Original Paper

Development and Evaluation of a Natural Language Processing Annotation Tool to Facilitate Phenotyping of Cognitive Status in Electronic Health Records: Diagnostic Study

Ayush Noori[1*]; Colin Magdamo[1*], BSc; Xiao Liu[1], MSc; Tanish Tyagi[1]; Zhaozhi Li[1], MSc; Akhil Kondepudi[1]; Haitham Alabsi[1,2], MD; Emily Rudmann[1,3], BSc; Douglas Wilcox[1,2], MD, PhD; Laura Brenner[2,4], MD; Gregory K Robbins[2,5], MD; Lidia Moura[1,2], MD, PhD; Sahar Zafar[1,2], MD; Nicole M Benson[2,6,7], MD, MBI; John Hsu[2,6], MD, PhD; John R Dickson[1,2], MD, PhD; Alberto Serrano-Pozo[1,2], MD, PhD; Bradley T Hyman[1,2], MD, PhD; Deborah Blacker[2,8], MD, PhD; M Brandon Westover[1,2], MD, PhD; Shibani S Mukerji[1,2,5], MD, PhD; Sudeshna Das[1,2], PhD

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States

[2]Harvard Medical School, Boston, MA, United States

[3]Vaccine and Immunotherapy Center, Division of Infectious Disease, Boston, MA, United States

[4]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, United States

[5]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, United States

[6]Mongan Institute, Massachusetts General Hospital, Boston, MA, United States

[7]McLean Hospital, Belmont, MA, United States

[8]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Sudeshna Das, PhD
Department of Neurology
Massachusetts General Hospital
55 Fruit Street
Boston, MA, 02114
United States
Phone: 1 617 726 2000
Email: SDAS5@mgh.harvard.edu

Abstract

Background: Electronic health records (EHRs) with large sample sizes and rich information offer great potential for dementia research, but current methods of phenotyping cognitive status are not scalable.

Objective: The aim of this study was to evaluate whether natural language processing (NLP)–powered semiautomated annotation can improve the speed and interrater reliability of chart reviews for phenotyping cognitive status.

Methods: In this diagnostic study, we developed and evaluated a semiautomated NLP-powered annotation tool (NAT) to facilitate phenotyping of cognitive status. Clinical experts adjudicated the cognitive status of 627 patients at Mass General Brigham (MGB) health care, using NAT or traditional chart reviews. Patient charts contained EHR data from two data sets: (1) records from January 1, 2017, to December 31, 2018, for 100 Medicare beneficiaries from the MGB Accountable Care Organization and (2) records from 2 years prior to COVID-19 diagnosis to the date of COVID-19 diagnosis for 527 MGB patients. All EHR data from the relevant period were extracted; diagnosis codes, medications, and laboratory test values were processed and summarized; clinical notes were processed through an NLP pipeline; and a web tool was developed to present an integrated view of all data. Cognitive status was rated as cognitively normal, cognitively impaired, or undetermined. Assessment time and interrater agreement of NAT compared to manual chart reviews for cognitive status phenotyping was evaluated.

Results: NAT adjudication provided higher interrater agreement (Cohen κ=0.89 vs κ=0.80) and significant speed up (time difference mean 1.4, SD 1.3 minutes; P<.001; ratio median 2.2, min-max 0.4-20) over manual chart reviews. There was moderate agreement with manual chart reviews (Cohen κ=0.67). In the cases that exhibited disagreement with manual chart reviews, NAT

adjudication was able to produce assessments that had broader clinical consensus due to its integrated view of highlighted relevant information and semiautomated NLP features.

Conclusions: NAT adjudication improves the speed and interrater reliability for phenotyping cognitive status compared to manual chart reviews. This study underscores the potential of an NLP-based clinically adjudicated method to build large-scale dementia research cohorts from EHRs.

(J Med Internet Res 2022;24(8):e40384) doi: 10.2196/40384

KEYWORDS

chart review; cognition; cognitive status; dementia; diagnostic; electronic health record; health care; natural language processing; research cohort

Introduction

In recent years, electronic health records (EHRs) have become increasingly common in US health care facilities; they provide a wealth of information on patient demographics, medical history, clinical data, and health system interactions. EHRs offer an unprecedented opportunity to improve clinical care and examine a broad variety of scientific, health care utilization, and heath policy questions [1-3]. An important first step in conducting EHR research is accurately identifying patients with a certain health condition, event, or disease, which is known as phenotyping [1,4]. The identified patient sample is subsequently leveraged for a wide range of purposes, such as providing clinical decision support for health care delivery [5], conducting epidemiological research [4,6], and for the practice of precision medicine [7].

Phenotyping cognitive status (ie, distinguishing between normal cognition and any stage of cognitive impairment) in EHR is a major challenge since dementia is underrecognized, underdiagnosed, and underreported in claims data [8-12], leading to inaccurate identification of dementia cases in many studies based on claims or EHR data [13-15]. Informative missingness, errors, and biases in EHR may further exacerbate the challenges of defining dementia outcomes [16]. Yet another challenge of phenotyping arises from complex, subjective, loosely-defined diagnostic criteria as well as the format—that is, structured (eg, diagnosis codes and medications) versus unstructured (eg, clinical notes and images)—in which the information is stored [4]. Previous studies have demonstrated that information on cognitive status is often found only in free text [17-19]. Clinicians may chart symptoms of cognitive problems in clinical notes but may not make a formal diagnosis, refer to a specialist, or prescribe medication for multiple reasons including clinical role, lack of time or expertise, patient resistance, or limited treatment options [20-22]. Thus, accurately phenotyping cognitive status requires the combined use of both structured data, such as diagnosis codes, medications, and laboratory test results, as well as unstructured clinical notes.

