Cognitive Function Characterization Using Electronic Health Records Notes

Adrienne Pichon, MPH[1#], Betina Idnay, RN[2,3,4#], Karen Marder, MD, MPH[3,4], Rebecca Schnall, PhD, MPH, RN[2], Chunhua Weng, PhD[1]
[1] Department of Biomedical Informatics, [2] School of Nursing, [3] Department of Neurology, [4]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York, USA

Abstract

Cognitive impairment is a defining feature of neurological disorders such as Alzheimer's disease (AD), one of the leading causes of disability and mortality in the elderly population. Assessing cognitive impairment is important for diagnostic, clinical management, and research purposes. The Folstein Mini-Mental State Examination (MMSE) is the most common screening measure of cognitive function, yet this score is not consistently available in the electronic health records. We conducted a pilot study to extract frequently used concepts characterizing cognitive function from the clinical notes of AD patients in an Aging and Dementia clinical practice. Then we developed a model to infer the severity of cognitive impairment and created a subspecialized taxonomy for concepts associated with MMSE scores. We evaluated the taxonomy and the severity prediction model and presented example use cases of this model.

Introduction

There are an estimated 5.8 million individuals in the United States (US) age 65 and older living with Alzheimer's disease (AD), with a projected increase to 13.8 million by 2050[1]. As the sixth-leading cause of death in the US, AD is one of the most significant unmet medical needs of our time[2]. The Food and Drug Administration (FDA) recently approved a disease-modifying treatment, aducanumab, based on the expected drug's effect on the surrogate endpoint – 18 years since the last FDA-approved treatment [3]. Clinical trials are the gold standard for providing evidence on the potential harms and benefits of an investigational treatment, but they are time-consuming and expensive. On average, the development of a disease-modifying treatment for AD requires 13 years and costs $5.7 billion[4]. This highlights how successful AD clinical trials are crucial.

Eligibility prescreening of potential participants is a major bottleneck to successful AD clinical trial recruitment, even though prescreening has resulted in decreased costs incurred by screen failures due to ineligibility and has helped in strategizing recruitment efforts[4]. One of the challenges is the determination of the level of cognitive impairment, which is a critical component for determining a potential participant's eligibility to participate in an AD clinical trial[5]. The 30-item Folstein Mini-Mental State Examination (MMSE) is the most common measure of cognitive function used in AD clinical trials to define the severity of dementia[6]. Any score of 24 or more (out of 30) indicates normal cognition. Below this, scores can indicate mild dementia (19–23 points), moderate dementia (10–18 points), or severe dementia (≤9 points)[7]. However, recent MMSE score (i.e., within one year) is not always readily available in the electronic health record (EHR) and when documented is only found in unstructured clinical notes, rendering it difficult to determine without manual inspection[8,9]. Given the complexity and the long range of trajectory of changes of AD pathological process over time, determining the patient's potential eligibility to a clinical trial is challenging without a recent MMSE score. This warrants the research team to consider other documented cognitive symptoms (e.g., increased forgetfulness, worsening word finding difficulty) in lieu of a recent MMSE score, which can result in an inaccurate representation of the patient's level of cognitive impairment due to research staff's subjective interpretation of the clinical notes[10]. Hence, a more efficient way to characterize cognitive status for AD clinical trials is needed.

Unsupervised learning approaches and automated search algorithms have been developed to identify clinical subtypes of AD using EHR narrative[11-14]. Subspecialized terminology based on keywords and phrases from narrative text were also constructed to classify cognitive impairment[11]. There are also ontologies available to formally represent AD such as the Alzheimer's Disease Ontology[15], the AD Map Ontology[16], and the AlzFuzzyOnto[17]. The Common Alzheimer's Disease Research Ontology was developed for AD research working to enable integration and comparative analysis of AD research[18]. The Semantic Web Application in Neuromedicine was developed to build applications for bench scientists initially for, but not limited to, AD research[19]. However, to the best of our knowledge, there has been no mapping from cognitive function concepts to MMSE score. **Our goal is to develop a subspecialized taxonomy for cognitive status characterization and a model for MMSE prediction using related concepts**. In this paper, we identified the concepts pertinent to cognitive function measurement in relation to MMSE from clinical notes and mapped them to the Unified Medical Language System (UMLS). We then developed a model to infer the severity of cognitive impairment and used this to construct a novel taxonomy. This paper reports on a computational method for

# Contributed equally, co-first authors

phenotyping cognitive impairment using EHR narratives that has the potential to support assorted downstream tasks, such as identifying patients for recruitment of prospective clinical trials, constructing and describing cohorts for retrospective observational studies, and implementation of supportive tools for providers embedded in EHRs and clinical workflows[20]. These advancements could facilitate a personalized approach to care (therapies and supportive services according to the progression of disease) and provide insight into underlying mechanisms of disease to advance precision medicine. This method could be applied in other contexts with clinically relevant score-based proxies.

Methods

Data source and sample selection

This study was approved by the Columbia University Irving Medical Center (CUIMC) Institutional Review Board (#AAAD1873). We purposely selected 150 clinical visit notes, each containing an MMSE score, from 118 distinct patients diagnosed with prodromal (amnestic mild cognitive impairment (aMCI)) or probable AD and seen by CUIMC Aging and Dementia clinicians between February 1, 2020 and November 15, 2020. We extracted the following information from their EHR data: *diagnosis* (aMCI or AD); *the date of visit*; *type of visit* (initial or follow up; in-person or telehealth); *MMSE during the visit*; *language used to administer the MMSE*; *chief complaint; history of present illness* (for initial visits); *interval history* (for follow up visits); *neuropsychiatric symptoms; functional abilities assessment; impression;* and *plans*. Initial visit note is included only if a follow up visit note is available that indicates a diagnosis of AD or aMCI. MMSE score (0-30) was used as the label for this analysis. The eligibility of each patient for each visit was determined solely based on the MMSE score for three AD clinical trial protocols, representing phase 1 (NCT03822208), phase 2 (NCT03282916), and phase 3 (NCT03887455) studies respectively and covering a broad range of MMSE inclusion criterion thresholds (i.e., 16-28, 18-30, and 22-30). The workflow is outlined in Figure 1.

