THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Medical Sciences
cite as: *J Gerontol A Biol Sci Med Sci*, 2022, Vol. 77, No. 3, 524–530
doi:10.1093/gerona/glaa275
Advance Access publication October 30, 2020

OXFORD

Research Article

Ascertainment of Delirium Status Using Natural Language Processing From Electronic Health Records

Sunyang Fu, MHI,[1,2] Guilherme S. Lopes, PhD,[1] Sandeep R. Pagali, MD, MPH,[3] Bjoerg Thorsteinsdottir, MD,[3] Nathan K. LeBrasseur, PhD, MS,[4,5] Andrew Wen, MS,[1] Hongfang Liu, PhD,[1] Walter A. Rocca, MD, MPH,[1] Janet E. Olson, PhD,[1] Jennifer St. Sauver, PhD,[1] and Sunghwan Sohn, PhD[1,*]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota. [2]University of Minnesota, Minneapolis. [3]Department of Medicine, Mayo Clinic, Rochester, Minnesota. [4]Department of Physical Medicine & Rehabilitation, Mayo Clinic, Rochester, Minnesota. [5]Department of Physiology & Biomedical Engineering, Mayo Clinic, Rochester, Minnesota.

*Address correspondence to: Sunghwan Sohn, PhD, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905. E-mail: Sohn.Sunghwan@mayo.edu

Received: May 13, 2020; Editorial Decision Date: October 20, 2020

Decision Editor: Anne B. Newman, MD, MPH, FGSA

Abstract

Background: Delirium is underdiagnosed in clinical practice and is not routinely coded for billing. Manual chart review can be used to identify the occurrence of delirium; however, it is labor-intensive and impractical for large-scale studies. Natural language processing (NLP) has the capability to process raw text in electronic health records (EHRs) and determine the meaning of the information. We developed and validated NLP algorithms to automatically identify the occurrence of delirium from EHRs.

Methods: This study used a randomly selected cohort from the population-based Mayo Clinic Biobank (N = 300, age ≥65). We adopted the standardized evidence-based framework confusion assessment method (CAM) to develop and evaluate NLP algorithms to identify the occurrence of delirium using clinical notes in EHRs. Two NLP algorithms were developed based on CAM criteria: one based on the original CAM (NLP-CAM; delirium vs no delirium) and another based on our modified CAM (NLP-mCAM; definite, possible, and no delirium). The sensitivity, specificity, and accuracy were used for concordance in delirium status between NLP algorithms and manual chart review as the gold standard. The prevalence of delirium cases was examined using International Classification of Diseases, 9th Revision (ICD-9), NLP-CAM, and NLP-mCAM.

Results: NLP-CAM demonstrated a sensitivity, specificity, and accuracy of 0.919, 1.000, and 0.967, respectively. NLP-mCAM demonstrated sensitivity, specificity, and accuracy of 0.827, 0.913, and 0.827, respectively. The prevalence analysis of delirium showed that the NLP-CAM algorithm identified 12 651 (9.4%) delirium patients, the NLP-mCAM algorithm identified 20 611 (15.3%) definite delirium cases, and 10 762 (8.0%) possible cases.

Conclusions: NLP algorithms based on the standardized evidence-based CAM framework demonstrated high performance in delineating delirium status in an expeditious and cost-effective manner.

Keywords: Confusion assessment method, Delirium, Electronic health records, Natural language processing

Delirium is a syndrome with symptoms that present as confusion and is characterized by an acute change in mental status, fluctuating course, lack of attention, and disorganized thinking or altered level of consciousness (1). Delirium is common in hospitalized older adults (2,3), with prevalence in the postoperative setting ranging between 21% and 35% (4), and in intensive care unit patients between 60% and 85% (5). Delirium has been associated with multiple predisposing and precipitating risk factors, including infections, use of specific medications, and the presence of a wide range of other chronic conditions (6,7). In particular, persons with dementia are at high risk of delirium, and delirium has been associated with subsequent cognitive impairment (8) and additional adverse outcomes (9),

including longer hospital stays, increased likelihood of nursing home placement, and an increased risk of death (3,10–13).

