Identifying Symptoms of Delirium from Clinical Narratives Using Natural Language Processing

Aokun Chen†
Department of Health Outcomes and Biomedical Informatics
University of Florida
Gainesville, FL, USA
chenaokun1990@ufl.edu

Daniel Paredes†
Department of Health Outcomes and Biomedical Informatics
University of Florida
Gainesville, FL, USA
dparedespardo@ufl.edu

Zehao Yu
Department of Health Outcomes and Biomedical Informatics
University of Florida
Gainesville, FL, USA
zehao.yu@ufl.edu

Xiwei Lou
Department of Health Outcomes and Biomedical Informatics
University of Florida
Gainesville, FL, USA
lou.x@ufl.edu

Roberta Brunson
UF Health Shands Hospital
Gainesville, FL, USA
brunsr@ufl.edu

Jamie N. Thomas
UF Health Shands Hospital
Gainesville, FL, USA
thomjn@ufl.edu

Kimberly A. Martinez
UF Health Shands Hospital
Gainesville, FL, USA
handki@shands.ufl.edu

Robert J. Lucero
School of Nursing
University of California Los Angeles
Los Angeles, CA, USA
rlucero@sonnet.ucla.edu

Tanja Magoc
UF Clinical and Translational Science Institute
University of Florida
Gainesville, FL, USA
tmagoc@ufl.edu

Laurence M. Solberg
North Florida/South Georgia Veterans Health Service
Geriatrics Research, Education, and Clinical Center (GRECC)
Gainesville, FL, USA
lmsolberg@ufl.edu

Urszula A. Snigurska
College of Nursing
University of Florida
Gainesville, FL, USA
usnigurska@ufl.edu

Sarah E. Ser
Department of Epidemiology
University of Florida
Gainesville, FL, USA
sser@ufl.edu

Mattia Prosperi
Department of Epidemiology
University of Florida
Gainesville, FL, USA
m.prosperi@ufl.edu

Jiang Bian
Department of Health Outcomes and Biomedical Informatics
University of Florida
Gainesville, FL, USA
bianjiang@ufl.edu

Ragnhildur I. Bjarnadottir*
College of Nursing
University of Florida
Gainesville, FL, USA
rib@ufl.edu

Yonghui Wu*
Department of Health Outcomes and Biomedical Informatics
University of Florida
Gainesville, FL, USA
yonghui.wu@ufl.edu

2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-8373-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHI61247.2024.00046

***Abstract*—*Delirium is an acute decline or fluctuation in attention, awareness, or other cognitive function that can lead to serious adverse outcomes. Despite the severe outcomes, delirium is frequently unrecognized and uncoded in patients' electronic health records (EHRs) due to its transient and diverse nature. Natural language processing (NLP), a key technology that extracts medical concepts from clinical narratives, has shown great potential in studies of delirium outcomes and symptoms. To assist in the diagnosis and phenotyping of delirium, we formed an expert panel to categorize diverse delirium symptoms, composed annotation guidelines, created a delirium corpus with diverse delirium symptoms, and developed NLP methods to extract delirium symptoms from clinical notes. We compared 5 state-of-the-art transformer models including 2 models (BERT and RoBERTa) from the general domain and 3 models (BERT_MIMIC, RoBERTa_MIMIC, and GatorTron) from the clinical domain. GatorTron achieved the best strict and lenient F1 scores of 0.8055 and 0.8759, respectively. We conducted an error analysis to identify challenges in annotating delirium symptoms and developing NLP systems. To the best of our knowledge, this is the first large language model-based delirium symptom extraction system. Our study lays the foundation for the future development of computable phenotypes and diagnosis methods for delirium.*

***Keywords*—*delirium, medication information extraction, named entity recognition, deep learning, clinical natural language processing.*

I. Introduction

Delirium, defined as any episode of acute fluctuation in attention, awareness, and any other cognitive function(s), is one of the most common and costly iatrogenic conditions, particularly among older adults.[1]–[5] Among hospitalized older adult patients, 11-15% were estimated to have prevalent delirium while another 29-31% would develop incident delirium during the hospital stay.[6], [7] Delirium is associated with serious adverse outcomes, including a higher likelihood of