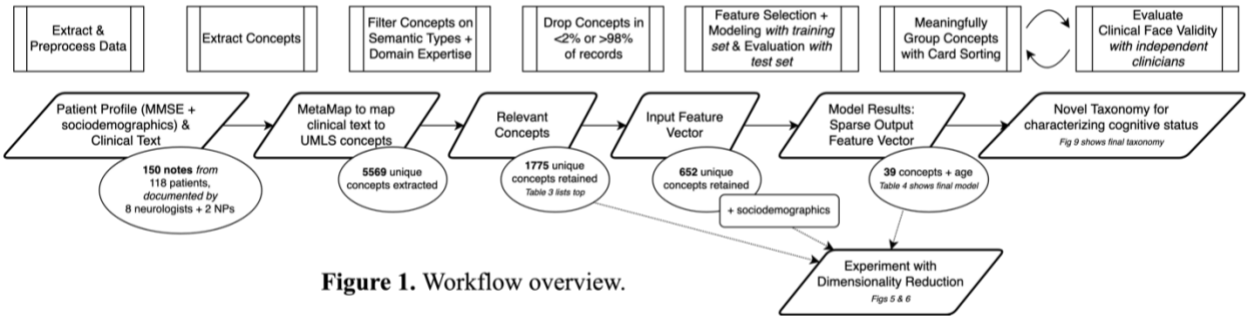

Figure 1. Workflow overview.

Data preparation and concept extraction

Python 3.6 was used for data processing and analysis. R 4.0.2 was used for descriptive statistical computations on demographic characteristics and the heatmap (using ggplot2). The corpus of extracted clinical text was imported and light text preprocessing was performed. A local version of MetaMap along with pymetamap[21] were used to parse the clinical text and extract concepts. The objects that are extracted with pymetamap contain information about the terms tagged, including the UMLS Concept Unique Identifier (cui), semantic types (semtypes), the phrase that triggered the mapping (trigger), negation (1/0 last digit of trigger), scoring information on the quality of the concept match (score), and location codes (location, pos_info, tree_codes). An example MetaMap Concept Object is shown below:

```
ConceptMMI(index='5', mm='MMI', score='5.18', preferred_name='Ability to Drive', cui='C4050139',
semtypes='[inpr]', trigger='["Driving"-tx-3-"driving"-verb-0]', location='TX', pos_info='137/7', tree_codes='')
```

Feature selection

After concepts were extracted, we experimented with different methods of selecting concepts for inclusion in subsequent modeling. Filtering based on semantic types was a particularly fruitful approach, specifically for excluding irrelevant or off-topic categories and ultimately selecting the ones used in the final model. Clinical expertise was also applied to review the remaining corpus and manually remove any CUIs that were not relevant (e.g., “wellplate”). We decided not to filter based on MetaMap score because even low scores (within relevant semantic types) matched well when manually compared in the patient chart. Finally, concepts occurring in >98% of notes or <2% of notes were omitted for their low salience or low prevalence.

Feature vector

MMSE score (0-30) was used as the outcome variable (label) for the regression model, and each concept feature (extracted from the EHR via MetaMap) was represented as 1 if it was present in the clinical note, 0 if absent, and -1

if present but negated. Several sociodemographic features were also included in the feature vector, specifically: age (standardized), language (Spanish), and sex (male).

Model training and evaluation

Split dataset. The full dataset was split 80/20 into a training and test dataset. The training dataset (n = 120) was used to train the model, and the testing dataset (n = 30) was held out and after training was complete, the model was evaluated on this unseen testing set.

Parameter tuning. Regression with Lasso is useful for feature selection and modeling at the same time, producing a sparse matrix of features that are relevant in predicting an outcome of interest[22]. The Lasso penalizes each additional model parameter, driving the coefficients towards either 0 or inclusion in the model. In training the model, we experimented with different alpha parameters between 0.05 and 0.4. These parameters resulted in inclusion of more or fewer concept features in the final model, with varying performance on metrics across the data. In the end, an alpha parameter of 0.25 resulted in a reasonable model that was sparse and not overfit or underfit.

Table 1. Sample Characteristics (n = 150 notes)

Age, years (mean (SD) [range])		74.3 (8.3) [54-94]
Women (n (%))		94 (63)
Race (n (%))		
White		75 (50)
Black/African American		10 (6.67)
Other combinations not described		10 (6.67)
Native Hawaiian/Other Pacific Islander		2 (1.33)
American Indian/Alaskan		2 (1.33)
Asian		2 (1.33)
Unknown		49 (32.67)
Ethnicity (n (%))		
Not Hispanic/Latino/Spanish Origin		79 (52.67)
Hispanic/Latino/Spanish Origin		19 (12.67)
Unknown		52 (34.67)
Note Type (n (%))	Initial	41 (27)
	Follow up	109 (73)
Visit Type (n (%))	In-person	49 (33)
	Telehealth	101 (67)
Primary Diagnosis (n (%))	AD	117 (78)
	aMCI	33 (22)
MMSE Score (mean (SD) [range])		20.2 (7.1) [0-30]
Language of MMSE administration (n (%))		
	English	122 (82)
	Spanish	27 (18)

Model parameters and evaluation. The magnitude and direction of association for final model parameters are returned from the model training and can be used to calculate predicted MMSE scores for a set of feature vectors. The concepts selected by the model were inspected to determine broad categories for important terms, and to examine positive or negative associations between concepts and cognitive status and the magnitude of this association. The R-squared value and Root Mean Square Error (RMSE) were used to evaluate the model, comparing performance on both the training and held-out test data. Sanity checks were performed across a subset of the clinical notes, to verify if concepts extracted and included as important were indeed present in the clinical text and used in the way assumed from the mapping.