Delirium is underdiagnosed in clinical practice and is not routinely coded for billing (14). A study of patients undergoing elective surgery indicated that delirium was given an International Classification of Diseases, 9th Revision (ICD-9) code in only 3% of patient records (2). Causes of delirium are multi-factorial, but it has been estimated that 30%–40% of cases may be preventable (3). However, further research is necessary to identify the best ways to detect, prevent, and manage delirium. Such research is currently limited by challenges in identifying persons with delirium from electronic health records (EHRs). Inouye et al. (2) have developed methods for identifying persons with delirium from chart review of medical records. However, manual chart review is time-consuming and costly to extract information from clinical notes for large patient populations. Natural language processing (NLP) has been adopted to computationally extract clinical information from EHRs for a wide range of applications ranging from advancing EHR-based clinical research (15,16) to supporting clinical decision making (17,18). Different NLP frameworks have been developed to convert clinical narratives into structured data, including MedLEE (19), MetaMap (20), KnowledgeMap (21), MedTagger (22), and cTAKES (23). In this study, we adopted the MedTagger framework with domain-specific customizability to develop and validate an NLP algorithm to identify the occurrence of delirium using clinical notes derived from the Mayo Clinic EHR.

Materials and Methods

Study Population

This study was approved by the Mayo Clinic Institutional Review Board and the Olmsted Medical Center Institutional Review Board. The study population consisted of participants of the Mayo Clinic Biobank; details regarding this population have been previously published (24). Briefly, the Mayo Clinic Biobank is an institutional resource comprised of volunteers who have donated biological specimens, provided risk factor data, and have given permission to access clinical data from their EHRs for clinical research studies. Participants were contacted as part of a prescheduled medical examination at Mayo Clinic sites between April 2009 and September 2015. All participants were 18 years or older at the time of consent. Approximately 57 000 participants have been enrolled, and 24 224 of these participants were 65 years of age or older at the time of consent. Among these participants, we identified all persons who received an ICD-9 code for delirium or alteration of consciousness after date of enrollment (N = 731; ICD-9 codes: 290.3, 290.41, 291.0, 292.81, 293.0, 293.1, 348.30, and 780.09). We randomly sampled persons (N = 300) from this population for the annotation guideline development and gold standard identification phases of the study (details below). Among 300 randomly selected individuals aged 65 and older, 48.3% were female. A total of 615 visit records were noted for these 300 individuals of which 247 were inpatient records, 55 observation records, 186 emergency records, and 127 outpatient records. Among the 247 inpatient records, 89 visits were ICU records. Half of the persons from the sample population (n = 150) were used to develop NLP algorithms to identify occurrences of delirium, and the remaining sample (n = 150) was set aside to evaluate the performance of the NLP algorithm. Figure 1 showed a population flow process including all samples during the cohort screening and sampling.

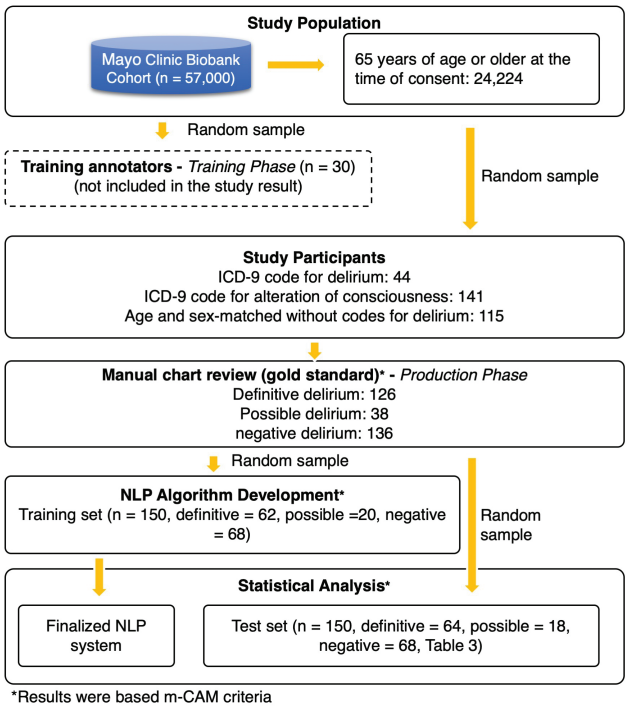

Figure 1. Workflow of cohort screening and sampling for the corpus annotation and NLP development. NLP = natural language processing. Full color version is available within the online issue.