Several algorithms have been developed for phenotyping cognitive status; some studies used structured data, such as diagnosis codes, missed appointments, or health care utilization patterns [15,23], whereas others have applied natural language processing (NLP) to unstructured notes [18,19,24]. None of these prior efforts combined both structured and unstructured input modalities, and manual annotation by clinical experts is limited by the lack of available tools to facilitate efficient chart review [25]. Thus, we hypothesized that the best approach for phenotyping cognitive status is a semiautomated one in which automated NLP is applied to clinical notes and presented in an integrated view to the clinical expert for final manual adjudication of cognitive status.

We developed NAT, a semiautomated NLP-powered annotation tool, to facilitate adjudication of cognitive status. The tool extracts and processes data from EHRs and then ranks clinical notes based on a deep learning NLP algorithm (Macro F_1=0.92) that classifies whether a note indicates normal cognition, cognitive impairment, or has no pertinent information [26]. It highlights key information and presents a summarized view to the annotator. We evaluated NAT in two EHR data sets: (1) Medicare beneficiaries from the Mass General Brigham (MGB) Accountable Care Organization (ACO) who were labeled in another study using manual chart reviews [15] and (2) MGB patients with laboratory confirmed SARS-CoV-2 (a case-control study to investigate the effects of COVID-19 on people with and without HIV was used as an exemplar of a research cohort that requires labeling of cognitive status). We evaluated interrater agreement in the first data set and compared it to interrater agreement in Epic—the EHR system used at MGB since 2015. The second data set was used to compare timings of manual to NAT adjudication, as the timing of manual adjudication was not available in the first data set.

By addressing the gaps in current chart review methods and leveraging existing NLP methods, we demonstrate that NAT increases both the efficiency and the interrater reliability of phenotyping cognitive status in EHR (relative to manual chart reviews) to build future research cohorts.

Methods

Clinical Settings and Data Sources

This diagnostic study was conducted at MGB—formerly Partners Healthcare—a private nonprofit organization comprising two major academic hospitals, community hospitals, and community health centers in the Boston metropolitan area. Data were sourced from the MGB Enterprise Data Warehouse that stores data from Epic. We evaluated NAT adjudication for phenotyping cognitive status on two distinct data sets. The first one included EHR data from January 1, 2017, to December 31, 2018, of 100 patients randomly selected from a larger data set that was expert-annotated via manual Epic chart reviews and reported elsewhere [15]. Specifically, this manually expert-annotated data set contained 1002 Medicare beneficiaries

from the MGB ACO who were classified into (1) normal cognition, (2) borderline of normal cognition and mild cognitive impairment (MCI), (3) MCI, (4) borderline of MCI and dementia, or (5) dementia [15]. The experts graded their confidence in the adjudication as low, medium, moderate, or high. The 100 patients were randomly sampled from these 5 classes with 20 from each class, ensuring that each class had a similar distribution of confidence scores. The second data set included 527 MGB patients with a laboratory confirmed SARS-CoV-2 infection based on polymerase chain reaction testing between March 1 and December 31, 2020. The data set was created for a case-control study to investigate the effects of COVID-19 on people with and without HIV; EHR data up to 2 years prior to and any time after the index positive polymerase chain reaction test were used to investigate the performance of NAT adjudication.

Ethics Approval

This study was approved by the MGB Institutional Review Board (2015P001915).

Definition of Cognitive Impairment

In this study, to phenotype cognitive status, patients were annotated with three labels: (1) cognitively normal (CN), (2) cognitively impaired (CI), and (3) undetermined. Patients were labeled as CI if there was any documented suspicion or concern of memory or cognitive decline, whether based on symptoms, observations, or objective testing. This ranged from any dementia-related International Classification of Diseases (ICD) codes or medicines in the patients' charts to cognitive concerns—relayed by patients, family members or friends, or providers in the notes and phone logs—as these concerns often reflect an underlying change in cognition even if a cognitive evaluation is normal (in which case they prompt a diagnosis of subjective cognitive decline [27]). Conversely, to be annotated as CN, at least implicit evidence of no cognitive concerns was required (eg, the patient continued to work, clearly managed their own care or hobbies, and followed complicated instructions, or they had annual wellness or specialist notes with multisystem assessment and no mention of a cognitive concern). The strongest evidence for a CN annotation was a cognition test performed with an explicit note of intact cognition. If there was conflicting evidence of both cognitive impairment and evidence of no cognitive impairment in a patient's chart, the latest evidence or specialist notes (if any were available) informed the adjudication. Finally, patients were marked as "undetermined" if the EHR did not have sufficient information.

Data Preparation

Data query, preparation, and preprocessing steps are described in Multimedia Appendix 1. For each patient, the following EHR data from the relevant time period were extracted from the Enterprise Data Warehouse: (1) patient demographic information, including name, medical record number, birth date, sex, ethnic group, marital status, and educational level; (2) all clinical notes, including reason for visit, history, note text, encounter type, and MGB provider (including provider department, specialty, and qualifications); (3) current primary care provider; (4) patient care coordination note; (5) medication history and current medications; (6) magnetic resonance imaging and computerized tomography orders; (7) laboratory orders and results; (8) problem list, including ICD diagnoses and diagnosis codes; and (9) visit cancellations.

Several features were engineered from the EHR to facilitate assessment of cognitive status. Dementia-related medications and ICD codes (medications: galantamine, donepezil, rivastigmine, and memantine; ICD-9 codes: 290.X, 294.X, 331.X, and 780.93; ICD-10 codes: G30.X and G31.X) and laboratory tests (eg, vitamin B12, folate, and thyroid-stimulating hormone) related to assessment of cognitive status were identified and highlighted. The numbers of cancellations, no-shows, and refill requests, relative to the total number of encounters, were computed.

Finally, NLP was applied to the clinical notes. We curated two lists of regular expressions or keywords related to the presence or absence of both (1) cognitive impairment and (2) the functional impairment of activities of daily living (ADLs) or independent ADLs, respectively (Multimedia Appendices 2 and 3). We identified regular expression matches and highlighted these within the text of the notes with different colors for each category (eg, cognition vs ADLs) to facilitate their identification by the clinician. We applied a previously developed NLP model [26] to generate classification probabilities of the following classes for each note: CI, no CI, or neither. The notes were ranked based on these classification probabilities, and notes that the model predicted as indicative of CI were displayed at the top.

