Digital Biomarkers

Research Reports - Research Article

Digit Biomark 2023;7:115–123
DOI: 10.1159/000533188

Received: May 3, 2023
Accepted: July 2, 2023
Published online: August 31, 2023

Validation of an Automated Speech Analysis of Cognitive Tasks within a Semiautomated Phone Assessment

Daphne ter Huurne[a] Nina Possemis[a] Leonie Banning[b] Angélique Gruters[c] Alexandra König[d, e] Nicklas Linz[e] Johannes Tröger[e] Kai Langel[f] Frans Verhey[a, b] Marjolein de Vugt[a, b] Inez Ramakers[a, b]

[a]Alzheimer Center Limburg, School for Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands; [b]Maastricht University Medical Center+ (MUMC+), Maastricht, The Netherlands; [c]Catharina Hospital, Medical Psychology, Eindhoven, The Netherlands; [d]National Institute for Research in Computer Science and Automation (INRIA), Sophie Antipolis, France; [e]ki elements, Saarbrücken, Germany; [f]Janssen Clinical Innovation, Beerse, Belgium

Keywords

Phone assessment · Automated speech analysis · Mild cognitive impairment · Memory · Fluency

Abstract

Introduction: We studied the accuracy of the automatic speech recognition (ASR) software by comparing ASR scores with manual scores from a verbal learning test (VLT) and a semantic verbal fluency (SVF) task in a semiautomated phone assessment in a memory clinic population. Furthermore, we examined the differentiating value of these tests between participants with subjective cognitive decline (SCD) and mild cognitive impairment (MCI). We also investigated whether the automatically calculated speech and linguistic features had an additional value compared to the commonly used total scores in a semiautomated phone assessment. ***Methods:*** We included 94 participants from the memory clinic of the Maastricht University Medical Center+ (SCD N = 56 and MCI N = 38). The test leader guided the participant through a semiautomated phone assessment. The VLT and SVF were audio recorded and processed via a mobile application. The recall count and speech and linguistic features were automatically extracted. The diagnostic groups were classified by training machine learning classifiers to differentiate SCD and MCI participants. ***Results:*** The intraclass correlation for inter-rater reliability between the manual and the ASR total word count was 0.89 (95% CI 0.09–0.97) for the VLT immediate recall, 0.94 (95% CI 0.68–0.98) for the VLT delayed recall, and 0.93 (95% CI 0.56–0.97) for the SVF. The full model including the total word count and speech and linguistic features had an area under the curve of 0.81 and 0.77 for the VLT immediate and delayed recall, respectively, and 0.61 for the SVF. ***Conclusion:*** There was a high agreement between the ASR and manual scores, keeping the broad confidence intervals in mind. The phone-based VLT was able to differentiate between SCD and MCI and can have opportunities for clinical trial screening.

karger@karger.com
www.karger.com/dib

Karger
OPEN ACCESS

Correspondence to:
Inez Ramakers, i.ramakers@maastrichtuniversity.nl

Introduction

A neuropsychological assessment (NPA) plays an important role in the diagnosis of dementia and its prodromal stage (mild cognitive impairment [MCI] [1]) by investigating the severity of cognitive impairment and exploring the individual profile of cognitive strengths and weaknesses as a start for differential diagnostics [2]. An accurate and easy-to-use remote way to assess neuropsychological performances would be of added value to follow the disease course over time, mainly in medical deserts, and could be used for screening purposes in clinical trial designs.

Recently developed tools, such as videoconference or computer-based NPAs, have already been validated as a good alternative to face-to-face assessments [3]. However, most of these alternative NPA options require an internet connection and a computer to support video or teleconferencing platforms [3–6]. This may be difficult for the elderly population or those with cognitive impairments, as technical interactions would be required [3–6]. Therefore, there is a need for an easy method for remote NPA, such as phone-based testing. In this setting, the person would only need to pick up the phone, and no technological interactions are required [7, 8]. Previous research has suggested that this could be a valid way to assess or screen patients for clinical trials. Phone-based assessments have already been used to perform a screening task, such as the Mini-Mental State Examination (MMSE), or more extensive cognitive tasks such as memory or language tasks [9, 10], which can be easily recorded via the phone.

