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

Inez Y Oh  
Washington University School of Medicine in St. Louis
Suzanne E Schindler  
Washington University School of Medicine in St. Louis
Nupur Ghoshal  
Washington University School of Medicine in St. Louis
Albert M Lai  
Washington University School of Medicine in St. Louis
Philip R O Payne  
Washington University School of Medicine in St. Louis

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.

Recommended Citation
Oh, Inez Y; Schindler, Suzanne E; Ghoshal, Nupur; Lai, Albert M; Payne, Philip R O; and Gupta, Aditi, "Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing." JAMIA Open. 6, 1. ooad014 (2023).  
https://digitalcommons.wustl.edu/oa_4/1467

This Open Access Publication is brought to you for free and open access by the Open Access Publications at Digital Commons@Becker. It has been accepted for inclusion in 2020-Current year OA Pubs by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Research and Applications

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


1Institute for Informatics, Washington University School of Medicine, St. Louis, Missouri, USA, 2Department of Neurology, Washington University School of Medicine, St. Louis, Missouri, USA, 3Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, USA and 4Division 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
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, 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, 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, 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. Structural imaging showing atrophy of specific brain regions indicating neurodegeneration can be used to diagnose AD, and when tracked over time, assess disease progression. Meanwhile infarcts noted in structural brain imaging, 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. 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. 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. in 2014, no single solution has been established, despite the plethora of NLP-based tools for clinical phenotype extraction that currently exist.

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. 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. The existing literature shows a distinct bias toward 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 package.

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, behavior 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”.

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.
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 (e.g., AD, vascular dementia) and a Genetic Relations ontology term (e.g., mother, brother, grandfather) occurring in any order within 5 words of each other and no other Disease or Symptom ontology term (e.g., 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. 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).

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).

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, widely used for staging dementia severity in clinical and research settings, including clinical trials. 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.

Table 1. AD cohort demographics extracted from the EHR

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

Other includes Native Hawaiian or Other Pacific Islander, Other, Unknown, Declined, or unreported.

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
the notes used in the manual annotation revealed that family history of dementia would be captured within this section. However, because of frequent documentation in other parts of the notes besides the “Family Hx” section, which relies on identification of family history of dementia, also improved (refined F1 score 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 at similar rates in structured and unstructured 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 2. Availability of phenotypes in structured data vs. unstructured clinical notes

<table>
<thead>
<tr>
<th>Category</th>
<th>Phenotype</th>
<th>Unique patients represented in structured data only (EHR source)</th>
<th>Unique patients represented in unstructured notes only</th>
<th>Unique patients represented in structured or unstructured data (% of structured only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral indicators</td>
<td>Mislacing</td>
<td>0/2680</td>
<td>1434/2680</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Repeating</td>
<td>0/2680</td>
<td>1687/2680</td>
<td>NA</td>
</tr>
<tr>
<td>Comorbidities/personal</td>
<td>Hypertension</td>
<td>500/2680 (diagnoses)</td>
<td>142/2680</td>
<td>1477/2680 (295%)</td>
</tr>
<tr>
<td>medical history</td>
<td>Depression</td>
<td>515/2680 (diagnoses)</td>
<td>1652/2680</td>
<td>1694/2680 (329%)</td>
</tr>
<tr>
<td>Family history</td>
<td>Family history of dementia</td>
<td>0/2680</td>
<td>1350/2680</td>
<td>NA</td>
</tr>
<tr>
<td>Neurobehavioral tests/</td>
<td>Mini Mental Status Exam</td>
<td>1853/2680 (labs)</td>
<td>1329/2680</td>
<td>1901/2680 (103%)</td>
</tr>
<tr>
<td>ratings</td>
<td>Clinical dementia rating</td>
<td>1078/2680 (labs)</td>
<td>905/2680</td>
<td>1460/2680 (135%)</td>
</tr>
<tr>
<td>Neuroimaging findings</td>
<td>Atrophy</td>
<td>248/2680 (diagnoses)</td>
<td>666/2680</td>
<td>848/2680 (342%)</td>
</tr>
<tr>
<td></td>
<td>Infarct</td>
<td>122/2680 (diagnoses)</td>
<td>198/2680</td>
<td>279/2680 (229%)</td>
</tr>
<tr>
<td>Biomarker test results</td>
<td>Total tau and phosphorylated tau</td>
<td>0/2680 (labs)</td>
<td>89/2680</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Numbers reflect second iteration of query.