Development of an Annotation Tool

We designed and developed a web-based chart review and annotation tool, using the Python-based open-source Django web development framework with a SQLite database. We established data models for patient-level demographic and clinical data, encounter-level clinician notes, user account creation and authentication, and patient assignment to individual or multiple annotators (Multimedia Appendix 4). We created several user interfaces (ie, pages) to present the various data modalities in an integrated fashion for annotation.

Statistical Analysis

We evaluated NAT adjudication using three metrics: agreement with manual Epic chart reviews, assessment time, and interrater agreement. We evaluated agreement between manual Epic chart reviews and NAT adjudication as well as interrater agreement for NAT adjudication using Cohen κ, whereas assessment time in minutes was compared using a paired samples Wilcoxon test (also known as the Wilcoxon signed-rank test). There were no missing data for these variables. All analyses were conducted using the R statistical software (version 4.1.2; R Core Team).

Results

Patient Characteristics

The patient characteristics of the two data sets are shown in Table 1. The ACO data set comprised 100 patients (63/100, 63.0% were women; mean age 78.8, SD 7.4 years; 7/100, 7% racial or ethnic minorities, 1 missing; 51/100, 51.0% with a

college degree or more, 3 missing; and 50/100, 50.0% were married). The COVID-19 data set comprised 527 patients (226/527, 42.9% women; mean age 52.6, SD 15.0 years; 318/527, 60.35% racial or ethnic minorities, 21 missing; 160/527, 30.4% college education or more, 62 missing; and 195/527, 37.0% married, 16 missing).

Table 1. Characteristics of Accountable Care Organization (ACO) and COVID-19 data sets used for NLP[a] annotation tool (NAT) evaluation.

Characteristics	Patients (N=627)	
	ACO data set (n=100)	COVID-19 data set (n=527)
Sex, n (%)		
Male	37 (37)	301 (57.1)
Female	63 (63)	226 (42.9)
Age (years), mean (SD)	78.8 (7.4)	52.6 (15)
Minorities, n (%)		
Black	4 (4)	163 (30.9)
Hispanic	2 (2)	138 (26.2)
Asian	1 (1)	16 (3)
Indigenous	0 (0)	1 (0.2)
College education, n (%)	51 (51)	160 (30.4)
Married, n (%)	50 (50.0)	195 (37)
Clinical characteristics		
Number of encounters, median (min-max)	164 (8-858)	106 (1-2474)
PCP[b] visit, n (%)	71 (71)	423 (80.3)
Dementia ICD[c] code and medication, n (%)	51 (51)	166 (5.3)

[a]NLP: natural language processing.

[b]PCP: primary care provider.

[c]ICD: International Classification of Diseases.

Features of NAT

Upon logging in to our annotation tool, an authenticated user is presented with a dashboard listing the patient IDs, ages, and sexes of their assigned patients (Figure 1). In addition, the total number of notes, the sequences within the notes that match a cognition or ADL keyword (Multimedia Appendices 1 and 2), and the number of notes for each predicted class (ie, cognition and ADL) are also presented. After annotation, the patient's label (CN, CI, or undetermined) is displayed with background colors reflecting the patient's annotated cognitive status.

Selecting a patient navigates the user to an annotation view summarizing the patient's demographic and clinical information (Figure 2A). Engineered features, including the total number of notes, encounters, no shows, cancellations, and refill requests, along with the patient care coordination note (if any), diagnosis ICD codes, and medications, are displayed (Figure 2B). Brain imaging and relevant laboratory tests, such as thyroid-stimulating hormone or vitamin B12, allow annotators to consider systemic causes of cognitive changes (Figure 2C). Finally, notes sorted by the predicted probability and with highlighted keywords are presented to expedite the review of the entire chart history during the relevant period for the clinical adjudication of cognitive status. Examples of the three predicted classes of notes (CN, CI, and undetermined) are shown in Figure 2D.

Figure 1. NAT dashboard: screenshot of the NAT dashboard displaying the current workload and assigned patients. A summary of patient information is displayed in each row, and the background reflects the cognitive status assigned to the patient. NAT: NLP annotation tool; NLP: natural language processing.