Taxonomy development and evaluation. The final concepts selected by the model were then used to generate the novel taxonomy by card sorting[23] iteratively among authors (BI and AP) until consensus was achieved, led by BI who has domain experience in this clinical setting. The face validation of the final taxonomy was evaluated by two independent clinicians, a nurse practitioner who sees individuals with AD in clinical and research settings and a lead clinical research coordinator of an AD research center.

Results

Study sample and MMSE-based eligibility

The study sample (Table 1) included 118 distinct patients corresponding to 150 clinical visits notes (63% female, 50% White, and 53% Non-Hispanic) with mean (SD) age of 74.3 (8.3) years. Of these clinical notes, 49 (33%) visits were conducted in person and 101 (67%) visits were completed via telehealth. A majority of the clinical notes were from follow-up visits (73%). The sample includes a total of 117 (78%) notes that indicate a diagnosis of AD, while 33 (22%) indicate aMCI. Mean (SD) MMSE score was 20.2 (7.1), representing the full range of possible scores. A majority of the MMSE tests were administered in English (82%). Eligible patients based on MMSE are 112 (75%), 106 (71%) and 76 (51%) respectively for the phases 1-3 trials in order. The sample includes 118 clinical visit notes of patients who were deemed eligible to participate in one, two, or all of the research protocols during that particular visit, showing many patients are eligibility for more than one study. Of these notes, 61% (n=71) indicates that the patient was eligible in all the three studies. A total of 32 clinical visits notes document visits with patients who are too cognitively impaired (i.e., MMSE < 16) to be considered for any of the trials.

Final model and concepts selection

A total of 5569 total unique concepts were extracted from the clinical notes for the initial corpus. After expert inspection by BI, a total of 18 UMLS sematic types were included to identify and select relevant concepts (Table 2). After filtering, 1775 unique concepts remained in the corpus. Table 3 shows concepts that were mentioned more than 90 times. The final model (Table 4) includes 40 features, including 39 concepts and age (standardized), with an alpha parameter of 0.25.

Table 2. UMLS Semantic types used to filter relevant concepts.

Event	Language	Physiologic Function
Family Group	Mental Process	Self-help or Relief Organization
Finding	Mental or Behavioral Dysfunction	Social Behavior
Functional Concept	Pathologic Function	Sign or Symptom
Health Care Activity	Physical Object	Spatial Concept
Individual Behavior	Phenomenon or Process	Therapeutic or Preventive Procedure

Table 3. Most frequently used UMLS concepts extracted after filtering irrelevant concepts via semantic types.

CUI	Preferred Name	Frequency	CUI	Preferred Name	Frequency
C1518422	Negation	340	C1527305	Feelings	126
C2584313	Discussion (communication)	210	C0085639	Falls	125
C1261322	Evaluation procedure	200	C0700287	Reporting	124
C1512346	Patient Visit	184	C0015576	Family	115
C1301732	Planned	184	C0150312	Present	110
C0392747	Changing	154	C2004062	History of previous events	105
C0700327	Memory observations	151	C0019665	Historical aspects qualifier	105
C0025260	Memory	148	C0018524	Hallucinations	104
C0011011	Daughter	145	C0442519	Home environment	102
C1515187	Take	142	C0242664	husband	101
C0242665	wife	142	C4553314	Hallucinations, CTCAE	98
C0589120	Follow-up status	136	C1299586	Has difficulty doing (qualifier value)	98
C1522577	follow-up	136			
C0332257	Including (qualifier)	132	C1299581	Able (finding)	97
C0262926	Medical History	130	C0686904	Patient need for (contextual qualifier)	92

The final input feature vector included 652 concepts and 3 sociodemographic features. There was an average of 108 concepts per record, the minimum was 21 concepts, and maximum was 335 concepts (Figure 2). We extracted a number of concepts from the notes that represents the full range of MMSE score. Figure 3 demonstrates that having a higher or lower MMSE score is not associated with having significantly more or fewer concepts extracted from the note, which may further indicate the existence of different weights among the concepts. Figure 4 shows the distribution of filtered concepts across the 150 notes. Filtered concepts appear an average of 20 times and at most appeared 145 times.

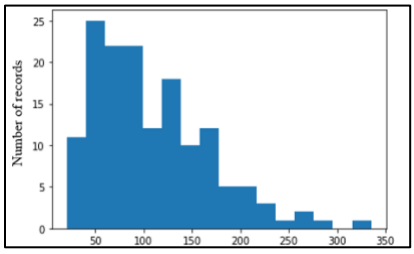

Figure 2. Number of concepts per note (n=150)

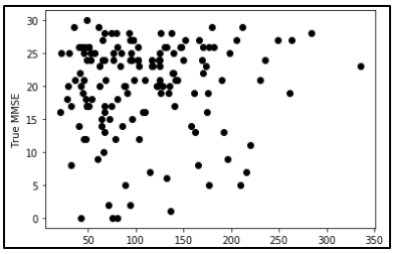

Figure 3. Count of concepts by MMSE among notes (n = 150)

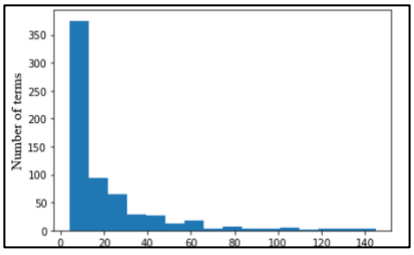

Figure 4. Count of filtered concepts appear in the notes (n=150)

Exploration with dimensionality reduction

We intended to use the concepts extracted from the regression with Lasso modeling to generate the novel taxonomy through unsupervised clustering methods. Before doing so, we inspected potential patterns that may structurally exist in the data by mapping the data from a high-dimensional space to a two-dimensional space with points colored by

outcomes of interest (severity, study eligibility, provider, and language) using Uniform Manifold Approximation and Projection (UMAP)[24]. The dimensionality reduction embedding of the feature vector returned from the Lasso (i.e., the final model parameters) is not useful in discriminating between any of the useful outcomes of interest specific to clinical trial eligibility for a particular protocol (Figure 5). In fact, the plots that show the dimensionality reduction embedding using just the sparse feature vector returned from the model did not show any discernable clustering at all.