Automatic speech recognition (ASR) software can score cognitive tests through recordings, in which speech features can also be extracted automatically. Moreover, ASR may be helpful in differentiating people with subjective cognitive decline (SCD) from MCI/dementia in a clinical face-to-face setting (Possemis et al., unpubl. data) [11, 12]. In addition, comparing manual and automatic scoring in a face-to-face setting showed a high level of agreement, demonstrating that ASR can provide accurate results. This accuracy and differentiating value have not yet been studied in a phone-based setting.

In this study, we examined the utility of a semiautomated phone-based NPA, guided by the test leader, in a Dutch memory clinic sample of participants with SCD or MCI. We investigated the accuracy of the ASR software compared to manually derived scores of a verbal episodic memory test and a verbal fluency test. Furthermore, we examined the differentiating value of these tasks between participants with SCD and MCI. In addition, we investigated whether the automatically extracted speech and linguistic features had an additional value compared to the commonly used total scores.

Methods

Participants

As part of the DeepSpA (Deep Speech Analysis for Cognitive Assessment in Clinical Trials) project, 140 participants were included via the BioBank Alzheimer Centre Limburg (BBACL) study [11]. Out of the 140 participants, 94 (56 SCD, 38 MCI) completed the semiautomated phone assessment (see Table 1 for participant characteristics). Reasons for the 46 participants who did not participate are as follows: 15 participants were excluded from follow-up because people diagnosed with dementia at baseline were not invited for a follow-up due to study design. The other 31 participants did not participate due to multiple reasons, namely: refusal of phone assessment (*N* = 4), refusal due to illness (*n* = 4), general refusal (*N* = 12), technical malfunction (*N* = 3), no response (*N* = 5), and deceased (*n* = 3).

BBACL is an ongoing prospective cohort study that includes consecutive patients from the memory clinic of the Maastricht University Medical Center+ (MUMC+) in the Netherlands. The inclusion criteria were a Mini-Mental State Examination (MMSE) total score of ≥20 and a Clinical Dementia Rating scale (CDR) total score of ≤1. Exclusion criteria were non-neurodegenerative neurological diseases, a recent history of severe psychiatric disorders (such as major depression), the absence of a reliable informant, and the clinical judgment that a follow-up assessment after 1 year will not be feasible.

Baseline diagnoses were based on the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV-TR, DSM-V) criteria for MCI (cognitive disorder not otherwise specified [NOS] in DSM-IV-TR; mild neurocognitive disorder in DSM-V) [13–15]. Participants without objective cognitive disorders were classified as SCD [16].

Semiautomated Phone Assessment

At baseline, each participant underwent a face-to-face NPA at the hospital as part of a clinical routine [11]. After 6 months, participants underwent a semiautomated phone assessment, in which a well-trained test leader guided the participant through the phone assessment. The baseline CDR score was used to measure disease severity [17, 18], and the baseline MMSE was used to measure global cognition [19, 20].

Within the semiautomated phone assessment, participants were called by the test leader via a mobile application. The participants were instructed to be in a room without distractions and were not allowed to use paper and pen. The test leader guided the participants through the tasks and started each task separately. The test instructions were prerecorded in the mobile application, so the assistance of the test leader was only necessary if the participant had a specific question. Within the assessment, the 15-verbal learning task (VLT) [21] and the 1-min semantic verbal fluency (SVF) task [22] were administered. The VLT is an episodic memory task in which fifteen unrelated monosyllabic words were named in five trials, and after each presentation, the participants

Table 1. Characteristics and comparisons between SCD and MCI participants (n = 94)