† Equally contributed
* Corresponding author

2575-2634/24/$31.00 ©2024 IEEE
DOI 10.1109/ICHI61247.2024.00046

hospital mortality, increased length of hospital stays, greater risk of one-year mortality after discharge, functional decline, and increased caregiver burden.[2]–[4] However, despite its severe sequelae, delirium is frequently missed and under-coded.[8]–[12] In fact, one study found that less than 3% of patients with delirium had the International Classification of Diseases, Ninth Revision, (ICD-9) codes for delirium in their electronic health record (EHR) charts.[12] This was partially due to the transient nature of delirium and the difficulty in its identification.[13] As real-time identification is difficult, the prognostic study of delirium and its interventions has become extremely important. Nevertheless, training a computational model to recognize delirium using only patients with certain diagnostic codes, as opposed to all patients who have delirium, is jeopardized by the error from under-coded cases. Thus, there is an urgent need for developing and deploying natural language processing (NLP) systems to identify symptoms of delirium to help identify delirium cases and inform clinicians of the interventions' effectiveness that ultimately improve the diagnosis methods for delirium.

Identifying symptoms of delirium is a typical clinical concept extraction task, which is a fundamental NLP task to identify concepts of important clinical meaning from clinical narratives. Previously, both rule-based and machine learning-based approaches have been applied. Rule-based solutions are good at recognizing concepts with fixed patterns and fewer variations such as dates and numeric sizes (e.g., tumor size), while machine learning models have good generalizability to recognize free-text concepts with diverse documentation variations such as symptoms. Clinical NLP systems such as cTAKES[14], MetaMap[15], [16], and MedTagger[17] have been developed to extract general clinical concepts defined by the Unified Medical Language System (UMLS). CLAMP is a machine learning-based NLP system to facilitate the full life cycle of clinical NLP development from annotation to model training. Recent progress in NLP has greatly improved clinical concept extraction from clinical narratives. Deep learning-based large language models (LLM) trained using transformer architecture have become the state-of-the-art solution for many NLP tasks, including named-entity recognition, relation extraction, natural language inference, and question answering. Bidirectional Encoder Representations from Transformers (BERT) is one of the popular transformer structures.[18] In transformer-based models, the training process was split into pre-training with large, unlabeled data, and fine-tuning, where a small set of labeled task-specific data was involved. This structure enables the transfer learning ability: - one transformer-based model can be applied to many NLP tasks.

Previous studies have applied NLP in studies of delirium diagnosis and outcomes. Wang et al. developed an NLP system to detect the sentiment from radiology reports to help identify delirium cases and reported that NLP-derived sentiment information improved the identification rate of delirium cases.[19] Fu et al. built two rule-based NLP systems, i.e., NLP-CAM and NLP-mCAM, based on the Confusion Assessment Method (CAM) and the modified Confusion Assessment Method (mCAM), to identify delirium patients from electronic health records.[20] Later, Pagali et al. further utilized the NLP-CAM to assess the analysis of delirium within COVID-19 patients.[13] Ge et al. explored SVM, CNN-LSTM, and BERT in identifying delirium-related sentences from clinical narratives.[21] They used a set of keywords to label delirium-related sentences using regular expressions. Up until now, there has been no NLP solution to systematically identify and categorize delirium symptoms from clinical narratives.

To assist in and improve the diagnosis of delirium, we composed annotation guidelines, created a corpus, and developed NLP methods to extract delirium symptoms from clinical narratives. We systematically explored 5 transformer-based large language models, i.e., BERT,[18] BERT_MIMIC, RoBERTa,[22] RoBERTa_MIMIC, and GatorTron[23]. To the best of our knowledge, this is the first LLM-based delirium symptom extraction NLP system. Our study lays the foundation for the future development of computable phenotypes and diagnosis methods for delirium.

II. Methods

A. Data Source

We extracted patient EHR data from the University of Florida (UF) Health Integrated Data Repository (IDR), including UF Health Shands in Gainesville, FL and UF Health Jacksonville in Jacksonville, FL. The IDR is a clinical data warehouse that collects and aggregates information from various clinical and administrative information systems across UF Health clinical and research enterprises, including the Epic EHR system. The IDR contains more than 2 billion observational facts pertaining to more than 2 million patients. We identified a cohort of 170,868 patients using the following criteria: adults (i.e., ≥ 18 years of age) who were admitted to one of the 21 medical-surgical units at the UF Health hospitals (including both UF Health Gainesville and Jacksonville) from 2012 to 2021. This study was approved by the UF Institutional Review Board (IRB #201900208).

TABLE I. Demographics of the cohort.