*aUnion of patients represented in structured and unstructured data.

*bUnion of patients represented in structured and unstructured data/Number of patients in structured data *100.

Table 3. Inter-annotator agreement of manual annotations

<table>
<thead>
<tr>
<th>Phenotype category</th>
<th>Target phenotype</th>
<th>Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral indicators</td>
<td>Mislacing</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Repeating</td>
<td>0.90</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>Hypertension</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Depression</td>
<td>0.77</td>
</tr>
<tr>
<td>Family history</td>
<td>Family history of dementia</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Specific relation: mother</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Specific relation: father</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>Specific relation: sister</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Specific relation: brother</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Specific relation: grandmother</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Specific relation: grandfather</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Specific relation: aunt</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Specific relation: uncle</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Specific relation: cousin</td>
<td>1.00</td>
</tr>
<tr>
<td>Neurobehavioral test scores</td>
<td>Date of CDR assessment</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>CDR</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Date of MMSE assessment</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>MMSE</td>
<td>0.93</td>
</tr>
<tr>
<td>Neuroimaging findings</td>
<td>Atrophy</td>
<td>0.81</td>
</tr>
<tr>
<td>Biomarkers</td>
<td>Total tau measurement</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Total tau concentration</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Phosphorylated tau measurement</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Phosphorylated tau concentration</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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).
Neuroimaging findings 8. Atrophy 0.89 0.99 0.89 0.94

Furthermore, conditions used for hospital reimbursement, and thus documented minimally to systematically and easily stored, accessed, analyzed, and compared.41,42 However, in the United States, ICD codes are primarily 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 Furthermore, 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

**Table 4. Average performance metrics**

<table>
<thead>
<tr>
<th>Category</th>
<th>Phenotype</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral indicators</td>
<td>1. Repeat</td>
<td>0.82</td>
<td>1.00</td>
<td>0.80</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>2. Misplace</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>3. Hypertension</td>
<td>0.86</td>
<td>0.96</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>4. Depression</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>Family history</td>
<td>5a. Family hx of dementia?</td>
<td>Original</td>
<td>0.55</td>
<td>1.00</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Refined</td>
<td>0.68</td>
<td>0.76</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>5b. Specific relation</td>
<td>Original</td>
<td>0.54</td>
<td>0.77</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Refined</td>
<td>0.61</td>
<td>0.74</td>
<td>0.59</td>
</tr>
<tr>
<td>Neurobehavioral tests score (with matched test date)</td>
<td>6. MMSE score</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>7. CDR score</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Neuroimaging findings</td>
<td>8. Atrophy</td>
<td>0.89</td>
<td>0.99</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>9. Infarct</td>
<td>Original</td>
<td>0.39</td>
<td>0.30</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Refined</td>
<td>0.81</td>
<td>0.71</td>
<td>0.60</td>
</tr>
<tr>
<td>Biomarkers</td>
<td>10. Presence of Tau measurement (total or phosphorylated)</td>
<td>0.99</td>
<td>1.00</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>10a. Total tau concentration</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>10b. Phosphorylated tau concentration</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
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) or incorporate higher-level semantic processing, (eg, cTAKES) 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, spawning derivatives such as Clinical BERT and BioBert that are optimized for clinical and biomedical texts, 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.0vtr4b8s3g, or on the Linguamatics Community webpage (https://community.linguamatics.com/queries), accessible with the creation of a free account.

REFERENCES