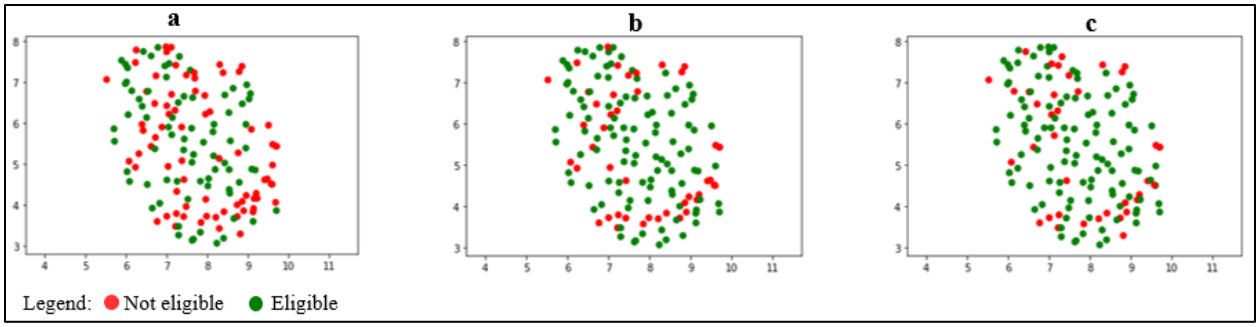

Figure 5. UMAP Projection of all feature concepts returned from the Lasso by clinical trial protocol eligibility: (a) phase 1 study (NCT03822208); (b) phase 2 study (NCT03282916); and, (c) phase 3 study (NCT03887455).

To explore further, embeddings were generated and visualized for not only the feature vector returned from the model, but also the vector of all relevant features and the input feature vector from the modeling step (Figure 6). A distinct cluster is noticeable when all features are included, and this is still the case for the embedding using only the features that go into the Lasso regression (i.e., after filtering based on semantic type, the cluster still emerged in the embedding). However, this distinct clustering is gone when only features from the final regression with Lasso model were included, suggesting that our model was able to smooth out this underlying structure (and remove concepts that might confound the results).

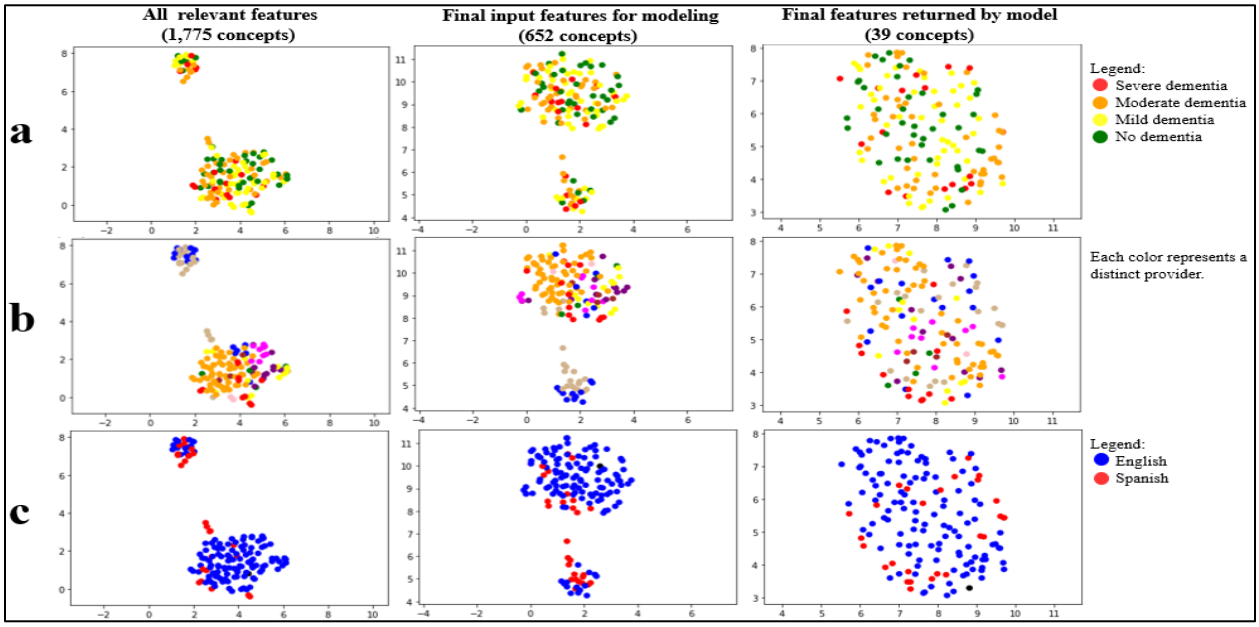

Figure 6. UMAP Projection of feature concepts by (a) dementia severity, (b) provider, and (c) language of MMSE administration.

After further investigation, the observed difference probably relates to the difference in provider (Figure 6b). The embedding plots suggest that the blue and the tan providers are outliers; it turns out that these two providers work together in one of the AD centers in the department where they see many of the Spanish-speaking patients (Figure 6c). These differences in concepts extracted by provider were not clear upon manual review of the charts; in fact, the charts seemed on the surface to resemble all of the others. When the two providers (who work together) were removed from the dataset, the natural clustering in the dimensionality reduction disappears. However, we decided to keep the two providers in the final analysis because this better represents real-world practice.

Evaluation metrics

The plots below (Figure 7) show the true MMSE score vs the MMSE score predicted by the model, for both training and test data. A perfectly accurate model would show dots along the red reference line y=x. Indeed, at higher MMSE scores, the model is fairly accurate and clustered around this reference line. At lower MMSE scores, the model is much further off with the prediction. Considering the R-squared value, the final model explains 60.1% of the variance in the training data and 31.8% of the variance in the test data.

