Journal of the American Medical Informatics Association, 26(8-9), 2019, 787–795
doi: 10.1093/jamia/ocz093
Advance Access Publication Date: 2 July 2019
Research and Applications

Research and Applications

Identifying vulnerable older adult populations by contextualizing geriatric syndrome information in clinical notes of electronic health records

Tao Chen,[1] Mark Dredze,[2] Jonathan P. Weiner,[3] and Hadi Kharrazi[3,4]

[1]Center for Language and Speech Processing, Johns Hopkins Whiting School of Engineering, Baltimore, Maryland, USA, [2]Department of Computer Science, Johns Hopkins Whiting School of Engineering, Baltimore, Maryland, USA, [3]Center for Population Health IT, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA, and [4]Division of Health Sciences Informatics, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

Corresponding Author: Tao Chen, PhD, Center for Language and Speech Processing, Johns Hopkins Whiting School of Engineering, Hackerman Hall 226, 3400 North Charles Street, Baltimore, MD 21218, USA (taochen.nus@gmail.com)

Received 11 January 2019; Revised 12 May 2019; Editorial Decision 16 May 2019; Accepted 17 May 2019

ABSTRACT

Objective: Geriatric syndromes such as functional disability and lack of social support are often not encoded in electronic health records (EHRs), thus obscuring the identification of vulnerable older adults in need of additional medical and social services. In this study, we automatically identify vulnerable older adult patients with geriatric syndrome based on clinical notes extracted from an EHR system, and demonstrate how contextual information can improve the process.

Materials and Methods: We propose a novel end-to-end neural architecture to identify sentences that contain geriatric syndromes. Our model learns a representation of the sentence and augments it with contextual information: surrounding sentences, the entire clinical document, and the diagnosis codes associated with the document. We trained our system on annotated notes from 85 patients, tuned the model on another 50 patients, and evaluated its performance on the rest, 50 patients.

Results: Contextual information improved classification, with the most effective context coming from the surrounding sentences. At sentence level, our best performing model achieved a micro-F_1 of 0.605, significantly outperforming context-free baselines. At patient level, our best model achieved a micro-F_1 of 0.843.

Discussion: Our solution can be used to expand the identification of vulnerable older adults with geriatric syndromes. Since functional and social factors are often not captured by diagnosis codes in EHRs, the automatic identification of the geriatric syndrome can reduce disparities by ensuring consistent care across the older adult population.

Conclusion: EHR free-text can be used to identify vulnerable older adults with a range of geriatric syndromes.

Key words: geriatric syndrome, vulnerable geriatric population, electronic health records, clinical notes, natural language processing, deep neural network, sentence classification

INTRODUCTION

Vulnerable older adult populations are at increased risk for a wide range of medical and social conditions. A variety of factors affecting vulnerable geriatric populations can lead to health disparities that go unrecognized by medical professionals. Some of these factors are termed geriatric syndromes, which are a set of complex symptoms with high prevalence in older adults that do not fit specific disease categories.[1] Geriatric syndromes such as falls, incontinence, lack of social support, and frailty, are often associated with increased morbidity and poor outcomes, which can substantially diminish the quality of life among vulnerable older adults.[2,3]

While identifying and studying vulnerable older adults are of great interest to health disparity researchers,[4] geriatric syndromes are difficult to study due to their complex nature and poor representation in diagnosis codes (eg, International Classification of Diseases [ICD]).[5–7] Coding challenges limit research opportunities and create disparities between groups of patients where geriatric syndromes are more difficult to track. While many of these symptoms are contained in the free-text of EHRs,[5–7] the lack of structured data may lead to clinicians and researchers being unaware of ongoing issues affecting health equity, such as identifying vulnerable patients, setting inclusion and exclusion criteria in clinical trials, and aligning provider-driven population health efforts with public health goals and policies.[8–11]

To address the challenges in identifying vulnerable patients with geriatric syndromes, we automatically discover geriatric syndromes from the free-text of EHRs using machine learning algorithms. The automatic identification of these syndromes can help assure that consistent care is delivered to a medically complex and heterogeneous elderly population. We focus on 10 common geriatric syndrome constructs: falls (FL), malnutrition (ML), dementia (DE), severe urinary control issues (UC), absence of fecal control (BC), visual impairment (VI), walking difficulty (WD), pressure ulcers (PU), lack of social support (SS), and weight loss (WL).

We use information extraction (IE) techniques to identify patients that exhibit geriatric syndromes using EHR free-text. IE is a natural language processing (NLP) task to transform free-text into structured output. In this setting, we seek to identify any of the 10 geriatric syndrome labels (ie, constructs) based on a clinician's note in an EHR. We create an IE system using supervised machine learning, whereby labeled textual examples of the 10 constructs are used to train a statistical NLP model.

Traditionally, IE systems analyze 1 sentence at a time, meaning that each sentence in the clinical note is independently analyzed to determine if it expresses a syndrome for a patient. This technique works well for common clinical IE tasks, such as identifying disorders[12] or medications[13] whose presence can be determined by examining only the immediate context around the mention. However, a key challenge of geriatric syndrome identification is the ambiguity exhibited in the local context within the sentence.[5–7] Consider the sentence "patient has lost a few pounds since May." Losing weight could be either unintentional (a geriatric syndrome construct) or intentional (not a geriatric syndrome construct). Thus, a single sentence can ambiguously describe a geriatric syndrome, while the disambiguating context is out of reach of traditional IE systems.