	SCD (N = 56)	MCI (N = 38)	Total (N = 94)	p value
Age, years	62.7 (10.3)	68.8 (9.10)	65.2 (10.3)	0.004
Sex (% male)	66	66	64	0.745
Education (% low/mid/high)	25/39/36	39/32/29	31/36/33	0.329
CDR SOB	0.7 (0.8)	1.9 (1.5)	1.2 (1.3)	<0.001
MMSE	28.7 (1.2)	26.9 (1.8)	28.0 (1.7)	<0.001
VLT immediate recall manual	40.4 (11.3)	30.4 (9.7)	36.4 (11.7)	<0.001
VLT delayed recall manual	8.1 (3.3)	4.6 (3.2)	6.7 (3.7)	<0.001
SVF manual	19.3 (7.2)	15.0 (6.5)	17.5 (7.2)	0.004
VLT immediate recall ASR	35.0 (11.4)	22.5 (8.7)	30.0 (12.0)	<0.001
VLT delayed recall ASR	6.9 (3.2)	3.5 (2.9)	5.5 (3.5)	<0.001
SVF ASR	16.7 (6.9)	12.2 (5.9)	14.9 (6.8)	0.001

Data are presented as mean (SD), unless otherwise specified. SCD, subjective cognitive decline; MCI, mild cognitive impairment; CDR, Clinical Dementia Rating scale; SOB, sum of boxes; MMSE, Mini-Mental State Examination; VLT, verbal learning task; ASR, automatic speech recognition; SVF, semantic verbal fluency.

had to recall as many words as they could remember. After 20 min, during which nonverbal tasks were administered, participants had to recall as many words as possible (delayed recall) [21]. Within the SVF, the participant had to name as many supermarket items as possible within 60 s [22].

Speech Data Processing

The VLT and SVF were manually scored by the test leader. In addition, both tasks were audio recorded, scored, and processed by a mobile application from ki elements GmbH (iOS iPad version) [23]. The application records participant's voice responses, while the assessments were conducted. The application used the microphone of the participants' phones to record. After the speech sample was recorded, it was sent to ki elements backend for preprocessing (cutting speech into relevant parts and audio transformation), ASR, and feature calculation. The ASR tool was Google Speech API. The total word count of both the VLT and SVF and the speech and linguistic features were automatically extracted. Examples of speech and linguistic features of the VLT are "primacy count," "serial clustering," and "recency count" and for the SVF, examples are "semantic clustering," "temporal clustering," and "mean word frequency" [8, 11, 12, 24, 25] (for the full list of the speech and linguistic features, see online suppl. Supplement 1; for all online suppl. material, see https://doi.org/10.1159/000533188). These speech and linguistic features were automatically extracted from the automatically derived ASR transcript.

Statistical Analyses

The data were analyzed with IBM SPSS Statistics Mac (version 27) and Python 3.9.7 [26]. Differences between groups were analyzed with independent t tests for continuous variables and with χ^2 tests for categorical variables (or comparable nonparametric tests in case data were not normally distributed).

To examine the agreement between the manual and the ASR score of both the VLT and SVF, the intraclass correlation coefficient (ICC) was used. The ICC is used as a measure of inter-rater agreement for continuous variables [27]. The ICC was based on a mean rating (k = 2), absolute agreement, 2-way mixed-effects model.

Machine learning models (classification method "extra trees") were trained to differentiate between SCD and MCI using the sklearn Python package [28]. Due to the limited sample size, no holdout test set could be maintained. Instead, models were evaluated using leave-one-out cross-validation, a procedure in which a dataset is split into a training set and a testing set, using all but one sample as part of the training set. This procedure was repeated for each sample, and average of the model's performance was calculated. Area under the receiver operating characteristic curve, which allows visualization of multiple different potential trade-offs between sensitivity and specificity, were created for three models, namely: (1) ASR total word count; (2) ASR total word count and age; and (3) ASR total word count, age, and the automatically derived speech and linguistic features (see online suppl. Supplement 1 for the features used within model 3 per task, for all online suppl. material, see https://doi.org/10.1159/000533188). This was performed separately for the VLT immediate recall, VLT delayed recall, and SVF. An AUC of 0.5 means no discriminative value, 0.6–0.7 is considered poor, 0.7–0.8 is considered acceptable, 0.8–0.9 is considered excellent, and 0.9 and higher is considered outstanding [29].

Results

Participant Characteristics

The characteristics of the SCD and MCI participants are presented in Table 1. MCI participants were significantly older and had higher scores on the CDR and lower scores on the MMSE, VLT, and SVF than SCD participants. The groups did not differ significantly for sex and education level.

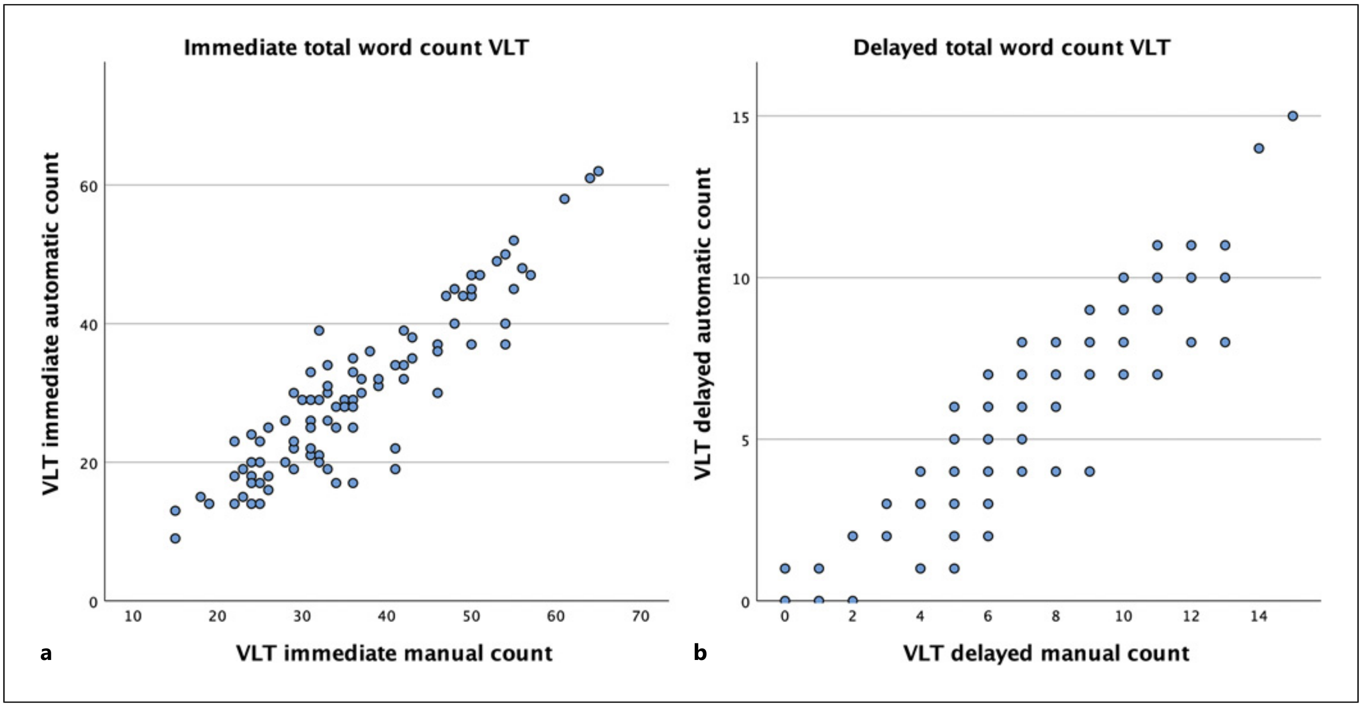

Fig. 1. Scatterplot of the manually and automatically derived (automatic speech recognition [ASR]) score immediate (**a**) and delayed (**b**) verbal learning task (VLT) total word count for the semiautomated phone assessment.

Agreement between Manual and Automatic Total Word Count

The ICC for describing the agreement between the manual and ASR total word count for the immediate recall of the VLT was 0.89 (95% CI 0.09–0.97; Fig. 1a) and for the delayed recall was 0.94 (95% CI 0.68–0.98; Fig. 1b). The ICC of the SVF task was 0.93 (95% CI 0.56–0.97; Fig. 2). The mean difference between the manual score and automatic score for the VLT immediate recall was 6.4 (SD = 4.8) with a range of −7 to 22 words, with one outlier in which the ASR missed 22 words. Without the outlier, the ICC remained identical, and the mean difference remained almost the same, namely, 6.2 words difference. For the VLT delayed recall, this was 1.2 (SD = 1.3) with a range of −1 to 5 words, without any outliers. Finally, the mean difference for the SVF was 2.7 (SD = 2.5) with a range from −3 to 13 words, with one outlier where the ASR missed 13 items. Without the outlier, the ICC remained the same and the mean difference remained almost the same, namely, 2.6 words difference.

Diagnosis Classification

The ROC curves differentiating between SCD and MCI within the semiautomated phone assessment are shown in Figures 3–5. For the immediate recall of the

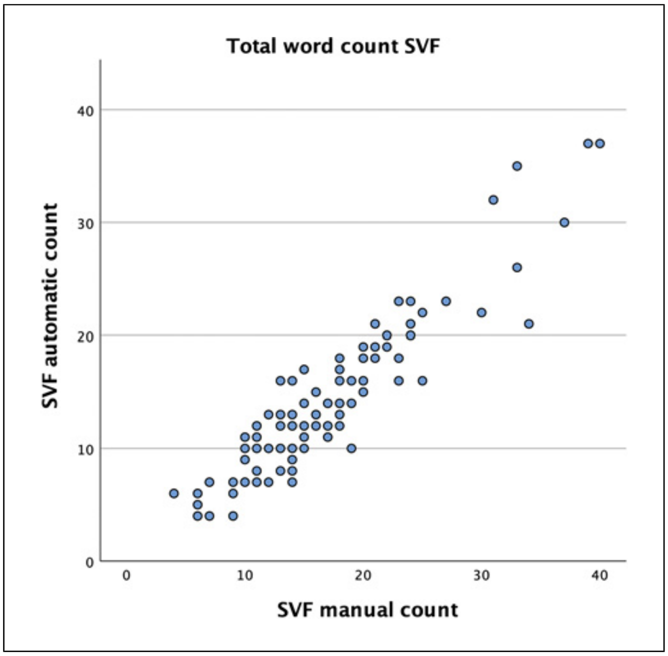

Fig. 2. Scatterplot of the manually and automatically derived (automatic speech recognition [ASR]) score semantic verbal fluency (SVF) task total word count for the semiautomated phone assessment.

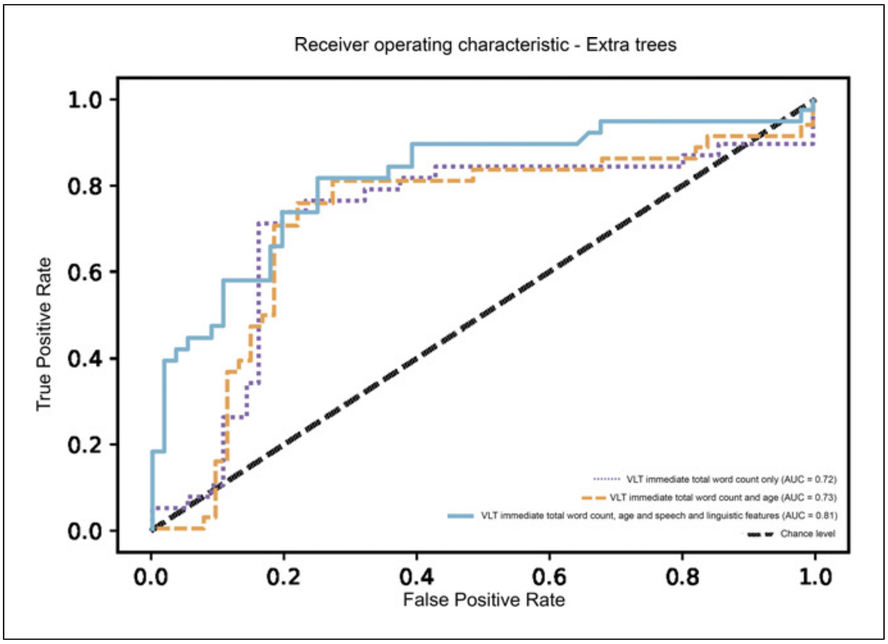

Fig. 3. Receiver operating characteristic curve for the verbal learning task (VLT) immediate recall differentiating between subjective cognitive decline (SCD) and mild cognitive impairment (MCI).

