

Original Paper

# Extracting Critical Information from Unstructured Clinicians' Notes Data to Identify Dementia Severity Using a Rule-Based Approach: Feasibility Study

Ravi Prakash<sup>1</sup>, MS; Matthew E Dupre<sup>2,3</sup>, PhD; Truls Østbye<sup>2,4</sup>, MPH, MD; Hanzhang Xu<sup>4,5,6,7</sup>, PhD, RN

<sup>1</sup>Thomas Lord Department of Mechanical Engineering and Materials Science, Pratt School of Engineering, Duke University, Durham, NC, United States

<sup>2</sup>Department of Population Health Sciences, School of Medicine, Duke University, Durham, NC, United States

<sup>3</sup>Department of Sociology, Trinity College of Arts & Sciences, Duke University, Durham, NC, United States

<sup>4</sup>Department of Family Medicine and Community Health, School of Medicine, Duke University, Durham, NC, United States

<sup>5</sup>School of Nursing, Duke University, Durham, NC, United States

<sup>6</sup>Center for the Study of Aging and Human Development, Duke University, Durham, NC, United States

<sup>7</sup>Health Services and Systems Research (HSSR), Duke-NUS Medical School, Singapore, Singapore

**Corresponding Author:**

Hanzhang Xu, PhD, RN

Department of Family Medicine and Community Health

School of Medicine

Duke University

2100 Erwin Rd

Durham, NC, 27710

United States

Phone: 1 9196849465

Email: [hanzhang.xu@duke.edu](mailto:hanzhang.xu@duke.edu)

## Abstract

**Background:** The severity of Alzheimer disease and related dementias (ADRD) is rarely documented in structured data fields in electronic health records (EHRs). Although this information is important for clinical monitoring and decision-making, it is often undocumented or “hidden” in unstructured text fields and not readily available for clinicians to act upon.

**Objective:** We aimed to assess the feasibility and potential bias in using keywords and rule-based matching for obtaining information about the severity of ADRD from EHR data.

**Methods:** We used EHR data from a large academic health care system that included patients with a primary discharge diagnosis of ADRD based on *ICD-9 (International Classification of Diseases, Ninth Revision)* and *ICD-10 (International Statistical Classification of Diseases, Tenth Revision)* codes between 2014 and 2019. We first assessed the presence of ADRD severity information and then the severity of ADRD in the EHR. Clinicians' notes were used to determine the severity of ADRD based on two criteria: (1) scores from the Mini Mental State Examination and Montreal Cognitive Assessment and (2) explicit terms for ADRD severity (eg, “mild dementia” and “advanced Alzheimer disease”). We compiled a list of common ADRD symptoms, cognitive test names, and disease severity terms, refining it iteratively based on previous literature and clinical expertise. Subsequently, we used rule-based matching in Python using standard open-source data analysis libraries to identify the context in which specific words or phrases were mentioned. We estimated the prevalence of documented ADRD severity and assessed the performance of our rule-based algorithm.

**Results:** We included 9115 eligible patients with over 65,000 notes from the providers. Overall, 22.93% (2090/9115) of patients were documented with mild ADRD, 20.87% (1902/9115) were documented with moderate or severe ADRD, and 56.20% (5123/9115) did not have any documentation of the severity of their ADRD. For the task of determining the presence of any ADRD severity information, our algorithm achieved an accuracy of >95%, specificity of >95%, sensitivity of >90%, and an  $F_1$ -score of >83%. For the specific task of identifying the actual severity of ADRD, the algorithm performed well with an accuracy of >91%, specificity of >80%, sensitivity of >88%, and  $F_1$ -score of >92%. Comparing patients with mild ADRD to those with more advanced ADRD, the latter group tended to contain older, more likely female, and Black patients, and having received their

diagnoses in primary care or in-hospital settings. Relative to patients with undocumented ADRD severity, those with documented ADRD severity had a similar distribution in terms of sex, race, and rural or urban residence.

**Conclusions:** Our study demonstrates the feasibility of using a rule-based matching algorithm to identify ADRD severity from unstructured EHR report data. However, it is essential to acknowledge potential biases arising from differences in documentation practices across various health care systems.

(*JMIR Aging* 2024;7:e57926) doi: [10.2196/57926](https://doi.org/10.2196/57926)

## KEYWORDS

electronic health record; EHR; electric medical record; EMR; patient record; health record; personal health record; PHR; unstructured data; rule based analysis; artificial intelligence; AI; large language model; LLM; natural language processing; NLP; deep learning; Alzheimer's disease and related dementias; AD; ADRD; Alzheimer's disease; dementia; geriatric syndromes

## Introduction

More than 6 million Americans aged 65 years and older are currently living with Alzheimer disease and related dementias (ADRD), constituting about 11% of the total American population aged 65 years and older [1]. This number is projected to double by 2060, reaching 13.8 million individuals affected by ADRD [1]. Despite the absence of a cure, timely identification of ADRD can significantly improve the quality of life of patients and better prepare their families with essential support resources [2]. Early identification of ADRD will also allow health care professionals and policy makers to develop adequate care programs for both patients and their families. Furthermore, the recent US Food and Drug Administration approval for ADRD treatment, lecanemab-irmb, has demonstrated promising advancements in the pharmacological management of the disease [3]. However, it is important to note that most of these treatments are only targeted at patients with mild cognitive impairment or early-stage dementia. Therefore, identifying ADRD at an early stage of the disease has become even more critical.

Early identification of ADRD poses significant challenges [4,5]. Even among patients with diagnosed ADRD, determining disease severity remains complex from a clinical perspective. Information regarding the presence and severity of ADRD is often limited to structured fields of electronic health records (EHRs) and is more likely to be stored within unstructured EHR sections, such as clinical reports. Moreover, wide variability exists in documentation practices and data structures across different health care systems, specialties, and even among clinicians within the same clinic [6]. Over the past decade, there has been a substantial increase in leveraging information contained in the EHR to improve diagnostic precision [7]. In this context, natural language processing (NLP) has emerged as a promising approach to extract relevant information from EHR data, bridging the gap between structured and unstructured clinical information.