This study improves the identification of geriatric syndrome constructs by expanding the context considered by the IE system. We evaluate methods for incorporating 3 types of contexts into the IE system: sentences adjacent to the sentence under consideration, the entire clinical document, and diagnosis codes (ie, ICD9 codes) extracted from both structured and unstructured data. We frame the task of identifying geriatric syndromes as sentence classification: "which of the 10 geriatric syndromes, if any, are exhibited by this sentence?". We build on recent work using deep neural networks for general NLP[14] and clinical NLP[15] tasks to build a sentence classification system. We then propose a novel end-to-end neural architecture that incorporates the 3 types of contexts. Our experiments show that the addition of contexts significantly improves the identification of geriatric syndromes.

OBJECTIVE

We propose a method to automatically identify vulnerable older adults with geriatric syndromes from unstructured free-text from EHRs. We introduce a deep learning system for sentence classification that incorporates contextual information from surrounding sentences, the entire document, and structured diagnostic codes. We demonstrate that contextual information improves the accuracy of identifying geriatric syndromes.

MATERIALS AND METHODS

Data collection and annotation

The anonymized EHR data used in this study were provided by a large multispecialty medical group in Massachusetts, United States for a cohort of elderly patients enrolled in a regional Medicare Advantage health maintenance organization. We utilized a cohort of 18 341 members aged 65 or older who received continuous medical and pharmacy benefits coverage for at least 24 months from Jan 1, 2011 to Dec 31, 2013. The EHR data included both structured fields and unstructured free-text (eg, clinical notes). All data used in this research were stored on a secure network approved by the institutional review board of Johns Hopkins School of Public Health (IRB #6196).

To enable our study, we further constructed a data set with geriatric syndrome constructs/labels. We randomly assigned a sample of 185 patients from the larger cohort of 18 341 members,[5,6] resulting in 8442 clinical notes. We then used the clinical Text Analysis and Knowledge Extraction System (cTAKES)[16] to segment the notes into sentences. The sentence detector of this system extends OpenNLP's[17] supervised sentence detector to the medical domain and predicts whether end-of-line characters (eg, period, question mark, exclamation mark, new line, tab) indicate the end of a sentence. We obtained 150 947 sentences in total.

Three physicians carefully examined all 8442 notes to determine the mentions of geriatric syndrome constructs for each sentence and also to identify the words/phrases that indicate the constructs. Before the formal annotation, the physician annotators were trained using a shared guideline and coded a similar text to ensure an acceptable consensus.[5,6] Due to the considerable annotation workload (150 974 sentences), we did not ask all annotators to label all sentences. Each sentence was annotated by 1 of the physicians and the annotations took around 240 person-hours in total. As sentences were split among annotators, we were unable to calculate inter-rater agreement for the entire annotated text.

In the annotated data set, only 3.4% of sentences were identified to contain at least 1 of the 10 constructs. Our study results are based on the annotated data set (representing 185 patients) while the unlabeled notes (representing 18 156 patients) were used to train unsupervised embeddings that enhance our models (detailed in the following section). Table 1 shows a few sample sentences from our data set. We have provided additional examples in the Supplementary Material.

Proposed model

Our analysis of the labeled data found that manual labeling of clinical notes for geriatric syndromes is a challenging task. While broad agreement occurred on which sentences contain a construct, significant differences existed between the specific words selected by each annotator that indicated a construct. For example, some annotators excluded words they deemed unimportant, while others included

Table 1. Example sentences that contain a geriatric syndrome construct

Geriatric Syndrome Construct	**Example Sentence**[a]
Absence of fecal control (BC)	She has also been experiencing urinary incontinence and a few **episodes of fecal incontinence** too.
Dementia (DE)	Patient **has dementia** and daughter feels as though it has worsened since Labor Day.
Falls (FL)	She suffered **a fall** this past Tuesday and then was complaining of left shoulder pain.
Malnutrition (ML)	**Inadequate energy intake** as evidenced by weight loss.
Pressure ulcers (PU)	She has 2 **intragluteal decubitus**.
Lack of social support (SS)	She **is alone** at home much of the day.
Severe urinary control issues (UC)	She **has a suprapubic catheter** in (placed under interventional radiology at…) because she was having pain on urination.
Visual impairment (VI)	Has been seen by vision rehab and **is registered with of blind**.
Walking difficulty (WD)	**Ambulates slowly**, uses Vital signs as above.
Weight loss (WL)	Sed rate had been mildly elevated except the last one over 70 but in setting of acute illness and **weight loss**.

[a]Phrases annotated as geriatric syndrome constructs are bolded.

them (eg, "with a walker," "walks with a walker," or "walker" are parts of the same sentence tagged by different annotators for walking difficulty). Using the same data set, our prior work experimented with regular expressions for geriatric syndrome identification[5,6]; however, that study focused on evaluating precision and not recall thus lacking a test set which we could use to assess our current approach. In other words, the inconsistencies made it challenging to rely on statistical information extraction with a sequence tagger, in which each word must be correctly identified as part of a construct.[7] Since our goal is to identify patients and records—not individual phrases—we instead formulated the task as sentence classification: that is, whether the sentence indicates the presence of a geriatric syndrome.

We construct a multi-class sentence classification model, as sentences with more than 1 construct are extremely rare (0.02% in our data set). Sentence classification systems are widely used across various tasks in NLP, including sentiment analysis,[18,19] opinion detection,[20,21] and question type classification.[22] Prior work has utilized various architectures such as a convolutional neural network,[23–25] long short-term memory (LSTM) recurrent neural networks,[26–28] and, recently, Bidirectional Encoder Representations from Transformers (BERT).[29] Each approach learns a representation of the input sentence, and utilizes that representation for making classification decisions. In our work, we develop a deep neural network to approach the sentence classification task, and adopt an LSTM to learn a representation for the target sentence in our base model.

