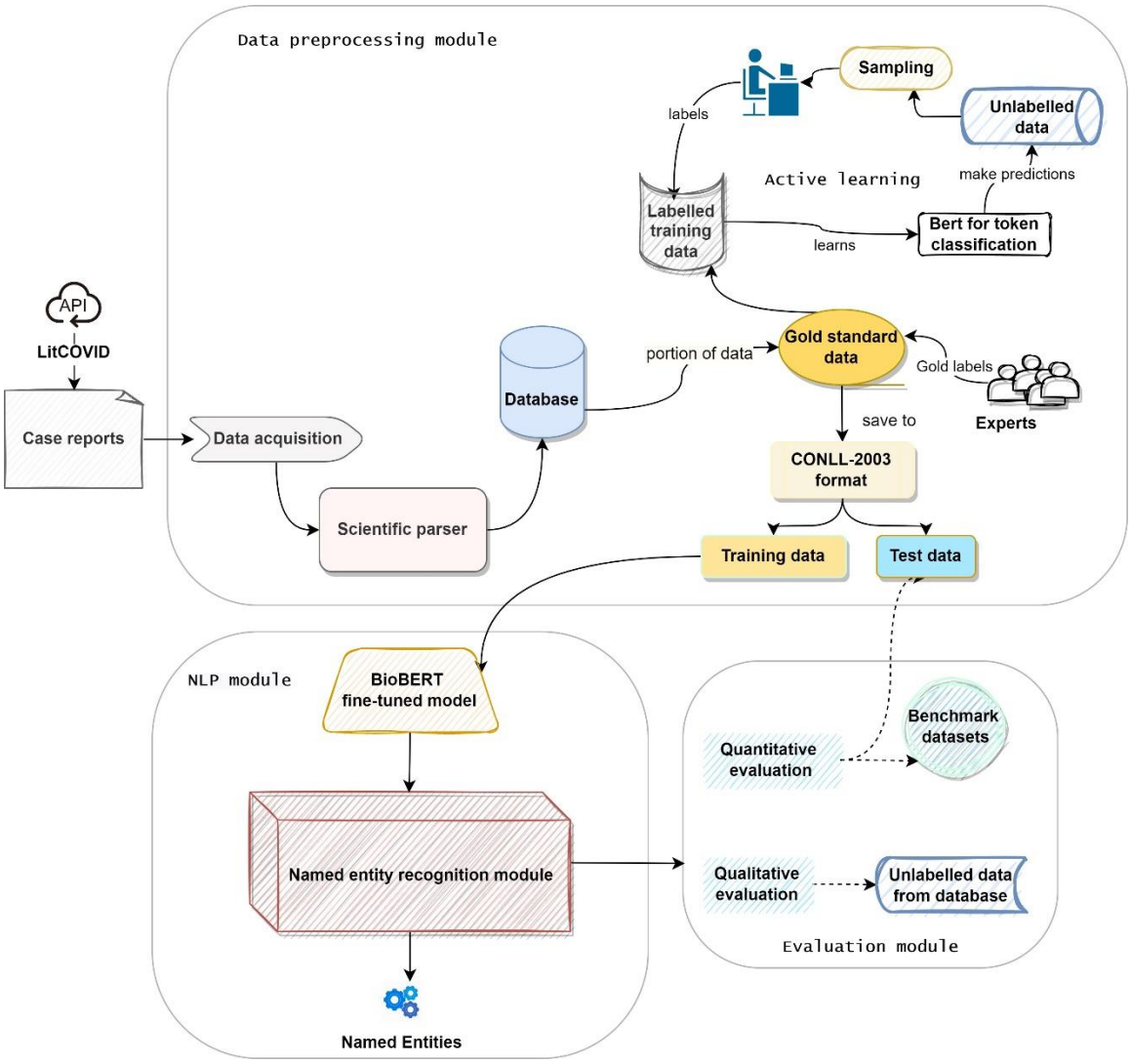


Figure 1: SDOH NLP proposed architecture

The active learning technique [21] is used to alleviate the label shortage problem by strategically selecting which samples to annotate. BERT [38] for token classification is trained on the annotated data and run on the unlabeled data. This active learning loop begins with approximately 450 sentences (from 200 case reports) and we continue to add batches of 100 samples (reports) until it reaches approximately 500 samples,

with an accuracy of approximately 94.18%. We prepared and stored this information in CONLL-2003 format[22], a prototypical format for the NER tasks [22]. By the end of this task, we had approximately 25K sentences with 280K named entities. We train our NER model (discussed in the NLP layer) using this annotated data. A portion of the annotated data (30%) is reserved for evaluation purposes.

Natural language processing (NLP) module

This layer performs the NER task framed as a sequence classification task where every token in the input sequence is assigned to a label. For example, given the sentence "The patient has COVID-19", the output should be "O, I-PERSON, O, I-DISEASE" in which "O" indicates a non-entity type and "I-DISEASE" indicates a single-token DISEASE type.