VLT, the full model of the VLT including total word count, age, and speech and linguistic features had an AUC of 0.81, which is classified as “excellent” differentiation 29 between SCD and MCI participants (Fig. 3). Here the full model had a higher AUC in comparison to the age-corrected total score (AUC = 0.73, Fig. 3) and word count alone (AUC = 0.72). The full model of the VLT could differentiate "acceptably 29 between SCD and MCI participants (AUC = 0.77, Fig. 4), and this was slightly higher than the model correcting for age (AUC = 0.76) and for word count alone (AUC = 0.75). Lastly, the full model of the SVF including the total word count, age and speech, and linguistic features had an AUC of 0.61, which is classified as “poor” differentiation 29 between SCD and MCI participants (AUC = 0.61, Fig. 5). The AUC of the full model including the speech and linguistic features was not higher than the model for the age-corrected total score (AUC = 0.64) or the total word count alone (AUC = 0.63).

Discussion

In this study, we examined an automated analysis of the VLT and SVF in a semiautomated phone assessment. Results showed that the ASR total word count of both the VLT and SVF were comparable to the manually retrieved total word count with high ICCs, although the confidence intervals showed a broad range for all tasks. Moreover, the automatically derived speech and linguistic features for the VLT immediate recall and delayed recall had a high diagnostic discriminative power between SCD and MCI participants, while for the SVF, the discriminative power was low.

Previous research has already shown good comparability of manual and automatic scoring of the VLT and SVF in a face-to-face setting for both the French (for the SVF) [12] and Dutch language (for the SVF and the VLT) [11] (Possemis et al., unpubl. data). This is in line with our results of the semiautomated phone assessment. The ICC of the SVF in the face-to-face setting (ICC = 0.84) was lower than in the semiautomated phone assessment (ICC = 0.93) [11]. This difference could not be explained by the different microphones used in both settings because the ICCs of the VLT were comparable between face-to-face [11] (Possemis et al., unpublished) and semiautomated phone. Although the comparability between the manual and ASR score of the VLT and the SVF was found to be good, wide confidence intervals were found, and the mean difference between the manual and ASR score was relatively high. For example, within the VLT immediate recall, the ASR missed 6 words. This means that the use of ASR leads to a systematic underestimation of the performance of the participant for the VLT immediate recall. To explore the clinical impact of these mean differences, we translated the mean

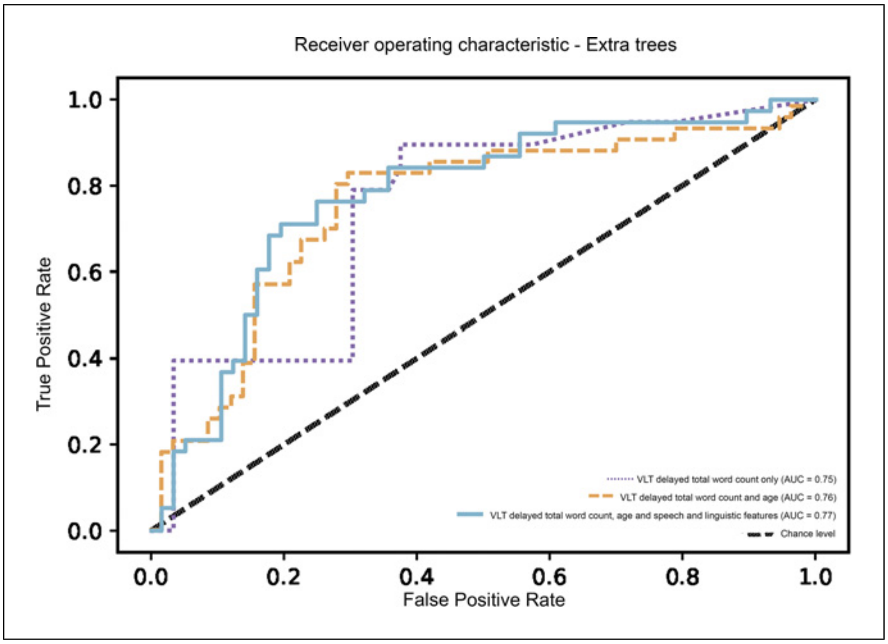

Fig. 4. Receiver operating characteristic curve for the verbal learning task (VLT) delayed recall differentiating between subjective cognitive decline (SCD) and mild cognitive impairment (MCI).