We leverage context by augmenting a sentence classification model with learned representations of the context. We consider 3 types of contexts: (1) the surrounding sentences, (2) the document as a whole, and (3) the diagnosis codes (ie., ICD9 codes) mentioned in the free-text of the note as well as the structured field of the encounter associated with the note.

Figure 1 illustrates our proposed model architecture. The model consists of 4 modules: a sentence classification component representing the base model (ie, target sentence), and 3 optional advanced modules that represent the contextual information (ie, surrounding sentences, whole document, and diagnostic codes). The modules used in our NLP architecture include (Figure 1) the following:

Target Sentence (T): This component learns a representation of the sentence to classify. We use a Bidirectional Long Short-Term Memory (BiLSTM) to learn a representation. The input to the BiLSTM are word embeddings after applying dropout. We pretrained word embeddings on the unlabeled notes of 18 156 patients (ie, representing the whole population of 18 341 patients but excluding the 185 annotated patients; hereafter, *large–unlabeled*) using the skip-gram model from *Word2vec*.[30] We then used an attention mechanism[31] to produce a single representation of the sentence that aggregates BiLSTM outputs after dropout for individual words. In the base model, this representation is then fed to a fully connected layer followed by a dropout and a softmax to produce a classification for geriatric syndromes.

Surrounding Sentences (S): We define the surrounding sentences as those adjacent to the target sentence in a fixed-size window (ie, window size K is a tunable hyperparameter). To represent these sentences as a vector, we leveraged *Paragraph2Vec*,[32] an unsupervised algorithm that learns a fixed-length feature representation from variable-length pieces of texts, such as sentences and documents. We trained *Paragraph2Vec* on all sentences in our large-unlabeled data set, and applied the model to the labeled data set. Each of the learned sentence embeddings are passed to an attention layer[31] to learn a single fixed-length representation for the surrounding sentences. The attention layer aims to capture the importance of each surrounding sentence.

Document (D): We trained *Paragraph2Vec*[32] on all documents of the large-unlabeled data set (ie, 18 156 patients), and applied the model to infer the embeddings for documents in the labeled data set. This vector is regarded as the document representation of the clinical notes.

ICD9 Codes (I): The diagnosis codes used in our data set are ICD9 codes. The ICD9 code typically appears in the structured field of the encounter associated with the clinical note, but it can also be mentioned in the note as free-text. Thus, we extracted ICD9 codes from both sources. We employed *Med2vec*,[33] an unsupervised algorithm to learn a code representation on the large-unlabeled data set. Med2vec uses the same concept of Word2vec's skip-gram[34] to model the co-occurrences of ICD9 codes within a patient's visit and the co-occurrences of a patient's visits in a context window. Since each note may have multiple ICD9 codes, we used a max-pooling layer to combine these codes' representations (after dropout) to form a fixed-length vector.

We concatenated each of the aforementioned learned representations into a single vector. This vector is provided to a fully connected layer, followed by a dropout and a softmax which predicts 1 of the possible 11 labels (ie, 10 geriatric syndrome constructs plus no construct). We assessed all combinations of context modules, as well as the standard target sentence model (detailed in Table 3). All models were implemented in Google's Tensorflow[35] neural network library.

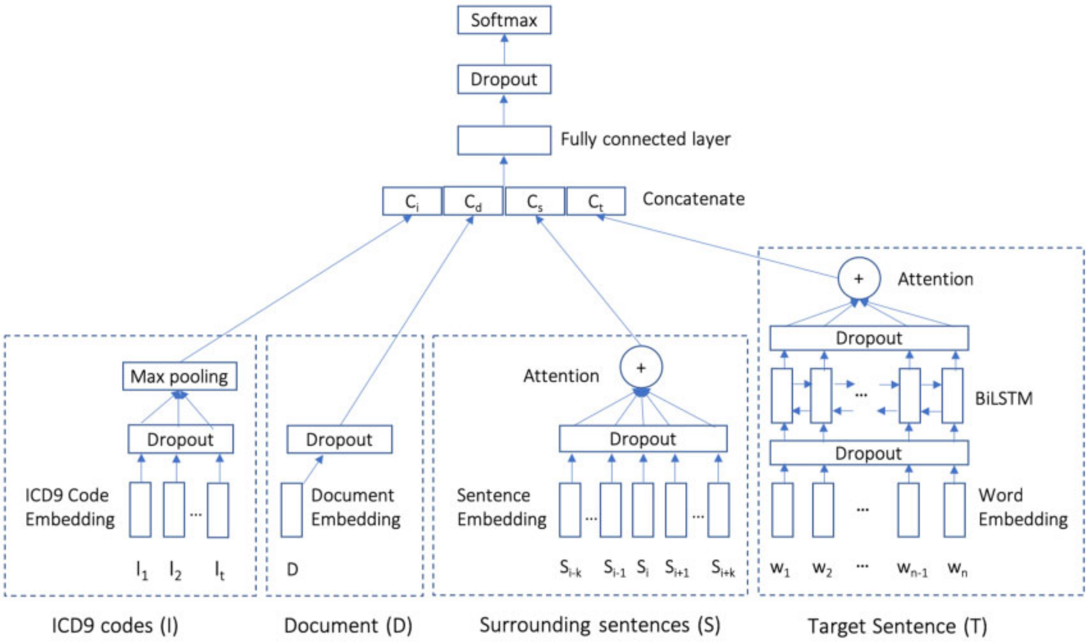

Figure 1. Our proposed context-aware geriatric syndrome identifier model. Context modules (S, D, and I) are optional.