In recent years, EHR data and NLP have been used in various ways to improve ADRD care, such as identifying correlated symptoms [8] and common description of cognitive impairment used by clinicians [9], establishing consensus on cognitive test scores [10], phenotyping of cognitive status [11], and predicting the onset of cognitive decline [12]. However, while structured EHR data have traditionally been used in previous research, they fall short of adequately documenting the severity of ADRD

[13]. To address this gap, the unstructured component of EHR becomes critical for capturing essential symptoms and severity indicators related to ADRD [11]. The extent to which unstructured EHR data can be used to determine the severity of ADRD remains largely unknown.

In this study, we assess the feasibility and potential bias of a rule-based matching algorithm for extracting information on ADRD severity in patients with a primary discharge diagnosis of ADRD. Specifically, we develop an algorithm that acquires cognitive test scores and identifies distinct mentions of the presence and severity of ADRD from the primary discharge diagnosis. As each health care system may encode clinical information in the EHR differently, we propose a general framework that health care systems can adopt to tailor their needs and reduce irrelevant “noise” in the EHR—any unwanted irrelevant information.

## Methods

### Data Sources and Study Population

This is a retrospective cohort study that used data from Duke University Health System. Due to the sensitive nature of the data, qualified researchers trained in human subject confidentiality protocols may send requests to access the data that support the findings of this study to the corresponding author.

EHR data were extracted using Duke Enterprise Data Unified Content Explorer, a data extraction system based on Epic (Maestro Care) that identifies patient cohorts and provides access to clinical data stored in the organizational data warehouse [14]. According to previous research, we identified patients with ADRD as those who had at least one clinical encounter at Duke University Health System with at least one principal discharge diagnosis of ADRD based on *ICD-9 (International Classification of Diseases, Ninth Revision)* or *ICD-10 (International Statistical Classification of Diseases, Tenth Revision)* codes [9,15,16]. The list of *ICD-9/10* codes was based on an established algorithm from previous literature [10,17]. A total of 9115 patients aged 40 years or older, diagnosed with ADRD between January 1, 2014, and December 31, 2019, were included, totaling 65,576 patient records.

## Data Management

As most of the content present in a patient record is not directly related to the severity of ADRD, we used a keyword list that contains common ADRD terminologies to flag patient records and sections of the record that include words directly related to ADRD. The initial list was generated based on previous work [9] and was further modified to include synonyms and additional keywords corresponding to *ICD-9/10* codes related to ADRD. An expert panel of clinicians provided input on the list through an iterative process. The final list (Table S1 in [Multimedia Appendix 1](#)) included 38 unique ADRD keywords, consisting of terminologies used by clinicians at Duke University Health System to document ADRD-related information in patient records. This step helped us extract the context in which these words and phrases were mentioned to reduce noise (ie, text not related to the severity of ADRD).

## NLP Algorithms

Unstructured EHR data, characterized by its absence of standardized writing patterns, often manifests inconsistencies in both quality and content. This includes the presence of spelling errors, typographical inaccuracies, formatting inconsistencies, uncommon abbreviations, and other customary challenges inherent in note-taking practices [18]. In response to these challenges, progress in NLP capabilities has helped to filter out important and relevant information from EHR data automatically. This can be done through a rule- or learning-based approach. Previous research has suggested that learning-based methods struggle with interpretability [19] and require a considerable amount of labeled text data, which can hinder scalability. Therefore, in our study, we used a rule-based approach, allowing us to interpret our results clearly and quantify the scope of our algorithm.

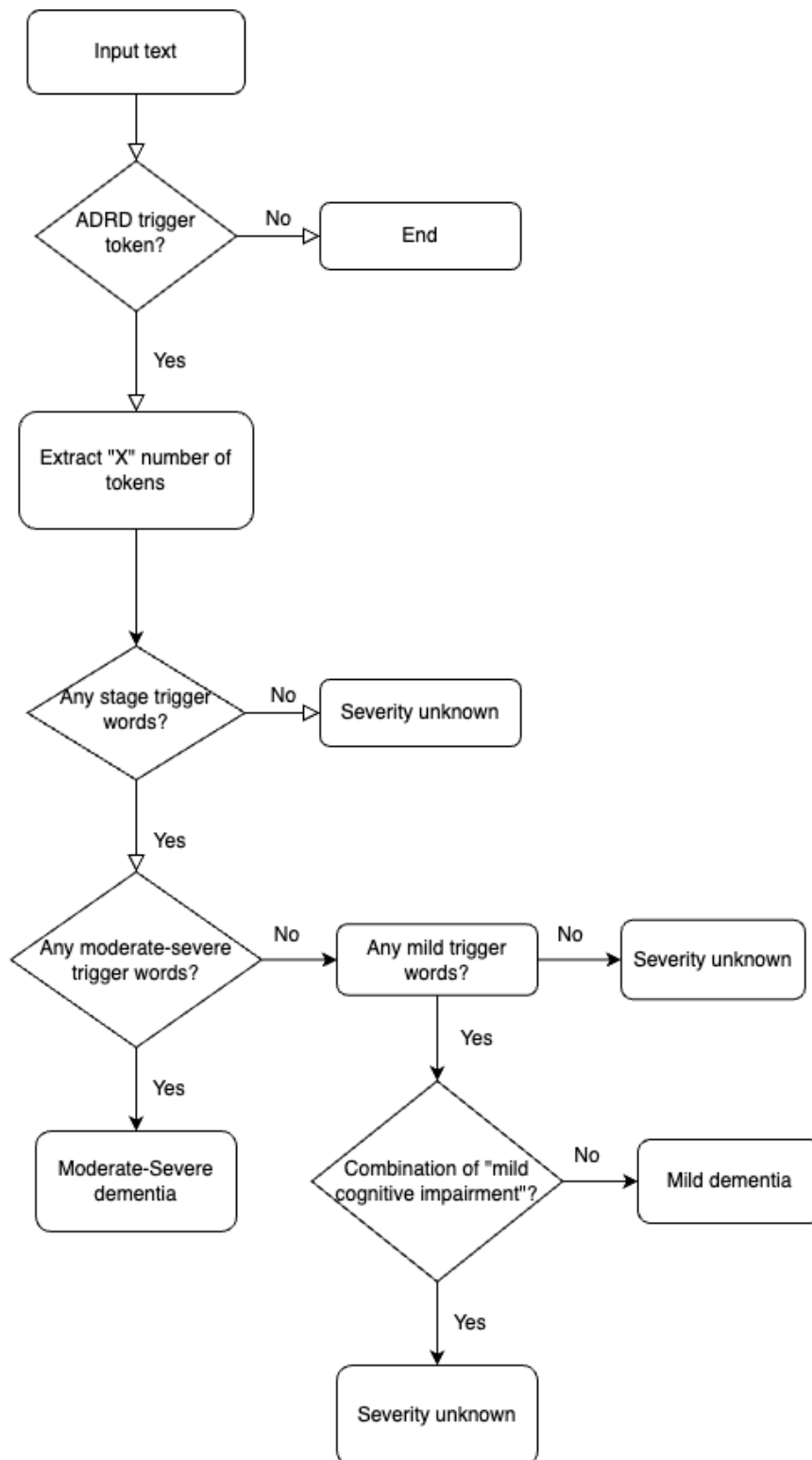
Specifically, our algorithm worked by sequentially narrowing down the target tokens (parts of the sentence) to extract information pertaining to ADRD textual mentions (ie, ADRD Text) and cognitive scores (ie, ADRD Cognitive Score). Then the algorithm determined the severity of ADRD based on either the direct mention of keywords pertaining to ADRD severity, or cognitive test scores. We categorized the severity of ADRD