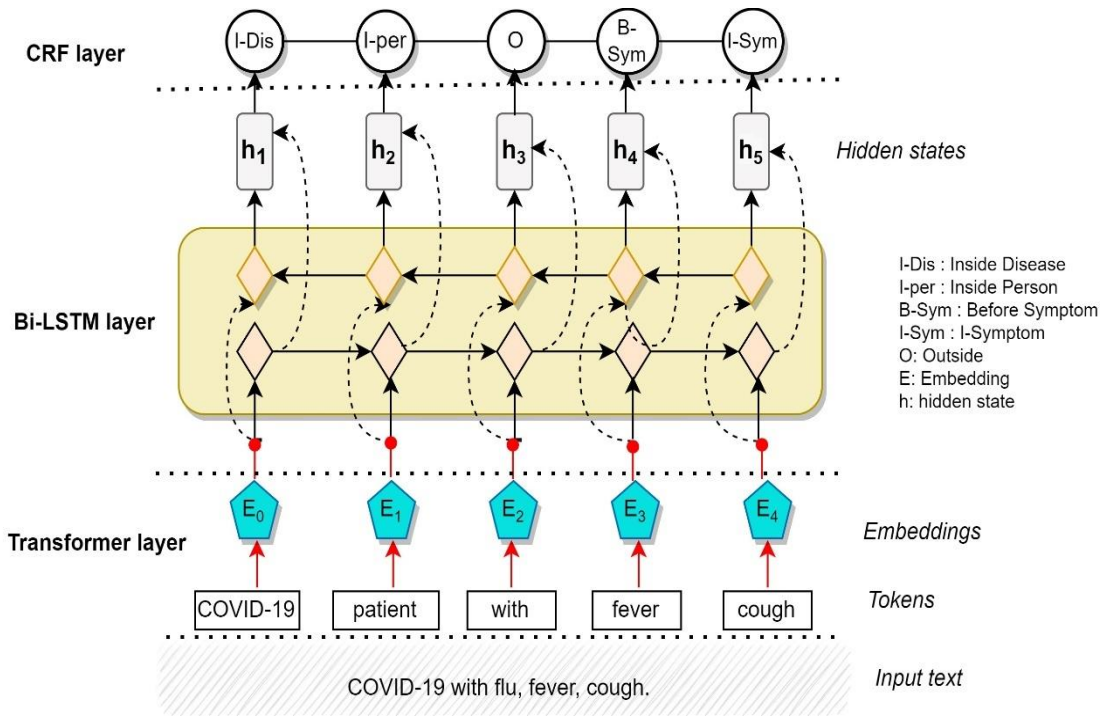


Figure 2: Proposed framework for named entities task.

The proposed NER model is composed of a Transformer layer, a Bidirectional Long Short-Term Memory (BiLSTM), and a Conditional Random Field (CRF) layer [23] as shown in Figure 2 . The novelty of the proposed approach lies in its subtle blend: the Transformer module to convert the input sequence into a comprehensive embedding; the BiLSTM layer to learn context features, and the CRF layer module to obtain hidden constraint rules of tags when training data.

The first layer is the *Transformer layer* that is BioBERT[24] fined-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 that task-agnostic pre-trained models and suffice our purpose. Following 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 the hidden information in the previous text and the following text. The output sequence of BiLSTM layer is the input of the *CRF layer* which is the last layer in the NLP pipeline. The CRF model's function is to capture 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[25].

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

We adopt a two-fold evaluation protocol. To determine the reliability and effectiveness of our NER approach, we evaluated its performance quantitatively and qualitatively. For quantitative evaluation, the performance of our approach has been compared against the state-of-the-art baseline models on commonly used biomedical datasets listed in Table 1. Additionally, to measure how good our model is in extracting SDOH entities, we evaluated its performance on our test set which is 30% of our annotated data. The goal of this evaluation scheme is to measure the correctness of different methods in identifying the named entities.

In the qualitative evaluation, we evaluate the effectiveness of our proposed approach to the COVID-19 case reports and show a detailed analysis. The idea of this evaluation scheme is to show the pandemic surveillance and provide a deeper understanding of the proposed approach to case reports data.

We use the following benchmark datasets for NER task, as given in Table 1.

Table 1: Benchmark datasets used.

Corpus	Entity types	Data size
NCBI-Disease [47]	Diseases	793 Pubmed abstracts
BC5CDR [48]	Diseases and diseases	1500 Pubmed articles
Species800 [49]	Species (organisms)	800 Pubmed abstracts
tmVar [50]	Disease-related mutation	500 PubMed articles
BC4CHEMD [51]	Chemicals	10,000 Pubmed abstracts
BC2GM [52]	Gene/Proteins	20,000 sentences
JNLBPA [53]	Genes, proteins	2404 abstracts
i2b2-Clinical [54]	Problem, Treatment, and Test.	426 discharge summaries

The train-test split information from the publication of each dataset is used. If there is no official train-test split in the original datasets, the 5-Folds stratified cross-validation strategy is used. A test set of 30% of our annotated data with ground truth labels is also set aside. The benchmark methods for this work are listed below. The precision, recall, F1-score, and macro-average F1 are used as evaluation metrics, following the practice as in previous works[26,27].