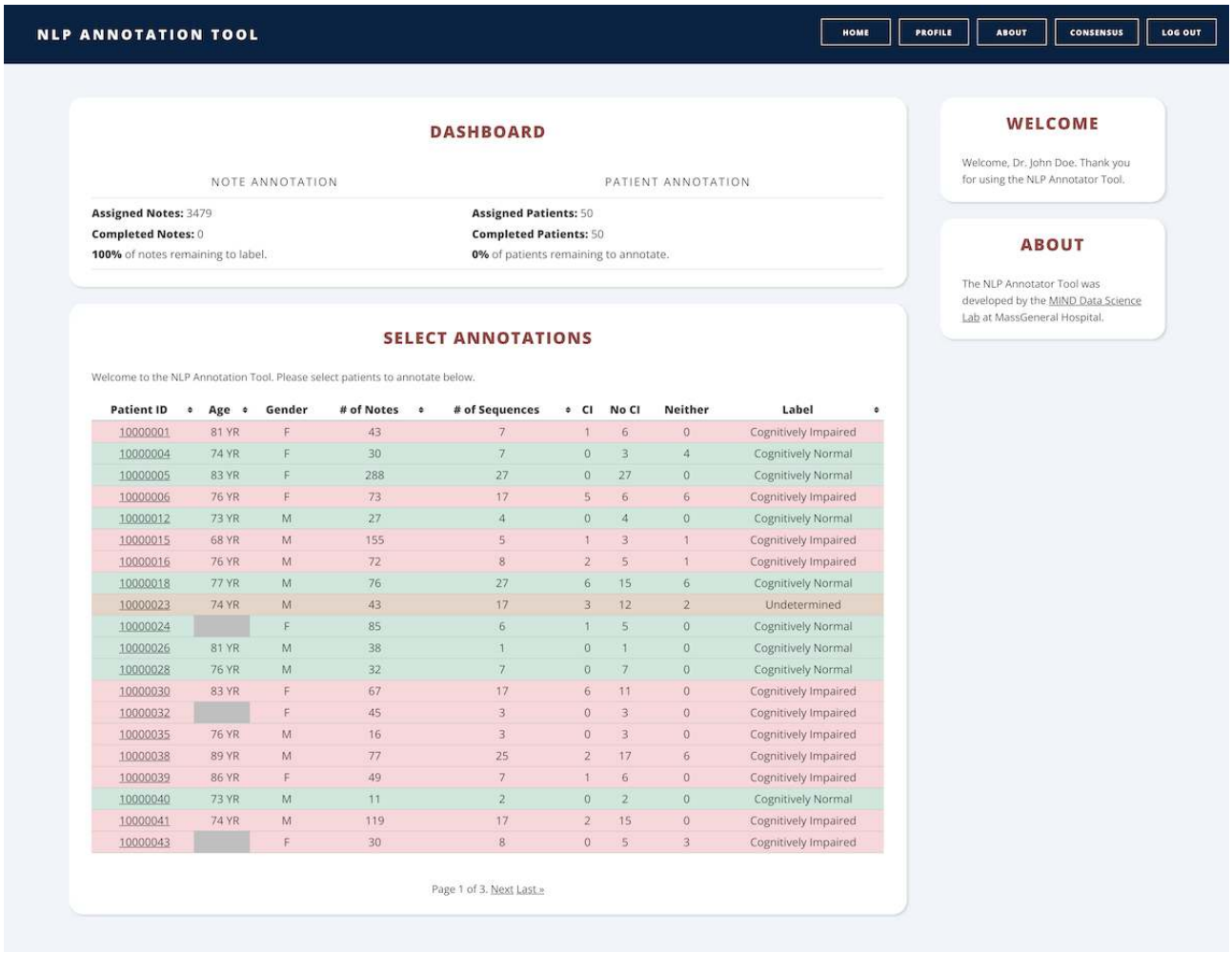
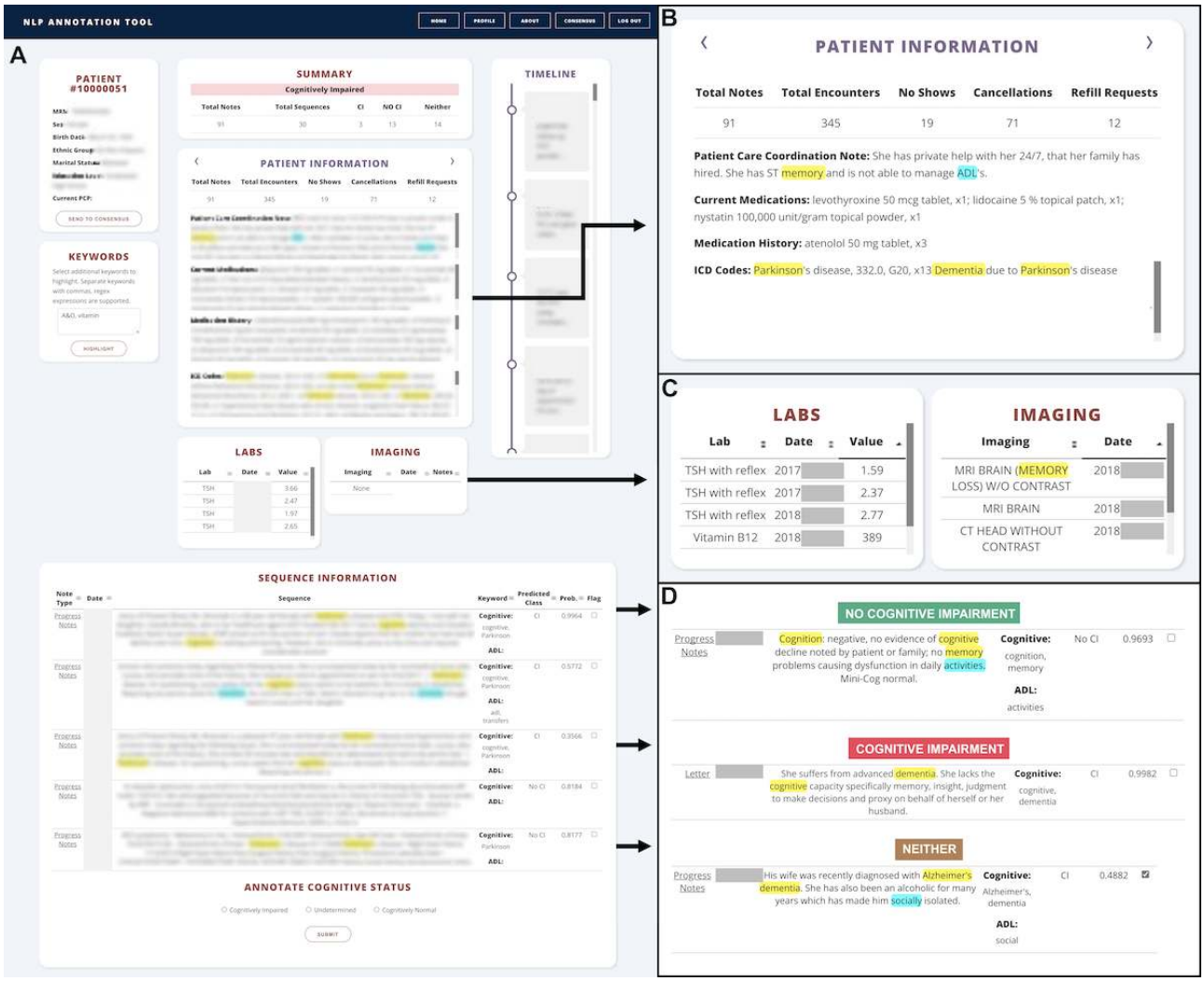


Figure 2. Annotation view: (A) patient view displaying summary information at the top and sequences from clinical notes at the bottom; (B) the Patient Information box summarizes health care interaction, patient care coordination notes, current medications, and diagnosis codes; (C) laboratory tests and imaging conducted on the patient; (D) sample sequences from notes with dementia and activities of daily living (ADLs) keywords highlighted. Each sequence is classified as cognitive impairment (CI), no CI, or neither, with a probability, and allows annotators to flag incorrect classifications.