into three categories that include (1) mild, (2) moderate-to-severe (ie, advanced stage of ADRD), and (3) no severity indicated. Keywords corresponding to each severity stage are included in Table S2 in [Multimedia Appendix 1](#). In the absence of any direct mention of the severity keywords in patient records, the severity of ADRD was defined based on cognitive test scores (Table S3 in [Multimedia Appendix 1](#)). Separate lists of keywords for the severity and cognitive scores were generated to aid the NLP algorithm. For patients whose records contained multiple severity levels from the same source (eg, all from cognitive scores), the more severe ADRD was assigned to the patient. As patients are likely to progress from mild to more advanced dementia over time, the more severe disease information was more likely to be the most updated information. In cases where there were discrepancies between the severity indicated by the explicit keywords and the cognitive scores, we defined the severity of ADRD based on the keywords, as providers may characterize the severity of the condition of the patients comprehensively based on additional assessments besides cognitive tests.

## ADRD Text

We used the Python package spaCy's pattern-matching function to identify occurrences of words listed in the ADRD keyword list (ie, trigger words). Once a word was identified, we assigned it a positional value of 0. We then extracted 5 tokens before and after the identified word (-5,5), creating a variable with 11 tokens, including the trigger word. The decision to extract this specific number of tokens was based on language constraints and observations from chart reviews, which revealed that most keywords related to the severity of ADRD appeared near the trigger words (eg, "mild dementia"). Next, we performed another pattern search using the severity keywords listed in Table S2 in [Multimedia Appendix 1](#), but this time only on the previously extracted 11 tokens. If a match was confirmed, we extracted the keyword defining the severity of ADRD. The extracted keyword was then assigned a severity category based on the Montreal Cognitive Assessment (MoCA) and Mini Mental State Examination (MMSE) score ranges as outlined in Table S3 in [Multimedia Appendix 1](#). The flowchart for this method is illustrated in [Figure 1](#).

**Figure 1.** Flowchart to determine the severity of Alzheimer disease and related dementias based on Alzheimer disease and related dementias trigger words. ADRD: Alzheimer disease and related dementias.



### ADRD Cognitive Score

Based on previous literature and input from clinicians, we focused on the cognitive scores from MMSE and MoCA to define the severity of ADRD [16,20]. These 2 cognitive tests were widely used in clinical practice with established cutoffs to determine the severity of cognitive impairment [16,20]. Through chart review and consultation with clinicians, we

observed the following patterns for reporting the scores in patient records. For MMSE, the scores were reported either as variations of "AB/30" or "AB". For MoCA, the scores were reported as variations of "AB/30," "AB," or a special case, "score AB," where  $A \in (0,3)$ ,  $B \in (0,9)$ . Some exemplars are present in Table 1. For each patient record, a pattern-matching search was performed to identify occurrences of MoCA and MMSE trigger words. Once a keyword was identified, it was

assigned a positional value of 0, and 10 tokens before and after the word were extracted together (-10,10) into a variable. We then used regular expression (Regex) to extract only the numerical score value from the variables with extracted tokens.

The above method was used for all the cases mentioned above, with MoCA scores including a special case, due to a slight deviation in terms of extracted tokens, belonging to (0,50) as seen in [Figure 2](#).