Demographics	Total (N = 170,868)
Age, mean (SD)	
Mean (SD)	56.8 (17.2)
Sex	
Female (%)	83,415 (48.82%)
Male (%)	87,449 (51.18%)
Unknown (%)	4 (<0.1%)
Race-Ethnicity	
Non-Hispanic White (%)	113,370 (66.35%)
Non-Hispanic Black (%)	42,266 (24.74%)
Non-Hispanic Other (%)	2,100 (1.23%)
Hispanic (%)	7,313 (4.28%)
Unknown (%)	5,819 (3.41%)

TABLE I summarizes the demographics of this cohort. The average age of this cohort was 56.8 years, and the cohort had slightly more male patients (male: 51.18%). The majority of patients were non-Hispanic White (66.35%), and the second largest racial-ethnic group was non-Hispanic Black (24.74%). We identified a total of 6.9 million clinical notes from this cohort.

B. Annotation

Identify applicable funding agency here. If none, delete this text box.

As not all nurses' notes have documented signs and symptoms of delirium, we applied a snowball sampling strategy to curate a set of keywords that are related to delirium signs and symptoms, and that can be, therefore, used to identify a subset of relevant notes. Specifically, domain experts (RIB, LMS, UAS) first initiated a list of keywords of common delirium signs and symptoms. We then iteratively reviewed batches of 30 notes to identify new keywords to extend the keyword list and re-identify delirium-related notes, until there were no new keywords identified. A total of 45 keywords were identified using this snowball strategy. Then, we filtered 6.9 million notes using the 45 keywords to identify notes with at least 3 delirium keywords and randomly selected 600 notes for manual annotation. We formed an expert panel of NLP researchers (YW, AC, DP), clinical researchers (RIB, LMS, UAS, RJL), and practicing nurses (RB, JNT, KAM) to develop annotation guidelines. The expert panel reviewed delirium signs and symptoms and categorized them into 8 categories:

(1) **Disturbed attention**: defined as reduced ability to direct, focus, shift, or sustain attention, or reduced orientation to the environment (e.g., "Disoriented", "unable to follow directions", "confused"). Altered mental status is also included.

(2) **Disturbed perception**: defined as reduced ability to identify, organize, and interpret sensory information. Examples include hallucinations, illusions, misinterpretations, and various delusions (e.g., "trying to poison me", "stealing from me").

(3) **Psychomotor activity**: defined as problematic behaviors and restless physical activity arising from mental tension; often purposeless and unintentional, or slow physical activity arising from inhibition of mental activity. For example, "pulling off tubes", "combative", "restless", "spitting".

(4) **Fluctuations**: defined as changes in symptoms during the course of the day and night; for example, "becomes more agitated", "progressively", "increasingly".

(5) **Memory deficit**: defined as reduced ability to encode, store, and retrieve information when needed. For example, inability to remember recent events (e.g., taking medications) or instructions (e.g., to call for help), forgetfulness, forgetting, can't recall/can't remember, no recollection, memory loss (e.g., "forgetfulness", "did not know why he was brought here", "short term memory loss").

(6) **Consciousness level**: defined as waking state (wakefulness); a condition of awareness of one's surroundings, generally coupled with an ability to communicate with others or to signal understanding of what is being communicated by others. Examples include drowsy, lethargic, obtunded, stuporous, coma, etc. (e.g., "unable to stay awake", "unresponsive", "lethargic").

(7) **Disturbed sleep**: defined as disturbed sleep-wake cycles circadian state characterized by partial or total suspension of consciousness, voluntary muscle inhibition, and relative insensitivity to stimulation. For example, daytime sleepiness and nighttime agitation (e.g., "trouble falling asleep", "poor sleep", "minimal sleep").

(8) **Disorganized thinking**: defined as a disrupted form or structure of thinking; manifested in disorganized speech. For example, rambling or irrelevant conversation (tangentiality), a large amount of nonessential information (circumstantiality), unclear or illogical flow of ideas (loose associations), unpredictable switching from subject to subject (derailment), mumbling, incoherent, illogical, rambling (e.g., "unable to clearly verbalize", "fixated on", "screaming incoherent words").

We also used an 'Other' category to capture "hard" cases not covered by the annotation guidelines. The categories were created and pre-defined based on the diagnostic criteria which were described in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V). Following standard NLP development practice, we recruited three annotators and conducted training sessions to train the annotators until a good agreement score was achieved. The three annotators manually reviewed the 546 notes to identify eight categories of delirium symptoms. We monitored the annotation agreement and periodically collected the discrepancies from annotators and instances annotated as 'Other' and discussed them in expert panel meetings. We also improved the annotation guidelines as necessary to cover new cases and solve discrepancies as needed.