Figure 8 shows the difference between the predicted and true MMSE score, and we again see that lower true MMSE scores were not predicted very well by the model for either the training or test data, but higher scores were not as far off. If the scores were fully accurate, the dots would line up on the red y=0 reference line.

It is to be expected that for both training and test data, lower scores would be predicted higher and higher scores would be predicted lower when compared to the ground truth. The training data has a Root Mean Square Error (RMSE) of 4.6 and the test data returns an RMSE of 5.0. While still a wide margin, a predicted score that is about 5 points off in either direction would still be helpful in prescreening records to accomplish the clinical task.

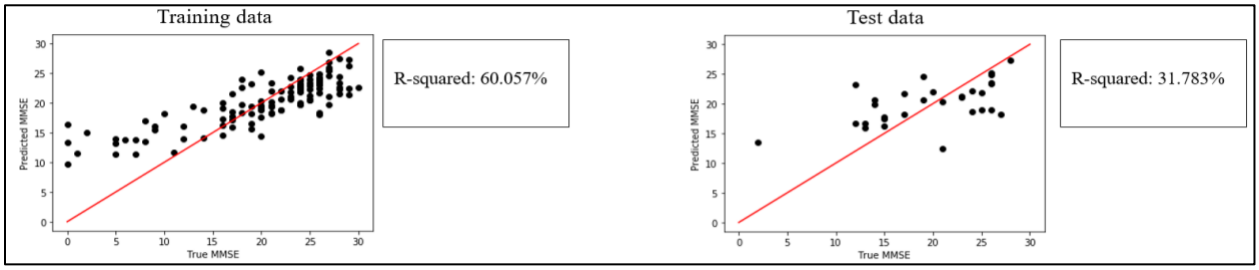

Figure 7. True and predicted MMSE score, with Regression + Lasso

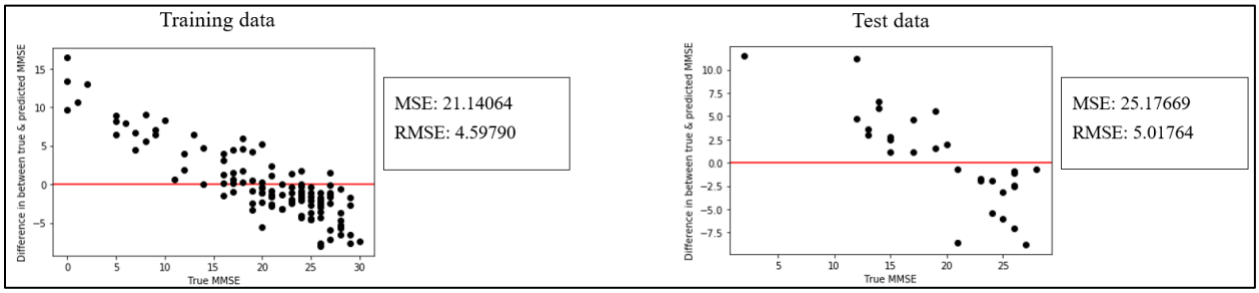

Figure 8. True MMSE score and the difference between true and predicted MMSE score. (MSE: Mean Squared Error; RMSE: Root Mean Square Error)

Final model concepts and novel taxonomy

We constructed a novel taxonomy (Figure 9) by using iterative card sorting of the concepts identified in our model (Table 4) and evaluated its face validity among clinicians in aging and dementia practice . The taxonomy includes concepts within the domains of the MMSE (i.e., orientation, concentration, working memory, memory recall, language, and visuospatial)[5], and concepts representing domains outside of MMSE that are incredibly important and relevant for determining cognitive status functioning decline in AD such as agitation and home environment. The final classes include memory and cognition, activities of daily living, mood and behavior, medical, and descriptors.

The negative coefficient means that if a concept is present, the MMSE score is lower, and a larger magnitude is associated with a larger drop in MMSE score. In our quality check, the following selected concepts were substantiated by an excerpt from clinical notes: concept "Usually Need Help from Another Person for Eating" from excerpt "*would not let him (son) feed her*"; concept "Unable to Feed Self" from "*discussed feeding tube and palliative care*"; concept "Has difficulty doing" from "*recently having difficulty with calculation*"; and concept "Better than Others" from "*he thinks his memory is better than his peers"*. Further, the standardized age is associated with lower MMSE, which is expected[25]. There are two concepts in the descriptors category that fall under "Spatial Concept" sematic type (i.e., local and adjacent), two concepts under the "Finding" semantic type (i.e., unable and better than others), one concept under the "Therapeutic or Preventive Procedure" semantic type (i.e., change – procedure), and the rest under the "Functional Concept" semantic type.

Further analysis of the data via visualization revealed that the use of concepts, both present and negated, may vary across the levels of cognitive impairment (Figure 10). Concepts pertaining to activities of daily living (e.g., grooming)

were used more in clinical text of patients with severe cognitive impairment. Interestingly, "restlessness" was used across the level of cognitive impairment but increasingly so as the cognitive impairment progresses. Further, the negated concepts were mostly observed in the clinical texts of patients with no to mild cognitive impairment.

Table 4. Final model (output features and weights) used to generate subspecialized taxonomy for cognitive status.