Table 2. Data set statistics

Construct	Training set[a]		Validation set[b]		Test set[c]	
	Sentence # (%)	Patient # (%)	Sentence # (%)	Patient # (%)	Sentence # (%)	Patient # (%)
BC	40 (0.05)	12 (14.12)	46 (0.15)	4 (8.0)	8 (0.02)	3 (6.0)
DE	222 (0.3)	15 (17.65)	85 (0.27)	9 (18.0)	127 (0.28)	10 (20.0)
FL	379 (0.51)	37 (43.53)	79 (0.25)	21 (42.0)	189 (0.42)	23 (46.0)
ML	84 (0.11)	9 (10.59)	6 (0.02)	4 (8.0)	33 (0.07)	6 (12.0)
PU	512 (0.69)	53 (62.35)	348 (1.12)	30 (60.0)	425 (0.94)	30 (60.0)
SS	222 (0.3)	16 (18.82)	21 (0.07)	4 (8.0)	92 (0.2)	7 (14.0)
UC	92 (0.12)	16 (18.82)	38 (0.12)	6 (12.0)	119 (0.26)	13 (26.0)
VI	590 (0.79)	56 (65.88)	355 (1.14)	26 (52.0)	383 (0.85)	34 (68.0)
WD	99 (0.13)	21 (24.71)	87 (0.28)	14 (28.0)	237 (0.52)	19 (38.0)
WL	42 (0.06)	8 (9.41)	33 (0.11)	5 (10.0)	161 (0.36)	12 (24.0)
No construct	72 391 (96.94)	–	30 028 (96.47)	–	43 374 (96.07)	–

Abbreviations: BC, absence of fecal control; DE, dementia; FL, falls; ML, malnutrition; PU, pressure ulcers; SS, lack of social support; UC, severe urinary control issues; VI, visual impairment; WD, walking difficulty; WL, weight loss.

[a]85 patients and 74 673 sentences.

[b]50 patients and 31 126 sentences.

[c]50 patients and 45 148 sentences.

Baselines

We compare 2 baseline systems that only consider the target sentence with our proposed context-enhanced classification system. Both baseline systems use a BiLSTM to learn a representation of the target sentence. The first baseline constructs a single sentence representation using max pooling over the hidden states (BiLSTM-Max, Figure 2 left). The second baseline uses an attention layer[31] to combine the hidden states (BiLSTM-Att, Figure 2 right). Both models feed the sentence vector into a softmax layer. Both models use word embeddings initialized by the same skip-gram model used in our context model. Baseline models do not use a fully connected network before the softmax output.

Experimental setting

We randomly split our labeled data set of 185 patients into 85 patients as training, 50 as validation, and 50 as test. This approach ensures that the system is assessed on both clinical notes and patients that were unseen during training. For the very few sentences with multiple constructs (0.02%), we replicated sentences and paired them with each construct.

Table 2 details the construct distribution for both sentences and patients. The data set has 2 key characteristics: First, the majority of sentences (eg, 96.94% in training set) do not have a construct. Second, constructs exhibit an imbalanced distribution. In the training set, the 3 most common constructs are visual impairment (VI; 0.79% of sentences and 65.88% of patients), pressure ulcers (PU;

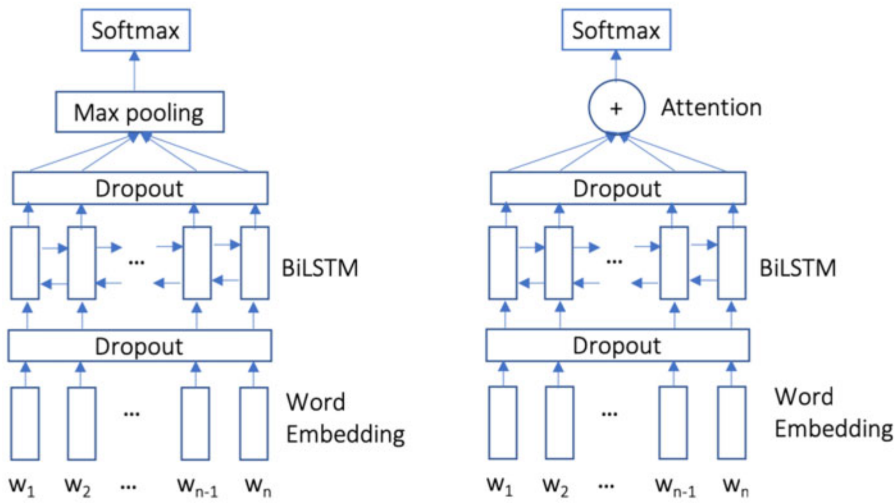

Figure 2. Two baselines models of BiLSTM-Max (left) and BiLSTM-Att (right) that incorporate the target sentence via BiLSTM.

Table 3. Results on test set. Micro-F_1 is the metric used to tune model hyperparameters.

Model[a]	Micro-averaged Sentence-level			Macro-averaged Sentence-level			Micro-averaged Patient-level			Macro-averaged Patient-level		
Measure	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
1 BiLSTM-Max	0.623	0.530	0.573	0.662	0.524	0.585	0.739	0.847	0.789	0.729	0.803	0.764
2 BiLSTM-Att*	0.582	0.576	0.579	0.631	0.564	0.595	0.741	0.892	0.809	0.712	0.832	0.767
3 T*	0.577	0.585	0.581	0.600	0.588	0.594	0.701	0.879	0.780	0.652	0.828	0.729
4 T+S*	0.666	0.554	**0.605**	0.716	0.553	0.624	0.819	0.834	0.826	0.816	0.770	0.792
5 T+D*	0.688	0.499	0.579	0.610	0.496	0.547	0.755	0.762	0.758	0.782	0.822	0.801
6 T+I**	0.629	0.542	0.582	0.657	0.545	0.596	0.805	0.815	0.810	0.816	0.767	0.791
7 T+SI**	0.615	0.571	0.592	0.633	0.558	0.593	0.778	0.873	0.823	0.728	0.818	0.771
8 T+SID*	0.654	0.546	0.595	0.726	0.529	0.612	0.846	0.841	**0.843**	0.833	0.781	0.806

Abbreviations: D, document; I, ICD9 codes; P, precision; R, recall; S, surrounding sentences; T, target sentence.
[a]McNemar's test was used to measure the difference between the results of BiLSTM-Max and other approaches.
***, **, and * indicate that p value is smaller than .001, .01, and .05.