C. Delirium symptom extraction using transformer models

We approached the delirium symptom extraction as a clinical concept extraction (or named-entity recognition [NER]) task and adopted the standard beginning-inside–outside (BIO) annotation format. We applied tokenization, sentence boundary detection, and BIO format transformation using a preprocessing pipeline from a previous study.[24] Then, we applied transformer models to identify delirium symptoms. Specifically, we generated distributed representations of text using the transformer models and calculated probability scores for each BIO category using a linear layer with a softmax activation function. The cross-entropy loss was used for fine-tuning.

D. Transformer-based deep learning models

We explored 5 pretrained LLMs, including 2 transformer models (BERT and RoBERTa) for general English domain and 3 transformer models (BERT_MIMIC, RoBERTa_MIMIC, and GatorTron) for clinical domain.

BERT and BERT_MIMIC: Following the introduction of the transformer model by Vaswani et al.[25], Devlin et al. improved it with Bidirectional Encoder Representations from Transformers (BERT)[18]. BERT used bidirectional representations and an encoder structure that improved the performance of fine-tuning the pre-trained model. BERT_MIMIC followed the same structure of BERT but pre-trained with clinical notes from the Medical Information Mart for Intensive Care (MIMIC) dataset. In our experiment, we adopted the BERT model implemented in Hugginface.

RoBERTa and RoBERTa_MIMIC: Liu et al. optimized the training strategies of BERT and created RoBERTa.[22] RoBERTa introduced new strategies including dynamic masking, full sentence sampling, large mini-batches, large byte level encoding, and removed next sentence prediction loss. RoBERTa_MIMIC utilized the same optimization as RoBERTa but trained over the MIMIC dataset. In our experiment, we explored the RoBERTa model implemented in Hugginface.

GatorTron model: We adopted the GatorTron model in the delirium symptom extraction system. GatorTron is a BERT-style large clinical language model. We pretrained GatorTron with >90 billion words of text, including >80 billion words from >290 million notes identified at the UF Health system covering patient records from 2011–2021 from over 126 clinical departments and ~50 million encounters. These clinical narratives covered healthcare settings including but not limited to inpatient, outpatient, and emergency department visits. We used the GatorTron model with 345 million parameters for this study.[23]

E. Training strategies

For delirium symptom extraction and classification, we adopted the standard NER training procedure to recognize and classify delirium symptoms using a training set and a development set. We trained the models using the training set (train) of 381 notes and monitored the performance on the development set (dev) of 55 notes. The best of each transformer model was selected based on the validation performance on the development set.

F. Experiment and evaluation

We reused the pretrained models from the public GitHub repository for two transformer models from the general domain, including BERT and RoBERTa. For the two clinical transformer models, we adopted the BERT_MIMIC and RoBERTa_MIMIC models developed by fine-tuning the general models using clinical text from the MIMIC III database in our previous study.[24] The GatorTron model was developed by training from scratch using >90 billion words of text (including >82 billion words of de-identified clinical text from UF Health) in our previous studies.[23] We evaluated our delirium symptom extraction system with the test set of 110 notes on both strict (i.e., exact boundary surface string match and entity type) and lenient (i.e., partial boundary match over the surface string) precision, recall, and F1-score. The evaluation scores were calculated with the evaluation script from a pipeline implemented in a previous study.[24]

III. Results

After annotation, we excluded 54 notes that either were duplicated or without valid delirium mentions. We annotated a total of 2,496 concepts for 8 categories of delirium symptoms from a total of 546 clinical notes. The annotation agreement measured by F1-score among the 3 annotators improved from 29.2% to 97.1% among the different batches. All the notes were annotated independently by at least two annotators. **TABLE II.** shows the summary statistics for this dataset.

The top 3 most frequently mentioned categories of symptom concepts are psychomotor activity (train vs. dev vs. test = 45.53% : 39.91% : 43.71%), disturbed attention (23.51% : 23.25% : 22.81%), and consciousness level (12.62% : 24.12% : 14.93%). Comparing train, dev, and test set, training set had slightly more disturbed attention concepts (23.51% : 23.25% : 22.81%), psychomotor activity concepts (45.53% : 39.91% : 43.71%), and fluctuation concepts (6.89% : 6.14% : 6.61%) than the dev or test set, while also having noticeably more disturbed perception concepts (3.56% : 0.44% : 2.99%), especially compared to the dev set, while dev set had significantly more consciousness level mentions (12.62% : 24.12% : 14.93%), disorganized thinking mentions (2.22% : 3.07% : 1.49%), and slightly more disturbed sleep mentions (1.72% : 2.19% : 2.13%) than train and test set. Finally, the test set contained significantly more memory deficit concepts (3.95% : 0.88% : 5.33%) than the train and dev sets.