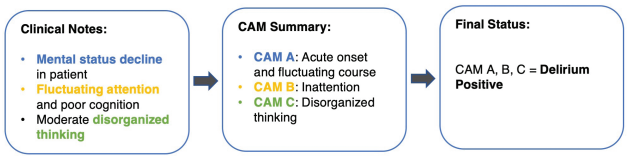

Figure 2. Comparison of original CAM and the modified CAM. CAM = confusion assessment method. Full color version is available within the online issue.

Annotation Guideline Development

Delirium is diagnosed based on a constellation of established clinical symptoms. Therefore, our primary criterion for identification of delirium was an explicit mention of a delirium episode (eg, "Patient experienced acute post-operative delirium this morning") documented in the clinical notes. If there was no clear diagnosis or mention of delirium, then we identified delirium based on whether symptoms documented in the clinical notes satisfied the "Confusion Assessment Method" (CAM) criteria (25).

Briefly, CAM has 4 features that are used to facilitate the diagnosis of delirium, including (A) acute onset and fluctuating course, (B) inattention, (C) disorganized thinking, and (D) altered level of consciousness. Each feature was further represented by a list of specific delirium-related concepts, such as deteriorating mental status, drowsiness, mumbling gibberish, impaired orientation, and encephalopathy (25). For example, the CAM feature "Disorganized thinking" was represented by several expressions, including "mumbling gibberish," "rambling speech," "unclear flow of ideas," etc. Figure 2 summarizes CAM features.

In our study, we developed 2 versions of criteria: the original CAM and the modified CAM (mCAM) (Table 1). For the original CAM, we operationalized the results as "definitive delirium" or "no delirium." "Definitive delirium" status was achieved when the

Table 1. Confusion Assessment Method (CAM)

A: acute onset and fluctuating	**B**: inattention
course Do the abnormal behaviors? • Come and go • Fluctuate during the day • Increase/decrease in severity	Does the patient: • Have difficulty focusing attention • Become easily distracted • Have difficulty keeping track of what is said
C: disorganized thinking Is the patient's thinking? • Disorganized • Incoherent	**D**: altered level of consciousness What is the patient's level of consciousness? • Alert (normal) • Vigilant (hyper-alert) • Lethargic (drowsy but easily roused) • Stuporous (difficult to rouse) • Comatose (unrousable)
Original CAM: Definitive: A and B and (C or D)	*Modified CAM*: Definitive: At least 3 unique CAM criteria Possible: Any 2 criteria and does not meet the definitive criteria as above

medical records described symptoms that match criteria A and B and either C or D of the CAM criteria within 1 month. The average duration of time from the first symptom to the last symptom was 13.8 days (range: 0–28 days). Most persons with delirium met the criteria within 48 hours (38%).

Two Mayo geriatricians and one palliative care physician (S.P., Z.X., B.T.) helped to define the review criteria and noted that symptoms of delirium may be poorly documented (missing information related to delirium features and concepts) (26). Therefore, we created mCAM to address this issue. Using the mCAM criteria, definitive delirium status was defined as when the medical records describe symptoms that matched 3 of 4 CAM criteria (eg, CAM B + C + D or A + C + D). Possible delirium status was defined as when symptoms matched exactly 2 CAM criteria. The comprehensive annotation guideline can be found in Supplementary Material 1.

Corpus Annotation

Corpus annotation is the process of manual chart review, marking interpretative linguistic (eg, syntax, negation) or predefined clinical information (eg, delirium-related concepts) to a corpus that can be used for NLP algorithm development and evaluation (27–29). There were 2 phases involved in our process: (a) training phase to be familiar with the annotation process and refine annotation guidelines and (b) production phase to create the gold standard for NLP algorithm development and evaluation.

Training phase

One geriatrician and one psychologist (S.P. and G.S.L.) annotated a random sample of records obtained from 15 patients who had an ICD-9 code for delirium or alteration of consciousness and another 15 patients who did not have code to identify documentation of delirium and/or keywords and terms related to the 4 CAM components. In the training phase, annotators reviewed the full medical records from 30 days prior to the date of the ICD-9 code through 30 days following receipt of the initial diagnosis code. They identified delirium-related terms previously described by Puelle et al. (30) for the identification of delirium from medical records as well as additional keywords associated with episodes of delirium observed in the sample records. Discrepancies between reviewers were discussed and resolved, and annotation criteria were updated.