0.69% and 62.35%), falls (FL; 0.51% and 45.33%), and the 3 least common constructs are absence of fecal control (BC; 0.05% and 14.12%), weight loss (WL; 0.06% and 9.41%), and malnutrition (ML; 0.11% and 10.59%).

While we train our system to recognize constructs in a sentence, we evaluate accuracy on both sentence and patient-level predictions. A patient is considered associated with a geriatric syndrome construct if any sentence in his/her clinical notes is predicted as that label. This allows the system to correctly assign a construct to a patient if even 1 sentence in the patient's record is correctly identified as exhibiting the construct. Since the data set exhibits a skewed label distribution, we adopt precision (positive predictive value), recall (true positive rate), and F_1 metric (harmonic mean of precision and recall) for both sentence and patient evaluation. We report both the micro-averaged (aggregate the contributions of all classes to compute the average metric) and macro-averaged (compute the metric independently for each class and then take the average) scores over all the construct labels. Since we had a skewed data set, micro-F_1 was deemed the most appropriate metric in this study, which was used to tune model hyperparameters.

We trained all models using an ADAM optimizer[36] and set the initial learning rate to 0.001. The dimensionality for all the embedding layers was 100. We used the validation set to tune model hyperparameters based on the sentence micro-F_1, such as: the dimension of BiLSTM hidden states with a selection from the set (50, 100); the dimension of the fully connected layer with a selection from (50, 100); dropout rates with a selection from (0, 0.1, 0.2, 0.3, 0.4, 0.5); and window size of surrounding sentences with a selection from 2 (1 sentence before and after the target sentence), 10 (5 sentences before and after), and 20 (10 sentences before and after). We did similar hyperparameter tuning for the 2 baselines. To prevent overfitting, we adopted an early-stop training strategy, in which we stopped model training when performance did not improve for 10 epochs on the validation set.

RESULTS

In our experiments, we carefully tuned the hyperparameters of each model on the validation set based on the sentence micro-F_1 score. We report the results obtained with the final chosen hyperparameters. The optimal surrounding sentences context window size was 10 (5 sentences before and after the target sentence). Window size of 2 (1 before and after) captured too small of a context, while 20 (10 before and after) captured a wide context that was often not relevant to the target sentence. For all of the models, the dimension of the BiLSTM hidden state in each direction was 100. The dropout rate of word embedding, sentence embedding, document embedding, and ICD9 code embedding were 0.2, 0.2, 0.3, and 0.3, respectively. The dropout rate of the attention layer was 0.5. Similar to prior work,[37] we found proper dropout rates were effective in preventing model overfitting.

Table 3 details the experimental results of each model with the best hyperparameter setting on the test set. We used McNemar's test,[38] a commonly used statistical test for classification models that are difficult to train (eg, neural models),[39] to measure the decision (ie, classification label) differences between the models, although McNemar's test does not necessarily reflect the performance (eg, micro-F_1) differences between models.

First, we found that attention was more effective than max pooling for the base model using only the target sentence (micro-F_1 of 0.579 vs 0.573 for sentence-level analysis; and 0.809 vs 0.789 for patient-level analysis). All of our models had statistically significant improvements over the BiLSTM-Max baseline. We also found that the BiLSTM-Att and the target sentence model performed similarly, with the BiLSTM-Att model generating better accuracy on the patient-level (micro-F_1 of 0.809 vs 0.780). We next evaluated how each context affected system accuracy. We considered adding context from the surrounding sentences, document context, and ICD9 codes. Adding the surrounding sentences consistently improved over the target sentence alone across all metrics (micro-F_1 of 0.605 vs 0.581 for sentence-level analysis and 0.826 vs 0.780 for patient-level analysis). By comparison, the ICD9 context helped modestly, and the document context impaired the recall but improved precision.

Finally, we considered using all 3 contexts in 1 model. We also experimented with other combinations of the 3 context modules, but the model with the 3 context modules worked best. Although adding document context alone decreased the overall F_1, incorporating it with the other 2 context modules added value (Table 3, rows 7 vs 8). Our final model with 3 contexts (Table 3, row 8) achieved the best performing patient-level model, yielding nearly a 4-point improvement over the context-free BiLSTM-Att baseline (micro-F_1 of 0.843 vs 0.809).

Table 4 shows model performance by construct for both sentence and patient levels for the best performing models. The performance varied widely for different constructs. At the sentence-level, 6 constructs BC, FL, DE, WD, VI, and SS obtained an F_1 score greater than 0.7, while the worst performing construct ML had an F_1 score as low as 0.184. At the patient level, all constructs except UC (F_1 = 0.571) obtained an F_1 score larger than 0.7, which shows the model's robustness in patient-level prediction.

Table 4. The results of our best performing model by construct on test set. The last 2 rows are the overall macro and micro-averaged results, respectively.