Training details: The experimental environment in this study is Google Colab Pro with cloud-based GPUs (K80, P100, or T4) and 32GB RAM. Google Colab is linked to Google Drive, so we have enough storage to complete the transfer learning. The general parameters used for implementing BiLSTM and Transformer-based architectures are given in Appendix 1: Table S2. The BERT encoder layers are implemented using the PyTorch BERT implementation available from Huggingface.co.

Results

Benchmarking against baselines

Table 2 provides a comparison of proposed NER approach to other benchmarks. We conducted a comprehensive comparison with a few Bi-LSTM-based and Transformer-based models. *BiLSTM-based models* include BiLSTM-CRF [25], BiLSTM-CNN-Char[26], BiLSTM-CRF-MTL[28], Doc-Att-BiLSTM-CRF[29], CollaboNet[30]. *Transformer-based NER models* include BLUE-BERT[31], BLUE-ELMO[31], BioBERT [24] with weights: BioBERT-Base v1.0, BioBERT-Base v1.1, BioBERT-Base v1.2.

Table 2: Test results in NER task (bold means best and italic as second-best) based on macro-averaged F1-scores. All the baselines are optimized to their optimal settings, and we report the best results for each method. For the methods, where there is no such result on a dataset, we experiment five times and report the best result for each method.

	NCBI	**BC5CDR**	**Species-800**	**tmVar**	**BC4CHEMD**	**BC2GM**	**JNLBPA**	**i2b2-clinical**	**Our dataset**
BiLSTM-CRF	85.81	86.18	69.28	86.28	89.48	81.08	74.29	85.66	87.24
BILSTM-CNN-Char	88.19	*88.58*	70.14	87.20	90.06	83.29	77.29	84.08	89.25
BiLSTM-CRF-MTL	88.85	84.93	65.39	83.35	89.42	82.12	77.03	83.25	86.78
Dic-Att-BiLSTM-CRF	88.61	87.33	68.28	83.31	88.2	81.80	76.23	85.18	86.94
CollaboNet	84.08	87.12	70.28	81.75	87.12	79.73	76.58	85.61	87.13
BLUE-BERT	86.37	87.62	72.98	87.24	90.19	82.93	75.27	86.09	88.10
BLUE-ELMO	84.28	87.12	71.04	84.28	89.01	82.13	74.69	85.23	87.98
BioBERT-Base v1.0	88.09	87.38	71.38	88.09	90.77	82.54	77.21	85.64	90.18
BioBERT-Base v1.1	88.30	87.67	*75.31*	88.30	90.78	84.41	*77.59*	85.53	90.39
BioBERT-Base v1.2	*89.71*	88.37	74.06	*89.10*	*91.08*	*84.72*	77.49	*86.73*	*92.34*
Our approach	**90.08**	**89.98**	**83.45**	**89.13**	**91.58**	**89.15**	**79.92**	**89.1**	**92.98**

We observe in Table 2 that our proposed method outperforms all the baseline methods on all the datasets. Our method performs better than the conventional BiLSTM-based methods and the Transformer based methods. We attribute the superiority of our method to its design, which combines the strength of sequence labeling based on a BiLSTM-CRF architecture with contextualized pre-trained embeddings from the Transformer architecture, resulting in an improvement of 1% to 5% over baseline methods.

We see a similar pattern of performance on the NCBI-disease, tmVar, Species-800, BC4CHEMD, BC2GM, and JNLBPA datasets. These results show that Transformer

layers do make an impact in inferring contextualized representations. On the BC5CDR dataset, our method obtains a macro-averaged F1-score of 89.98 % and BiLSTM-CNN-Char achieves a score of 88.58%. This probably demonstrates that while simpler deep neural networks are useful, contextualized embeddings (as in our model) increase the value of the extraction task. On the i2b2 dataset that is annotated with clinical named entities, our method and BioBERT method perform quite well. This result, suggests that entity types on which a method is preliminary trained, typically, perform well. Overall, we observe that our model applies to a variety of entity types. It performs best when trained on a dataset with a variety of entity types, and it also does best when trained on a different dataset with different entity types.

Pandemic surveillance

Effectiveness of the proposed approach on medical and social determinants of health

The fine-grained performance of our proposed NLP approach in extracting 10 SDOH entity classes from our test set is shown in Table 3.

Table 3: 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%).

Case study of the prediction results from different models

In this section we provide three running examples to further demonstrate the efficacy of our SDOH extraction approach (Table 4). A NER task presents three major challenges, as shown in Table 4: (i) identifying a long entity without breaking part, (ii) failing to identify an entity type, and (iii) misclassifying a non-entity

Table 4: Running examples, each running example contains a sentence extracted from a case report as well as ground-truth labels (or entities). For each sentence there are three sets of predicted labels (to be used as entities) generated by the top two best-performing models, BiLSTM-CNN-Char and BioBERT, as well as our proposed NLP model.