Production phase

A new sample of 44 patients with an ICD-9 code for delirium and 141 patients with an ICD-9 code for the alteration of consciousness was matched by age (±1 year) and sex to 115 patients without a code for delirium (Figure 1). All 300 patient records within ± 30 days anchored by ICD-9 delirium diagnosis (total 8761 documents) were double-annotated by 2 reviewers (G.S.L. and D.I.) using the final annotation guidelines. Interannotator agreement (IAA) was calculated at the patient-level delirium status (eg, definitive delirium, no delirium) and for each concept (eg, confusion). All conflicting cases were adjudicated by a geriatrician and palliative care physician with expertise in geriatrics (S.P. and B.T.). The results of the final adjudicated annotation served as the gold standard for the development and test of the NLP algorithm.

NLP Algorithm Development

The NLP algorithm was developed to automate EHR chart review to identify patients with delirium based on 2 definitions of delirium (CAM and mCAM). To identify a patient's delirium status, the NLP algorithm screens the clinical notes to extract direct mention of delirium (physician's diagnosis) and delirium-related clinical concepts that match the CAM criteria. Each concept was then normalized into a standard form based on the CAM instruments. For example, "unresponsiveness" and "decreased responsiveness" were normalized into "disconnected." Furthermore, the normalized CAM instruments were mapped into the CAM Features A, B, C, and D. As an example in Figure 3, a patient who experienced acute altered mental status (CAM Feature A), inattention (CAM Feature B), and disorganized thinking (CAM Feature C) was considered "definitive" delirium.

To implement the algorithm, we adopted the open-source NLP pipeline MedTaggerIE (22), an open-source unstructured information management architecture–based information extraction framework. This system separates task-specific NLP knowledge engineering (ie, CAM criteria) from the generic routine NLP, which enables words and phrases containing clinical information (ie, keywords relevant to CAM features) to be directly coded by subject matter experts. The tool has been utilized in various clinical NLP tasks and adopted by multiple studies of phenotyping algorithm development (15,31,32). Additionally, we utilized the Mayo Clinic big data NLP platform (33), a distributed parallel computing environment to support data sets of extremely large volume which integrates NLP with existing EHR data stores. This enabled us to execute the NLP algorithm without manually retrieving large sets of documents prior to execution.

Figure 4 shows the NLP workflow. The generic NLP includes sentence segmentation, tokenization, temporal status detection (eg, present, history), and assertion detection (eg, negation, possible, hypothetical). The task-specific NLP includes the detection of keywords relevant to delirium in the text using regular expressions and normalized to specific delirium concepts. The summarization component applies heuristic rules (ie, CAM criteria) for assigning the delirium status.

The NLP algorithms were developed in the following 3 steps: (a) prototype algorithm development based on CAM, (b) formative algorithm development using the training data after the acceptable performance was reached (accuracy >0.95), and (c) final algorithm evaluation on the independent test set. The algorithm was applied and

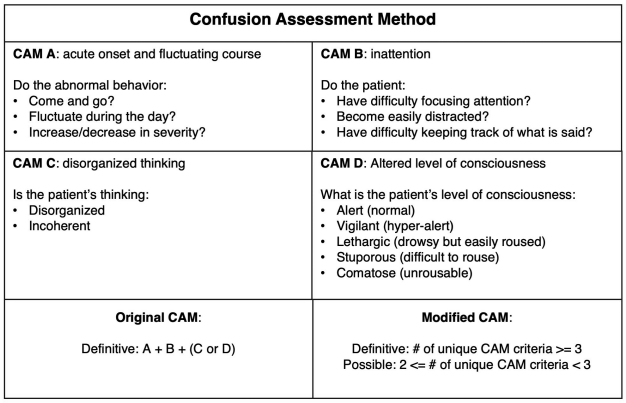

Figure 3. An example for detecting delirium status based on CAM. CAM = confusion assessment method.

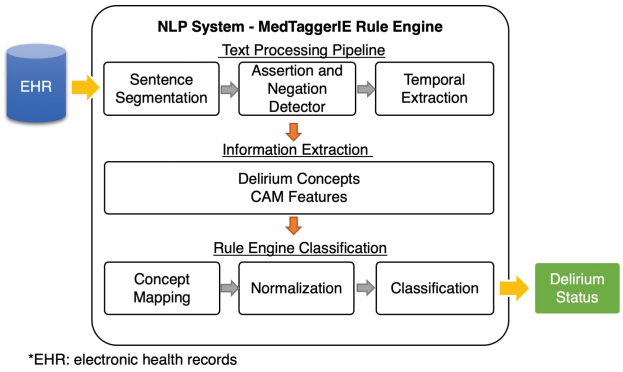

Figure 4. Architecture of the NLP. NLP = natural language processing. Full color version is available within the online issue.

refined on the training data. Incorrect cases were manually reviewed by 2 domain experts (G.S.L. and D.K.I.) and iteratively refined until all issues were resolved. The comprehensive algorithms can be found at https://github.com/OHNLP/AgingNLP/tree/master/delirium.

Statistical Analysis

The manual annotation of delirium status by 2 annotators was assessed by F1-score (34). F1-score (eqn (1)) is a well-established metric in the information retrieval and machine learning community. It measures both positive predictive value (precision) and sensitivity (recall) of the test object. The performance of the algorithm was assessed by using sensitivity (eqn (2)), specificity (eqn (3)), and accuracy (eqn (4)), to assess concordance between delirium status identified using the NLP algorithms and delirium status identified via manual chart review (gold standard).

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (1)$$

$$\text{Sensitivity} = \text{True positive}/(\text{True positive} + \text{False negative}) \quad (2)$$

$$\text{Specificity} = \text{True negative}/(\text{False positive} + \text{True negative}) \quad (3)$$

$$\text{Accuracy} = (\text{True positive} + \text{True negative})/(\text{Total positive} + \text{Total negative}) \quad (4)$$

Prevalence Analysis

To further compare the effectiveness between the NLP algorithm and ICD-9, we applied NLP-CAM and NLP-mCAM to screen all hospitalized patients who visited Mayo Clinic Rochester from April 2009 to September 2015. About 134 910 patients aged 65 years or older who were hospitalized at the Mayo Clinic in Rochester Minnesota were identified. For the purposes of comparison, we calculated the prevalence of positive delirium cases based on ICD-9, NLP-CAM, and NLP-mCAM.

Table 2. 2 × 2 Contingency Table of NLP-CAM

		Gold Standard		
		Delirium	No delirium	*Total*
NLP	Delirium	*57*	0	*57*
	No delirium	*5*	88	*93*
	Total	*62*	*88*	*150*
		Sensitivity = 0.919	Specificity = 1.000	Accuracy = 0.967

Note: NLP-CAM = natural language processing–confusion assessment method.

RESULTS

Interannotator Agreement

Among the 300 patients, 8761 clinical documents were double-reviewed, and 7515 delirium-related concepts were annotated. The IAA of patient-level delirium status (N = 300) between 2 annotators in F1-score was 0.94. Agreement between the 2 annotators at the concept level (ie, whether 2 annotators identify the same delirium-related terms, eg, Figure 3, box 1, mental status decline; N = 7515) in F1-score was 0.87. Overall, there were high agreements between the 2 annotators at both the delirium status and individual delirium concept levels.

Concordance in Delirium Status Between NLP Algorithms and Gold Standard

The NLP-CAM algorithm demonstrated a sensitivity, specificity, and accuracy of 0.919, 1.000, and 0.967, respectively, at identifying delirium compared to the gold standard (Table 2). The NLP-mCAM algorithm demonstrated a sensitivity, specificity, and accuracy of 0.827, 0.913, and 0.827, respectively, at identifying definite, possible, and no delirium compared to the gold standard (Table 3).

Prevalence Analysis

When the NLP-CAM algorithm was applied to the clinical notes of patients who were aged 65 or older and hospitalized at the Mayo Clinic in Rochester Minnesota between April 2009 and September 2015, 12 651 (9.4%) patients were identified as having delirium. The NLP-mCAM algorithm yielded 20 611 (15.3%) definite delirium cases and 10 762 (8.0%) possible cases. About 5490 (4.1%) of the sample population were identified as having delirium through ICD-9 screening. The results were consistent with the published literature, between 15% and 20% for patients with age greater or equal to 65 (4,35). These results reflect the ability of the NLP-based phenotyping algorithm to identify more likely delirium cases than conventional code-based screening methods.