Evaluation of NAT

Two teams of expert clinicians were randomly assigned patients and adjudicated the ACO data set, using NAT (team 1: LB, GKR, SSM; team 2: MBW and HA). We compared the phenotyping of cognitive status using NAT to manual chart reviews using Epic (labels were obtained from Moura et al [15]; patients who were not CN were grouped into the CI class). We removed patients annotated as "undetermined" in the set adjudicated using NAT, as they had little information in EHR to assess cognitive status and could not be directly compared to the labels obtained from Moura et al [15]. The agreement between NAT and manual Epic chart reviews was moderate for both team 1 (Cohen κ=0.68) and team 2 (Cohen κ=0.65) with a mean Cohen κ=0.67; the breakdown is shown in Figure 3A. Surprisingly, patients whose NAT label disagreed with the manual Epic chart reviews were annotated as CI using Epic and as CN using NAT. We manually reviewed the patients where the diagnostic labels disagreed; we found that NAT was able to highlight certain passages of text, such as "language, attention, and memory function are intact with good fund of knowledge"; the highlighted text facilitated the labeling of the patient as CN, whereas such phrases were easily missed in manual chart reviews. Moreover, if a patient had a transient cognitive deficit and was later evaluated as CN, for example, NAT presented all notes with highlighted evidence along with their dates in one view, making it easier to follow the sequence of events. The disagreements were mostly among patients annotated with a low confidence score in the Epic manually annotated data set [15] (Figure 3B). The interrater agreement of NAT adjudication between team 1 and team 2 was higher (Cohen κ=0.89) than the interrater agreement (Cohen κ=0.80) with manual Epic chart reviews reported in Moura et al [15].

Next, we compared the time required for phenotyping of cognitive status via NAT adjudication versus manual chart reviews in Epic. Four of the authors (DW, ER, HA, and SSM) adjudicated the full COVID-19 data set using NAT and recorded the annotation time for 129 patients. Two of the authors (HA and SSM) timed manual chart reviews in Epic for 32 randomly sampled patients. To ensure that a patient was not adjudicated using both methods by the same person, HA used Epic to perform chart reviews of patients adjudicated by SSM using NAT and vice versa. For most of the patients, the annotation time was substantially shorter with NAT as compared to manual

chart reviews in Epic (Figure 3C). Adjudications using NAT provided substantial speed-up of annotations compared to manual chart reviews in Epic (time difference mean 1.4, SD 1.3 minutes; $P<.001$; ratio median 2.2, min-max 0.4-20). Additionally, we observed that clinicians spent more time using NAT on the first half of patients compared to the second half. This "learning effect" was not observed with manual Epic chart reviews. The breakdown of the cognitive status for the COVID-19 data set is shown in Figure 4. Notably, the cognitive status for 21.1% (n=111) of patients was undetermined, suggesting that there was little information in EHR to determine their cognitive status.

Figure 3. Comparison of adjudication with natural language processing (NLP)–powered annotation tool (NAT) and manual Epic chart reviews: (A) contingency table displaying adjudication with NAT versus Epic by team 1 (top row) and team 2 (bottom row); (B) distribution of confidence scores assigned in Epic manual chart reviews (Moura et al [15]) for agreements and disagreements between the two methods; (C) annotation time comparisons between NAT versus Epic.

