Received: 28 September 2021 | Revised: 10 January 2022 | Accepted: 18 January 2022 | Published online: 23 February 2022

DOI: 10.1002/dad2.12294

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Automated semantic relevance as an indicator of cognitive decline: Out-of-sample validation on a large-scale longitudinal dataset

Gabriela Stegmann[1,2] | Shira Hahn[1,2] | Samarth Bhandari[2] | Kan Kawabata[2] | Jeremy Shefner[3] | Cayla Jessica Duncan[3] | Julie Liss[1,2] | Visar Berisha[1,2] | Kimberly Mueller[4,5,6]

[1] Arizona State University, Phoenix, Arizona, USA

[2] Aural Analytics, Scottsdale, Arizona, USA

[3] Barrow Neurological Institute, Phoenix, Arizona, USA

[4] Wisconsin Alzheimer's Disease Research Center, University of Wisconsin–Madison School of Medicine and Public Health, Madison, Wisconsin, USA

[5] Department of Communication Sciences and Disorders, University of Wisconsin–Madison, Madison, Wisconsin, USA

[6] Wisconsin Alzheimer's InstituteUniversity of Wisconsin–Madison School of Medicine and Public Health, Madison, Wisconsin, USA

Correspondence
Gabriela Stegmann, Aural Analytics, 1355 N Scottsdale Rd UNIT 110, Scottsdale, AZ 85257, USA.
E-mail: Gabriela.Stegmann@asu.edu

Julie Liss, Visar Berisha, and Kimberly Mueller share senior authorship.

Abstract

We developed and evaluated an automatically extracted measure of cognition (semantic relevance) using automated and manual transcripts of audio recordings from healthy and cognitively impaired participants describing the Cookie Theft picture from the Boston Diagnostic Aphasia Examination. We describe the rationale and metric validation. We developed the measure on one dataset and evaluated it on a large database (>2000 samples) by comparing accuracy against a manually calculated metric and evaluating its clinical relevance. The fully automated measure was accurate (r = .84), had moderate to good reliability (intra-class correlation = .73), correlated with Mini-Mental State Examination and improved the fit in the context of other automatic language features (r = .65), and longitudinally declined with age and level of cognitive impairment. This study demonstrates the use of a rigorous analytical and clinical framework for validating automatic measures of speech, and applied it to a measure that is accurate and clinically relevant.

KEYWORDS
algorithm, automatic, cognition, digital, language, longitudinal, speech

1 | OVERVIEW

The power of language analysis to reveal early and subtle changes in cognitive–linguistic function has been long recognized[1,2,3] but challenging to implement clinically or at scale because of the time and human resources required to obtain robust language metrics. This is particularly true of picture description tasks, which are regarded as rich data sources because they require a broad range of cognitive and linguistic competencies to successfully complete.[4] For example, the Boston Diagnostic Aphasia Examination (BDAE) includes an elicitation of the Cookie Theft picture description[5] and this task is widely used clinically and in research across a swath of clinical conditions, including cognitive decline and dementia.[6,7] The information extracted from transcripts of the picture descriptions provides insight to the likely sources of deficit and differential diagnosis. Yet, the burden of the analyses on human resources is prohibitively high for routine clinical use and impedes rapid dissemination of research findings. In this study, we demonstrate the feasibility of using an automated algorithm to measure an interpretable and clinically relevant language feature extracted from picture descriptions while dramatically reducing the human burden of manually assigning codes to features of interest.

A commonly extracted, high-yield metric for the characterization of cognitive–linguistic function in the context of dementia involves assessment of the relationship of the words in the transcribed picture description to the word targets in the picture. This measure has been described with varying terminology, including "correct information units,"[8] "content information units,"[9] and "semantic unit idea density." [1,10] All these terms encapsulate essentially the same concept: the ratio of a pre-identified set of relevant content words to the total words spoken. For example, in the Cookie Theft picture description, people are expected to use the words "cookie," "boy," "stealing," etc., corresponding to the salient aspects of the picture. We developed an automated algorithm to measure this relationship, called the Semantic Relevance (SemR) of participant speech. We chose to use this new term, "semantic relevance," to better frame the concept of the measure. SemR measures the proportion of the spoken words that are directly related to the content of the picture, calculated as a ratio of related words to total words spoken. Like its manual predecessor, "semantic unit idea density,"[1] the automated SemR metric provides an objective measure of the efficiency, accuracy, and completeness of a picture description relative to the target picture.

The goal of this study is two-fold. First, we completely automated the process for measuring SemR by transcribing recordings of picture descriptions using automatic speech recognition (ASR) and algorithmically computing SemR; done manually, this is a burdensome task and prohibitive at a large scale. Second, we use this study to illustrate a rigorous analytical and clinical validation[11] framework in which we evaluated SemR on a new, large (>2000 observations) database.

We first show that the SemR scores remain accurate at each step that the measure is automated. Next, we discuss our use of a large evaluation sample to show the accuracy achieved after automating the computation of SemR. We first show the accuracy achieved when the content units for calculating SemR are identified algorithmically rather than through manual coding. Second, we show the accuracy achieved when the transcripts are obtained through ASR instead of manually transcribing them. We then evaluated what happens when the data collection is done remotely and without clinical supervision. To do this, we compared the SemR scores between participants who provided picture descriptions in clinic supervised by a clinician and at home in an unsupervised setting. In the second part of the study, we demonstrate the relationship between SemR and cognitive function. We used the fully automated version of SemR and evaluated it for its clinical relevance computing its test–retest reliability, its association with cognitive function, its contribution to cognitive function above and beyond other automatically obtained measures of language production, and its longitudinal change for participants with different levels of cognitive impairment.

RESEARCH IN CONTEXT

1. **Systematic Review**: The authors conducted a literature review to identify studies that make use of content information units (CIU) extracted from picture description tasks as a predictor of cognitive impairment. Several articles and abstracts discussed changes in CIUs and other language parameters that reflect individuals' cognitive functioning.
2. **Interpretation**: Our study shows how a targeted language feature was developed based on previous studies and fully automated such that it can be algorithmically extracted from the Cookie Theft picture description recorded speech. We evaluated it on a large database (>2000 samples) and showed that it is clinically relevant, repeatable, and tracks changes in cognition.
3. **Future Directions**: The article illustrates a framework for rigorous validation of digitally extracted language features. This framework can be used in future work to validate other speech-based measures of cognition.

2 | METHODS

2.1 | Development dataset

We used a small dataset (25 participants, 584 descriptions of pictures) for developing the SemR algorithm. These participants had amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) of varying degrees of severity. The inclusion of participants with unimpaired speech along with speech impacted by dysarthria and cognitive impairment for developing the algorithm provided us with a rich dataset with samples that varied in the picture descriptions' length and content. Details are found in the supporting information. This dataset was used for the development of the algorithm, but not the clinical validation, and therefore this study does not claim that the clinical validation results generalize to ALS or FTD.

2.2 | Evaluation dataset

The sources of the evaluation data included the Wisconsin Registry for Alzheimer's Prevention (WRAP) study, DementiaBank,[12] and Amazon's Mechanical Turk. WRAP and DementiaBank conducted the data collection in clinic with supervision from a clinician, and were evaluated for their degree of cognitive impairment. The data collection through Mechanical Turk was conducted remotely; participants self-selected to participate in an online "speech task" study from their computers and were guided through the study via a computer application.

At each data collection, recorded descriptions of the Cookie Theft picture were obtained. The sample consisted of various

TABLE 1 Description of the evaluation sample

	Evaluation data			
Demographic	**CU**	**CU-D**	**MCI**	**Dementia**
Age mean (SD)	58.5 (10.5)	63.6 (6.0)	66.7 (6.4)	71.2 (8.6)
Sex (% female)	58% F	61% F	73% F	65% F
Race (%White)	93% W	84% W	78% W	97% W
Education (% less than high school, % completed high school, % more than high school)	1% < HS, 16% HS, 83% > HS	2% < HS, 10% HS, 88% > HS	12% < HS, 15% HS, 73% > HS	33% < HS, 31% HS, 38% > HS
Number of observations	2,610	327	64	311
Number of participants	1258	180	26	195

Abbreviations: CU, cognitively unimpaired; CU-D, cognitive unimpaired showing atypical decline; HS, high school; MCI, mild cognitive impairment; SD, standard deviation.

TABLE 2 Number of observations for each sample characteristic

Sample characteristics	**Number of observations**
Speech was manually transcribed	2716
Manual transcription was manually annotated to manually calculate SemR	2163
Speech was transcribed using ASR	2921
Speech was collected in clinic	2716
Speech was collected remotely	595
Speech sample was collected with paired MMSE	2564
Speech was collected in close temporal proximity (separated by ≈1 week)	319

Abbreviations: ASR, automatic speech recognition; MMSE, Mini-Mental State Examination; SemR, semantic relevance.

characteristics, including participants who provided repeated measurements over the course of years, participants who completed a paired Mini-Mental State Examination (MMSE),[13] participants who provided the picture descriptions in clinic supervised by a clinician, and participants who provided the picture descriptions from home. Additionally, the sample included transcripts that were manually transcribed, transcripts transcribed by ASR, and transcripts that were manually annotated by trained annotators to compute SemR. The WRAP participants were diagnosed according to a consensus conference review process as being cognitively unimpaired and stable over time (CU), cognitively unimpaired but showing atypical decline over time (CU-D), clinical mild cognitive impairment (MCI), and dementia (D). The DementiaBank[12] participants were described as healthy controls (coded here as CU) and as participants with dementia. Mechanical Turk participants self-reported no cognitive impairment (CU), absent clinical confirmation. Table 1 shows descriptive statistics of the sample for each diagnostic group. Table 2 shows the number of samples available for each type of data, for a total of 552 (DementiaBank), 2186 (WRAP), and 595 (Mechanical Turk).

In the following sections, unless otherwise specified, each analysis used all the data that was available given the required characteristics (e.g., when estimating the accuracy of the automatically computed SemR with the manually annotated SemR, all observations for which both sets of SemR scores were available were used for the analysis).

2.3 | Development of semantic relevance

We focused efforts on automation of the SemR measure because of the demonstrated clinical utility of picture description analysis, as well as its ability to provide insight into the nature of different deficit patterns and differential diagnosis.[1,3] The goal of the SemR measure is to gauge retrieval abilities, ability to follow directions, and ability to stay on task in a goal-directed spontaneous speech task. We used the complex picture description task from the BDAE,[5] in which participants were shown a picture of a complex scene and were asked to describe it. SemR is higher when the picture description captures the content of the picture and is lower when the picture description shows signs of word finding difficulty, repetitive content, and overall lack of speech efficiency. In other words, SemR measures the proportion of the picture description that directly relates to the picture's content.

The algorithm operates as follows: First, the speech is transcribed. Then, each word is categorized according to whether it is an element from the picture or not. For this, the algorithm requires a set of inputs that indicate what elements from the picture need to be identified. For the Cookie Theft picture, we chose the 23 elements indicated in Ahmed et al.[10] (e.g., boy, kitchen, cookie) and allowed the algorithm to accept synonyms (e.g., "young man" instead of "boy"). Finally, the total number of unique elements from the picture that a participant identifies is annotated and divided by the total number of words that the participant produced. Importantly, these keywords were fixed after development and were not modified during evaluation.

The supporting information contains an illustration of how SemR provides a window into speech production in cognitive impairment.

2.4 | ASR transcription

Google Cloud's[14] Speech-to-Text software transcribed the speech samples. The ASR algorithm was customized for the task by boosting the standard algorithm such that the words that are expected in the transcript have increased probability that they would be correctly recognized and transcribed. This was implemented in Python using Google's Python application programming interface.[14]

2.5 | Data analysis

The data analysis is split into three sections to evaluate: (1) accuracy of the automatic algorithm, (2) sensitivity of SemR to the administration method, and (3) clinical utility of SemR by measuring differences in SemR scores across levels of cognitive impairment, and within-participant longitudinal change.

2.5.1 | Evaluation of semantic relevance: removing the human from the SemR computation

In the manual implementation of SemR there are two steps that involve human intervention, including manually transcribing the participant's recorded picture description then manually annotating the content units mentioned. To establish the analytical validity of the automated SemR, we tested replacement of human intervention in two ways. First, we used manual transcriptions to compare performance of the manually annotated SemR to the algorithmically computed SemR. Second, we used ASR-generated transcripts to compare the automatically computed SemR scores with the manually transcribed and annotated SemR and manually transcribed and automatically computed SemR scores. The goal of this series of analyses was to show that the automated accuracy was maintained relative to ground truth (human intervention) at each step of transcription and calculation of SemR.

To measure the accuracy achieved at each step, we computed the correlation between each pair (using a mixed-effects model[15] given the repeated measurements per participant) and the mean absolute error (MAE) of the two.

2.5.2 | Evaluation of semantic relevance: removing the human from the data collection

Next, we evaluated the feasibility of automating the data collection to be done remotely, without supervision, instead of in clinic and supervised. We selected a sample of 150 participants matched on age and sex, half of whom provided data in clinic (WRAP, DementiaBank) and half at home (Mechanical Turk). We selected only in-clinic participants who were deemed CU by a clinician, and at-home participants who denied cognitive impairment. The final sample for this analysis consisted of 75 participants in clinic and 75 participants at home with average age 62 (standard deviation = 8.0) years old and with 42 women and 33 men in each group. A Welch's test (unequal variances) was conducted comparing the mean SemR scores of the two samples.

TABLE 3 Correlations and differences between the manually annotated, manually transcribed algorithmically computed, and ASR-transcribed algorithmically computed SemR values

Analysis	Correlation	MAE
Human-transcript-and-SemR versus		
Human-transcript-automatic-SemR	0.87	0.04
Human-transcript-automatic-SemR versus		
ASR-transcript-automatic-SemR	0.95	0.01
Human-transcript-and-SemR versus		
ASR-transcript-automatic-SemR	0.84	0.03

Abbreviations: ASR, automatic speech recognition; MAE, mean absolute error; SemR, semantic relevance.

2.5.3 | Evaluation of the clinical relevance of SemR

After establishing the accuracy and feasibility of fully automating the data collection and computation of SemR, we generated an ASR transcript and automatically computed SemR for each participant. We evaluated its clinical relevance by: (1) estimating the test–retest reliability using intra-class correlation (ICC), standard error of measurement (SEM), and coefficient of variation (CV); (2) estimating its association with cognitive function and its contribution to cognitive function above and beyond other automatically obtained measures of language production by fitting a model predicting MMSE and by classifying between disease groups (CU vs. the three disease groups); and (3) estimating the longitudinal within-person change of SemR for participants at different levels of cognitive impairment using a growth curve model (GCM). The supporting information provides a detailed description of the statistical analyses performed.

3 | RESULTS

3.1 | Evaluation of semantic relevance: removing the human from the semR computation

For the analytical validation of SemR, we compared the automatic SemR on manual transcripts, SemR calculated based on manual annotations on the manual transcripts, and automatic SemR on ASR transcripts. Figure 1 shows the plot for each comparison and Table 3 shows the correlations and MAE. All three versions of SemR correlated strongly[16] and had a small MAE, indicating that the automatic computation of SemR did not result in a substantial loss of accuracy.

3.2 | Evaluation of semantic relevance: removing the human from the data collection

Next, we evaluated the impact of the data collection method by comparing SemR scores of supervised (in-clinic) and unsupervised

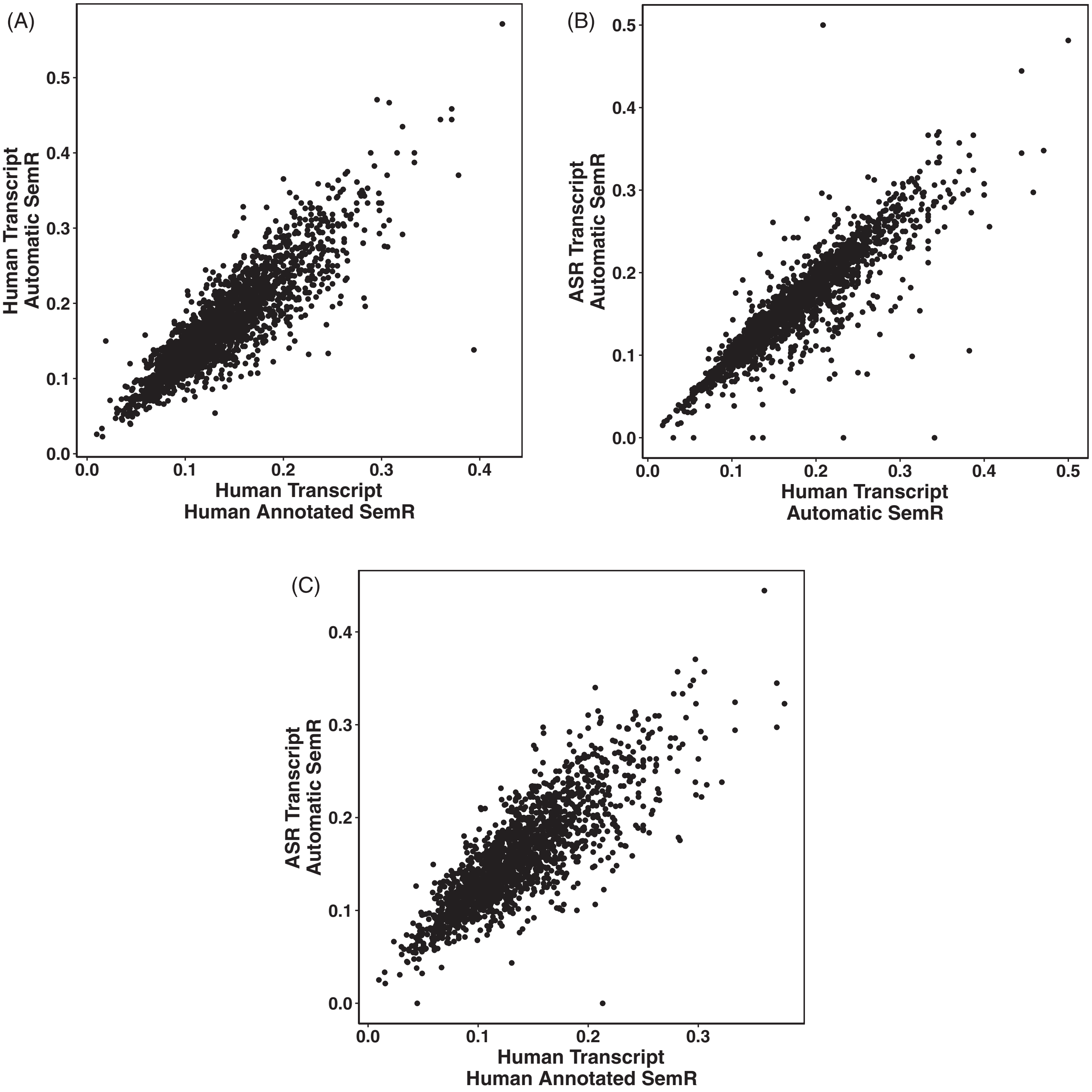

FIGURE 1 Scatterplots showing: (A) the manually annotated SemR values versus manually transcribed algorithmically computed SemR values, (B) manually transcribed algorithmically computed SemR values versus ASR-transcribed algorithmically computed SemR values, and (C) manually annotated SemR values versus ASR-transcribed algorithmically computed SemR values. SemR, semantic relevance

(at-home) participants. A Welch's test indicated that the mean SemR scores were significantly different between the two groups (at home = .21, in clinic = .18, t = 2.55, P = .01, Cohen's d = .43). However, Cohen's d = .43 indicated that the difference between the two groups was small. Figure 2 shows the boxplots with the SemR scores for the at-home and in-clinic samples.

3.3 | Evaluation of the clinical relevance of SemR

3.3.1 | Test–retest reliability

To evaluate the clinical validity of SemR, we first estimated the test–retest reliability. We found that ICC = .73, SEM = .04, CV = 19%.

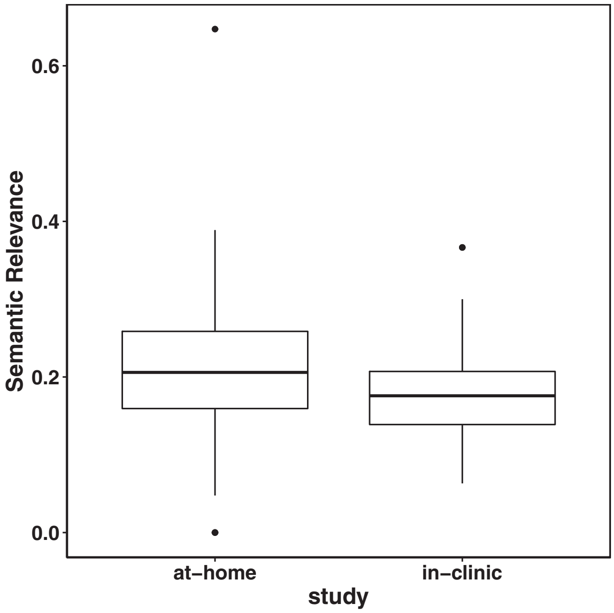

FIGURE 2 Boxplots of SemR scores for at-home (unsupervised) and in-clinic (supervised) samples. SemR, semantic relevance

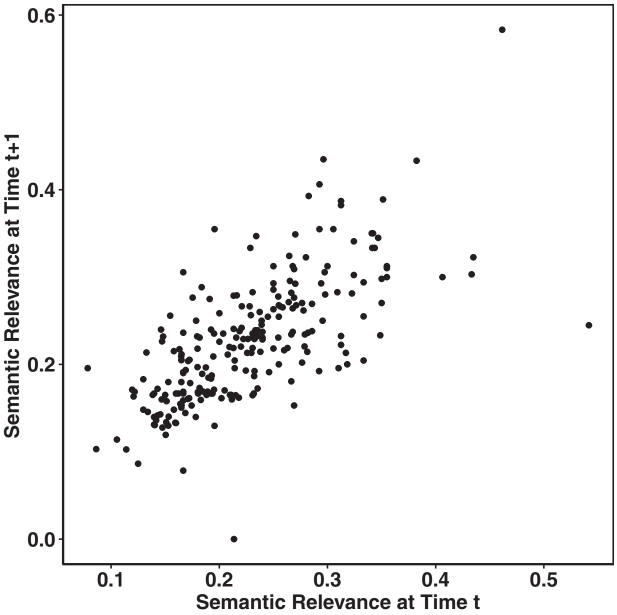

FIGURE 3 Test-retest reliability plot for SemR. SemR, semantic relevance

This was moderate[17] to good[18] reliability, which was considerably higher than most off-the-shelf language features extracted from text.[19] Figure 3 shows the test–retest plot.

3.3.2 | Cross-sectional relationship between SemR and cognitive impairment

We fit a series of models to evaluate how SemR was related to cognitive impairment. The final results showed that when using SemR alone, the correlation between SemR and MMSE was $r = .38$. When using the set of automatically computed language metrics (not including SemR), the correlation between the predicted and observed MMSE (using 10-fold cross-validation) was $r = .38$ with MAE = 4.4. Finally, when using SemR in addition to the set of metrics to predict MMSE, the correlation between the observed and predicted MMSE was $r = .65$ and MAE = 3.5. Finally, we evaluated SemR's ability to classify disease (CUs vs. the three clinical groups) above and beyond the MMSE alone, and found that the area under the curve (AUC) increased from AUC = .78 (MMSE alone) to AUC = .81 (MMSE and SemR). This indicated that SemR offered insight into one's cognition both as a stand-alone measure and above and beyond what was possible through other measures. Figure 4 shows the observed and predicted MMSE scores for the final model.

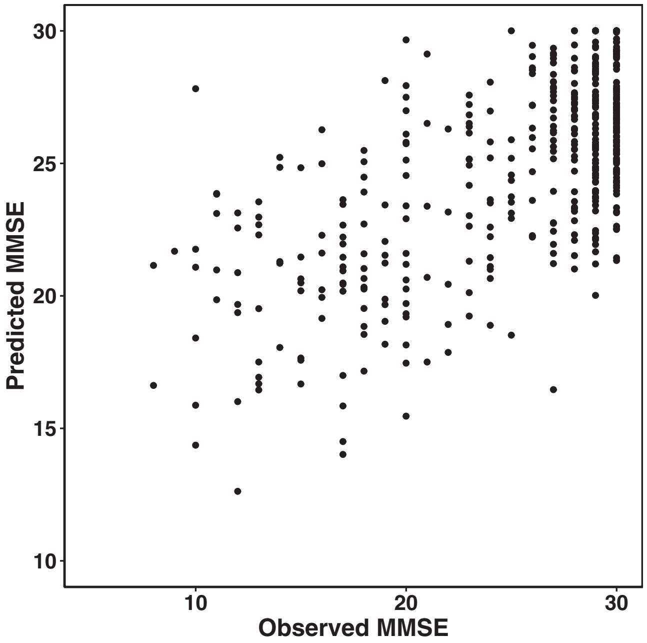

FIGURE 4 Scatterplot showing the predicted and observed Mini-Mental State Examination (MMSE) values

3.3.3 | Longitudinal trajectory of SemR

The longitudinal analyses showed that all groups had declining SemR scores. However, the CUs had slower-declining SemR scores than the impaired groups. Among the impaired groups, the results showed an apparent non-linear decline, in which the scores started at the highest point among the CU-D participants, followed by the MCI participants with intermediate scores and the steepest decline, finally followed by the dementia participants, who had the lowest SemR scores and whose trajectory flattened again. Table 4 shows GCM parameters for the four groups. Figures 5A and B show the expected longitudinal trajectories according to the GCM parameters for the healthy (A) and cognitively impaired (B) groups. Although all data were used for the analyses, for easier visualization of the results in the cognitively impaired groups we

TABLE 4 Parameter estimates for the GCMs for each cognitive group

Parameter	CU estimate (S.E.)	CU-D estimate (S.E.)	MCI estimate (S.E.)	Dementia estimate (S.E.)
Fixed effects				
Intercept (centered at age 65)	0.158 (.002)	0.167 (.004)	.163 (.01)	.132 (.005)
Slope	–.0004 (.0002)	–.0014 (.0006)	–.0026 (.0015)	–.0005 (.0004)
Random effects				
Participant intercepts SD	0.03	0.05	0.03	0.03
Residuals SD	0.04	0.04	0.04	0.05

Abbreviations: CU, cognitively unimpaired; CU-D, cognitive unimpaired showing atypical decline; GCM, growth curve model; MCI, mild cognitive impairment; SD, standard deviation; S.E., standard error.

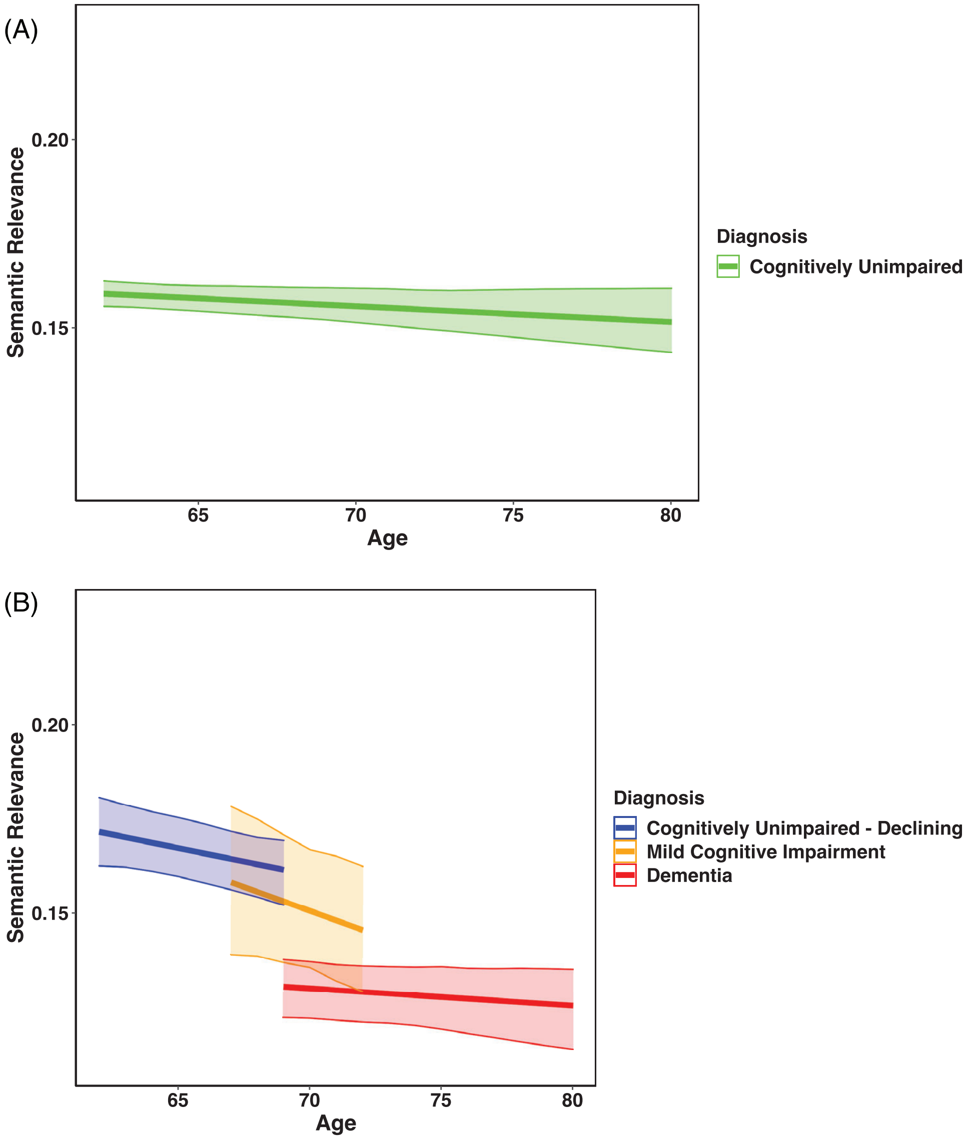

FIGURE 5 Longitudinal plots showing the SemR values as a function of age for (A) cognitively unimpaired participants and (B) cognitively unimpaired declining, mild cognitive impairment, and dementia participants. The dark solid lines are based on the fixed effects of the growth curve model, and the shaded areas show the 95% confidence bands

restricted the plots to the age range with the greatest density of participants in each group (approximately between Q1 and Q3 for each cognition group).

4 | DISCUSSION

The present study builds on the work of Mueller et al.,[1] which evaluated the contribution of connected language, including the Cookie Theft picture descriptions, to provide early evidence of mild cognitive–linguistic decline in a large cohort of participants. They used latent factor analysis to discover that longitudinal changes in the "semantic category" of measures were most associated with cognitive decline. Semantic relevance in this highly structured picture description task captures the ability to speak coherently by maintaining focus on the topic at hand. Some studies have shown that older adults tend to produce less global coherence (and more irrelevant information) in discourse than younger adults.[20] Furthermore, more marked discourse coherence deficits have been reported across a variety of dementia types including Alzheimer's disease (AD) dementia[21] and the behavioral variant of FTD.[22] The neural correlates of coherence measures are difficult to capture, because multiple cognitive processes contribute to successful, coherent language. However, the SemR measure is an ideal target for the cognitive processes known to be affected across stages of dementia. For example, in the case of AD dementia, lower semantic relevance could be the result of a semantic storage deficit,[23] search and retrieval of target words,[24] or inhibitory control deficits,[25] all of which can map onto brain regions associated with patterns of early AD neuropathology.

The development of the automated SemR metric in the present report was intended to mitigate the labor-intensive task of coding content units manually, in an effort to validate a tool that can expedite research and enhance clinical assessment in the context of pre-clinical detection of cognitive decline. The clinical validation of SemR yielded results that were consistent with previous research (e.g., declining scores for older and more cognitively impaired participants).

In addition to developing and thoroughly evaluating the automatically extracted language measure SemR, this article illustrates the use of a rigorous framework for analytical and clinical validation[11] for language features. There has been a great deal of recent interest in automated analysis of patient speech for assessment of neurological disorders.[26,27,28] In general, machine learning (ML) is often used to find "information" in this high-velocity data stream by transforming the raw speech samples into high-dimensional feature vectors that range from hundreds to thousands in number. The assumption is that these features contain the complex information relevant for answering the clinical question of interest. However, this approach carries several risks[29] and most measures of this type fail to undergo rigorous validation, both because large datasets containing speech from clinical groups are difficult to obtain, and because there is no way to measure the accuracy of an uninterpretable feature, for which there is no ground truth. The consequence is measures that vary widely in their ability to capture clinically relevant changes.[11] In contrast, we followed best practices for "fit for purpose" algorithm development, as put forth by the Digital Medicine Society.[11] First, the algorithm was developed on one set of data from participants with ALS and FTD, and then tested on a separate, large, out-of-sample dataset from a different clinical population (CU, MCI, and D), thus fully separating the development, freezing of the algorithm, and testing. During the testing of the algorithm, we showed how we evaluated for accuracy at each step of the automation. Finally, we validated SemR as a clinical tool, evaluating its reliability, association with cognitive function, and change over time.

5 | LIMITATIONS AND FUTURE DIRECTIONS

A surprising finding in the GCM (longitudinal) analyses was that CU-D participants had slightly higher mean SemR than CU participants (a mean SemR difference of $< .01$ between the two groups), and the difference remained even after controlling for age, sex, education, reading scores, and the number of sessions per participant. Because there are many factors that could be responsible for small differences in point-wise estimates in GCM models, between-group differences at any given point should be interpreted with caution. Rather, the GCM analysis should be used to visualize approximate longitudinal trends across groups. Evaluating point-wise trends at different points in time requires an age-matched sample and an analysis that controls for other variables that may impact language that are not accounted for in this study.

There are also several other ways in which SemR can be further validated. First, SemR was evaluated for its association with cognition by comparing it to MMSE scores. However, the MMSE is only a single measure of cognitive functioning with established ceiling effects. Therefore, further measures of cognitive function, including language measures or brain imaging biomarkers, should be tested to further extend the SemR validation.

Second, the effect of demographic characteristics on SemR was not evaluated. Although this study showed that SemR declined with cognitive impairment and age, cross-sectional effects may be confounded by education, intelligence, culture, generational differences, etc.

Third, in this study we compared the in-clinic and at-home scores in CU participants only. This comparison needs to be extended using the clinical populations of interest to determine whether cognitively impaired participants can perform the same tasks unsupervised.

Finally, SemR does not by itself completely characterize cognition. Ongoing work is needed for continuing the development and out-of-sample validation of complementary features that can assess other cognitive domains as accurately and reliably as the semantic relevance measure presented here.

ACKNOWLEDGMENTS

Grants: NIA-NIH R01 AG070940; NIA-NIH R01 AG027161; NIA-NIH R01 AG054059; NIDCD-NIH R01 DC006859; NSF SBIR 1853247.

CONFLICTS OF INTEREST

SH is PI on a grant (NSF SBIR 1853247) that in part funded the work reported on here. KM received funding from NIA-NIH R01 R01

AG070940. VB received funding from The National Science Foundation, National Institutes of Health, Office of Naval Research, Department of Defense, Boehringer Ingelheim Foundation. JL received funding from National Institutes of Health, National Science Foundation. JS received funding from NIH, ALS Association, Amylyx, Biogen, Biotie Therapies (now Acorda Therapeutics), Cytokinetics Incorporated, Mitsubishi Tanabe Pharma America, Alexion, Medicinova, Ionis, Alector, and Orphazyme. JL received book royalties from Cengage Publishing. JS received royalties as the neuromuscular section editor for *UpToDate*. VB received funding from Boehringer Ingelheim, Cytokinetics. JS is a consultant for Amylyx, Apic Biosciences, Neurosense, Cytokinetics, Denali, GSK, Mitsubishi Tanabe Pharma America, Orphazyme, Pinteon, RRD, Swanbio, Helixsmith, Novartis, Sanofi, EMD Serono. KM received UW Madison StartUp Discretionary Funds. VB received funding from Boehringer Ingelheim, National Science Foundation (Reviewer). JL is National Science Foundation reviewer and National Institutes of Health Chair, Review Panel. SH is an inventor on provisional US patent application numbers 63/169,069; 62/964,646; 62/964,642; 63/123,265. VB and JL's list of IP disclosures includes: Speech analysis algorithmic system and method for objective evaluation and/or disease detection; Speech analysis devices and methods for identifying migraine attacks; Objective assessment of patient response for calibration of therapeutic interventions; Tracking articulatory and prosodic development in children; Objective assessment of speech resulting from congestion or decongestion; Objective assessment of speech and respiration changes after smoking cessation; Systems and methods for estimation of forced vital capacity using speech acoustics; Systems and methods for assessing speech, language, and social skills; Systems and methods for digital speech-based evaluation of cognitive function. gs's ip disclosures include: systems and methods for digital speech-based evaluation of cognitive function; Systems and methods for estimation of forced vital capacity using speech acoustics. SB's IP disclosure includes: Systems and methods for digital speech-based evaluation of cognitive function. JS serves on DSMBs for Braingate, Kadimastem, NIH sponsored Fore DMD. VB is Director, Aural Analytics. JL is Chief Clinical Officer at Aural Analytics. JS is on Executive Committee for NEALS, Healey Center for ALS. SH holds stock options in Aural Analytics, the company that provided the app used for speech recordings in this study. GS holds stock options in Aural Analytics, the company that provided the app used for speech recordings in this study. VB and JL are co-founders and have equity in Aural Analytics. JS holds stock options at Aural Analytics. SH is an employee at Aural Analytics, the company that provided the app used for speech recordings in this study. GS was an employee at Aural Analytics, the company that provided the app used for speech recordings in this study. SB was an employee at Aural Analytics, the company that provided the app used for speech recordings in this study. KK is an employee at Aural Analytics, the company that provided the app used for speech recordings in this study. CJD has no relationships/activities/interests to disclose related to the content of this submission.

REFERENCES

1. Mueller KD, Koscik RL, Hermann BP, Johnson SC, Turkstra LS. Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin Registry for Alzheimer's prevention. *Front Aging Neurosci*. 2018;9:437. http://doi.org/10.3389/fnagi.2017.00437.
2. Bschor T, Kühl K-P, Reischies FM. Spontaneous speech of patients with dementia of the Alzheimer's type and mild cognitive impairment. *Int Psychogeriatr*. 2001;13(3):289-298. http://doi.org/10.1017/S1041610201007682.
3. Mueller KD, Koscik RL, Turkstra LS, et al. Connected language in late middle-aged adults at risk for Alzheimer's Disease. *J Alzheimer's Dis*. 2016;54(4):1539-1550. http://doi.org/10.3233/JAD-160252.
4. Cummings L. Describing the cookie theft picture: sources of breakdown in Alzheimer's dementia. *Pragmatics and Society*. 2019;10(2):153-176.
5. Goodglass H, Kaplan E, Barresi B. *Boston Diagnostic Aphasia Examination*. Lippincott, Williams & Wilkins; 2001.
6. Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J Clin Experiment Neuropsychol*. 2018;40(9):917-939.
7. Slegers A, Filiou R, Montembeault M, Brambati S. Connected speech features from picture description in Alzheimer's disease: a systematic review. *J Alzheimer's Dis*. 2018;65(2):519-542.
8. Nicholas L, Bookshire R. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *J Speech and Hear Res*. 1993;36:338-350.
9. Croisile B, Ska B, Brabant MJ, et al. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain Lang*. 1996;53(1):1-19.
10. Ahmed S, Haigh A-MF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*. 2013;136(12):3727-3737. http://doi.org/10.1093/brain/awt269.
11. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). *NPJ Digit Med*. 2020;3(1):55. http://doi.org/10.1038/s41746-020-0260-4.
12. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch Neurol*. 1994;51(6):585-594.
13. Folstein MF, Folstein SE, Mc Hugh PR. Mini mental state: a practical method for grading the cognitive state of patients for the clinician. *J Psychiatry Res*. 1975;12:189-198.
14. Google Cloud. 2021. https://cloud.google.com/speech-to-text/docs/boost
15. Lorah J. Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large Scale Assess Educ*. 2018;6:1-11.
16. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L. Erlbaum Associates; 1988.
17. Portney L, Watkins M. *Foundations of Clinical Research: Applications to Practice*. Davis Company; 2015.
18. Fleiss JL. *The Design and Analysis of Clinical Experiments*. Wiley; 1999.
19. Stegmann G, Hahn S, Liss J, et al. Repeatability of commonly used speech and language features for clinical applications. *Digital Biomarkers*. 2020;4(3):109-122. http://doi.org/10.1159/000511671.
20. Long MR, Horton WS, Rohde H, Sorace A. Individual differences in switching and inhibition predict perspective-taking across the lifespan. *Cognition*. 2018;170:25-30.
21. Glosser G, Deser T. Patterns of discourse production among neurological patients with fluent language disorders. *Brain Lang*. 1991;40(1):67-88.

22. Ash S, Moore P, Antani S, McCawley G, Work M, Grossman M. Trying to tell a tale: discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology*. 2006;66(9):1405-1413.
23. Rofes A, de Aguiar V, Jonkers R, Oh SJ, DeDe G, Sung JE. What drives task performance during animal fluency in people with Alzheimer's disease?. *Front Psychol*. 2020;11:1485.
24. Weakley A, Schmitter-Edgecombe M. Analysis of verbal fluency ability in Alzheimer's disease: the role of clustering, switching and semantic proximities. *Arch Clin Neuropsychol*. 2014;29(3):256-268.
25. Rabi R, Vasquez BP, Alain C, Hasher L, Belleville S, Anderson ND. Inhibitory control deficits in individuals with amnestic mild cognitive impairment: a meta-analysis. *Neuropsychol Rev*. 2020;30(1):97-125.
26. Komeili M, Pou-Prom C, Liaqat D, Fraser KC, Yancheva M, Rudzicz F. Talk2Me: automated linguistic data collection for personal assessment. Greatorex Riches N, ed. *PLoS ONE*. 2019;14(3):e0212342. http://doi.org/10.1371/journal.pone.0212342.
27. Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. *Front Psychol*. 2021;12:620251. https://doi.org/10.3389/fpsyg.2021.620251.
28. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J Am Med Inform Assoc*. 2020;27(11):1784-1797. http://doi.org/10.1093/jamia/ocaa174.
29. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *NPJ Digi Med*. 2021;4(1):1-8.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Stegmann G, Hahn S, Bhandari S, et al. Automated semantic relevance as an indicator of cognitive decline: Out-of-sample validation on a large-scale longitudinal dataset. *Alzheimer's Dement*. 2022;14:e12294. https://doi.org/10.1002/dad2.12294