Table 3. 3 × 3 Contingency Table of NLP-mCAM

		Gold Standard			
		Definitive delirium	Possible delirium	No delirium	*Total*
NLP	Definitive delirium	60	4	2	*66*
	Possible delirium	1	8	10	*19*
	No delirium	3	6	*56*	*65*
	Total	*64*	*18*	*68*	*150*
		Sensitivity = 0.938	Sensitivity = 0.444	Sensitivity = 0.824	Accuracy = 0.827
		Specificity = 0.930	Specificity = 0.917	Specificity = 0.890	Sensitivity* = 0.827
					Specificity* = 0.913

Note: NLP-mCAM = natural language processing–modified confusion assessment method.
*Micro average: calculated by using global counts of all categories.

Discussion

In this study, we developed and evaluated 2 highly accurate NLP algorithms (NLP-CAM and NLP-mCAM) in the task of automatically identifying patients with delirium from clinical notes, even without a definitive diagnosis of delirium, using descriptive terminology consistent with and meeting standard CAM criteria. The implementation of both algorithms demonstrated excellent performance in accurately classifying delirium. In addition, when applied to a large group of randomly selected patients, the NLP algorithms were able to identify more patients with delirium compared to structured data (ICD codes). These results suggest that these algorithms may have high sensitivity to capture occurrences of delirium by mining relevant keywords and concepts in clinical free text. Compared with conventional manual chart review, NLP provides more systematic and scalable solutions in identifying clinical concepts. In general, human annotators have a lower test–retest reliability score (eg, missing true positives due to human error) and manual chart review is impractical to do on a large number of documents. NLP can solve these issues. However, the performance of NLP algorithms is affected by the quality of EHR documentation (eg, presence, absence, and consistency of information required for delirium) because NLP algorithms are based on the records in EHR documents.

We noticed that the accuracy of the algorithm was lower when classifying the delirium based on the modified CAM criteria compared with the original CAM criteria. This may be due to the introduction of the "possible" category. Our modified CAM changed the task from a 2-class to a 3-class classification problem. This implementation causes a relatively lower performance measure than NLP-CAM even though NLP-mCAM is able to identify more number of total delirium cases (definite and possible). During the NLP evaluation, we performed an error analysis to identify the most common causes of errors. We found that extracting "disorganized thinking" related concepts from clinical notes was a major challenge. The same concept can be expressed in many different ways. For example, the same concept can be expressed through different behaviors, such as speech, cognitive process, decision making, and patient reactions. Therefore, the NLP algorithms needed to be both accurate and generalizable in capturing these concepts. To address this problem, we used relaxed word distance, that is, allowing number of words between 2 anchor words, to help fine-tune the rules in multiple iterations (36). This process is, however, time-consuming. In the future, we will explore advanced machine learning techniques to capture the common meaning of the various expressions to aid the development effort.

Implications for Research

The process of manually abstracting clinical concepts for delirium ascertainment is time-consuming, costly, and non-scalable. NLP algorithms are distinctive in their ability to extract critical information from free text in EHRs. NLP techniques offer a sophisticated way of handling free text with high levels of accuracy, allowing efficient mining of unstructured data for broad applications. The NLP-CAM algorithm was developed strictly based on the original definition of CAM with the objective of achieving a high degree of precision. Researchers may find NLP-CAM to be helpful for identifying delirium with high confidence. However, we recognize that delirium symptoms are likely to be poorly documented in the clinical notes (26). Thus, strictly applying the original definition may not be sensitive enough to capture highly likely or possible cases. We therefore modified the criteria definition to adapt to the real-world EHR. Instead of strictly following the CAM criteria, we developed a modified definition for definite cases and added an additional "possible" category. The modified NLP-mCAM algorithm also had a good performance in identifying definite delirium cases and also identified a significant number of possible cases. Depending on the research study, investigators may wish to include only definite cases or may want to include possible cases. The NLP-CAM and NLP-mCAM algorithms therefore offer investigators the flexibility to apply either algorithm depending on the needs of the study.

Implications for Clinical Practice

The NLP algorithms also have many potential clinical applications especially when it comes to proactively identify patients at high risk of delirium based on prior history, or flagging hospitalized patients, who per clinical documentation are showing signs of delirium in real time. Because delirium is underreported and not all patients have a formal assessment for delirium diagnosis, the use of NLP algorithms on routine EHRs can facilitate the early detection of delirium. This can be achieved by integrating the NLP algorithms into clinical workflow through application programming interface technologies, which allows outputs from NLP to be delivered to clinicians by mobile applications or EHR system (eg, EPIC). Such clinical decision tools could facilitate the implementation of preventive measures to reduce the incidence of delirium (through risk factors) and institute early intervention strategies to avoid escalating symptoms and associated complications of delirium.