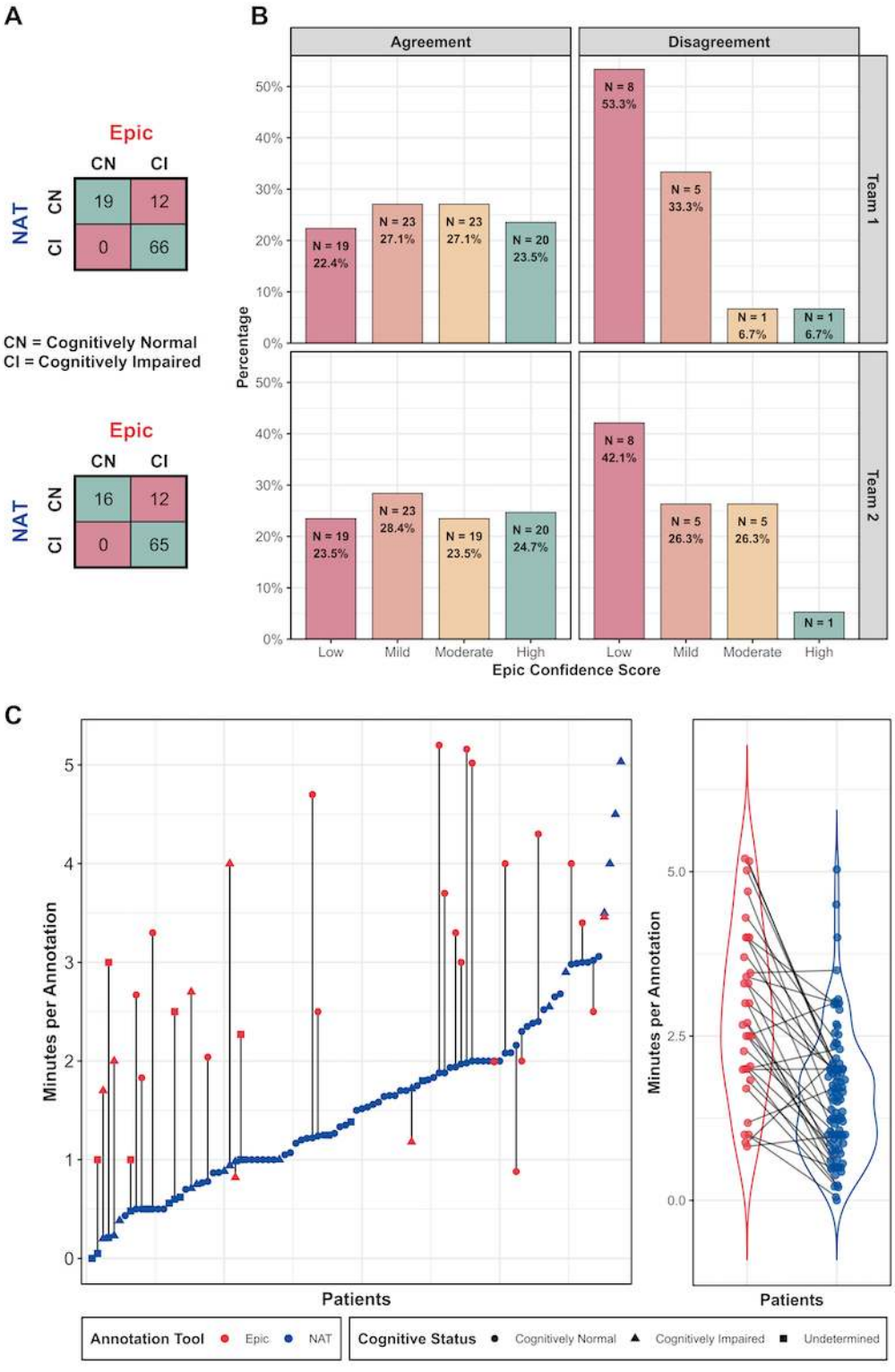
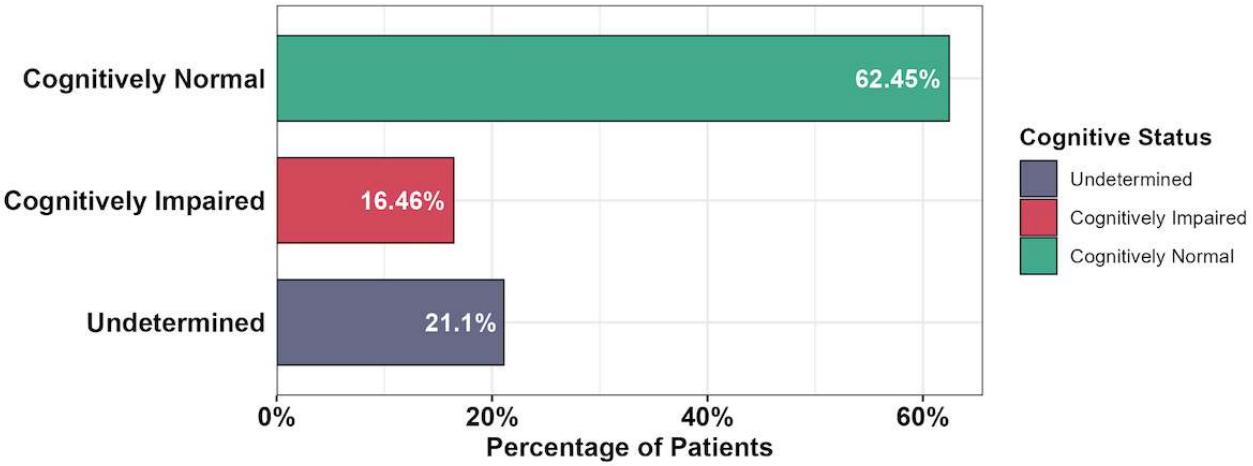


Figure 4. COVID-19 data set cognitive scores and distribution of cognitive scores in the COVID-19 data set.



Discussion

Principal Findings

In this study, we developed and evaluated a novel semiautomated NLP-powered annotation tool, NAT, to facilitate phenotyping of cognitive status. Clinical experts adjudicated the cognitive status of 627 patients at MGB health care using NAT or traditional chart reviews. NAT improves the efficiency and interrater reliability of chart review as compared to manual adjudication.

Strengths

Phenotyping methods have been applied to EHR to successfully identify patients with autism [28], diabetes [29], immunological diseases [30], and several chronic diseases [16]. EHR has been extensively used for dementia research, but the outcomes are typically defined by diagnosis codes or specialist diagnoses. Although phenotyping tools using NLP have been developed to detect cognitive impairment [18,19,24], they have been limited by their performance. In this study, we propose a novel semiautomated approach that combines NLP outputs with manual adjudication.

We selected this approach as it combines the automation of an NLP tool and the expert review required for phenotyping cognitive status. Phenotyping cognitive status requires the input from both structured (eg, diagnosis codes and medications) and unstructured (eg, clinical notes and images) data, and currently, there are no machine learning tools that integrate multiple data modalities. The approach has several advantages over manual chart reviews. Cognitive concerns are often subjective, and a significant amount of information is required to confidently ascertain the correct diagnosis. Since diagnoses are staged across months or years, individual notes across time must be evaluated together—NAT filters data for the period of interest and thus facilitates the adjudication process. Next, the absence of cognitive deficits is often difficult to adjudicate with confidence. In these cases, the annotator needs to review all notes to ensure there were no signs of cognitive impairment. NAT improves the efficiency of such tasks, as it automatically flags notes with signs of cognitive impairment as well as those with information on normal cognition and ranks them in order of importance. In addition, clinicians often use a wide variety of terms and phrases in clinical notes that can easily be missed in manual reviews. NAT, on the other hand, highlights all cognition-related patterns and phrases, decreasing the likelihood that the annotator might miss any information relevant to the decision-making task. Finally, NAT streamlines an established adjudication protocol and thus improves interrater agreement. NAT can, in principle, be extended to local hospitals and clinics that have digitized data but not an EHR system.