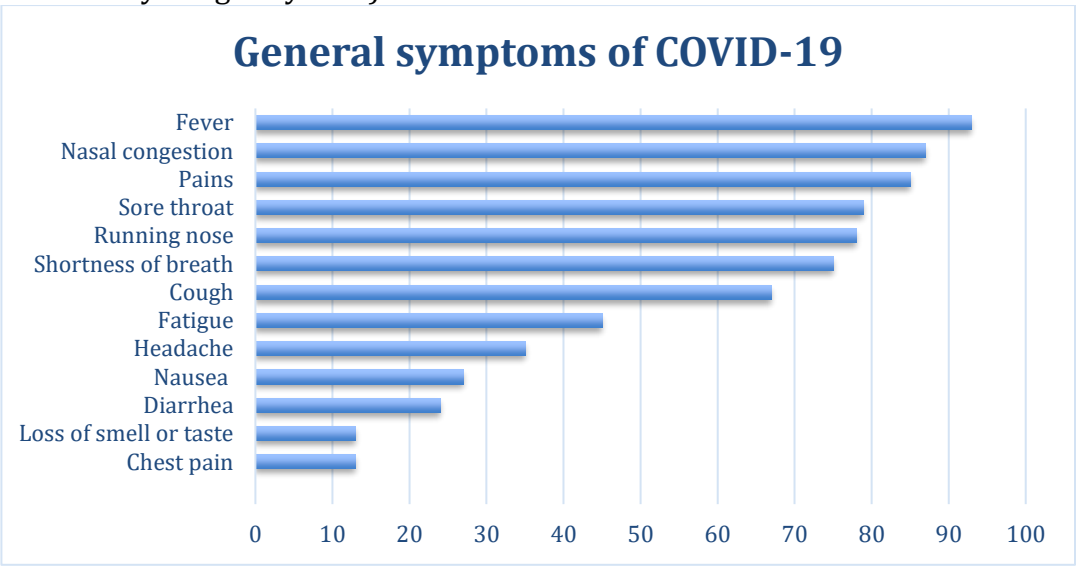
Diseases/Symptoms		
Case 1	True label	We report a 23-year-old man who developed, after asymptomatic COVID-19, a complex symptomatology characterized by fatigue, episodic shortness of breath, nocturnal tachycardia, and chest pain.
	BILSTM-CNN-Char	We report a 23-year-old man who developed, after asymptomatic 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%'.)

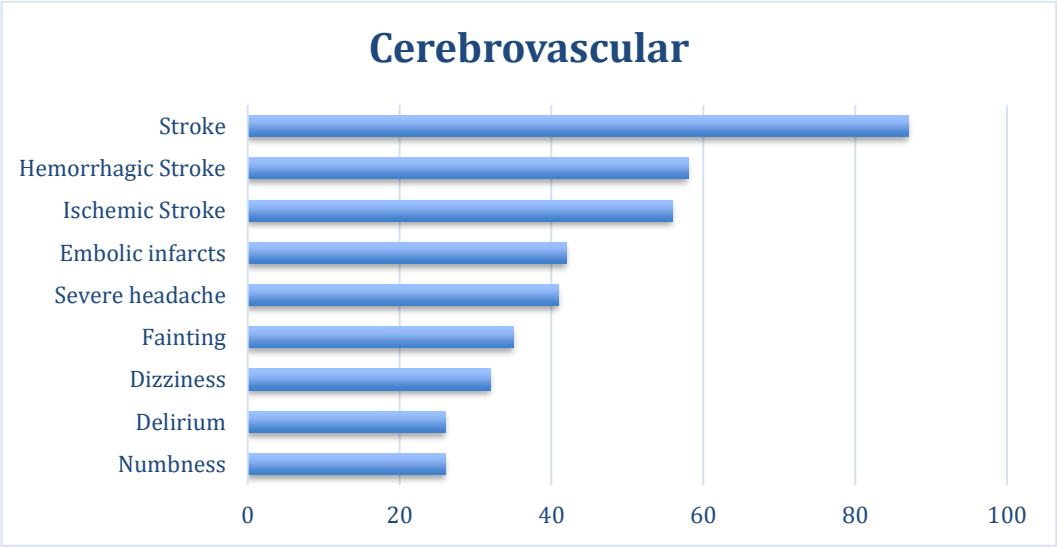
Overall, we see in these examples that BILSTM-CNN-Char and BioBERT missed a few named entities, but BioBERT did not misclassify any. Our model did not misclassify or miss any part from the long entities. However, in one case, our model failed to predict one named entity which is due to the lack of full context [57].