TABLE II. Summary of delirium symptom concepts.

Symptom concepts	Train (%)	Dev (%)	Test (%)
Disturbed attention	423 (23.51%)	53 (23.25%)	107 (22.81%)
Disturbed perception	64 (3.56%)	1 (0.44%)	14 (2.99%)
Psychomotor activity	819 (45.53%)	91 (39.91%)	205 (43.71%)
Fluctuations	124 (6.89%)	14 (6.14%)	31 (6.61%)
Memory deficit	71 (3.95%)	2 (0.88%)	25 (5.33%)
Consciousness level	227 (12.62%)	55 (24.12%)	70 (14.93%)
Disturbed sleep	31 (1.72%)	5 (2.19%)	10 (2.13%)
Disorganized thinking	40 (2.22%)	7 (3.07%)	7 (1.49%)

TABLE III. compares the performance of 5 transformer models for delirium symptom extraction. GatorTron achieved the best strict and lenient F1-scores of 0.8055 and 0.8759, second by BERT_MIMIC (strict: 0.7901, lenient: 0.8724). The RoBERTa model achieved the best recall. All clinical transformer models, except RoBERTA_MIMIC, outperformed general transformer models on strict F1-score.

TABLE III. Comparison of 5 transformer models using the overall performance.

Model	Strict			Lenient		
	Pre	Rec	F1	Pre	Rec	F1
BERT	0.7520	0.8022	0.7763	0.8453	0.8883	0.8663
BERT_MIMIC	0.7705	0.8107	0.7901	0.8537	0.8920	0.8724
RoBERTa	0.7508	**0.8265**	0.7868	0.8265	**0.9078**	0.8626
RoBERTa_MIMIC	0.7728	0.8010	0.7867	0.8590	0.8799	0.8693
GatorTron	**0.7993**	0.8119	**0.8055**	**0.8696**	0.8823	**0.8759**

*Pre: precision; Rec: recall; F1: f1-score.

**Best Strict and Lenient precision, recall, and F1-scores are highlighted in bold.

TABLE IV. further breaks down the performance of GatorTron on different delirium symptoms. GatorTron achieved the best extraction performance on consciousness level and disturbed sleep with a strict F1-score over 0.9. For most of the symptoms, GatorTron demonstrated good performance with strict F1-scores between 0.73 and 0.88.

TABLE IV. Detailed performance of GatorTron for each category of delirium symptoms.

Symptoms	Strict			Lenient		
	Pre	Rec	F1	Pre	Rec	F1
Disturbed attention	0.7960	0.8556	0.8247	0.8657	0.9305	0.8969
Disturbed perception	0.7200	0.7826	0.7500	0.7600	0.8261	0.7917
Psychomotor activity	0.8191	0.7918	0.8052	0.9040	0.8715	0.8874
Fluctuations	0.7755	0.7037	0.7379	0.8367	0.7593	0.7961
Memory deficit	0.8293	0.9189	0.8718	0.8537	0.9459	0.8974

Consciousness level	0.9286	0.8835	0.9055	0.9694	0.9223	0.9453
Disturbed sleep	0.9333	0.8750	0.9032	0.9333	0.8750	0.9032
Disorganized thinking	0.2857	0.4000	0.3333	0.4762	0.6667	0.5556

*Pre: precision; Rec: recall; F1: f1-score.

TABLE V. shows the strict precision, recall, and F1-score of the best performed transformer model for each delirium symptom. The best model determined by the micro-average F1-score over all categories, GatorTron, achieved the best performance in 4 out of 8 delirium symptoms (i.e., disturbed attention, psychomotor activity, consciousness level, disturbed sleep), followed by RoBERTa on 2 categories of symptoms (memory deficit, disorganized thinking), and BERT and BERT MIMIC on the remaining 2 categories of symptoms (fluctuations, disturbed perception).

TABLE V. THE BEST PERFORMANCE AND MODEL FOR EACH DELIRIUM CATEGORY.