Geriatric Syndrome	Sentence (T + S)			Patient (T + SID)		
Measure	P	R	F_1	P	R	F_1
BC	1.000	0.750	0.857	1.000	0.667	0.800
DE	0.667	0.740	0.701	0.714	1.000	0.833
FL	0.685	0.794	0.735	0.786	0.957	0.863
ML	0.708	0.106	0.184	0.842	0.842	0.842
PU	0.750	0.455	0.566	0.800	0.667	0.727
SS	0.647	0.600	0.623	0.935	0.967	0.951
UC	0.455	0.543	0.495	0.571	0.571	0.571
VI	0.891	0.479	0.623	0.889	0.615	0.727
WD	0.689	0.601	0.642	0.906	0.853	0.879
WL	0.669	0.460	0.545	0.889	0.667	0.762
Macro	0.716	0.553	0.624	0.833	0.781	0.806
Micro	0.666	0.554	0.605	0.846	0.841	0.843

Abbreviations: BC, absence of fecal control; D, document; DE, dementia; FL, falls; I, ICD9 codes; ML, malnutrition; P, precision; PU, pressure ulcers; R, recall; S, surrounding sentences; SS, lack of social support; T, Target sentence; UC, severe urinary control issues; VI, visual impairment; WD, walking difficulty; WL, weight loss.

DISCUSSION

Measuring geriatric syndromes to identify vulnerable and potentially underserved patients at a population level is of great interest to health providers and researchers who are seeking to address health equity challenges among older adults. Due to the complex nature of geriatric syndromes, however, they are poorly captured by diagnosis codes, yet they are present in the clinical text. Such coding challenges significantly limit research opportunities and create difficulties to track vulnerable older adults with geriatric syndromes.

Our work creates new opportunities for health equity research by improving and expanding the identification of vulnerable older adults in need of additional medical and social services. We aimed to extract geriatric syndromes from the free-text of EHRs. Our best model performed well at a patient level, achieving a micro-F_1 score of 0.843. Our model can be used to identify geriatric syndrome constructs from EHR notes, which could expand the coverage of geriatric syndrome in EHR systems. Additionally, our system can ensure that, despite a lack of coding for these syndromes,[5–7] all relevant cases are tracked across patients thereby improving the inclusion of vulnerable older adults in research (eg, clinical trials),[8,9] alignment of specific population health management efforts (eg, access to nursing home and assisted living),[40–42] and potentially impacting public health interventions.[11,43,44]

To the best of our knowledge, our study is the first to apply machine learning for extracting geriatric syndromes from EHR free-text to identify vulnerable older adults, and potentially addressing functional and social disparities among the geriatric population. We demonstrate a model that effectively incorporates context from the document and patient in information extraction decisions. Since most prior work on information extraction uses the sentence alone,[12,23,45,46] our model may benefit other IE tasks in identifying health disparity markers, such as social determinants of health, that are often not coded in EHRs.[44] Moreover, our model does not require any task-specific feature engineering as it relies fully on learned representations of the text.

EHR vendors have recently started to roll out specific built-in modules to collect social determinants of health as structured data at the point of care; however, common terminologies are yet to be adopted to encode such coded information properly in EHRs.[47,48] Given the lack of standardized structured social determinants of health (including a number of geriatric syndromes), deploying statistical NLP techniques (which are superior to pattern matching techniques) will enable health care providers to efficiently prescreen patients for potential underlying health disparities, narrow the denominator of vulnerable patients needed to go through other confirmatory means (eg, surveys and interviews), and effectively align social service resources.[49]

In summary, despite the increased adoption of EHRs among providers, some providers (mainly serving rural and lower socioeconomic regions) may not be able to fully mature their EHRs in the near future.[50–52] Lack of advanced EHR functionalities to identify underlying social determinants of health, including social constructs of the geriatric syndrome, may consequently limit the ability of value-based providers to address health disparities among various patient populations.[48,53–56] As EHRs are becoming a major source of risk stratification for providers,[41,57–60] incorporating advanced NLP methods to extract risk factors of social determinants of health (eg, lack of social support) can propel value-based providers to leverage EHRs to identify and adjust for potential disparities within their population health management efforts[42,61] in addressing the needs of vulnerable populations such as older adults.[40]

Technical limitations and future work

Our work focused on 3 types of contexts for improving information extraction: surrounding sentences, the entire document (clinical note), and the diagnosis codes. Other types of contexts may be beneficial, such as the containing paragraph or section. Specifically, we are interested in ways to include the entire document as context but allow the model to learn and emphasize text in closer proximity to the target sentence. Additionally, we are interested in models that would allow us to directly train on patient-level labels, instead of individual sentences. Finally, we expect that different information extraction tasks would benefit from different types of contexts. We plan to explore this by considering our model for other complex IE tasks such as identifying patients with other social determinants of health needs (eg, housing instability, food insecurity).

Our neural model is based on learning contextual representations using recurrent neural networks. For several years, recurrent neural networks, and specifically LSTM-based models, have represented the state of the art in NLP. Recently, these models have given way to new contextual representations based on Transformers,[62] including the new BERT model which has achieved high performance on several different NLP tasks.[29] We plan to explore how BERT performs in detecting social determinants of health, and how it can be augmented with the types of contexts-aware models that we have proposed in this work.

CONCLUSION

Structured data of EHRs provide an incomplete picture of geriatric syndromes and potential disparities among older adults. To identify vulnerable older adults, we presented a statistical NLP model for extracting geriatric syndromes from EHR clinical notes. We proposed a deep neural network model that incorporated context from the clinical notes and patient records to improve construct extraction. Our final model achieved a micro-F_1 of 0.843 for patient-level determination of geriatric syndrome constructs, significantly improving traditional models using target sentences alone (0.789). This NLP methodology can be adapted and used to identify other functional or social markers, such as housing instability and food insecurity, in EHR's free-text to address health equity issues among older adults.

FUNDING

This work was funded by Atrius Health and the Center for Population Health IT, Johns Hopkins University.

AUTHOR CONTRIBUTORS

All authors were involved in the conceptualization of the research. TC and MD lead the technical development and evaluation of the NLP methodology. JW and HK evaluated the findings relevance to the identification of vulnerable older adults. TC, MD, JW, and HK drafted the manuscript. All authors reviewed and commented on the final manuscript before submission.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGEMENT

We acknowledge the support of Dr Joe Kimura and Leilani Hernandez from Atrius Health in preparing and extracting the underlying data set used in this study. We also acknowledge Dr Fardad Gharghabi and Mr Tom Richards (Johns Hopkins School of Public Health) for preparing the study database and scrubbing the free-text, and the support of the administrative staff at the Center for Population Health Information Technology for enabling us to coordinate and complete this study.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Olde Rikkert MGM, Rigaud AS, van Hoeyweghen RJ, *et al*. Geriatric syndromes: medical misnomer or progress in geriatrics? *Neth J Med* 2003; 61 (3): 83–7.
2. Ironside PM, Tagliareni ME, McLaughlin B, *et al*. Fostering geriatrics in associate degree nursing education: an assessment of current curricula and clinical experiences. *J Nurs Educ* 2010; 49 (5): 246–52.
3. Inouye SK, Studenski S, Tinetti ME, *et al*. Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept. *J Am Geriatr Soc* 2007; 55 (5): 780–91.
4. Hazra NC, Dregan A, Jackson S, *et al*. Differences in health at age 100 according to sex: population-based cohort study of centenarians using electronic health records. *J Am Geriatr Soc* 2015; 63 (7): 1331–7.
5. Anzaldi LJ, Davison A, Boyd CM, *et al*. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr* 2017; 17: 248.
6. Kharrazi H, Anzaldi LJ, Hernandez L, *et al*. The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc* 2018; 66 (8): 1499–507.
7. Chen T, Dredze M, Weiner JP, *et al*. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. *JMIR Med Inform* 2019; 7 (1): e13039.
8. Welch VA, Norheim OF, Jull J, *et al*. CONSORT-Equity 2017 extension and elaboration for better reporting of health equity in randomised trials. *BMJ* 2017; 359: j5085.
9. Jull J, Whitehead M, Petticrew M, *et al*. When is a randomised controlled trial health equity relevant? Development and validation of a conceptual framework. *BMJ Open* 2017; 7 (9): e015815.
10. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. *J Gen Intern Med* 2014; 29 (7): 976–8.
11. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. *Yearb Med Inform* 2018; 27: 199–206.
12. Elhadad N, Pradhan S, Gorman S, *et al*. SemEval-2015 task 14: analysis of clinical text. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*; June 4–5, 2015. Denver, Colorado, USA. Stroudsburg, PA: Association for Computational Linguistics; 2015: 303–10.
13. Xu H, Stenner SP, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010; 17 (1): 19–24.

14. Young T, Hazarika D, Poria S, *et al*. Recent trends in deep learning based natural language processing [review article]. *IEEE Comput Intell Mag* 2018; 13 (3): 55–75.
15. Shickel B, Tighe PJ, Bihorac A, *et al*. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2018; 22 (5): 1589–604.
16. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
17. Apache: Open NLP. https://opennlp.apache.org. Accessed June 9, 2019.
18. Socher R, Perelygin A, Wu J, *et al*. Recursive deep models for semantic compositionality over a sentiment tree bank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*; October 18–21, 2013; Stroudsburg, PA. 1631–42.
19. Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics; June 25 – 30, 2005; Ann Arbor, Michigan, USA. 115–24.
20. Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. 21–26 July , 2004; Barcelona, Spain. Stroudsburg, PA: Association for Computational Linguistics; 2004.
21. Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. *Lang Res Eval* 2005; 39 (2–3): 165–210.
22. Li X, Roth D. Learning question classifiers. In: *Proceedings of the 19th International Conference on Computational Linguistics;* August 24 - September 01, 2002; Taipei, Taiwan—*Volume 1*. Stroudsburg, PA: Association for Computational Linguistics; 2002: 1–7.
23. Kim Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; October 25–29, 2014; Doha, Qatar. Stroudsburg, PA: Association for Computational Linguistics. doi: 10.3115/v1/d14-1181.
24. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; June 22–27, 2014; Baltimore, Maryland, USA. doi: 10.3115/v1/p14-1062. Stroudsburg, PA: Association for Computational Linguistics.
25. dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. August 23–29, 2014; Dublin, Ireland. Stroudsburg, PA: Association for Computational Linguistics; 2014: 69–78.
26. Wang X, Liu Y, Chengjie SUN, *et al*. Predicting polarities of tweets by composing word embeddings with long short-term memory. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Beijing, China; July 27 – 31, 2015. Stroudsburg, PA:Association for Computational Linguistics. doi: 10.3115/v1/p15-1130.
27. Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Beijing, China July 27 – 31, 2015. Stroudsburg, PA: Association for Computational Linguistics. doi: 10.3115/v1/p15-1150.
28. Lin Z, Feng M, dos Santos CN, *et al*. A structured self-attentive sentence embedding. In: *The 5th International Conference on Learning Representations (ICLR 2017)*; Toulon, France. April 24 – 26, 2017; OpenReview.-net.
29. Devlin J, Chang M-W, Lee K, *et al*. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019; June 2 - 7, 2019; Minneapolis, Minnesota, USA; Stroudsburg, PA: Association for Computational Linguistic, 2019: 4171–4186.
30. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2*. December 05 – 10, 2013; Lake Tahoe, Nevada, USA. Red Hook, NY, USA: Curran Associates Inc; 2013: 3111–9.
31. Yang Z, Yang D, Dyer C, *et al*. Hierarchical attention networks for document classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Rochester, New York, USA. April 23–25, 2007. Stroudsburg, PA: Association for Computational Linguistics; 2016: 1480–9.
32. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. Beijing, China. June 21–26, 2014. jmlr.org. 2014: 1188–96.
33. Choi E, Bahadori MT, Searles E, *et al*. Multi-layer representation learning for medical concepts. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16*. San Francisco, CA, USA; Aug 13 – 17, 2016. New York, NY, USA: ACM. doi: 10.1145/2939672.2939823.
34. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, *et al*., eds. *Advances in Neural Information Processing Systems 26*. Red Hook, NY, US: Curran Associates; 2013: 3111–9.
35. Abadi M, Barham P, Chen J, *et al*. TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. November 02 – 04, 2016; Savannah, GA, USA Berkeley, CA: USENIX Association; 2016: 265–83.
36. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *The 3rd International Conference for Learning Representations (ICLR)*. May 7 – 9, 2015. San Diego, CA, USA. JMLR.org. http://arxiv.org/abs/1412.6980.
37. Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15: 1929–58.
38. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; 12 (2): 153–7.
39. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998; 10 (7): 1895–923.
40. Kan HJ, Kharrazi H, Leff B, *et al*. Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. *Med Care* 2018; 56 (3): 233–9.
41. Kharrazi H, Chi W, Chang H-Y, *et al*. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med Care* 2017; 55 (8): 789–96.
42. Kharrazi H, Lasser EC, Yasnoff WA, *et al*. A proposed national research and development agenda for population health informatics: summary recommendations from a national expert workshop. *J Am Med Inform Assoc* 2017; 24 (1): 2–12.
43. Sadana R, Blas E, Budhwani S, *et al*. Healthy ageing: raising awareness of inequalities, determinants, and what could be done to improve health equity. *Geront* 2016; 56 Suppl 2: S178–93.
44. Hatef E, Weiner JP, Kharrazi H. A public health perspective on using electronic health records to address social determinants of health: the potential for a national system of local community health records in the United States. *Int J Med Inform* 2019; 124: 86–9.
45. Pradhan S, Elhadad N, Chapman W, *et al*. SemEval-2014 task 7: analysis of clinical text. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland. August 23–24, 2014. Stroudsburg, PA: Association for Computational Linguistics; 2014: 54–62.

46. Hughes M, Li I, Kotoulas S, *et al*. Medical text classification using convolutional neural networks. *Stud Health Technol Inform* 2017; 235: 246–50.
47. Bazemore AW, Cottrell EK, Gold R, *et al*. Community vital signs': incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc* 2016; 23 (2): 407–12.
48. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. *Health Aff (Millwood)* 2018; 37 (4): 585–90.
49. Vest JR, Grannis SJ, Haut DP, *et al*. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform* 2017; 107: 101–6.
50. Kharrazi H, Gonzalez CP, Lowe KB, *et al*. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. *J Med Internet Res* 2018; 20 (8): e10458.
51. Chan KS, Kharrazi H, Parikh MA, *et al*. Assessing electronic health record implementation challenges using item response theory. *Am J Manag Care* 2016; 22: e409–15.
52. Adler-Milstein J, DesRoches CM, Kralovec P, *et al*. Electronic health record adoption in US hospitals: progress continues, but challenges persist. *Health Aff* 2015; 34 (12): 2174–80.
53. Hughes LS, Phillips RL Jr, DeVoe JE, *et al*. Community vital signs: taking the pulse of the community while caring for patients. *J Am Board Fam Med* 2016; 29 (3): 419–22.
54. Hatef E, Lasser EC, Kharrazi HHK, *et al*. A population health measurement framework: evidence-based metrics for assessing community-level population health in the global budget context. *Popul Health Manag* 2018; 21: 261–70.
55. Hatef E, Kharrazi H, VanBaak E, *et al*. A state-wide health IT infrastructure for population health: building a community-wide electronic platform for Maryland's all-payer global budget. *Online J Public Health Inform* 2017; 9: e195.
56. Kharrazi H, Hatef E, Lasser E, *et al*. A guide to using data from Johns Hopkins electronic health record for behavioral, social and systems science research: identifying research needs and assessing the availability of behavioral and social variables. Johns Hopkins Institute for Clinical and Translational Research; 2018. https://ictr.johnshopkins.edu/wp-content/uploads/Phase2.Epic_.Social.Guide_2018.06.30_final.pdf. Accessed April 10, 2019.
57. Chang H-Y, Richards TM, Shermock KM, *et al*. Evaluating the impact of prescription fill rates on risk stratification model performance. *Med Care* 2017; 55 (12): 1052–60.
58. Lemke KW, Gudzune KA, Kharrazi H, *et al*. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. *Am J Manag Care* 2018; 24 (6): e190–5.
59. Kharrazi H, Chang H-Y, Heins SE, *et al*. Assessing the impact of body mass index information on the performance of risk adjustment models in predicting health care costs and utilization. *Med Care* 2018; 56 (12): 1042–50.
60. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. *Med Care* 2018; 56 (2): 202–3.
61. Hatef E, Searle KM, Predmore Z, *et al*. The impact of social determinants of health on hospitalization in the veterans health administration. *Am J Prev Med* 2019; 56 (6): 811–18.
62. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, *et al*., eds. *Advances in Neural Information Processing Systems 30*. Red Hook, NY, US: Curran Associates; 2017: 5998–6008.