**Table 1.** Examples of potential causes of misclassification and representative phrases.

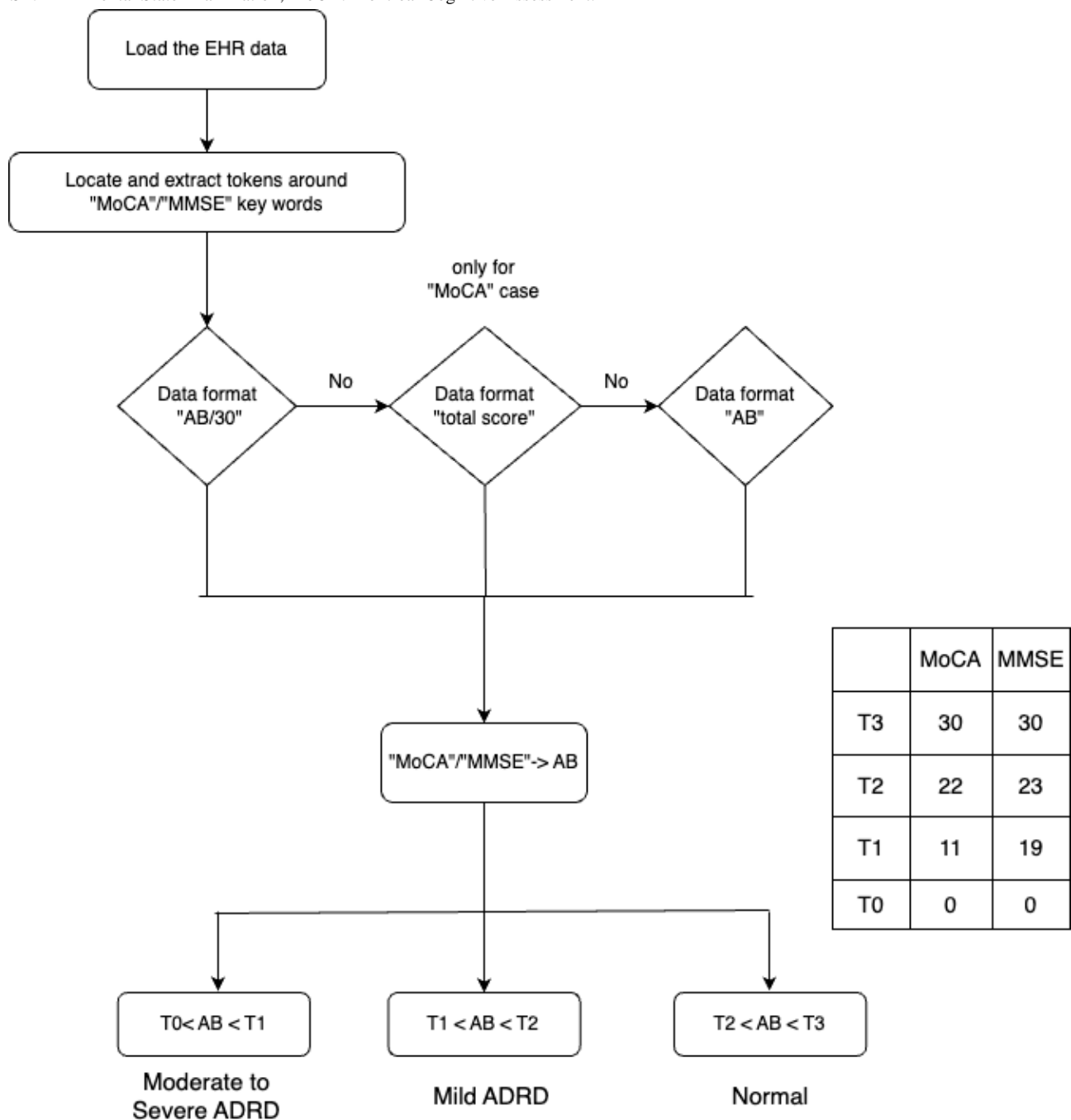
Category and cause of error	Sample sentence	Explanation
<b>ADRD<sup>a</sup> text</b>		
Incorrect textual representation	<ul style="list-style-type: none"> <li>Patient demonstrates decreased function secondary to decreased activity tolerance; cognitive deficits; medical status limitations</li> <li>['contrast. \r\n\r\n indication: dementia \r\n\r\n findings: \r\n no', ' - - - - - ']</li> </ul>	<ul style="list-style-type: none"> <li>Token delimiter missing</li> </ul>
Lack of contextual relationship between texts	<ul style="list-style-type: none"> <li>Lost 2 points for recalling only 1 out of 3 words after 3 minutes. Her neuropsychological testing concluded that she had evidence of a "mild cognitive disorder, nos."</li> <li>Mini-cognitive total scoring 1-2 recall and normal cdt: negative for cognitive impairment</li> <li>Vascular dementia with a superimposed severe delirium</li> </ul>	<ul style="list-style-type: none"> <li>Presence of "mild" and "cognitive" words in close vicinity</li> <li>Failure to identify negative context.</li> <li>Presence of "dementia" and "severe" in close vicinity.</li> </ul>
<b>ADRD cognitive score</b>		
Particular pattern of reporting MoCA <sup>b</sup> score	<ul style="list-style-type: none"> <li>MoCA XX/XX/20XX trails 1 cube 1 clock 2 naming 3 digit span 1 letter a 1 serial 7s 3 sentence repetition 2 fluency 1 abstraction 2 orientation 6 memory 0 education level 0 total score 23</li> </ul>	<ul style="list-style-type: none"> <li>Consistent pattern has been addressed in the algorithm.</li> </ul>
Complex score reporting format	<ul style="list-style-type: none"> <li>Montreal Cognitive Assessment by on 8/22 (scored 19/30, normal is 26-30/30)</li> </ul>	<ul style="list-style-type: none"> <li>Multiple scores in AB/30 format.</li> </ul>
Human error	<ul style="list-style-type: none"> <li>MMSE<sup>c</sup> 36/50</li> </ul>	<ul style="list-style-type: none"> <li>Out of bounds for MMSE score and spelled wrong.</li> </ul>

<sup>a</sup>ADRD: Alzheimer disease and related dementias.

<sup>b</sup>MoCA: Montreal Cognitive Assessment

<sup>c</sup>MMSE: Mini Mental State Examination.

**Figure 2.** Flowchart to determine Alzheimer disease and related dementias (ADRD) stage based on cognitive test score. EHR: electronic health record; MMSE: Mini Mental State Examination; MoCA: Montreal Cognitive Assessment.



**Sampling, Testing, and Analysis**

The rule-based algorithm was fine-tuned on a set of 200 patient records. To test the performance of the algorithm, we generated 2 additional sample files with 200 records each. The index of the first record was chosen using a random number generator to avoid bias. The annotation and validation were done at the level of each patient visit record. A certified clinician reviewed each EHR and determined the severity of ADRD. This manual annotation method served as our gold standard. All discrepancies were reviewed manually by both the data scientist and the clinician and were discussed on a case-by-case basis for resolution. Table 1 was also reviewed by 2 additional team members: 1 data scientist and 1 clinical scientist. We assessed

the model performance by comparing the results generated from the algorithm against the gold standard based on the following parameters:  $F_1$ -score (ie, the predictive power of the algorithm), accuracy (ie, correct classification), sensitivity, and specificity. In addition, we compared patient characteristics between patients with and without documented dementia severity using Mann - Whitney U and Pearson chi - square tests for continuous and categorical variables, respectively. Among patients with documented severity of dementia, we further compared patient characteristics between those with mild ADRD and those with moderate-to-severe ADRD.

## Ethical Considerations

This study was approved by the DUHS Institutional Review Board (Pro00104990).

## Results

A total of 9115 eligible patients were included in the study with over 65,000 records. The median age of the patient population was 78 (IQR 70-84) years. Approximately 60% (5547/9115) of the patients were female, 22.9% (2087/9115) were non-Hispanic Black, and about half were diagnosed by an ADRD specialist (neurologists, neuropsychiatrists, geriatricians, etc). On average, each record contained 931.15 words.

Among all included patients, less than half (3992/9115, 43.8%) had documented information on the stage of their dementia in the EHR. Specifically, about 35% (3190/9115) of the records (3172/9115, 34.8%) included explicit terms that indicated dementia severity, whereas about one-third of the notes (n=2977, 32.7%) included scores from cognitive tests that indicated the severity of patients' dementia. We found no differences between patients with and without their dementia severity documented with regard to sex ( $P=.45$ ) and race ( $P=.31$ ). However, patients who were older at the time of diagnosis ( $P=.01$ ) and those who were diagnosed at an ADRD specialty clinic or an in-hospital setting ( $P<.001$ ) were more likely to have the severity of their dementia documented in the EHR.

Among those with documented dementia severity, less than 25% (n=920, 23.0%) were determined only based on cognitive

scores. In our data, approximately half of these patients (n=1902, 47.7%) were documented to have moderate-to-severe ADRD. Compared with patients with mild ADRD, patients with moderate-to-severe ADRD were more likely to be older.

Table 1 shows potential causes of misclassification and representative phrases from the data sets used. Common causes of these challenges include (1) incorrect textual representation, (2) lack of contextual relationship between texts, (3) either particular or complex patterns of documenting cognitive scores, and (4) human data entry error.

Table 2 presents the performance of our algorithm across 3 sets of data. Overall, the algorithm can identify the presence of information on ADRD severity with high levels of accuracy ( $F_1$ -score=0.94, accuracy=0.97, sensitivity=0.94, and specificity=0.98) for training data (set 1). The overall accuracy across the 4 matrices in the 2 testing data sets (sets 2 and 3) was greater than 0.91, except for the  $F_1$ -score for set 3. To evaluate the performance of identifying the severity of ADRD, we defined a binary metric with "moderate-to-severe" diagnosis being positive and "mild" being negative. The ability of the algorithm to identify ADRD severity is comparable, if not better than identifying the presence of ADRD severity ( $F_1$ -score=0.94, accuracy=0.96, sensitivity=0.88, and specificity=1.0) for training data (set 1). The algorithm had accuracy greater than 0.91 with sensitivity of 1.00 in 2 testing data. The specificity for the 2 testing sets was greater than 0.80.

**Table 2.** Evaluation of ADRD<sup>a</sup> stage labeling algorithm for patient stage identification and severity of the diagnosis from unstructured EHR data.

Category	Set 1		Set 2		Set 3	
	ADRD information	ADRD severity	ADRD information	ADRD severity	ADRD information	ADRD severity
TP <sup>b</sup>	50	16	21	11	26	15
FP <sup>c</sup>	3	1	3	2	8	0
FN <sup>d</sup>	3	2	1	0	3	0
TN <sup>e</sup>	144	28	171	8	163	11
Accuracy	0.97	0.96	0.98	0.91	0.95	1.0
Sensitivity	0.94	0.88	0.95	1.0	0.90	1.0
Specificity	0.98	1.0	0.98	0.8	0.95	1.0
$F_1$ -score	0.94	0.94	0.91	0.92	0.83	1

<sup>a</sup>ADRD: Alzheimer disease and related dementias.

<sup>b</sup>TP: True positive.

<sup>c</sup>FP: False positive.

<sup>d</sup>FN: False negative.

<sup>e</sup>TN: True negative.

## Discussion

### Principal Findings

In this study, we developed and successfully implemented a rule-based algorithm to identify the severity of ADRD from unstructured EHR data. We detailed the steps to be taken for

extracting the relevant information from EHR data and highlighted the challenges associated with it due to heterogeneity in textual representation. We find a lack of access to specialty facilities may impede timely diagnosis and the possibility of treatment at early stages of ADRD progression. As the severity of dementia is critical for health care providers to prescribe

appropriate treatment and link resources to patients and their caregivers, our 2-pronged approach to search for relevant information presents a parsimonious yet effective way to make the disease severity information readily available across disciplines and care settings.

Similar to previous research [21-24], we developed the algorithm using a rule-based approach. Starting with an initial list of keywords or phrases based on previous work and contextual clinical knowledge, the list is refined iteratively to identify the target information. This step is followed by sampling the reports from the data set and dividing them into training and testing sets for gold-standard comparison and evaluation. Previous research applied a rule-based approach to identify caregiver availability [21], a record of mild cognitive impairment or Alzheimer disease [22], documentation of cognitive tests [23], and social determinants of health for patients with ADRD [24]. Unlike previous work, where either the rule definition step specific to the medical system [21], included *ICD (International Classification of Diseases)* codes [22], or had additional biomarker and cognitive tests information [23]. In comparison, our method shows robustness by using common occurring keywords and points toward the need for defining a minimum number of umbrella rules that have the potential to be generalizable for the entire data set and have better performance. Our developed algorithm is independent of the health care systems and provides clinicians with the flexibility to either use it without any modification or adapt it to their needs. A recent systematic review paper has suggested that rule-based NLP algorithms had similar performance compared with those using more sophisticated methods when the information is scarce in the EHR [25]. In our case, information on the severity of ADRD was presented using a few words, in less than half of the data. To evaluate the performance of our algorithm, we divided the tests into two categories that are (1) identifying the “presence of information on ADRD severity” and (2) “severity of ADRD.” We found our algorithm to be highly accurate in extracting documented information on ADRD severity from the EHR. The performance on sensitivity and specificity also indicates that our algorithm was able to correctly extract stage information where present and reduce false positive results. Taken together, these results support the clinical use of our simplified and generalizable approach to identify the severity of ADRD. Furthermore, compared with previous work, our algorithm showed an improved average  $F_1$ -score in identifying the ADRD severity of the condition of the patient [26]. The better performance of our algorithm compared with previous work could be attributed to the use of unstructured EHR data instead of structured EHR and defining clear umbrella rules by identifying recurring patterns in our data set for ADRD severity categorization.

The performance of our algorithm is slightly diminished in identifying the severity of dementia from the records. Upon inspection of the wrongly labeled cases, it can be attributed to (1) the test scores not reflecting the correct severity compared with the clinician’s evaluation, (2) lack of contextual understanding of the sentence, and (3) noise in EHR note (irrelevant information), also noted in previous work [21]. A majority of the abovementioned issues were mitigated by

defining subrules [27]. However, any further inclusion would have come at the cost of reduced performance and the need for increased clinician oversight, which limits its generalizability.