Limitations

This study has several limitations. First, NAT does not link to brain images, which may contain information relevant to brain function. Second, although NAT improves the efficiency of adjudicating cognitive status compared to manual chart reviews, it is not scalable to large data sets of thousands of patients. To scale to such sample sizes, fully automated machine learning algorithms that replicate the adjudication process are required. In the future, we plan to use NAT to create gold-standard data sets for training and validation of such machine learning algorithms for phenotyping cognitive status. Third, NAT adjudication was evaluated on data from a single health care system. Whether the cognition and ADL-related keywords apply to other health care settings is yet to be confirmed. The performance of the NLP tool [26] also needs to be evaluated with external data. Fourth, adjudicators were not blinded to identifiable information in EHR, which may have introduced biases in their labels. Tools, such as Philter, could be used in the future to remove protected health information in NAT [31]. Finally, research studies using EHR-based data sets are limited by the information available within the health care system and miss records of care outside the system. Such patients with missing information were labeled as “undetermined” in this study, but studies that use diagnosis codes for phenotyping of cognitive status may incorrectly label such patients as CN instead of distinguishing them as patients with insufficient information. Our study highlights the issue of missing information when phenotyping cognitive status in EHR, and consequently, the need for future work to minimize biases if such patients are excluded in a research study.

Conclusions

Although there is no substitute for a longitudinal cohort with formal cognitive evaluations to study Alzheimer disease and related dementias, leveraging EHR data with NLP holds promise. In this diagnostic study, we developed and evaluated a semiautomated NLP-powered annotation tool, NAT, to facilitate the phenotyping of cognitive status in EHRs. Expert clinicians adjudicated cognitive status of 627 patients from two distinct data sets; NAT had a high interrater agreement and improved the speed of annotations compared to manual chart reviews. Using NAT to adjudicate cognitive status would likely increase the feasibility and scalability of building gold-standard data sets for machine learning algorithms and research cohorts to study cognitive decline.

Acknowledgments

This study was supported by funding from National Institute on Aging awards (K08AG053380, R01AG073410, and P30AG062421), a National Institute of Mental Health award (K23MH115812), the James S McDonnell Foundation, and the Rappaport Fellowship. The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Authors' Contributions

ASP, BTH, DB, MBW, SSM, and SD conceptualized and designed the study. Acquisition, analysis, or interpretation of data was conducted by AN, CM, XL, TT, AK, HA, ER, DW, LB, GKR, LM, SZ, NMB, JH, and JD. The manuscript was drafted by AN, CM, and SD. All authors contributed to critical revision of the manuscript. Statistical analysis was performed by AN, CM, and XL. Funding was obtained by BTH, SSM, and SD. The study was supervised by SSM and SD.

Conflicts of Interest

SD, BTH, and ASP report research funding from Abbvie Inc. JD serves on a scientific review board for I-Mab Biopharma. NMB volunteers for the Epic Behavioral Health Subspecialty Steering Board.

Multimedia Appendix 1

Data query, preparation, and preprocessing steps.
[PDF File (Adobe PDF File), 253 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Regular expressions of dementia-related keywords.
[XLSX File (Microsoft Excel File), 7 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Regular expressions of activities of daily living (ADLs) keywords.
[XLSX File (Microsoft Excel File), 6 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Data model.
[PDF File (Adobe PDF File), 650 KB-Multimedia Appendix 4]

References

1. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191-200 [FREE Full text] [Medline: 32477638]
2. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine learning in health care: a critical appraisal of challenges and opportunities. EGEMS (Wash DC) 2019 Jan 24;7(1):1 [FREE Full text] [doi: 10.5334/egems.287] [Medline: 30705919]
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 Jun;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
4. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. Curr Epidemiol Rep 2018 Dec;5(4):331-342 [FREE Full text] [doi: 10.1007/s40471-018-0165-9] [Medline: 30555773]
5. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
6. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc 2019 Dec;14(12):3426-3444 [FREE Full text] [doi: 10.1038/s41596-019-0227-6] [Medline: 31748751]

7. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2020 Dec;2(12):e677-e680 [FREE Full text] [doi: 10.1016/S2589-7500(20)30200-4] [Medline: 33328030]
8. Alzheimer's Association. 2021 Alzheimer's disease facts and figures. Alzheimers Dement 2021 Mar;17(3):327-406. [doi: 10.1002/alz.12328] [Medline: 33756057]
9. Amjad H, Roth DL, Sheehan OC, Lyketsos CG, Wolff JL, Samus QM. Underdiagnosis of dementia: an observational study of patterns in diagnosis and awareness in US older adults. J Gen Intern Med 2018 Jul;33(7):1131-1138 [FREE Full text] [doi: 10.1007/s11606-018-4377-y] [Medline: 29508259]
10. Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. Alzheimer Dis Assoc Disord 2009;23(4):306-314 [FREE Full text] [doi: 10.1097/WAD.0b013e3181a6bebc] [Medline: 19568149]
11. Taylor, Jr. DH, Østbye T, Langa KM, Weir D, Plassman BL. The accuracy of Medicare claims as an epidemiological tool: the case of dementia revisited. JAD 2009 Jul 23;17(4):807-815. [doi: 10.3233/jad-2009-1099]
12. Kotagal V, Langa KM, Plassman BL, Fisher GG, Giordani BJ, Wallace RB, et al. Factors associated with cognitive evaluations in the United States. Neurology 2014 Nov 26;84(1):64-71. [doi: 10.1212/wnl.0000000000001096]
13. Ostbye T, Taylor DH, Clipp EC, Scoyoc LV, Plassman BL. Identification of dementia: agreement among national survey data, medicare claims, and death certificates. Health Serv Res 2008 Mar;43(1 Pt 1):313-326 [FREE Full text] [doi: 10.1111/j.1475-6773.2007.00748.x] [Medline: 18211532]
14. Chen Y, Tysinger B, Crimmins E, Zissimopoulos JM. Analysis of dementia in the US population using Medicare claims: insights from linked survey and administrative claims data. Alzheimers Dement (N Y) 2019 Jun 06;5(1):197-207 [FREE Full text] [doi: 10.1016/j.trci.2019.04.003] [Medline: 31198838]
15. Moura LMVR, Festa N, Price M, Volya M, Benson NM, Zafar S, et al. Identifying Medicare beneficiaries with dementia. J Am Geriatr Soc 2021 Aug 26;69(8):2240-2251 [FREE Full text] [doi: 10.1111/jgs.17183] [Medline: 33901296]
16. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
17. Wei W, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016 Apr;23(e1):e20-e27 [FREE Full text] [doi: 10.1093/jamia/ocv130] [Medline: 26338219]
18. Gilmore-Bykovskyi AL, Block LM, Walljasper L, Hill N, Gleason C, Shah MN. Unstructured clinical documentation reflecting cognitive and behavioral dysfunction: toward an EHR-based phenotype for cognitive impairment. J Am Med Inform Assoc 2018 Sep 01;25(9):1206-1212 [FREE Full text] [doi: 10.1093/jamia/ocy070] [Medline: 29947805]
19. Reuben DB, Hackbarth AS, Wenger NS, Tan ZS, Jennings LA. An automated approach to identifying patients with dementia using electronic medical records. J Am Geriatr Soc 2017 Mar;65(3):658-659. [doi: 10.1111/jgs.14744] [Medline: 28152164]
20. Yarnall KSH, Pollak KI, Østbye T, Krause KM, Michener JL. Primary care: is there enough time for prevention? Am J Public Health 2003 Apr;93(4):635-641. [doi: 10.2105/ajph.93.4.635] [Medline: 12660210]
21. Boustani M, Perkins AJ, Fox C, Unverzagt F, Austrom MG, Fultz B, et al. Who refuses the diagnostic assessment for dementia in primary care? Int J Geriat Psychiatry 2006 Jun;21(6):556-563. [doi: 10.1002/gps.1524]
22. Fowler NR, Frame A, Perkins AJ, Gao S, Watson DP, Monahan P, et al. Traits of patients who screen positive for dementia and refuse diagnostic assessment. Alzheimers Dement (Amst) 2015 Jun;1(2):236-241 [FREE Full text] [doi: 10.1016/j.dadm.2015.01.002] [Medline: 26258162]
23. Barnes DE, Zhou J, Walker RL, Larson EB, Lee SJ, Boscardin WJ, et al. development and validation of eRADAR: a tool using EHR data to detect unrecognized dementia. J Am Geriatr Soc 2020 Jan;68(1):103-111. [doi: 10.1111/jgs.16182] [Medline: 31612463]
24. Amra S, O'Horo JC, Singh TD, Wilson GA, Kashyap R, Petersen R, et al. Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records. J Crit Care 2017 Feb;37:202-205. [doi: 10.1016/j.jcrc.2016.09.026] [Medline: 27969571]
25. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010 Sep 21;7(9):e1000326 [FREE Full text] [doi: 10.1371/journal.pmed.1000326] [Medline: 20877712]
26. Tyagi T, Magdamo C, Noori A, Li Z, Liu X, Deodhar M, et al. Using deep learning to identify patients with cognitive impairment in electronic health records. ArXiv Preprint posted online on Nov 13, 2021 arXiv:2111.09115 [cs.CL]. [doi: 10.48550/arXiv.2111.09115]
27. van Harten AC, Mielke MM, Swenson-Dravis DM, Hagen CE, Edwards KK, Roberts RO, et al. Subjective cognitive decline and risk of MCI. Neurology 2018 Jun 29;91(4):e300-e312. [doi: 10.1212/wnl.0000000000005863]
28. Leroy G, Gu Y, Pettygrove S, Galindo MK, Arora A, Kurzius-Spencer M. Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders: development, evaluation, and application. J Med Internet Res 2018 Nov 07;20(11):e10497 [FREE Full text] [doi: 10.2196/10497] [Medline: 30404767]

29. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017 Dec;97:120-127 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.09.014] [Medline: 27919371]
30. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
31. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med 2020 Apr 14;3(1):57 [FREE Full text] [doi: 10.1038/s41746-020-0258-y] [Medline: 32337372]

Abbreviations

ACO: Accountable Care Organization
ADL: activities of daily living
CI: cognitively impaired
CN: cognitively normal
EHR: electronic health record
ICD: International Classification of Diseases
MCI: mild cognitive impairment
MGB: Mass General Brigham
NAT: NLP annotation tool
NLP: natural language processing

Edited by G Eysenbach, C Basch; submitted 17.06.22; peer-reviewed by J Walsh, L Hopper; comments to author 13.07.22; revised version received 29.07.22; accepted 31.07.22; published 30.08.22

Please cite as:
Noori A, Magdamo C, Liu X, Tyagi T, Li Z, Kondepudi A, Alabsi H, Rudmann E, Wilcox D, Brenner L, Robbins GK, Moura L, Zafar S, Benson NM, Hsu J, R Dickson J, Serrano-Pozo A, Hyman BT, Blacker D, Westover MB, Mukerji SS, Das S
Development and Evaluation of a Natural Language Processing Annotation Tool to Facilitate Phenotyping of Cognitive Status in Electronic Health Records: Diagnostic Study
J Med Internet Res 2022;24(8):e40384
URL: https://www.jmir.org/2022/8/e40384
doi: 10.2196/40384
PMID: