JAMIA Open, 6(1), 2023, ooad014
https://doi.org/10.1093/jamiaopen/ooad014
Research and Applications

Research and Applications

Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing

Inez Y. Oh [1], Suzanne E. Schindler[2], Nupur Ghoshal[2,3], Albert M. Lai [1], Philip R. O. Payne [1], and Aditi Gupta[1,4]

[1]Institute for Informatics, Washington University School of Medicine, St. Louis, Missouri, USA, [2]Department of Neurology, Washington University School of Medicine, St. Louis, Missouri, USA, [3]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, USA and [4]Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, USA

Corresponding Author: Inez Y. Oh, Institute for Informatics, Washington University School of Medicine, 660 S. Euclid Ave, Campus Box 8132, St Louis, MO 63110, USA; iyoh@wustl.edu

Received 18 November 2022; Revised 27 January 2023; Editorial Decision 1 February 2023; Accepted 10 February 2023

ABSTRACT

Objectives: There is much interest in utilizing clinical data for developing prediction models for Alzheimer's disease (AD) risk, progression, and outcomes. Existing studies have mostly utilized curated research registries, image analysis, and structured electronic health record (EHR) data. However, much critical information resides in relatively inaccessible unstructured clinical notes within the EHR.

Materials and Methods: We developed a natural language processing (NLP)-based pipeline to extract AD-related clinical phenotypes, documenting strategies for success and assessing the utility of mining unstructured clinical notes. We evaluated the pipeline against gold-standard manual annotations performed by 2 clinical dementia experts for AD-related clinical phenotypes including medical comorbidities, biomarkers, neurobehavioral test scores, behavioral indicators of cognitive decline, family history, and neuroimaging findings.

Results: Documentation rates for each phenotype varied in the structured versus unstructured EHR. Interannotator agreement was high (Cohen's kappa = 0.72–1) and positively correlated with the NLP-based phenotype extraction pipeline's performance (average F1-score = 0.65–0.99) for each phenotype.

Discussion: We developed an automated NLP-based pipeline to extract informative phenotypes that may improve the performance of eventual machine learning predictive models for AD. In the process, we examined documentation practices for each phenotype relevant to the care of AD patients and identified factors for success.

Conclusion: Success of our NLP-based phenotype extraction pipeline depended on domain-specific knowledge and focus on a specific clinical domain instead of maximizing generalizability.

Key words: natural language processing, Alzheimer's disease, electronic health records, routinely collected health data, information retrieval

LAY SUMMARY

There is much interest in understanding risk factors and predicting the clinical trajectory of Alzheimer disease (AD) dementia, for which there is substantial variability in the rate of clinical decline. Electronic health record data collected over the course of routine medical care contains vast amounts of patient data that could be useful for this purpose. In our dataset, we found that the richest source of AD-relevant information is the clinical notes. However, the unstructured nature of the clinical note poses a significant challenge to extracting information in a format useful for predictive analyses. Natural language processing was used to extract information from clinical notes relevant to the clinical care of an AD patient, and the success of this method was determined by comparing the accuracy of the information extracted to the information manually annotated by 2 AD clinical experts. The 2 clinical experts generally agreed, and our method performed well compared to their annotations. Accurate information retrieval from unstructured clinical notes will improve understanding of a patient's medical history and overall health, and thus the ability to predict AD risk and progression.

BACKGROUND AND SIGNIFICANCE

Dementia caused by Alzheimer disease (AD) typically progresses slowly, but the clinical trajectory of disease progression, specifically rate of cognitive decline and the particular functional domains impaired, varies substantially. There is much interest in understanding factors that drive this variation. Electronic health records (EHR) data, collected over the course of routine care, captures a targeted set of indicators of patient health and disease status over time, including demographics, medical problems, treatments received, and treatment outcomes,[1] and is therefore valuable for studying the clinical trajectory of AD dementia. Some studies have used phenotypes present within structured EHR data to predict the risk of developing AD and other dementias,[2–4] but most predictive modeling for AD has used data collected from research participants outside the clinic. Notably, as observed in our study dataset, important prognostic phenotypes relevant to AD such as neuropsychological test scores, brain imaging data, genetic data, and fluid biomarkers are often absent from the structured EHR. Neuropsychological evaluations such as the Clinical Dementia Rating® (CDR®) and Mini-Mental State Exam (MMSE) assessed over time track decline in cognitive function and thus disease progression,[5–8] as can the appearance of behaviors such as misplacing items or repetitive speech where they did not previously exist, indicating a progressive decline in the ability to perform daily activities.[9–11] Structural imaging showing atrophy of specific brain regions indicating neurodegeneration can be used to diagnose AD, and when tracked over time, assess disease progression.[6,12,13] Meanwhile infarcts noted in structural brain imaging,[14] as well as comorbidities such as hypertension and depression could indicate alternative causes or additional factors resulting in cognitive decline that may have implications for the treatment approach (reviewed in reference 15) Family history of dementia, as well as genetic variants that are risk factors for AD, cannot be untethered from each other, and have been associated with increased risk of AD as well as more rapid disease progression (reviewed in reference 16) Finally, elevated CSF total tau and phosphorylated tau concentration are useful biomarkers that together indicate the severity of neuronal damage and serve as a reliable diagnostic indicator of AD.[17] Often, these indicators, amongst others, are considered together in order to diagnose AD dementia, and subsequently track disease progression, and identify possible interventions to slow the rate of cognitive decline. Therefore, these data must be extracted to fully realize the potential of the EHR in studying the trajectory of AD dementia.

Unstructured clinical notes, an underutilized and relatively inaccessible part of the EHR, contain a wealth of information including cognitive concerns, behavioral changes, and personal or family medical history.[18–20] Clinicians synthesize this information to rate dementia severity, formulate differential diagnoses, and recommend appropriate testing, treatment, and management. However, in order to use the valuable clinical phenotypes embedded within clinical notes in computational models to predict disease outcomes, automated phenotype extraction techniques, particularly natural language processing (NLP), are needed. As summarized in Shivade et al[21] in 2014, no single solution has been established, despite the plethora of NLP-based tools for clinical phenotype extraction that currently exist.[22–27]

In the dementia domain, NLP has been used to retrospectively analyze information stored in an EHR to predict risk of subsequent dementia diagnosis, and one such study found a cognitive symptom measure that stratified risk of developing dementia up to 8 years before diagnosis.[28] Another study found that combining phenotypes extracted from unstructured text with phenotypes from structured EHR data fields (eg, demographics, other diagnoses, vital signs, and laboratory values) improved the performance of machine learning (ML) models that predicted dementia risk in patients.[29] However, the existing literature shows a distinct bias towards predicting future onset of dementia; an open research question is the use of NLP to extract clinical phenotypes from unstructured sources for models that guide clinical decision making after diagnosis of dementia, specifically dementia resulting from Alzheimer disease, such as understanding the factors that differentiate fast progressors from slow progressors after diagnosis.

OBJECTIVES

Recognizing the vast quantity of valuable yet relatively inaccessible information within unstructured clinical notes, we developed an NLP-based pipeline to automate the extraction of AD-relevant clinical phenotypes. These phenotypes could inform ML algorithms for the purpose of predicting disease outcomes, such as the trajectory of AD progression, and identifying the risk factors influencing these outcomes. We evaluated the performance of our NLP-based phenotype extraction pipeline by comparing the output to gold-standard data annotations by AD subject matter experts (SMEs). Here, we describe the results of the automated phenotype extraction and discuss the challenges of extracting information from clinical narratives.

MATERIALS AND METHODS

Dataset

This was a retrospective study of electronic health records (EHR) data extracted from the Washington University in St. Louis Research

Data Core (RDC), a repository of patient clinical data from BJC HealthCare and Washington University Physicians. This study was approved by the Washington University Institutional Review Board (#201905161) and granted a waiver of HIPAA Authorization for the use of Protected Health Information (PHI). The study cohort included adult patients (≥18 years) defined using ICD-9 (331.0) and ICD-10 (G30.1, G30.8, and G30.9) diagnosis codes for AD, not including other nonspecific dementias or mild cognitive impairments in order to be sure we were gathering information from a cohort that would ultimately be diagnosed with AD. The dataset, originating from Allscripts TouchWorks, included office visits from June 1, 2013 to May 31, 2018. This timeframe was selected to avoid data harmonization issues due to a transition in the EHR system starting on June 1, 2018. The dataset consisted of clinical notes associated with office visits, and corresponding metadata such as patient identifier, author, encounter identifier, and encounter date. The dataset also included structured data for the same patients, namely demographics, diagnoses, laboratory results, medications, procedures, and vital signs. Comorbidities and neuroimaging findings were identified using ICD-9 and ICD-10CM codes (detailed in Supplementary Table S1).

Development of an NLP-based pipeline

An NLP-based phenotype extraction pipeline was built to automatically extract AD-relevant clinical phenotypes from clinical notes. Next, the output was compared to manual annotation by clinical dementia specialists, our SMEs. Finally, the pipeline was modified to improve its performance with respect to that of the SMEs. A customized approach was applied to extract each target phenotype; some phenotypes only required pattern-based matching while others required a knowledge-based approach (Figure 1).

Preprocessing

The clinical notes extracted from EHR were in rich text format (RTF) contained within tab-delimited files (TXT) alongside metadata such as the patient medical record number, author, and date authored. These were preprocessed before being analyzed by the NLP-based phenotype extraction pipeline. This entailed converting the TXT files to comma-separated files (CSV), accounting for additional tab, quote, and newline characters present, and stripping the RTF formatting. These steps were performed using the Python Pandas[30] and striprtf[31] (version 0.0.10) packages.

Identify clinical phenotypes relevant to AD

We piloted our NLP-based phenotype extraction pipeline using 10 phenotypes of interest. We first surveyed the literature in order to identify known clinical predictors of AD progression, then searched for them in the structured and unstructured EHR data, consulting with the SMEs to confirm that the selected phenotypes were important for AD risk prediction. The selection criteria were as follows: (1) whether it was documented in the unstructured data, (2) how extensively it was documented in the structured EHR, (3) its importance for clinical assessment of an AD patient, (4) the SMEs' interest in determining the accuracy of an automated extraction of that phenotype, and (5) the final list of phenotypes should represent a variety in the types of information. The final 10 phenotypes included medical comorbidities, biomarkers, neurobehavioral test scores, behavioral indicators of cognitive decline, family history, and neuroimaging findings.

Develop the pipeline to extract each clinical phenotype

Separate NLP-based modules were developed to extract each selected phenotype. There were 2, nonmutually exclusive, approaches to identifying each target phenotype. The first was pattern-based logic which relies on regular expressions to match occurrences of a phenotype term. The second was a knowledge-based approach relying on precurated ontologies to extend the search for specific phenotype terms to closely related groups of concepts. An example of pattern-based logic underlying the module for extraction of misplacing behavior in patients is described as follows (Figure 2A):

1. Search for the word "misplace", allowing for spelling errors and morphology (eg, misplacing, misplaced).
2. Exclude results where a negation (eg, "does not", "denies") appears right before "misplace".

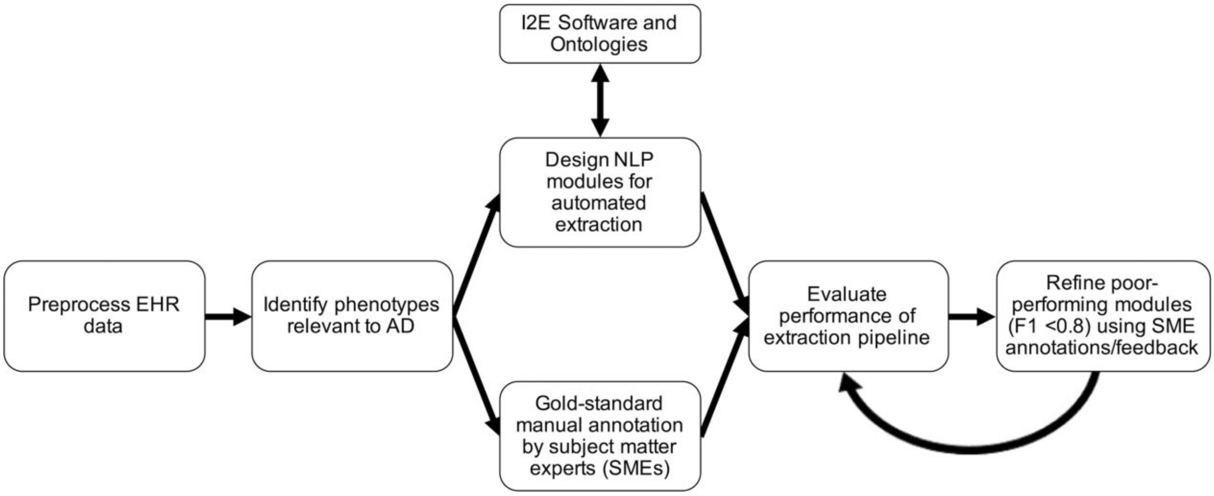
I2E Software and Ontologies
Design NLP modules for automated extraction
Preprocess EHR data
Identify phenotypes relevant to AD
Gold-standard manual annotation by subject matter experts (SMEs)
Evaluate performance of extraction pipeline
Refine poor-performing modules (F1 <0.8) using SME annotations/feedback

Figure 1. Study workflow. Unstructured notes from the EHR were preprocessed for use with the NLP platform. Clinical phenotypes relevant to AD were identified and NLP modules leveraging the I2E platform were built to extract these target phenotypes. In parallel, SMEs independently annotated a subset of notes, against which the results of the automated pipeline were compared. Modules performing poorly ($F1 < 0.8$) were refined with input from SMEs.

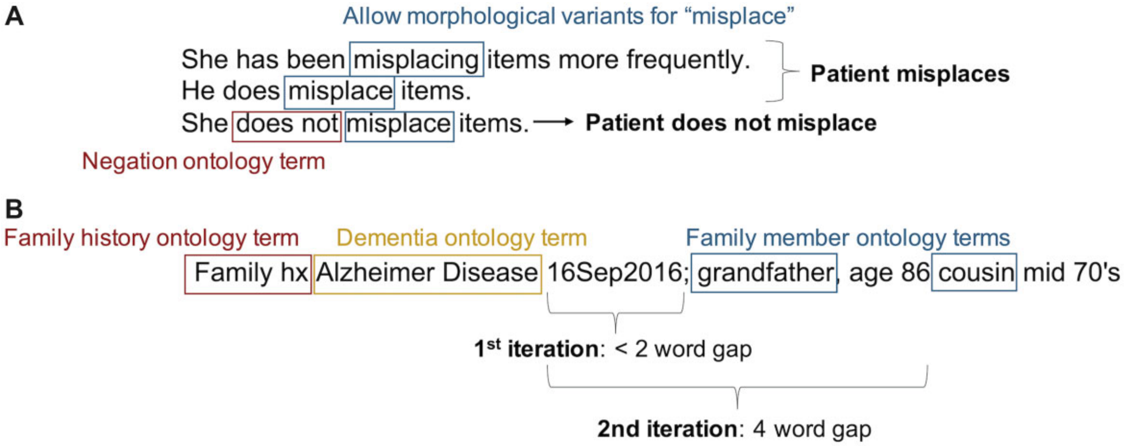
A
Allow morphological variants for "misplace"
She has been misplacing items more frequently.
He does misplace items.
Patient misplaces
She does not misplace items.
Patient does not misplace
Negation ontology term
B
Family history ontology term
Dementia ontology term
Family member ontology terms
Family hx Alzheimer Disease 16Sep2016; grandfather, age 86 cousin mid 70's
1st iteration: < 2 word gap
2nd iteration: 4 word gap

Figure 2. Example schema of NLP modules for (A) misplacing, which did not require refinement, and (B) family history of dementia, which was improved by increasing the word gap.

Knowledge-based search was utilized for some phenotypes. To illustrate this, the logic underlying the module for extracting family history of dementia is as follows (Figure 2B):

1. Define the phrase containing the target information—a phrase containing a Dementia ontology term (eg, AD, vascular dementia) and a Genetic Relations ontology term (eg, mother, brother, grandfather) occurring in any order within 5 words of each other and no other Disease or Symptom ontology term (eg, hypertension, stroke) within a 5-word space.
2. Identify the family history section of the note—Search for "Family hx" which marks the start of the section.
3. Determine if the phrase defined in (1) occurs after "Family hx".
   a. If yes, the patient has a family history of dementia.
   b. The word matching the Genetic Relations ontology term is extracted as the family member who had AD.
4. Account for negations—Exclude clinical notes which contain the following:
   a. "Denied Alzheimer Disease"
   b. Phrase containing a Negation ontology term, Family History ontology term, and Dementia ontology term within 2 words of each other

We used Linguamatics I2E, an NLP-based text-mining platform that combines text and pattern-recognition with semantic search capabilities based on curated domain knowledge, to implement our phenotype extraction pipeline.[32,33] After preprocessing, the clinical notes and their accompanying metadata were uploaded onto the I2E server and indexed using ontologies within I2E that were built on domain-specific knowledge including publicly available biomedical and healthcare terminologies.[34–37] While our modules were developed using the I2E platform, the same logic can be implemented using other NLP platforms or packages such as the open-source Natural Language Toolkit (NLTK), and publicly available ontologies such as those represented in the Unified Medical Language Systems (UMLS).[38]

Data annotation

The results of the NLP-based phenotype extraction pipeline were evaluated against a gold-standard manual annotation of 100 clinical notes conducted by our SMEs, 2 clinical dementia specialists, both board-certified neurologists specializing in memory disorders who regularly evaluate AD patients in the clinic. The 100 notes were selected via a semiautomated process to maximize the amount of information present for the target phenotypes, while ensuring that some notes contained no information regarding each phenotype in order to assess recall. A Research Electronic Data Capture (REDCap) form was created to collect responses from the SMEs who were not involved in developing the pipeline beyond identifying the phenotypes of interest. Each SME independently annotated the set of 100 notes, including copying evidence from the clinical note which guided their interpretation or choice into a free-text field.

Evaluate the performance of pipeline

For the 10 target phenotypes, results of the pipeline's output were compared independently against each SME's manual annotations and the performance metrics (precision, recall, and F1-score) averaged. Interannotator agreement (Cohen's kappa), precision, recall, and F1-score for weighted averages were calculated using the scikit-learn package (version 0.21.2).[39]

Pipeline refinement

NLP-based phenotype extraction modules which performed poorly (F1-score < 0.8) were identified and refined, using the evidence recorded in the manual annotations to analyze false positive and false negative scenarios to identify changes needed to improve the NLP module. For example, we altered the family history of dementia module as follows:

1. Problem: Low recall—Discussion of family history pertinent to AD was not limited to the "Family hx" section of notes.
   Solution: Remove limitations on search space.
2. Problem: False positives introduced when search space expanded.
   Solution: Exclude results coming from the phrase "families and patients dealing with Dementia" which referenced educational material not specific to the patient.
3. Problem: When several family members were listed as having a history of dementia, those listed later were missed.
   Solution: Expand the allowable word gap to 6 words.
4. Problem: Phrases such as "maternal grandmother" were interpreted as "mother" and "grandmother", resulting in false

positives.
Solution: Exclude the terms "maternal" and "paternal" when paired with another Genetic Relations ontology term. This sacrifices the additional detail of knowing which side of the patient's family is affected, but retains information about degree of relatedness.

RESULTS

Cohort demographics

The cohort included 2680 patients with a median age at first encounter of 79 years. 61% of the patients were female; 83% were White and 14% were Black or African American; 97% were Non-Hispanic or Latino, while 1.3% were Hispanic or Latino (Table 1).

Phenotypes extracted from clinical notes augment data available in structured EHR

The NLP-based pipeline was used to extract clinical phenotypes relevant to the prediction of AD progression from unstructured clinical notes in our dataset. The phenotypes included neurobehavioral test scores (CDR and MMSE) and their corresponding test dates, comorbidities (hypertension and depression), neuroimaging findings (presence of atrophy or infarct), behavioral indicators of dementia (repeating and misplacing), biomarker levels (total and phosphorylated tau protein levels), and family history (whether there was a family history of dementia, and if yes, which family member(s)). The availability of each phenotype within the notes versus structured EHR was noted, as measured by the number of unique patients for whom the phenotype was documented at least once (Table 2).

Documentation patterns in the structured versus unstructured EHR differed for each type of information. Neurobehavioral test scores were better documented in the structured data than unstructured notes. Valid Mini Mental State Exam (MMSE) scores with values were available for 1329 unique patients in the unstructured clinical notes and 1853 unique patients in the structured laboratory results table. 1281 patients had MMSE scores documented in both sources. However, addition of MMSE scores from unstructured and structured sources resulted in only a 3% increase in the number of patients (1901 unique patients) for whom MMSE score was available, relative to the structured source.

Table 1. AD cohort demographics extracted from the EHR

Variable	Total
Number of patients	2680
Age at first encounter, median (IQR), years	79.1 (73.6–84.5)
Sex, *N* (%)	
Female	1644 (61.3)
Race, *N* (%)	
White	2212 (82.5)
Black or African American	384 (14.3)
Asian	27 (1.0)
Other[a]	56 (2.1)
Ethnicity, *N* (%)	
Non-Hispanic or Latino	2598 (96.9)
Hispanic or Latino	36 (1.3)
Unknown	46 (1.7)

[a] *Other* includes Native Hawaiian or Other Pacific Islander, Other, Unknown, Declined, or unreported.

On the contrary, including unstructured notes resulted in a 35% increase in patients for whom CDRs with values were available. The CDR is an important functional measure of AD severity and progression,[5,6] widely used for staging dementia severity in clinical and research settings, including clinical trials.[7] CDR scores were present within unstructured notes throughout the study timeframe, sporadically documented in the structured labs table in February 2015, then frequently after July 2016. The later inclusion of CDR scores in the structured labs table suggests recognition of its utility as a standardized value that should be stored in a structured, easily retrievable format.

Extracting comorbidities and neuroimaging findings from unstructured clinical notes also added significantly to the structured data. Compared to structured data alone, these were documented for 3 times as many patients upon including unstructured data (Table 2). Behavioral indicators of cognitive decline and family history of dementia were not found in the structured data but were documented in the unstructured clinical notes for approximately half of the cohort (50%–63%). Biomarker test results were not found in the structured labs table, but were identified in clinical notes for 89 unique patients.

Evaluation of NLP pipeline

The accuracy of the NLP-based phenotype extraction pipeline was evaluated by comparing extracted phenotypes to the SMEs' annotation of 100 notes (Supplementary Table S2).

The Cohen's kappa metric, measuring interannotator agreement for each target phenotype, ranged from 0.72 to 1 (Table 3). The annotations for neurobehavioral test scores, behavioral indicators, and biomarker measurements were in strong agreement, as indicated by high kappa values. We noted that these phenotypes were documented in a consistent manner (Figure 2A). The phenotypes for which the kappa was <0.8 were presence of hypertension, depression, infarct on neuroimaging, and family history of dementia (Figure 2B).

The average performance metrics for the pipeline's extraction of the 10 phenotypes compared to the gold-standard annotations are shown in Table 4. The pipeline performed similarly against the 2 independent sets of manual annotations and generally delivered better precision (0.30–1.00) than recall (0.16–1.00). The pipeline performed well when extracting behavioral indicators, comorbidities, neurobehavioral test scores, and biomarkers, producing F1 scores ranging from 0.87–1.00. The module targeting neuroimaging finding of brain atrophy also performed well ($F1 = 0.94$) after refinement; the initial version had a low F1 (0.44), affected by the low precision (0.30) relative to the high recall (0.84), contrary to the general trend. The initial module targeting family history also performed poorly, producing low F1 scores of 0.35 and 0.26 for presence of family history and specific relation with dementia respectively, due to the low recall (0.21 and 0.16, respectively) relative to the precision (1 and 0.77, respectively).

Pipeline refinement

After reviewing low-performing phenotype extractions (F1-score $<$ 0.8), we refined our NLP pipeline. The module to identify brain infarct performed poorly (original $F1 = 0.44$). While a preliminary review of the notes did not reveal negated mentions of "infarcts," the notes selected for manual annotation did contain a significant number of negations. Thus, a second iteration of the NLP module targeting infarct was built, improving the pipeline's F1 score to

Table 2. Availability of phenotypes in structured data vs. unstructured clinical notes

Category	Phenotype	Unique patients represented in structured data only (EHR source)	Unique patients represented in unstructured notes only	Unique patients represented in structured OR unstructured data[b] (% of structured only[c])
Behavioral indicators	Misplacing	0/2680	1434/2680	NA
	Repeating	0/2680	1687/2680	NA
Comorbidities/personal medical history	Hypertension	500/2680 (diagnoses)	1425/2680	1477/2680 (295%)
	Depression	515/2680 (diagnoses)	1652/2680	1694/2680 (329%)
Family history	Family history of dementia	0/2680	1350/2680[a]	NA
Neurobehavioral tests/ ratings	Mini Mental Status Exam (MMSE)	1853/2680 (labs)	1329/2680	1901/2680 (103%)
	Clinical dementia rating® (CDR®)	1078/2680 (labs)	905/2680	1460/2680 (135%)
Neuroimaging findings	Atrophy	248/2680 (diagnoses)	666/2680	848/2680 (342%)
	Infarct	122/2680 (diagnoses)	198/2680[a]	279/2680 (229%)
Biomarker test results	Total tau and phosphorylated tau	0/2680 (labs)	89/2680	NA

[a] Numbers reflect second iteration of query.

[b] Union of patients represented in structured and unstructured data.

[c] Union of patients represented in structured and unstructured data/Number of patients in structured data *100.

Table 3. Inter-annotator agreement of manual annotations

Phenotype category	Target phenotype	Cohen's kappa
Behavioral indicators	Misplacing	0.93
	Repeating	0.90
Comorbidities	Hypertension	0.76
	Depression	0.77
Family history	Family history of dementia	0.72
	Specific relation: mother	0.95
	Specific relation: father	0.91
	Specific relation: sister	0.93
	Specific relation: brother	0.82
	Specific relation: grandmother	1.00
	Specific relation: grandfather	1.00
	Specific relation: aunt	0.86
	Specific relation: uncle	1.00
	Specific relation: cousin	1.00
Neurobehavioral test scores	Date of CDR assessment	0.86
	CDR	0.94
	Date of MMSE assessment	0.85
	MMSE	0.93
Neuroimaging findings	Atrophy	0.81
	Infarct	0.75
Biomarkers	Total tau measurement	1.00
	Total tau concentration	1.00
	Phosphorylated tau measurement	1.00
	Phosphorylated tau concentration	1.00

0.65. Excluding negated occurrences of "infarct" decreased the recall (0.84–0.60), suggesting that this iteration overcompensated with stringent exclusion rules, but increased the precision (0.30–0.71) and accuracy (0.39–0.81) of the results.

Family history of dementia also performed poorly (original F1 score = 0.35). Initial review of the notes found that many had a "Family Hx" section where family history of various diseases could be documented. Therefore, the initial strategy assumed that family history of dementia would be captured within this section. However, the notes used in the manual annotation revealed that family history was frequently documented in other parts of the notes besides the "Family Hx" section, thus explaining the poor performance. A second iteration of the NLP-based module removed this restriction, resulting in a slight decrease in accuracy and precision, but a large increase in recall, leading to an overall improved F1 score (0.35–0.66). Relatedly, the performance for the extraction of the specific relation with dementia, which relies on identification of family history of dementia, also improved (refined F1 = 0.66, compared to original F1 = 0.26).

DISCUSSION

Informative features are necessary for the success of ML predictive models. Here, we developed an automated NLP-based pipeline to extract clinical phenotypes from the EHR for an AD cohort, with the intention that these phenotypes could be used for ML predictions of AD risk, progression, and outcomes. During this process, we first examined documentation practices for various clinical phenotypes relevant to the care of AD patients; then, we extracted information from the unstructured clinical notes for target phenotypes that were particularly informative and sparsely documented in the structured EHR.

In our dataset, we found the unstructured EHR, that is, clinical notes, to be the most comprehensive source of data pertaining to clinical care of AD patients. Behavioral indicators of cognitive decline, family history of dementia, and AD biomarker test results were documented solely in unstructured notes. Comorbidities and neuroimaging findings were identified in both structured and unstructured sources but were better documented in unstructured notes. Neurobehavioral test scores were identified at similar rates in structured and unstructured EHR sources. Overall, our NLP-based phenotype extraction pipeline performed well, and performance correlated with interannotator agreement for each target phenotype.

Comparing each phenotype's presence in the unstructured versus structured EHR, we noted that MMSE had higher rates of documentation within the structured EHR compared to CDR and decided to investigate this discrepancy. MMSE scores were often documented in a semistructured format as part of a neurobehavioral test battery,

Table 4. Average performance metrics

Category	Phenotype		Accuracy	Precision	Recall	F1
Behavioral indicators	1. Repeat		0.82	1.00	0.80	0.89
	2. Misplace		0.98	0.98	0.99	0.99
Comorbidities	3. Hypertension		0.86	0.96	0.86	0.91
	4. Depression		0.87	0.87	0.87	0.87
Family history	5a. Family hx of dementia?	Original	0.55	1.00	0.21	0.35
		Refined	0.68	0.76	0.58	0.66
	5b. Specific relation	Original	0.54	0.77	0.16	0.26
		Refined	0.61	0.74	0.59	0.66
Neurobehavioral tests score (with matched test date)	6. MMSE score		0.96	0.97	0.96	0.96
	7. CDR score		1.00	1.00	1.00	1.00
Neuroimaging findings	8. Atrophy		0.89	0.99	0.89	0.94
	9. Infarct	Original	0.39	0.30	0.84	0.44
		Refined	0.81	0.71	0.60	0.65
Biomarkers	10. Presence of Tau measurement (total or phosphorylated)		0.99	1.00	0.98	0.99
	10a. Total tau concentration		0.97	0.98	0.97	0.97
	10b. Phosphorylated tau concentration		0.99	0.99	0.99	0.99

easily obtained from structured data, and additional scores extracted from clinical notes did not markedly increase the number of patients for whom these scores were available. This suggests that these test scores can be accurately retrieved from unstructured clinical notes, based on local documentation practices, the relative utility of this approach is limited.

The greatest disparities between structured and unstructured sections of the EHR were observed for comorbidities and neuroimaging findings. The structured diagnosis table lists diagnoses with their corresponding International Classification of Disease (ICD) codes, which allow mortality and morbidity data collected globally to be systematically and easily stored, accessed, analyzed, and compared.[40] However, in the United States, ICD codes are primarily used for hospital reimbursement, and thus documented minimally to satisfy administrative requirements.[41,42] Furthermore, conditions diagnosed by external care providers may not enter the structured EHR of the hospital system from which the data were obtained, but likely communicated in notes or letters. Therefore, we cannot assume that structured diagnosis tables are a complete record of patient comorbidities; as we found, comorbidity information is more comprehensively documented within clinical notes. Comorbidities extracted from clinical notes, in addition to referencing current medical issues, also revealed historical and acute symptoms or diseases not found in the structured diagnosis table. This highlights the importance of integrating data from unstructured and structured sources to obtain a comprehensive understanding of a patient's overall health necessary to guide clinical decision-making.

Much of AD research has revolved around identification of biomarkers useful for early diagnosis and prognosis of AD, in particular Aβ protein, Tau protein, and the apolipoprotein E (*APOE*) gene (reviewed in references 43 and 44) However, our dataset contained little information regarding these biomarkers. Explanations for the paucity of these data include the invasiveness and expense of these tests resulting in low uptake and their being ordered only to resolve diagnostic uncertainty.[45,46] Further, while strongly associated with AD risk and outcomes in research studies, these biomarkers did not yet represent clinically actionable targets in the timeframe covered by the dataset, and thus possibly documented only in research databases. As more therapies to treat early-stage AD enter clinical trials and eventually practice,[47] and relatively inexpensive, minimally invasive, routine testing to identify early-warning biomarkers of AD is introduced,[48] the standard of care will evolve to identify AD patients early enough that treatment is indicated. This highlights the utility of EHR data as an important longitudinal data source.

While developing the NLP-based phenotype extraction pipeline, several factors were observed to influence success. Firstly, we found that limiting the source material to the most pertinent notes based on metadata (ie, care provider and healthcare facility) would reduce the need to account for variation in linguistic patterns, note structure, concepts, and conventions in documentation practices, thus simplifying the development of the NLP modules. Secondly, we consulted with SMEs to understand the clinical relevance of each target phenotype. For example, understanding the chronicity and expected range of neurobehavioral test scores allowed us to avoid spurious associations between scores and visit date that inaccurately reflect rate of disease progression. Lastly, we learned that it was necessary to specify clinical data domains to be considered during the target phenotype extraction; to illustrate, annotation conflicts for comorbidities arose because there was confusion about whether to include current, past, or well-controlled medical problems and whether medications would be used to infer the presence of a condition. For each target phenotype, performance of the extraction pipeline positively correlated with interannotator agreement suggesting that an NLP module that is difficult to optimize also reflects a more complex experience for a manual annotator trying to extract the target phenotype from the note (Supplementary Table S3, Supplementary Figure S1).

Like many studies using EHR data, one limitation of this study relates to quality and completeness of the EHR data, as well as the racial and socioeconomic composition of our clinic population, which could limit the variability in clinical phenotypes on which our analytic pipeline has been trained. Also, the effort needed to produce a high-quality gold-standard annotation precluded our ability to evaluate our pipeline refinements against an independent dataset within the scope of this work. However, we are optimistic that the dominant documentation patterns of the target phenotypes for this patient population were captured, and given the dynamic nature of language and medical knowledge, expect pipeline improvement to be iterative. Finally, another limitation is the use of a commercial software which precludes direct portability of our pipeline, although our observations and identified factors for success remain platform-agnostic. There exist several other open-source and commercial

NLP-based systems that aim to automate clinical phenotype extraction from unstructured clinical text (reviewed in reference 21) The algorithms used in these systems may map textual elements to standardized vocabularies or concepts (eg, MetaMap[22]) identify domain-specific named entities or keywords based on expert input such as the NLP-powered annotation tool (NAT) to facilitate phenotyping of cognitive status;[23] or incorporate higher-level semantic processing, (eg, cTAKES[24]) The I2E-based pipeline we have presented here functions along these lines. Recently, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a popular deep learning-based NLP model,[25] spawning derivatives such as Clinical BERT and BioBert that are optimized for clinical and biomedical texts,[26,27] and illustrating that domain-specificity remains essential for optimal performance of such NLP-based systems.

In future work, we plan to incorporate computable phenotypes extracted by our pipeline into ML models for AD-dementia progression and determine if their addition improves model performance, thereby justifying efforts to develop NLP-based pipelines such as the one presented here.

CONCLUSION

Success of our NLP-based phenotype extraction pipeline depended on access to domain-specific knowledge from SMEs and focus on a specific clinical domain rather than maximizing generalizability. Integrating structured EHR data with clinical phenotypes from unstructured clinical notes provide a more complete picture of a patient's medical history and overall health that should improve the accuracy of ML models seeking to predict AD risk, progression, or outcomes.

FUNDING

Centene Corporation contract (P19-00559) for the Washington University-Centene ARCH Personalized Medicine Initiative.

AUTHOR CONTRIBUTIONS

IO and AG conceived and planned the study. SS and NG performed the manual annotations and served as clinical subject matter experts. IO implemented the study, analyzed the results, and took the lead writing the manuscript. All authors contributed to the interpretation of results, provided critical feedback, and helped shape the research, analysis, and manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Linguamatics I2E query files (*.i2qy) and Enterprise Architect Simulation Library (EASL) code for each NLP module can be found in the Dryad Digital Repository, at https://dx.doi.org/10.5061/dryad.0vt4b8h3g, or on the Linguamatics Community webpage (https://community.linguamatics.com/queries), accessible with the creation of a free account.

REFERENCES

1. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. *Risk Manag Healthc Policy* 2011; 4: 47–55.
2. Nori VS, Hane CA, Martin DC, Kravetz AD, Sanghavi DM. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. *PLoS One* 2019; 14 (7): e0203246.
3. Grassi M, Rouleaux N, Caldirola D, *et al.*; Alzheimer's Disease Neuroimaging Initiative. A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. *Front Neurol* 2019; 10: 756.
4. Satone VK, Kaur R, Leonard H, *et al.* Predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning. bioRxiv 2019; 792432. doi:10.1101/792432. Accessed April 20, 2021.
5. Hughes CP, Berg L, Danziger WL, *et al.* A new clinical scale for the staging of dementia. *Br J Psychiatry* 1982; 140: 566–72.
6. Hughes CP, Gado M. Computed tomography and aging of the brain. *Radiology* 1981; 139 (2): 391–6.
7. O'Bryant SE, Lacritz LH, Hall J, *et al.* Validation of the new interpretive guidelines for the clinical dementia rating scale sum of boxes score in the National Alzheimer's Coordinating Center database. *Arch Neurol* 2010; 67 (6): 746–9.
8. Doody RS, Massman P, Dunn JK. A method for estimating progression rates in Alzheimer disease. *Arch Neurol* 2001; 58 (3): 449–54.
9. Merchant FM, Weiner RB, Rao SR, *et al.* In-hospital outcomes of emergent and elective percutaneous coronary intervention in octogenarians. *Coron Artery Dis* 2009; 20 (2): 118–23.
10. McGarrigle L, Howlett SE, Wong H, *et al.* Characterizing the symptom of misplacing objects in people with dementia: findings from an online tracking tool. *Int Psychogeriatr* 2019; 31 (11): 1635–41.
11. Cullen B, Coen RF, Lynch CA, *et al.* Repetitive behaviour in Alzheimer's disease: description, correlates and functions. *Int J Geriatr Psychiatry* 2005; 20 (7): 686–93.
12. Weiler M, Agosta F, Canu E, *et al.* Following the spreading of brain structural changes in Alzheimer's disease: a longitudinal, multimodal MRI study. *J Alzheimers Dis* 2015; 47 (4): 995–1007.
13. Pini L, Pievani M, Bocchetta M, *et al.* Brain atrophy in Alzheimer's disease and aging. *Ageing Res Rev* 2016; 30: 25–48.
14. Snowdon DA, Greiner LH, Mortimer JA, *et al.* Brain infarction and the clinical expression of Alzheimer disease: the Nun study. *JAMA* 1997; 277 (10): 813–7.
15. Silva MVF, Loures CdMG, Alves LCV, De Souza LC, Borges KBG, Carvalho MdG. Alzheimer's disease: risk factors and potentially protective measures. *J Biomed Sci* 2019; 26 (1): 33.
16. Loeffler DA. Modifiable, non-modifiable, and clinical factors associated with progression of Alzheimer's disease. *J Alzheimers Dis* 2021; 80 (1): 1–27.
17. Jack CR, Bennett DA, Blennow K, *et al.*; Contributors. NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement* 2018; 14 (4): 535–62.
18. Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011; 3 (79): 79re1.
19. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, *et al.* Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep* 2017; 7 (1): 12.
20. Wq W, Pl T HM, *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2015; 23: e20–7.
21. Shivade C, Raghavan P, Fosler-Lussier E, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014; 21 (2): 221–30.

22. Aronson AR. nih gov alansnlm. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001; 17. /pmc/articles/PMC2243666/?report=abstract. Accessed January 19, 2023.
23. Noori A, Magdamo C, Liu X, *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.
24. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
25. Devlin J, Chang MW, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf 2018; 1: 4171–86. doi:10.48550/arxiv.1810.04805.
26. Alsentzer E, Murphy JR, Boag W, *et al.* Publicly available clinical BERT embeddings. Published Online First: April 6, 2019. doi:10.48550/arxiv.1904.03323.
27. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
28. McCoy TH, Han L, Pellegrini AM, *et al.* Stratifying risk for dementia onset using large-scale electronic health record data: a retrospective cohort study. *Alzheimers Dement* 2020; 16 (3): 531–40.
29. Moreira LB, Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Comput Methods Programs Biomed* 2018; 165: 139–49.
30. McKinney W. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*, Volume 445; 2010: 56–61. doi: 10.25080/majora-92bf1922-00a.
31. joshy/striprtf: stripping rtf to plain old text. https://github.com/joshy/striprtf, Accessed April 20, 2021.
32. Bandy J, Milward D, McQuay S. Mining protein-protein interactions from published literature using Linguamatics I2E. *Methods Mol Biol* 2009; 563: 3–13.
33. Trivedi S, Gildersleeve R, Franco S, *et al.* Evaluation of a concept mapping task using named entity recognition and normalization in unstructured clinical text. *J Healthc Inform Res* 2020; 4 (4): 395–410.
34. Maglott D, Ostell J, Pruitt KD, *et al.* Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005; 33 (Database issue): D54–8.
35. Sioutos N, S de C, Haber MW, *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007; 40 (1): 30–43.
36. Hastings J, Owen G, Dekker A, *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016; 44 (D1): D1214–9.
37. Vreeman DJ, McDonald CJ, Huff SM. LOINC® - a universal catalog of individual clinical observations and uniform representation of enumerated collections. *Int J Funct Inform Personal Med* 2010; 3 (4): 273–91.
38. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–270.
39. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
40. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems*. tenth revision. 2nd ed. https://www.who.int/classifications/icd/ICD-10_2nd_ed_volume2.pdf. Accessed April 23, 2021.
41. Chute CG. Coding patient information, reimbursement for care, and the ICD transition. *Virtual Mentor* 2013; 15 (7): 596–9.
42. O'Malley KJ, Cook KF, Price MD, *et al.* Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; 40 (5 Pt 2): 1620–39.
43. Khoury R, Ghossoub E. Diagnostic biomarkers of Alzheimer's disease: a state-of-the-art review. *Biomark Neuropsychiatry* 2019; 1: 100005.
44. Zetterberg H, Bendlin BB. Biomarkers for Alzheimer's disease—preparing for a new era of disease-modifying therapies. *Mol Psychiatry* 2021; 26 (1): 296–308.
45. Shaw LM, Arias J, Blennow K, *et al.* Appropriate use criteria for lumbar puncture and cerebrospinal fluid testing in the diagnosis of Alzheimer's disease. *Alzheimers Dement* 2018; 14 (11): 1505–21.
46. Johnson KA, Minoshima S, Bohnen NI, *et al.*; Amyloid Imaging Taskforce. Appropriate use criteria for amyloid PET: a report of the Amyloid Imaging Task Force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association. *Alzheimers Dement* 2013; 9 (1): e-1-16.
47. FDA Grants Accelerated Approval for Alzheimer's Drug | FDA. https://www.fda.gov/news-events/press-announcements/fda-grants-accelerated-approval-alzheimers-drug. Accessed August 2, 2021.
48. Schindler SE, Bollinger JG, Ovod V, *et al.* High-precision plasma β-amyloid 42/40 predicts current and future brain amyloidosis. *Neurology* 2019; 93 (17): E1647–59.