Although the rule-based algorithm worked considerably well in most of the cases, it is limited by the patterns and rules defined by the developer. For cases where the token default token delimiter, in our case “space,” is changed or missing, the algorithm fails to extract information. One approach to solve this could be to have an alternative copy of the algorithm that includes other common delimiters (eg, “;” “,” “:”) to identify word tokens. In our evaluation, we only found a few outliers not following the default way of describing texts with “space” as the delimiter. The urge to include all the stray cases would lead to the hard coding of the algorithm and give rise to new challenges with considerable false positive results, making it difficult to comb through. With the development of large language models (LLMs), 1 potential solution might be to use LLM to shape the EHR data in the same format, such as converting all patient records to have equal spacing, removing random commas, etc, without changing textual content and then follow a rule-based approach as presented in this study. Our design philosophy has been to keep the algorithm general while including common patterns. One possible criticism of our study could be dichotomizing the severity of ADRD. The rationale for dichotomizing the severity of ADRD into mild versus moderate-to-severe dementia is related to clinical decision-making. Given that there are several types of ADRD, such as Alzheimer disease, vascular dementia, and Lewy body dementia, to name a few, the differential treatment and care plans are limited as a considerable number of patients have mixed dementia, and the gold standard for a definitive differential diagnosis is still based on autopsy studies. Therefore, for this study, we only focused on a dichotomized version of the severity of ADRD. Future studies should further investigate a more comprehensive classification of ADRD severity and possibly include other tests for diagnosis of ADRD outside MoCA and MMSE.

Overall, we demonstrate the ability of our rule-based algorithm to identify the severity of ADRD, where present, in the EHR and narrow it down to the location of occurrence in the EHR. This not only allows us to comb through valuable unstructured data with ease, but the sequential nature of the algorithm provides us with contextual data that has a high probability of containing information about the severity of dementia. The extracted data can be used in future work to train a machine learning (ML) model with rich and high-quality data. We expect to enhance our method to further identify and predict the progression of ADRD over time. As the performance of an ML model depends on the quality of the data set, following a segmented approach of using a rule-based algorithm for extracting relevant paragraphs from the EHR can be used first to enrich the data set and reduce noise (ie, nonrelevant information from the EHR) followed by model training on the data. Previous research has found that combining structured and unstructured data might be a viable approach to classify patients. With structured data containing useful demographic information and unstructured data containing contextual, patient,



and clinical notes, the path forward could be to leverage the qualities of both kinds of data [28] for the use of EHR.

Previous studies that used traditional ML models, such as logistic regression and support vector machine, often include only the structured EHR data due to the limited requirement for data management [6,12,13]. In recent years, increasing numbers of studies have applied deep learning approaches to classify patients for a given condition [6,29,30]. One of the strengths of the deep learning approach is its ability to incorporate relationships between words and a large amount of data in the analysis, which fits the need for using both structured and unstructured data together. Despite its strength, a common criticism against the implementation of deep learning approaches in clinical settings is the lack of interpretability [19]. The use of an interpretable rule-based approach has enabled us to highlight potential biases and pitfalls to be considered when using black-box deep learning models. One potential solution is to use rule-based pattern matching to highlight the trigger words and related neighborhood of words for added context and classify or label the patient record using deep learning techniques [31,32]. With the improvements in LLMs and their enhanced contextual and semantic understanding of texts, our rule-based method can be coupled with a pretrained LLM in pre- or postprocessing of the extracted texts [33]. Care must be taken while using LLMs due to the generative nature of text predictions in avoiding alteration of textual information and being limited to standardizing textual information. The proposed approach needs to be thoroughly evaluated through data privacy and model uncertainty lens before adoption.

We also found that the percentage of patients with missing information on the severity of ADRD is very high. Given the added significance of such information in recent times due to newly approved treatment, it is critical in current clinical practice to improve documentation of the severity of dementia to promote high-quality care. In addition, despite the missing information, our relatively simple algorithm approach has been successful in making previously inaccessible and hard-to-find information readily available to clinicians for a large number of patients. These patients would have otherwise not had this information available to their care team without our very practical approach.

Our study has a few limitations due to the algorithm of our choice and design decisions. First, as mentioned earlier, the rules have to be manually defined and fine-tuned based on a training set. This process, although simple in complexity, can be challenging as EHR can be very different based on health

care systems. As the rules are manually defined, it has room for human error. On the other hand, this approach gives the researchers the flexibility to adapt the algorithm structure easily to their health care systems and needs to be fine-tuned. Second, rule-based algorithm matching studies are limited by a lack of contextual understanding between text groups and fail to recognize connotations in sentences. For example, the presence of a negative test result may confuse the algorithm pattern identification process unless explicitly included in the algorithm definition. Third, our approach has been able to only include rules for patterns that are common throughout the data set, such as for the ADRD Cognitive Score function, we include every score defined in the format of AB/30, AB, and a special type of definition as mentioned before. Therefore, we might have missed out on some of the cases straying away from the common patterns. For our instance, attempts to include every unique case led to the results being very irrelevant as it started capturing a lot of unimportant information and the algorithm became very rigid. Fourth, the study determines dementia severity based on cognitive test scores and trigger words for ADRD and does not include medical prescriptions being used by the patient which can be a future direction to make our approach holistic.

Finally, even though we have attempted to keep our method generalizable, the data were extracted from 1 health care system, and the developed rules may not directly apply to data from other health care systems. We have described in detail the steps taken in designing the algorithm with the aim of serving as a baseline approach for research in identifying ADRD severity from unstructured EHR data. The algorithm does the job well for the criteria chosen, and the choice of criteria to include can be a decision of the study designer.

## Conclusion

Rule-based algorithms can provide an interpretable approach to process unstructured EHR data. This study demonstrates the value of unstructured EHR data in providing critical information about ADRD severity from patient records. Pattern-matching rule-based algorithms can be tuned and adapted to health care systems and study-specific needs. The proposed algorithm can serve as a baseline or initial point to shift through pages of EHR reports to identify the most relevant sections or regions. However, it is important to clearly identify the assumptions made, and their limitations while defining the rules. Differences in documentation may also introduce bias in the algorithm as it is fine-tuned. Overall, rule-based algorithms are powerful in handling unstructured EHR data while being transparent and interpretable.

---

## Acknowledgments

We would like to thank Dr Xiao Hu for their feedback during the initial phase of the study. A part of our work was presented at Alzheimer's Association International Conference 2023 by the authors.

---

## Conflicts of Interest

None declared.

---

## Multimedia Appendix 1

Description of the electronic health record data set.

[DOC File , 133 KB-Multimedia Appendix 1]

## References

1. NA. 2021 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2021;17(3):327-406. [doi: [10.1002/alz.12328](https://doi.org/10.1002/alz.12328)] [Medline: [33756057](https://pubmed.ncbi.nlm.nih.gov/33756057/)]
2. PAD 20/20 Work Group on Community-Based Detection of Cognitive Decline Dementia. Improving community health-care systems' early detection of cognitive decline and dementia. *Alzheimers Dement*. 2022;18(11):2375-2381. [doi: [10.1002/alz.12837](https://doi.org/10.1002/alz.12837)] [Medline: [36314503](https://pubmed.ncbi.nlm.nih.gov/36314503/)]
3. US Food and Drug Administration. FDA converts novel Alzheimer's disease treatment to traditional approval. URL: <https://www.fda.gov/news-events/press-announcements/fda-converts-novel-alzheimers-disease-treatment-traditional-approval> [accessed 2024-08-06]
4. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc*. 2016;23(5):1007-1015. [FREE Full text] [doi: [10.1093/jamia/ocv180](https://doi.org/10.1093/jamia/ocv180)] [Medline: [26911811](https://pubmed.ncbi.nlm.nih.gov/26911811/)]
5. McCoy TH, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: a retrospective cohort study. *Alzheimers Dement*. 2020;16(3):531-540. [FREE Full text] [doi: [10.1016/j.jalz.2019.09.084](https://doi.org/10.1016/j.jalz.2019.09.084)] [Medline: [31859230](https://pubmed.ncbi.nlm.nih.gov/31859230/)]
6. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care*. 2010;48(6 Suppl):S106-S113. [doi: [10.1097/MLR.0b013e3181de9e17](https://doi.org/10.1097/MLR.0b013e3181de9e17)] [Medline: [20473190](https://pubmed.ncbi.nlm.nih.gov/20473190/)]
7. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. [FREE Full text] [doi: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
8. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr*. 2017;17(1):248. [FREE Full text] [doi: [10.1186/s12877-017-0645-7](https://doi.org/10.1186/s12877-017-0645-7)] [Medline: [29070036](https://pubmed.ncbi.nlm.nih.gov/29070036/)]
9. Gilmore-Bykovskiy A, 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;25(9):1206-1212. [FREE Full text] [doi: [10.1093/jamia/ocy070](https://doi.org/10.1093/jamia/ocy070)] [Medline: [29947805](https://pubmed.ncbi.nlm.nih.gov/29947805/)]
10. Harding BN, Floyd JS, Scherrer JF, Salas J, Morley JE, Farr SA, et al. Methods to identify dementia in the electronic health record: comparing cognitive test scores with dementia algorithms. *Healthc (Amst)*. 2020;8(2):100430. [FREE Full text] [doi: [10.1016/j.hjdsi.2020.100430](https://doi.org/10.1016/j.hjdsi.2020.100430)] [Medline: [32553526](https://pubmed.ncbi.nlm.nih.gov/32553526/)]
11. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc*. 2018;66(8):1499-1507. [doi: [10.1111/jgs.15411](https://doi.org/10.1111/jgs.15411)] [Medline: [29972595](https://pubmed.ncbi.nlm.nih.gov/29972595/)]
12. Hane CA, Nori VS, Crown WH, Sanghavi DM, Bleicher P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Med Inform*. 2020;8(6):e17819. [FREE Full text] [doi: [10.2196/17819](https://doi.org/10.2196/17819)] [Medline: [32490841](https://pubmed.ncbi.nlm.nih.gov/32490841/)]
13. Ferrão JC, Janela F, Oliveira MD, Martins HMG. Using structured EHR data and SVM to support ICD-9-CM coding. 2013. Presented at: 2013 IEEE International Conference on Healthcare Informatics; 2013 September 09-11:511-516; Philadelphia, PA.
14. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE guided query tool: providing simplified access to clinical data for research and quality improvement. *J Biomed Inform*. 2011;44(2):266-276. [FREE Full text] [doi: [10.1016/j.jbi.2010.11.008](https://doi.org/10.1016/j.jbi.2010.11.008)] [Medline: [21130181](https://pubmed.ncbi.nlm.nih.gov/21130181/)]
15. Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med Inform Decis Mak*. 2019;19(1):128. [FREE Full text] [doi: [10.1186/s12911-019-0846-4](https://doi.org/10.1186/s12911-019-0846-4)] [Medline: [31288818](https://pubmed.ncbi.nlm.nih.gov/31288818/)]
16. Halpern R, Seare J, Tong J, Hartry A, Olaoye A, Aigbogun MS. Using electronic health records to estimate the prevalence of agitation in Alzheimer disease/dementia. *Int J Geriatr Psychiatry*. 2019;34(3):420-431. [doi: [10.1002/gps.5030](https://doi.org/10.1002/gps.5030)] [Medline: [30430642](https://pubmed.ncbi.nlm.nih.gov/30430642/)]
17. Taylor DH, Fillenbaum GG, Ezell ME. The accuracy of medicare claims data in identifying Alzheimer's disease. *J Clin Epidemiol*. 2002;55(9):929-937. [doi: [10.1016/s0895-4356\(02\)00452-3](https://doi.org/10.1016/s0895-4356(02)00452-3)] [Medline: [12393082](https://pubmed.ncbi.nlm.nih.gov/12393082/)]
18. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Stats*. Feb 14, 2021;13(6):e1549. [doi: [10.1002/wics.1549](https://doi.org/10.1002/wics.1549)]
19. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. *J Am Med Inform Assoc*. 2022;29(7):1292-1302. [doi: [10.1093/jamia/ocac058](https://doi.org/10.1093/jamia/ocac058)] [Medline: [35475536](https://pubmed.ncbi.nlm.nih.gov/35475536/)]
20. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 2005;53(4):695-699. [doi: [10.1111/j.1532-5415.2005.53221.x](https://doi.org/10.1111/j.1532-5415.2005.53221.x)] [Medline: [15817019](https://pubmed.ncbi.nlm.nih.gov/15817019/)]