Co-efficient	UMLS CUI	Times used	Concept	Co-efficient	UMLS CUI	Times used	Concept
-2.724	C4318483	19	Usually need help from another person for eating	-0.185	C0018524	26	Hallucinations
-2.270	C4050223	74	Progressive metabolic disease	-0.146	C1299586	65	Has difficulty doing (qualifier value)
-1.696	C0332287	12	In addition to	-0.115	C0686904	59	Patient need for (contextual qualifier)
-1.520	C0442519	31	Home environment	-0.083	C1272688	10	Status – in progress
-1.514	C1522577	103	follow-up	-0.029	C0205276	29	Local
-1.461	C0278329	37	Prescribed	-0.028	C0879626	73	Adverse effects
-1.296	C0442805	35	Increase	-0.005	C4319952	38	Change -- procedure
-1.271	C0596893	21	Medical appointment	0.000	C0037088	19	Signs and Symptoms
-1.106	C3887611	103	Restlessness	0.002	C1705914	6	Repeat object
-0.804	C0018249	48	Grooming	0.004	C0871703	27	Conversation
-0.648	C0040822	45	Tremor	0.023	C1524063	23	Use of
-0.618	C0566415	28	Unable to feed self	0.467	C0457083	91	Usage
-0.614	C0085631	17	Agitation	0.770	C4552810	29	Irritability, CTCAE
-0.600	C0557034	23	Assisting (procedure)	0.811	C0034770	17	Mental recall
-0.572	C0332148	14	Probable diagnosis	0.957	C4553821	25	Feels
-0.571	C0150369	10	Preventive monitoring	1.198	C4522046	18	Better than others
-0.51	C1299582	29	Unable	1.386	C0205117	21	Adjacent
-0.341	C0162340	18	Comprehension	2.149	C0013126	48	Intrinsic drive
-0.268	C2825141	64	Experimental finding	2.749	C1270972	41	Mild cognitive disorder
-0.220	C0589120	43	Follow-up status				

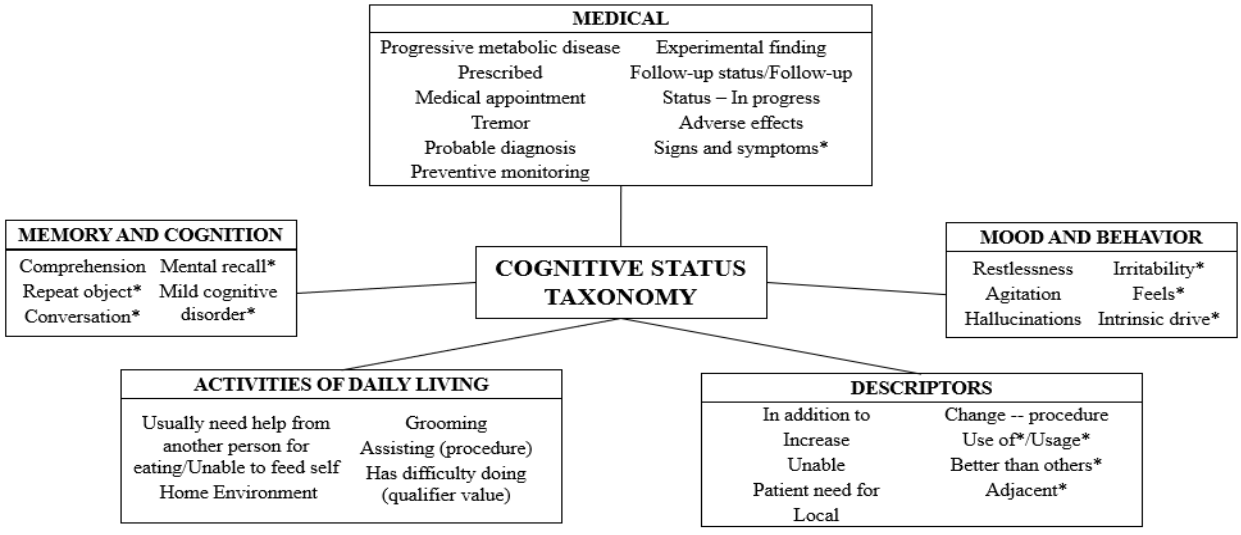

Figure 9. Subspecialized taxonomy for characterizing cognitive status of patients with aMCI or AD.

Discussion

The current study demonstrates that narrative text from outpatient clinical visit notes of patients diagnosed with aMCI and AD could be instrumental in indicating the level of cognitive impairment for AD patients. Consistent with the literature, we used a data-driven approach to derive five categories of concepts corresponding to MMSE scores: cognitive[26], functional[27], behavioral[8], medical[12], and descriptors. These findings provide an important, albeit preliminary, foundation for informing expansions of existing concepts related to a specific proxy (i.e., MMSE score).

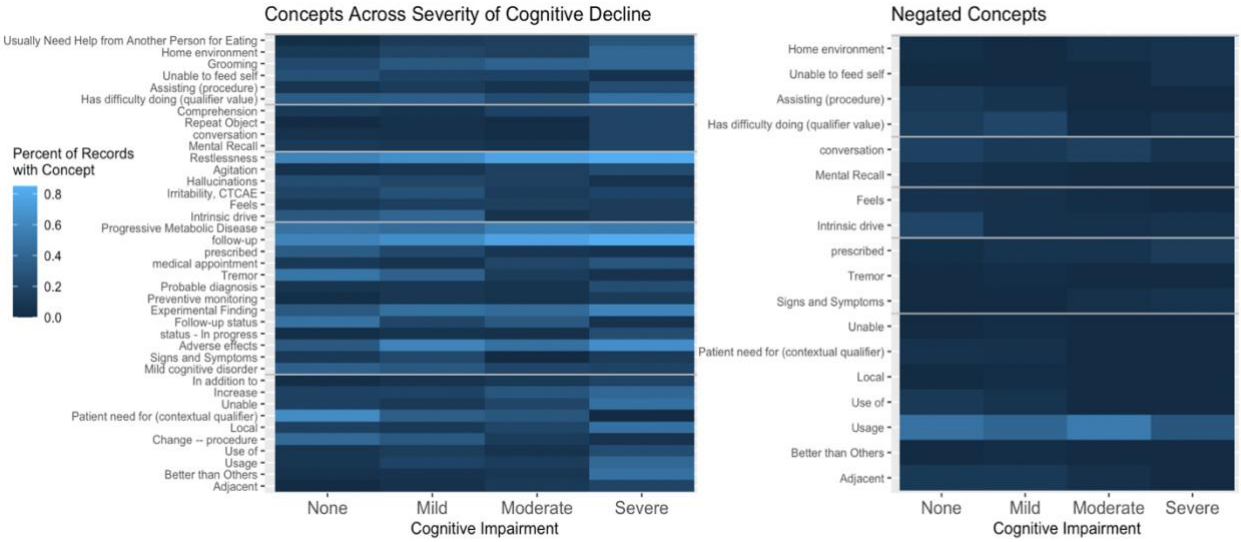

Figure 10. Percent of clinical text with concepts across the level of cognitive impairment, by taxonomy category.