Limitations

Our study has several limitations. The algorithms initially were developed after a review of clinical notes at a single institution, with the structure tailored to a specific EHR system. Although we have successfully demonstrated the external validity of other NLP algorithms on EHRs in another hospital setting (37,38), additional work is necessary to demonstrate the portability of these algorithms to other institutions and EHRs. The NLP algorithms do not intend to replace a formal delirium assessment. Instead, the NLP algorithms can be used to automate manual chart review based on CAM criteria. The system was developed based on manual chart review from EHRs, the success of NLP algorithms depends on the level of detail and accuracy of the medical records. If there is no documentation of the features and concepts about delirium (hypoactive and hyperactive), NLP cannot identify the cases. We also note that hyperactive delirium is more likely to be documented in the medical records than hypoactive delirium. Therefore, we expect to consistently underestimate the presence of hypoactive delirium.

Conclusions

We adopted the standardized evidence-based framework CAM to develop and evaluate NLP algorithms to identify the occurrence of delirium from EHRs. Our NLP algorithms demonstrated excellent performance in identifying patients with delirium using clinical notes in an expeditious and cost-effective manner. These algorithms represent a promising alternative to manual chart review used in EHR-based delirium research projects and artificial intelligence-based clinical decision support.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* online.

Funding

This work was supported by National Institute on Aging R21 AG58738, National Institute on Aging R01 AG34676, andNational Institute on Aging R01 AG068007.

Acknowledgments

The authors would like to gratefully acknowledge Donna M. Ihrke and Xin Zhang for case validation.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Author Contributions

S.F., J.S., and S.S. conceptualized and design the study. S.F. and S.S. wrote a draft of the manuscript and performed the analysis. G.S.L., B.T., and D.I. conducted manual annotation and B.T. and S.P. adjudicated discrepancies. S.F. and S.S. developed NLP algorithms. A.W. implemented algorithms to the Mayo big data environment. All authors participated in the interpretation of the data and contributed to manuscript editing and revisions.

References

1. Neufeld KJ, Thomas C. Delirium: definition, epidemiology, and diagnosis. *J Clin Neurophysiol*. 2013;30:438–442. doi:10.1097/WNP.0b013e3182a73e31
2. Inouye SK, Leo-Summers L, Zhang Y, Bogardus ST Jr, Leslie DL, Agostini JV. A chart-based method for identification of delirium: validation compared with interviewer ratings using the confusion assessment method. 2005;53:312–318. doi:10.1111/j.1532-5415.2005.53120.x
3. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet*. 2014;383:911–922. doi:10.1016/S0140-6736(13)60688-1
4. Ryan DJ, O'Regan NA, Caoimh RO, et al. Delirium in an adult acute hospital population: predictors, prevalence and detection. *BMJ Open*. 2013;3:e001772. doi:10.1136/bmjopen-2012-001772
5. Pun BT, Ely EW. The importance of diagnosing and managing ICU delirium. *Chest*. 2007;132:624–636. doi:10.1378/chest.06-1795
6. Wu X, Sun W, Tan M. Incidence and risk factors for postoperative delirium in patients undergoing spine surgery: a systematic review and meta-analysis. *Biomed Res Int*. 2019;2019:2139834. doi:10.1155/2019/2139834
7. Yang Y, Zhao X, Dong T, Yang Z, Zhang Q, Zhang Y. Risk factors for postoperative delirium following hip fracture repair in elderly patients: a systematic review and meta-analysis. *Aging Clin Exp Res*. 2017;29:115–126. doi:10.1007/s40520-016-0541-6
8. Parrish E. Delirium superimposed on dementia: challenges and opportunities. *Nurs Clin North Am*. 2019;54:541–550. doi:10.1016/j.cnur.2019.07.004
9. Diwell RA, Davis DH, Vickerstaff V, Sampson EL. Key components of the delirium syndrome and mortality: greater impact of acute change and disorganised thinking in a prospective cohort study. *BMC Geriatr*. 2018;18:24. doi:10.1186/s12877-018-0719-1
10. Raats JW, van Eijsden WA, Crolla RM, Steyerberg EW, van der Laan L. Risk factors and outcomes for postoperative delirium after major surgery in elderly patients. *PLoS One*. 2015;10:e0136071. doi:10.1371/journal.pone.0136071
11. Kiely DK, Marcantonio ER, Inouye SK, et al. Persistent delirium predicts greater mortality. *J Am Geriatr Soc*. 2009;57:55–61. doi:10.1111/j.1532-5415.2008.02092.x
12. Salluh JI, Wang H, Schneider EB, et al. Outcome of delirium in critically ill patients: systematic review and meta-analysis. *BMJ*. 2015;350:h2538. doi:10.1136/bmj.h2538
13. Witlox J, Eurelings LS, de Jonghe JF, Kalisvaart KJ, Eikelenboom P, van Gool WA. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA*. 2010;304:443–451. doi:10.1001/jama.2010.1013
14. Ritter SRF, Cardoso AF, Lins MMP, Zoccoli TLV, Freitas MPD, Camargos EF. Underdiagnosis of delirium in the elderly in acute care hospital settings: lessons not learned. *Psychogeriatrics*. 2018;18:268–275. doi:10.1111/psyg.12324
15. Fu S, Leung LY, Wang Y, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Med Inform*. 2019;7:e12109. doi:10.2196/12109
16. Liu F, Weng C, Yu H. Natural Language Processing, Electronic Health Records, and Clinical Research. In: Richesson R, Andrews J. eds., *Clinical Research Informatics. Health Informatics*. Springer; 2012. doi:10.1007/978-1-84882-448-5_16
17. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform*. 2009;42:760–772. doi:10.1016/j.jbi.2009.08.007
18. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assoc*. 2011;18(suppl 1):i150–i156. doi:10.1136/amiajnl-2011-000431
19. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. *Methods Inf Med*. 1998;37:334–344.
20. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17:229–236. doi:10.1136/jamia.2009.002733

21. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A III, eds. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2003.
22. Liu H, Bielinski SJ, Sohn S, et al. An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:149–153.
23. 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;17:507–513. doi:10.1136/jamia.2009.001560
24. Olson JE, Ryu E, Johnson KJ, et al., eds. The Mayo Clinic Biobank: a building block for individualized medicine. *Mayo Clinic Proceedings*. Elsevier; 2013. doi:10.1016/j.mayocp.2013.06.006
25. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med*. 1990;113:941–948. doi:10.7326/0003-4819-113-12-941
26. Hope C, Estrada N, Weir C, Teng CC, Damal K, Sauer BC. Documentation of delirium in the VA electronic health record. *BMC Res Notes*. 2014;7:208. doi:10.1186/1756-0500-7-208
27. Leech G. Corpus annotation schemes. *Lit Linguist Comput*. 1993;8:275–281.
28. Mollá D, Santiago-Martínez ME. Creation of a corpus for evidence based medicine summarisation. *Australas Med J*. 2012;5:503–506. doi:10.4066/AMJ.2012.1375
29. Fu SY, Leung LY, Raulli AO, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med Informatics Decis Mak*. 2020;20:1–12. doi:10.1186/s12911-020-1072-9
30. Puelle MR, Kosar CM, Xu G, et al. The language of delirium: keywords for identifying delirium from medical records. *J Gerontol Nurs*. 2015;41:34–42. doi:10.3928/00989134-20150723-01
31. Wyles CC, Tibbo ME, Fu SY, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *J Bone Joint Surg Am*. 2019;101:1931–1938. doi:10.2106/Jbjs.19.00071
32. McCarty CA, Chisholm RL, Chute CG, et al.; eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4:13. doi:10.1186/1755-8794-4-13
33. Wen A, Fu S, Moon S, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digit Med*. 2019;2:130. doi:10.1038/s41746-019-0208-8
34. Van Rijsbergen CJ. *The Geometry of Information Retrieval*. Cambridge University Press; 2004.
35. Kukreja D, Günther U, Popp J. Delirium in the elderly: current problems with increasing geriatric age. *Indian J Med Res*. 2015;142:655–662. doi:10.4103/0971-5916.174546
36. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. *J Biomed Inform*. 2020;109:103526. doi:10.1016/j.jbi.2020.103526
37. Sohn S, Wang Y, Wi CI, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *J Am Med Inform Assoc*. 2018;25:353–359. doi:10.1093/jamia/ocx138
38. Wi CI, Sohn S, Ali M, et al. Natural language processing for asthma ascertainment in different practice settings. *J Allergy Clin Immunol Pract*. 2018;6:126–131. doi:10.1016/j.jaip.2017.04.041