Symptoms	Model	Precision	Recall	F1-score
Disturbed attention	GatorTron	0.7960	0.8556	0.8247
Disturbed perception	BERT MIMIC	0.7826	0.7826	0.7826
Psychomotor activity	GatorTron	0.8191	0.7918	0.8052
Fluctuations	BERT	0.7593	0.7593	0.7593
Memory deficit	RoBERTa	0.8537	0.9459	0.8974
Consciousness level	GatorTron	0.9286	0.8835	0.9055
Disturbed sleep	GatorTron	0.9333	0.8750	0.9032
Disorganized thinking	RoBERTa	0.3500	0.4667	0.4000

IV. DISSCUSSION AND CONCLUSION

Identifying and categorizing delirium symptoms from clinical narratives is critical for the diagnosis and phenotyping of delirium. This study examined the documentation of delirium symptoms in clinical narratives and categorized them into 8 groups, created a corpus of delirium symptoms, and examined 5 transformer-based NLP models for extraction of delirium symptoms from clinical narratives. The best NLP model, GatorTron, achieved the best strict and lenient F1-scores of 0.8055 and 0.8759, respectively. All clinical transformers, except RoBERTa_MIMIC, outperformed the general-purpose transformers trained using general English text by a large margin, which is consistent with our previous observation that domain-specific transformers were outperforming general purpose transformer models.

Among the 8 categories of delirium symptoms, all transformers performed well for identifying memory deficit, consciousness level, and disturbed sleep symptoms, but struggled for disorganized thinking symptoms. We conducted an error analysis and generated a confusion matrix based on the best model, i.e., GatorTron. Figure 1 shows the confusion matrix, where the true labels are presented on the y-axis and the predicted labels are on the x-axis. As shown in **Fig. 1**, most errors in identifying "disorganized thinking" came from false negatives – our system missed the concepts labeled by human experts. Among the annotated delirium symptoms, "disorganized thinking" encompasses complicated situations, i.e., irrelevant conversation (tangentiality), a large amount of nonessential information (circumstantiality), unclear or illogical flow of ideas (loose associations), unpredictable switching from subject to subject (derailment), mumbling, incoherent, illogical, rambling. When combined with the lack of mentions in the training set (2.22% of all concepts), the impact on the performance of our clinical NLP system was much higher than the symptoms with simpler situations, e.g., "disturbed sleep" and "fluctuation". Upon first glance, this reveals one straightforward route to improve our current delirium symptom extraction system: increase the number of notes in our annotation. As shown in our results, the major bottlenecks are the concepts within disorganized thinking and disturbed sleep. Since there was a lack of annotated concepts in these categories (disorganized thinking: 2.22%, disturbed sleep: 1.72%), any additional annotated concepts would have great potential to improve model performance. However, the low ratio of these concepts in the notes also implies that adding even a few disorganized thinking or disturbed sleep concepts would require a large amount of manual annotation. In contrast, considering the labor-intensive nature of manual annotation, it becomes evident that enhancing the few-shot learning capability in Language Model-based Models (LLMs) presents a more promising avenue for long-term research. Such an advancement not only holds significance for delirium symptom extraction but also for addressing limitations in other clinical Natural Language Processing (NLP) tasks that suffer from insufficient concepts in annotation.

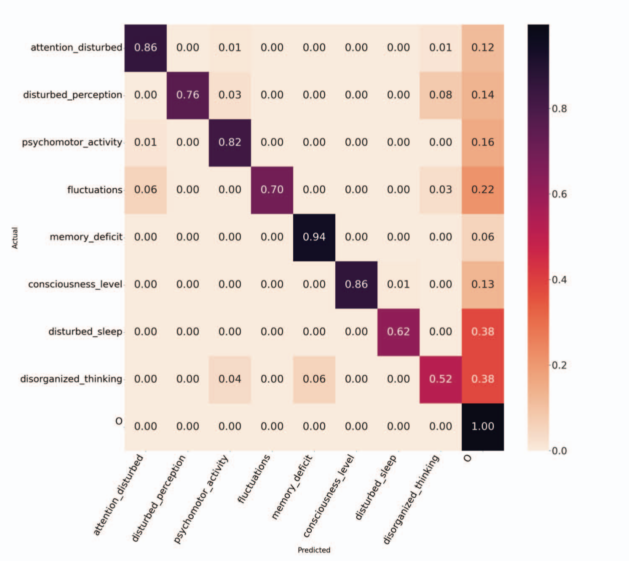

Fig. 1. Confusion matrix of the GatorTron model, where the true labels are presented on the y-axis and the predicted labels are on the x-axis.

Delirium is a complex condition that is difficult to diagnose, which is reflected in the annotation of our corpus.[26], [27] The judgement of whether a phrase is a delirium symptom or not highly depends on the context. For example, certain behaviors, such as ‘pacing’, ‘restlessness’, or ‘repeatedly asking questions’, could be symptoms of an altered mental state, or simply be a normal reaction to stress related to a healthcare issue. In addition, some well-known serious potential symptoms of delirium, such as hallucinations or delusions, were rarely documented in the clinical notes. Instead, clinicians often describe details of occurrences where patients were behaving erratically, often using direct quotes of irrational statements of patients. Therefore, our annotators annotated many long phrases (e.g., “claimed nurses were trying to kill him” – ‘disturbed perception’), in which they found it was difficult to be consistent on the boundaries. This was reflected by the gap between strict and lenient F1 scores. Another unique difficulty related to this annotation task was the importance of patients’ intent versus ability. For example, documentation of a patient being “unable to follow commands” was consistently annotated as a delirium symptom in many instances. However, a higher level of uncertainty emerged when a patient was described as “unwilling to follow commands” or “did not follow commands”. In those instances, more context from the document-level would be required to determine whether this was a potential symptom of delirium or not. There are also challenges when differentiating a clinical judgement of delirium symptoms and potential/historical risks. For example, clinical notes often include documentation of a history of symptoms, assessment if a patient is at risk of an event, and whether the patient exhibited certain behaviors or met specific conditions, e.g., “Pt has bilateral wrist restraints for risk of pulling out medical equipment.” and “Patient attempted to remove trach, DHT, and JP drain”.

Compared with previous NLP studies on delirium patients, our study classified delirium symptoms into 8 categories, which capture more detailed information about the progression of the disease; we also applied state-of-the-art clinical transformer models to ensure the accuracy of extraction. Our future work is two-fold. We will continue to improve the performance of the extraction system by improving LLMs on few shots learning, expanding concepts in annotation, and improving boundary consistency. We will also explore the NLP-extracted delirium symptoms for computable phenotyping of delirium to help develop efficient diagnosis methods.

Data Availiability

The datasets contain personal identifiable information and thus are not public available.

The computer codes to train GatorTron models are available from:

https://github.com/NVIDIA/Megatron-LM and https://github.com/NVIDIA/NeMo

The computer codes for preprocessing of text data are available from:

https://github.com/uf-hobi-informatics-lab/NLPreprocessing

https://github.com/uf-hobi-informatics-lab/GatorTron

The GatorTron models are available from:

https://huggingface.co/UFNLP/gatortron-base

https://huggingface.co/UFNLP/gatortronS

Acknowledgment

We would like to thank our practicing nurse annotators (RB, JNT, KAM) for their valuable contributions to this study despite the challenges posed by the COVID-19 pandemic. We gratefully acknowledge the support of NVIDIA Corporation and NVAITC with the donation of the GPUs used for this research.

Funding Statement

This study was partially supported by grants from the National Institutes of Health (NIH), National Institute on Aging, R33AG062884, R56AG069880, R01AG080624, R01AG080991, R01AG076234, a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), and the UF Clinical and Translational Science Institute (UL1TR001427). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

Contribution Statement

AC, DP, RIB and YW were responsible for the overall design, development, and evaluation of this study. AC, DP, and ZY developed the NLP systems. DP, RB, JNT, KAM, RJL, LMS, and UAS prepared and annotated the clinical texts. AC, DP, and YW did the bulk of the writing; XL, TM, LMS, UAS, SES, MP, RJL, RIB, and JB also contributed to the writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

References

[1] D. L. Leslie and S. K. Inouye, “The importance of delirium: economic and societal costs,” J. Am. Geriatr. Soc., vol. 59 Suppl 2, no. Suppl 2, pp. S241-3, Nov. 2011.

[2] S. K. Inouye, R. G. J. Westendorp, and J. S. Saczynski, “Delirium in elderly people,” Lancet, vol. 383, no. 9920, pp. 911–922, Mar. 2014.

[3] S. Wass, P. J. Webster, and B. R. Nair, “Delirium in the elderly: a review,” Oman Med. J., vol. 23, no. 3, pp. 150–157, Jul. 2008.

[4] T. G. Fong, S. R. Tulebaev, and S. K. Inouye, “Delirium in elderly adults: diagnosis, prevention and treatment,” Nat. Rev. Neurol., vol. 5, no. 4, pp. 210–220, Apr. 2009.

[5] “Neurocognitive Disorders,” in Diagnostic and Statistical Manual of Mental Disorders, in DSM Library. American Psychiatric Association Publishing, 2022.

[6] E. E. Vasilevskis, J. H. Han, C. G. Hughes, and E. W. Ely, “Epidemiology and risk factors for delirium across hospital settings,” Best Pract. Res. Clin. Anaesthesiol., vol. 26, no. 3, pp. 277–287, Sep. 2012.