Determining clinical and SDOH in COVID-19 patients

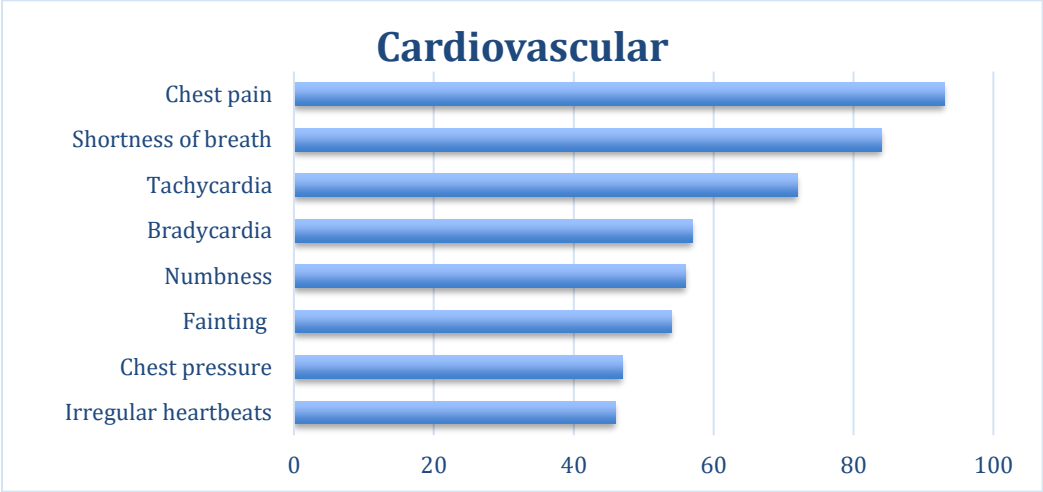
The prevalence of the most common COVID-19 symptoms reported in the case reports in our study is depicted in Figure 3. The symptoms are categorized into five groups of general symptoms, cerebrovascular, cardiovascular, pulmonary (based on the human body's organ system) and mental health issues.



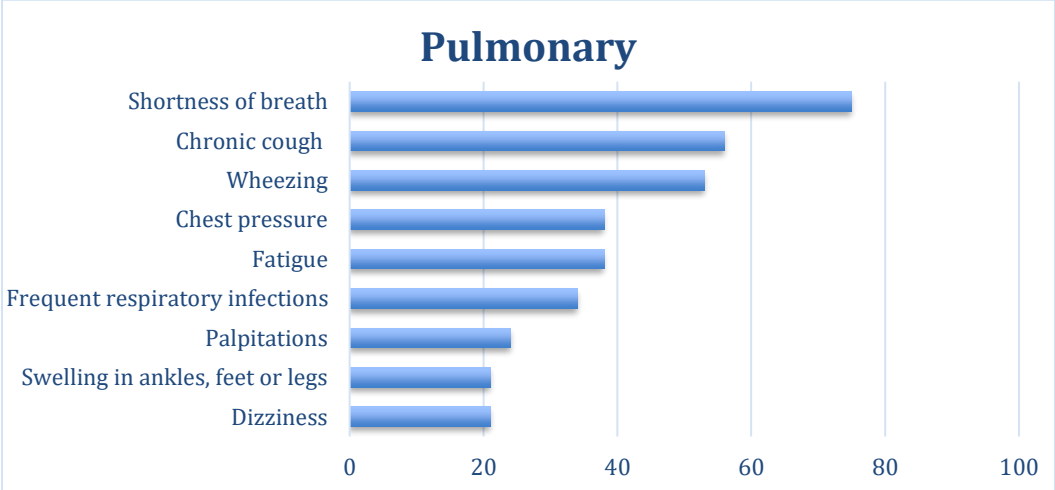
a.



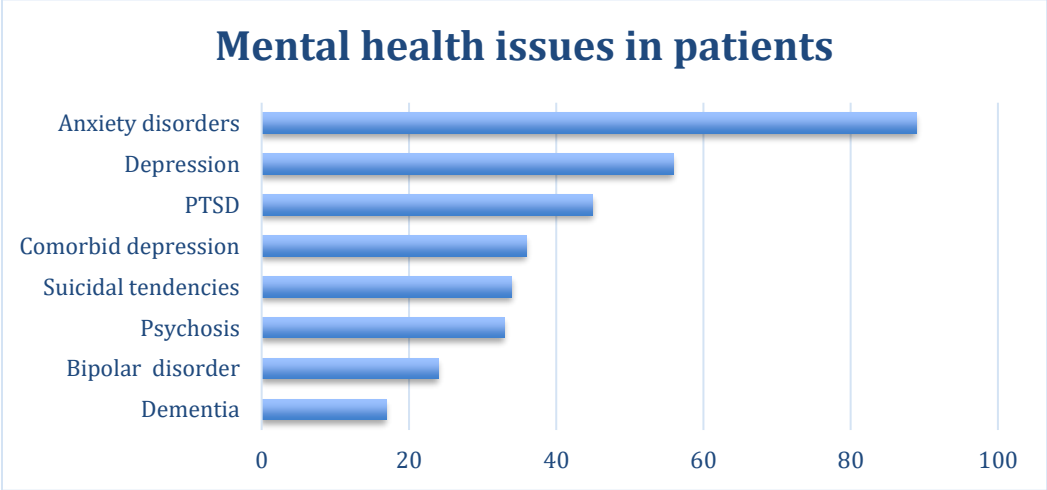
b.



c.