Our findings regarding the content and classification of cognitive impairment concepts provide a preliminary understanding of potential challenges for using EHR notes in automated prescreening approaches. First, there are concepts that are frequently used with different meanings in different context. For example, the term "local" is used to describe the proximity of clinical care (e.g., local neurologist) or a symptom (e.g., local tenderness). Examining documentation patterns across varying points of care (e.g., follow-up visit) and expanding the analysis of terms to other neighboring terms may provide additional information for the disambiguation of these concepts. For example, the term "feed" may not refer to the patient's ability to feed oneself but of forgetting to feed a pet, which is still critical in the classifying the impairment as memory and cognition or activities of daily living but also different than other uses. Further, the frequency of the terms used vary across the level of cognitive impairment progresses. For example, terms related to the concepts "grooming" and "restlessness" are more frequently used in clinical notes of patients with severe cognitive dementia; conversely, the use of terms related to the concepts of "hallucinations" and "tremor" decreased from normal cognition to sever dementia.

Provided the high prevalence, cost, and mortality associated with AD[1]—and the urgent goal of the National Alzheimer's Project Act's (Public Law 111-375) National Plan to Address Alzheimer's Disease to prevent and effectively treat AD by 2025[2]—meeting recruitment goals for AD clinical trials have important scientific, clinical, financial, ethical, and policy implications[28]. Study findings have the potential to improve recruitment rates for AD clinical research and subsequently accelerate further development of an efficacious disease-modifying treatment for AD. Availability of a specialized taxonomy commonly used in AD clinical care documentation has the potential to bolster eligibility prescreening approaches for clinical trial recruitment. Future work to further develop and refine the model across broader range of patients, providers, and institutions should focus on expanding the model's capacity to utilize clinical text with symbolic knowledge representation such as building into the model ways to include and infer from relevant broader and narrower concepts. For example, the concept "hallucinations" has "behavioral/psychiatric manifestations" as a parent concept and "sensory manifestations" as a grandparent concept, which are useful to interrogate for future inclusion in the model; on the other hand, "substance withdrawal severity" and "hypocalcemia severity" would not be relevant; finally, "general symptom" may be useful but is very broad. The present study derives potential concepts for inclusion in future ontologies and phenotypes, which can serve as the foundation of developing clinical decision support for clinicians and research teams to identify cohorts for AD clinical research and focus their time and effort into other aspects of research and clinical care such as recruitment and patient education[29]. Future work should examine the predictive accuracy of the terms used in clinical text and how these maps to UMLS concepts in determining AD clinical trial eligibility across different patient samples and EHRs. It would also be interesting to develop a longitudinal account of disease with the approach to computational phenotyping described in this study.

Limitations

The study has some limitations. The first limitation is the use of retrospective data of patients with two specific diagnoses (aMCI and AD) documented in EHR notes by ten specialists in a single subspecialized clinical practice in a quaternary academic medical center. This design strengthens internal validity, which is important in this early stage.

Moreover, there may be local variation in the terminology used within this single clinical practice, and these findings may not be generalizable to other settings or other EHR systems. As there is no mandate for MMSE testing at any particular visit, each provider perceives and documents different information (particularly across provider type), including the MMSE score; some providers may skew towards new patients, initial visits notes may be written by a clinical fellow or nurse practitioner under the provider's supervision, time spent with patients may vary widely, and the clinic was newly conducting telehealth visits due to the COVID-19 pandemic. Future research evaluating different data sources and settings is needed to understand whether and how documentation patterns specifying cognitive function vary across services and EHRs. Additionally, while the present study included important sociodemographic factors such as age and language, other relevant factors such as education and date of the previous MMSE were not included due to dispersed EHR phenotypes and fragmented EHR data[8]. As such, future research should investigate these factors using a much larger and more representative corpus including various clinical texts including discharge summaries and clinical notes.

As this study explored concepts in the clinical visit notes of patients diagnosed with aMCI or AD, these concepts may only represent concepts that clinicians who subspecialize in aging and dementia are more accustomed to using. Future research is needed to determine whether the concepts in this taxonomy are consistently used by clinicians across a different range of specialties (e.g., general neurology, primary care) and diagnoses. We could also expand the data corpus to include notes from general medicine and could consider using pretrained embedding models as an alternate approach. The present study focused on identifying terms used by clinicians to describe cognitive symptoms based on MMSE. However, other comorbidities such as stroke, epilepsy, or other neurodegenerative diseases may also affect cognitive symptoms in addition to AD. Future prospective research should aim to identify terminology that may be unique for AD or AD alongside other comorbidities. Further, we did not look at the presence or absence of AD biomarkers result to further confirm the probable diagnosis, and the medications that the patients were taking during the MMSE testing, which may affect the patient's performance. Finally, as this study represents an initial derivation of a subspecialized taxonomy from a gold-standard diagnostic group only, although the whole MMSE range was covered, it did not include matched controls. Further development and refinement of subspecialized taxonomy to characterize cognitive function will benefit from identification of common data features in the EHRs of patients with subjective cognitive complaints, which was not addressed in this study.

Conclusions

In this study, we introduced a subspecialized taxonomy based on concepts in clinical text to assist in characterizing cognitive impairment without using MMSE scores. Our work demonstrates the feasibility of subgrouping of patients using their EHR notes even when MMSE scores may not be directly available. By leveraging clinical narrative notes, a proxy of cognitive impairment was constructed using symbolic knowledge representation and computational modeling. This method could improve the efficiency and accuracy of cohort identification based on cognitive function for AD clinical research regardless of the presence of a recent MMSE score in the patient chart. We conclude that utilization of specialized taxonomy is a suitable approach to extract concepts from clinical notes and this approach may be more portable and generalizable than a purely computational approach and could possibly be used to infer the severity of cognitive impairment and provide interesting clinical insights. Future work is warranted to test how this approach may generalize to other domains for developing proxies for clinically relevant indicators or formal scores using narrative clinical notes and symbolic methods.

Acknowledgments

Research reported here was supported by the National Library of Medicine grants R01LM009886 (PI: Weng), 5T15LM007079 (PI: Hripcsak), the National Institute of Nursing Research grants T32 NR007969 (PI: Bakken) and K24NR018621 (PI: Schnall). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health. We thank Oliver Bear Don't Walk IV, Harry Reyes Nieva, and Michael Zietz for helping in debugging and improving our modeling pipeline; Wendy Gonzalez and Arlene Mejia for the face validation of the taxonomy; and Dr. Noémie Elhadad for her insightful comments on the manuscript.

References

1. 2020 Alzheimer's disease facts and figures. *Alzheimer's & Dementia.* 2020;16(3):391-460.
2. *National Plan to Address Alzheimer's Disease: 2019 Update.* U.S. Department of Health and Human Services;2019.
3. FDA/CEDR resources page. F ood and Drug Administration Web site. https://www.fda.gov/drugs/postmarket-drug-safety-information-patients-and-providers/aducanumab-marketed-aduhelm-information. Accessed June 15, 2021.

4. Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K. Alzheimer's disease drug development pipeline: 2019. *Alzheimer's & Dementia.* 2019;5:272-293.5. Kennedy RE, Cutter GR, Wang G, Schneider LS. Using baseline cognitive severity for enriching Alzheimer's disease clinical trials: How does Mini-Mental State Examination predict rate of change? *Alzheimer's & Dementia: Translational Research & Clinical Interventions.* 2015;1(1):46-52.
6. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12(3):189-198.
7. Crum RM, Anthony JC, Bassett SS, Folstein MF. Population-based norms for the Mini-Mental State Examination by age and educational level. *JAMA.* 1993;269(18):2386-2391.
8. 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.
9. Butler A, Wei W, Yuan C, Kang T, Si Y, Weng C. The Data Gap in the EHR for Clinical Research Eligibility Screening. *AMIA Jt Summits Transl Sci Proc.* 2018;2017:320-329.
10. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. *International Journal of Medical Informatics.* 2011;80(6):371-388.
11. Gilmore-Bykovskyi AL, Block LM, Walljasper L, Hill N, Gleason C, Shah MN. Unstructured clinical documentation reflecting cognitive and behavioral dysfunction: toward an EHR-based phenotype for cognitive impairment. *Journal of the American Medical Informatics Association.* 2018;25(9):1206-1212.
12. Alexander N, Alexander DC, Barkhof F, Denaxas S. Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer's Disease in Electronic Health Records. *Stud Health Technol Inform.* 2020;270:499-503.
13. Amra S, O'Horo JC, Singh TD, et al. Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records. *Journal of Critical Care.* 2017;37:202-205.
14. Reuben DB, Hackbarth AS, Wenger NS, Tan ZS, Jennings LA. An Automated Approach to Identifying Patients with Dementia Using Electronic Medical Records. *J Am Geriatr Soc.* 2017;65(3):658-659.
15. Malhotra A, Younesi E, Gündel M, Müller B, Heneka MT, Hofmann-Apitius M. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimers Dement.* 2014;10(2):238-246.
16. Henry V, Moszer I, Dameron O, Potier M-C, Hofmann-Apitius M, Colliot O. Converting Alzheimer's Disease Map into a Heavyweight Ontology: A Formal Network to Integrate Data. Paper presented at: Data Integration in the Life Sciences; 2019//, 2019; Cham.
17. Zekri F, Bouaziz R, Turki E. A fuzzy-based ontology for Alzheimer's disease decision support. Paper presented at: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); 2-5 Aug. 2015, 2015.
18. Refolo LM, Snyder H, Liggins C, et al. Common Alzheimer's Disease Research Ontology: National Institute on Aging and Alzheimer's Association collaborative project. *Alzheimers Dement.* 2012;8(4):372-375.
19. Ciccarese P, Wu E, Wong G, et al. The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics.* 2008;41(5):739-751.
20. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: Demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med.* 2016;71:57-61.
21. *pymetamap* [computer program]. GitHub Repository: GitHub; 2004.
22. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met.* 1996;58(1):267-288.
23. Righi C, James J, Beasley M, et al. Card sort analysis best practices. *J Usability Studies.* 2013;8(3):69–89.
24. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020.
25. Bravo G, Hébert R. Age- and education-specific reference values for the Mini-Mental and modified Mini-Mental State Examinations derived from a non-demented elderly population. *Int J Geriatr Psychiatry.* 1997;12(10):1008-1018.
26. Guerrero-Berroa E, Luo X, Schmeidler J, et al. The MMSE orientation for time domain is a strong predictor of subsequent cognitive decline in the elderly. *Int J Geriatr Psychiatry.* 2009;24(12):1429-1437.
27. Clark CM, Sheppard L, Fillenbaum GG, et al. Variability in annual Mini-Mental State Examination score in patients with probable Alzheimer disease: a clinical perspective of data from the Consortium to Establish a Registry for Alzheimer's Disease. *Arch Neurol.* 1999;56(7):857-862.
28. Huang GD, Bull J, Johnston McKee K, Mahon E, Harper B, Roberts JN. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. *Contemp Clin Trials.* 2018;66:74-79.
29. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort required in eligibility screening for clinical trials. *J Oncol Pract.* 2012;8(6):365-370.