21. Mahmoudi E, Wu W, Najarian C, Aikens J, Bynum J, Vydiswaran VGV. Identifying caregiver availability using medical notes with rule-based natural language processing: retrospective cohort study. *JMIR Aging*. 2022;5(3):e40241. [doi: [10.2196/40241](https://doi.org/10.2196/40241)] [Medline: [35998328](https://pubmed.ncbi.nlm.nih.gov/35998328/)]
22. Aguilar BJ, Miller D, Jasuja G, Li X, Shishova E, O'Connor MK, et al. Rule-based identification of individuals with mild cognitive impairment or Alzheimer's disease using clinical notes from the United States veterans affairs healthcare system. *Neurol Ther*. 2023;12(6):2067-2078. [FREE Full text] [doi: [10.1007/s40120-023-00540-2](https://doi.org/10.1007/s40120-023-00540-2)] [Medline: [37747662](https://pubmed.ncbi.nlm.nih.gov/37747662/)]
23. Chen Z, Zhang H, Yang X, Wu S, He X, Xu J, et al. Assess the documentation of cognitive tests and biomarkers in electronic health records via natural language processing for Alzheimer's disease and related dementias. *Int J Med Inform*. 2023;170:104973. [doi: [10.1016/j.ijmedinf.2022.104973](https://doi.org/10.1016/j.ijmedinf.2022.104973)] [Medline: [36577203](https://pubmed.ncbi.nlm.nih.gov/36577203/)]
24. Wu W, Holkeboer KJ, Kolawole TO, Carbone L, Mahmoudi E. Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records. *Health Serv Res*. 2023;58(6):1292-1302. [FREE Full text] [doi: [10.1111/1475-6773.14210](https://doi.org/10.1111/1475-6773.14210)] [Medline: [37534741](https://pubmed.ncbi.nlm.nih.gov/37534741/)]
25. Patra BG, Sharma MM, Vekaria V, Adekkanattu P, Patterson OV, Glicksberg B, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*. 2021;28(12):2716-2727. [FREE Full text] [doi: [10.1093/jamia/ocab170](https://doi.org/10.1093/jamia/ocab170)] [Medline: [34613399](https://pubmed.ncbi.nlm.nih.gov/34613399/)]
26. Vyas A, Aisopos F, Vidal M, Garrard P, Paliouras G. Calibrating Mini-Mental State Examination scores to predict misdiagnosed dementia patients. *Applied Sciences*. 2021;11(17):8055. [doi: [10.3390/app11178055](https://doi.org/10.3390/app11178055)]
27. Oh IY, Schindler SE, Ghoshal N, Lai AM, Payne PRO, Gupta A. Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing. *JAMIA Open*. 2023;6(1):ooad014. [FREE Full text] [doi: [10.1093/jamiaopen/ooad014](https://doi.org/10.1093/jamiaopen/ooad014)] [Medline: [36844369](https://pubmed.ncbi.nlm.nih.gov/36844369/)]
28. Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, et al. Predictive structured-unstructured interactions in EHR models: a case study of suicide prediction. *NPJ Digit Med*. 2022;5(1):15. [FREE Full text] [doi: [10.1038/s41746-022-00558-0](https://doi.org/10.1038/s41746-022-00558-0)] [Medline: [35087182](https://pubmed.ncbi.nlm.nih.gov/35087182/)]
29. Luo X, Gandhi P, Zhang Z, Shao W, Han Z, Chandrasekaran V, et al. Applying interpretable deep learning models to identify chronic cough patients using EHR data. *Comput Methods Programs Biomed*. 2021;210:106395. [doi: [10.1016/j.cmpb.2021.106395](https://doi.org/10.1016/j.cmpb.2021.106395)] [Medline: [34525412](https://pubmed.ncbi.nlm.nih.gov/34525412/)]
30. Magdamo CG, Hong ZM, Noori A, Sheu Y, Deodhar M, Ye EM, et al. Active deep learning to detect cognitive concerns in electronic health records. *Alzheimer's & Dementia*. 2021;17(S11). [doi: [10.1002/alz.055362](https://doi.org/10.1002/alz.055362)]
31. Magdamo CG, Hong ZM, Noori A, Sheu Y, Deodhar M, Ye EM, et al. Active deep learning to detect cognitive concerns in electronic health records. *Alzheimers Dement*. 2021;17(S11):e055362. [doi: [10.1002/alz.055362](https://doi.org/10.1002/alz.055362)]
32. Noori A, Magdamo C, Liu X, Tyagi T, Li Z, Kondepudi A, et al. 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. [FREE Full text] [doi: [10.2196/40384](https://doi.org/10.2196/40384)] [Medline: [36040790](https://pubmed.ncbi.nlm.nih.gov/36040790/)]
33. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, et al. LLMs Accelerate Annotation for Medical Information Extraction. In: *LLMs accelerate annotation for medical information extraction*. Proceedings of Machine Learning Research; 2023. Presented at: Machine Learning for Health (ML4H); 10th December, 2023:82-100; New Orleans, United States. URL: <https://proceedings.mlr.press/v225/goel23a/goel23a.pdf>

## Abbreviations

**ADRD:** Alzheimer disease and related dementias

**EHR:** electronic health record

**ICD:** *International Classification of Diseases*

**ICD-9:** *International Classification of Diseases, Ninth Revision*

**ICD-10:** *International Statistical Classification of Diseases, Tenth Revision*

**LLM:** large language model

**ML:** machine learning

**MMSE:** Mini Mental State Examination

**MoCA:** Montreal Cognitive Assessment

**NLP:** natural language processing

*Edited by AF Abdulai; submitted 29.02.24; peer-reviewed by S Matos, D Chrimes, W Wu, C-L Chi; comments to author 13.05.24; revised version received 08.07.24; accepted 24.07.24; published 24.09.24*

*Please cite as:*

*Prakash R, Dupre ME, Østbye T, Xu H*

*Extracting Critical Information from Unstructured Clinicians' Notes Data to Identify Dementia Severity Using a Rule-Based Approach: Feasibility Study*

*JMIR Aging 2024;7:e57926*

*URL: <https://aging.jmir.org/2024/1/e57926>*

*doi: [10.2196/57926](https://doi.org/10.2196/57926)*

*PMID: [39316421](https://pubmed.ncbi.nlm.nih.gov/39316421/)*

©Ravi Prakash, Matthew E Dupre, Truls Østbye, Hanzhang Xu. Originally published in JMIR Aging (<https://aging.jmir.org>), 24.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Aging, is properly cited. The complete bibliographic information, a link to the original publication on <https://aging.jmir.org>, as well as this copyright and license information must be included.