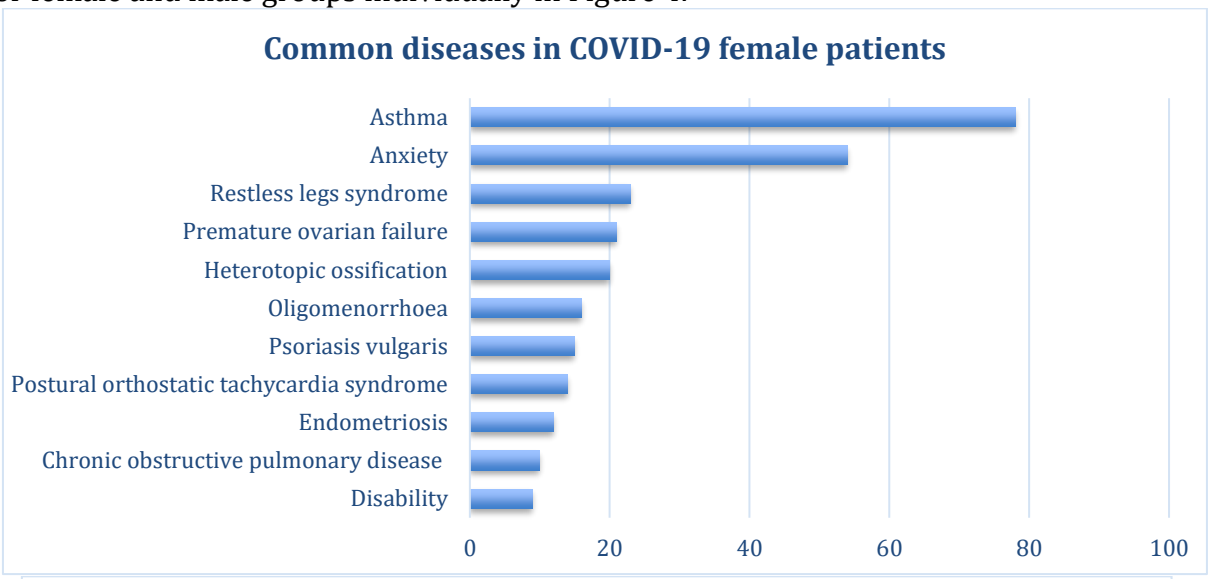
d.



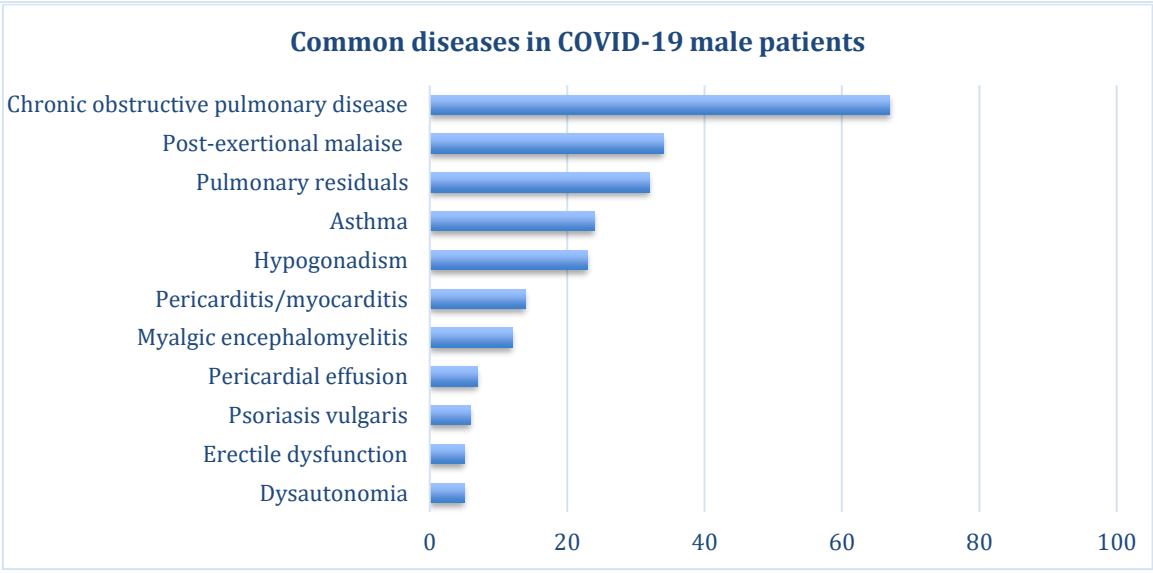
e.

Figure 3: Symptom prevalence estimates (occurring more than 70%) in COVID-19 patients, divided into general symptoms, body parts and mental health. Each bar represents the percentage of respondents who experienced that symptom.

Next, we show the prevalence of common disease disorders in COVID-19 patients for female and male groups individually in Figure 4.



a)

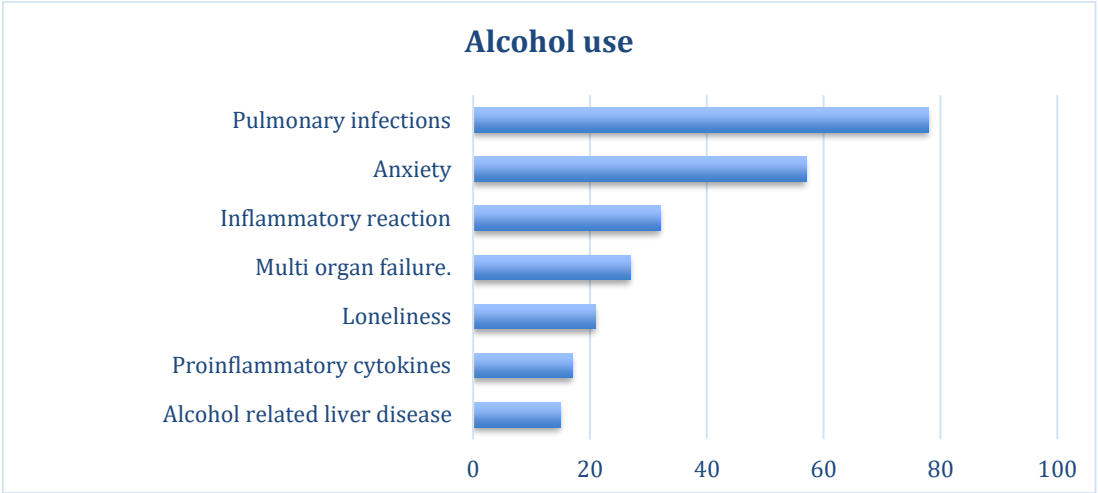


b)

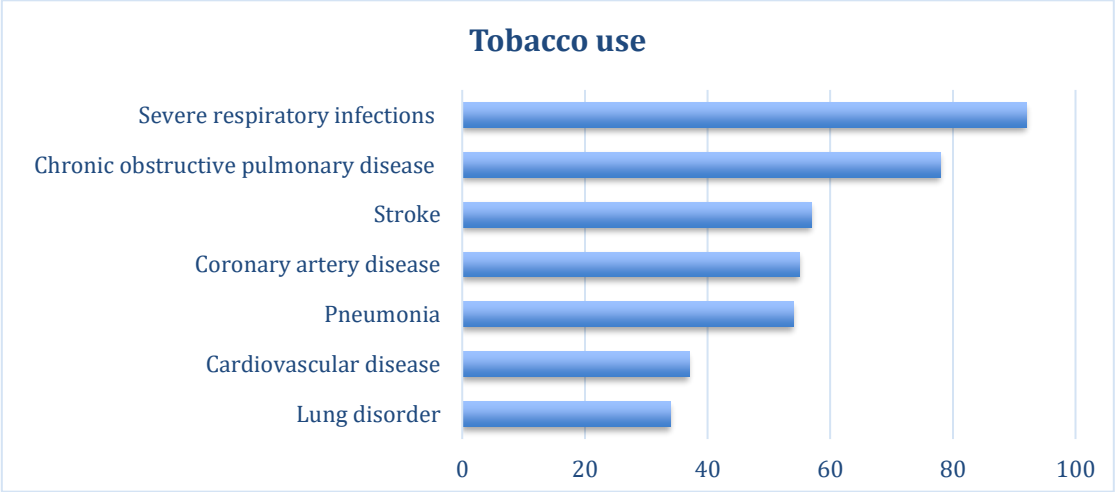
Figure 4: Common diseases in COVID-19 patients with respect to gender.

As shown in Figure 4(a), asthma, anxiety, and restless legs syndrome are common among female patients. Some diseases, such as premature ovarian failure, are also seen in female patients. Figure 4(b) depicts the prevalence of chronic pulmonary disease, post-exercise malaise, and asthma in male patients.

We show the prevalence of disease disorders in COVID-19 patients split across the two substances (alcohol, tobacco) use in Figure 5. We observe in Figure 5 (a) that pulmonary infections are most common disorder in COVID-19 patients with alcohol dependence. We also see in Figure 5(b) that severe respiratory infections are common among COVID-19 patients with tobacco use. This conclusion, however, is based solely on mentions of disease order with substance dependence in case reports.



a)



b)

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

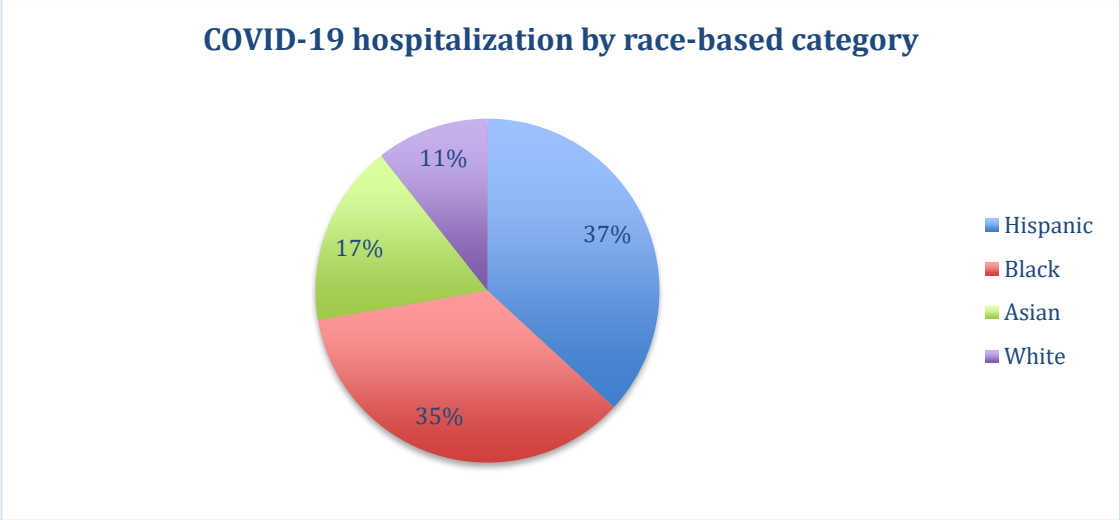


Figure 6: COVID-19 hospitalization by race and ethnicity.

Figure 6 depicts the percentage of COVID-19 patients across different race-based categories. We take the nominal race groups[32] 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 moratlity rate in COVID-19 patients with different age groups in Appendix 1, Figure S2.

Discussion

In previous works, SDOH have been extracted from clinical data using different methods. These state-of-the-art methods can be categorized into conventional methods like regular expressions, dictionaries or rule-based like cTAKES [33,34], or deep neural networks like CNN, LSTMs [35] and the latest Transformer based methods [34]. Surprisingly, in the large training sets, language model-based representations outperformed other trained neural nets. However, we also found that with enough training data, the BiLSTM-based neural networks (simple neural networks) can also perform very well, being only 1-5% worse than models trained using Transformer-based representations. However, with simpler neural network representations, it is important to use a model type with sufficient expressive power.

We acknowledge this study's limitations, which provide opportunities for future research**.** First, we anticipate that this dataset is not representative of all SDOH impacting patients with COVID-19. Publication bias favouring positive outcomes has been observed in studies of case reports [36,37]. For example, a review of 435 dental case reports[38] revealed that 99.1 % of the data indicate a positive outcome. This bias may impact both the types and frequencies of SDOH we observed. We hope to implement EHRs and clinical notes that are updated in real time in the future. As data privacy is a concern, we can mask the named entities associated with patients' personal information, such as named, location, and identifiers.

Second, in most cases, a case report will describe the patient's SDOH, but this is not always the case. 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. It is possible to avoid these limitations by extending the annotation scheme, defining the rules or semantics, and retraining the language models, albeit in a laborious manner.

Other future directions to extend this research are including a more extensive active learning approach [39], using prompt-based learning to avail the strengths of foundation models[40], and testing different model architectures and problem structures to further improve performance. By strengthening our models in this way, and addressing the limitations above, we believe this work will lead to an effective and general-purpose model. Such research allows the researchers to better integrate these factors into their analyses and further improve healthcare outcomes.

Conclusions

This study demonstrates that NLP-based methods can be used to identify clinical entities (diseases) and SDOH from large amounts of data. The proposed method uses a combination of neural networks to extract SDOH from the clinical texts. To assess the model's 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. 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.

Abbreviations

Supplementary material

Supplementary material is attached.

Funding Statement

This research was co-funded by the Canadian Institutes of Health Research's Institute of Health Services and Policy Research (CIHR-IHSPR) as part of the Equitable AI and Public Health cohort, and This work was funded through the Artificial Intelligence for Public Health (AI4PH) Health Research Training Platform (HRTP) supported by the Canadian Institutes of Health Research (CIHR).

Competing Interests Statement

The authors declare that they have no competing interests

Contributorship Statement

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.

Data Availability statement

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

Acknowledgements

The authors would like to thank Public Health Ontario scientists for their guidance on public health practices in COVID-19 research.

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 Raza S, Schwartz B, Rosella LC. CoQUAD: a COVID-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC Bioinformatics* 2022;**23**:210. doi:10.1186/s12859-022-04751-6

14 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

15 Rison RA, Shepphird JK, Kidd MR. How to choose the best journal for your case report. J. Med. Case Rep. 2017;**11**:1–9.

16 IMI. CARE case report guidelines. CARE Case Rep. Guidel. 2019.https://www.care-statement.org/

17 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

18 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.

19 Zhang Y, Zhang Y, Qi P, *et al.* Biomedical and clinical English model packages for the Stanza Python NLP library. *J Am Med Informatics Assoc* 2021;**28**:1892–9. doi:10.1093/jamia/ocab090

20 Patra BG, Sharma MM, Vekaria V, *et al.* Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Informatics Assoc* 2021;**28**:2716–27. doi:10.1093/JAMIA/OCAB170

21 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

22 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.

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

24 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
25 Lample G, Ballesteros M, Subramanian S, *et al.* Neural architectures for named entity recognition. *arXiv Prepr arXiv160301360* 2016.
26 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
27 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
28 Wang X, Zhang Y, Ren X, *et al.* Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* 2019;**35**:1745–52.
29 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.
30 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.
31 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.
32 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
33 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.
34 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
35 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.
36 Easterbrook PJ, Gopalan R, Berlin JA, *et al.* Publication bias in clinical research. *Lancet* 1991;**337**:867–72.
37 Nissen T, Wynn R. The clinical case report: a review of its merits and limitations. *BMC Res Notes* 2014;**7**:1–7.
38 Oliveira GJ, Leles CR. Critical appraisal and positive outcome bias in case reports published in Brazilian dental journals. *J Dent Educ* 2006;**70**:869–74.
39 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.
40 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