[7] K. Gibb et al., “The consistent burden in published estimates of delirium occurrence in medical inpatients over four decades: a systematic review and meta-analysis study,” Age Ageing, vol. 49, no. 3, pp. 352–360, Apr. 2020.

[8] V. L. Chuen, A. C. H. Chan, J. Ma, S. M. H. Alibhai, and V. Chau, “Assessing the Accuracy of International Classification of Diseases (ICD) Coding for Delirium,” J. Appl. Gerontol., vol. 41, no. 5, pp. 1485–1490, May 2022.

[9] P. Casey, W. Cross, M. W.-S. Mart, C. Baldwin, K. Riddell, and P. Dārziņš, “Hospital discharge data under-reports delirium occurrence: results from a point prevalence survey of delirium in a major Australian health service,” Intern. Med. J., vol. 49, no. 3, pp. 338–344, Mar. 2019.

[10] D. H. Kim et al., “Evaluation of algorithms to identify delirium in administrative claims and drug utilization database,” Pharmacoepidemiol. Drug Saf., vol. 26, no. 8, pp. 945–953, Aug. 2017.

[11] R. Katznelson et al., “Hospital administrative database underestimates delirium rate after cardiac surgery,” Can. J. Anaesth., vol. 57, no. 10, pp. 898–902, Oct. 2010.

[12] S. K. Inouye, L. Leo-Summers, Y. Zhang, S. T. Bogardus Jr, D. L. Leslie, and J. V. Agostini, “A chart-based method for identification of delirium: validation compared with interviewer ratings using the confusion assessment method,” J. Am. Geriatr. Soc., vol. 53, no. 2, pp. 312–318, Feb. 2005.

[13] S. R. Pagali, R. Kumar, S. Fu, S. Sohn, and M. Yousufuddin, “Natural Language Processing CAM Algorithm Improves Delirium Detection Compared With Conventional Methods,” Am. J. Med. Qual., vol. 38, no. 1, pp. 17–22, 2023.

[14] G. K. Savova et al., “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” J. Am. Med. Inform. Assoc., vol. 17, no. 5, pp. 507–513, Sep. 2010.

[15] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program,” Proc. AMIA Symp., pp. 17–21, 2001.

[16] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances,” J. Am. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, May-Jun 2010.

[17] M. Torii, Z. Hu, C. H. Wu, and H. Liu, “BioTagger-GM: a gene/protein name recognition system,” J. Am. Med. Inform. Assoc., vol. 16, no. 2, pp. 247–255, Mar-Apr 2009.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv [cs.CL], Oct. 11, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[19] L. Wang et al., “Boosting Delirium Identification Accuracy With Sentiment-Based Natural Language Processing: Mixed Methods Study,” JMIR Med Inform, vol. 10, no. 12, p. e38161, Dec. 2022.

[20] S. Fu et al., “Ascertainment of Delirium Status Using Natural Language Processing From Electronic Health Records,” J. Gerontol. A Biol. Sci. Med. Sci., vol. 77, no. 3, pp. 524–530, Mar. 2022.

[21] W. Ge et al., “Identifying Patients With Delirium Based on Unstructured Clinical Notes: Observational Study,” JMIR Form Res, vol. 6, no. 6, p. e33834, Jun. 2022.

[22] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv [cs.CL], Jul. 26, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[23] X. Yang et al., “A large language model for electronic health records,” NPJ Digit Med, vol. 5, no. 1, p. 194, Dec. 2022.

[24] X. Yang, J. Bian, W. R. Hogan, and Y. Wu, “Clinical concept extraction using transformers,” J. Am. Med. Inform. Assoc., vol. 27, no. 12, pp. 1935–1942, Dec. 2020.

[25] A. Vaswani et al., “Attention Is All You Need,” arXiv [cs.CL], Jun. 12, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[26] S. J. Moss, C. Hee Lee, C. J. Doig, L. Whalen-Browne, H. T. Stelfox, and K. M. Fiest, “Delirium diagnosis without a gold standard: Evaluating diagnostic accuracy of combined delirium assessment tools,” PLoS One, vol. 17, no. 4, p. e0267110, Apr. 2022.

[27] P. W. Lange, M. Lamanna, R. Watson, and A. B. Maier, “Undiagnosed delirium is frequent and difficult to predict: Results from a prevalence survey of a tertiary hospital,” J. Clin. Nurs., vol. 28, no. 13–14, pp. 2537–2542, Jul. 2019.