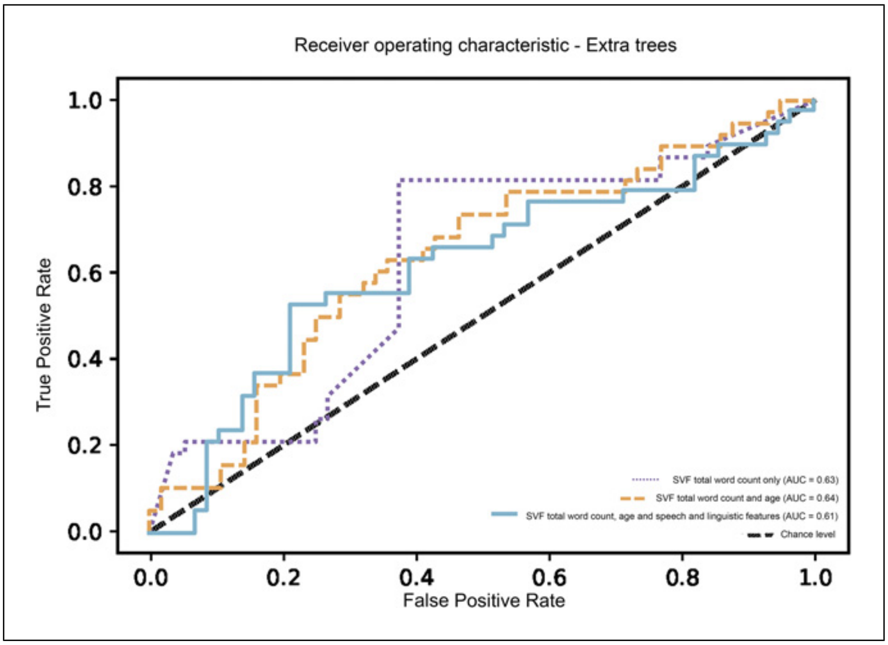

Fig. 5. Receiver operating characteristic curve for the semantic verbal fluency (SVF) task differentiating between subjective cognitive decline (SCD) and mild cognitive impairment (MCI).

difference into clinically used standardized scores, which are published and based on healthy controls [21, 22]. Characteristics of the current mean population were used (male, 65 years old, middle educational level). The differences in z-scores were 0.77 for the VLT immediate (z-score manual: −0.63, ASR: −1.40), 0.40 for the VLT delayed recall (z-score manual: −0.57, ASR: −0.97), and 0.52 for the SVF (z-score manual: −0.92, ASR: −1.44). This implies that participants with the ASR score are more often classified as cognitively impaired compared to the manual scoring, indicating that the ASR scores tend to have a higher sensitivity and a lower specificity. When listening to the recordings of the VLT and SVF, the participants with the greatest difference between manual and ASR spoke quietly, with a dialect, or very quickly. Moreover, some participants repeat words as an internal rehearsal strategy, and therefore these items were not meant to increase the total word count even though the ASR registered these as correct [30].

The differentiating value of the VLT (word count with correction for age) between SCD and MCI was high for both the immediate and delayed recall within the phone modality. This is in line with previous research stating that the VLT is able to differentiate between healthy controls or SCD and MCI in a face-to-face modality [31]. When the correction for age and speech and linguistic features were added to the VLT immediate recall, the differentiating power increased slightly. This increase should be interpreted with care, as the increase is limited and its clinical value is therefore uncertain. In previous studies, clustering appears to be impaired within MCI patients [32, 33] which in turn could help differentiate between SCD and MCI when adding these to the model. For the VLT delayed recall, no increase of differentiating power, between SCD and MCI, over the word count and correction for age was reported. An explanation could be a ceiling effect for the VLT delayed recall as this is already a very sensitive measure for cognitive disorders [31, 34].

The diagnostic differentiation between SCD and MCI was poor overall for the SVF. As the ICC is high, the ASR software can derive the items well, and this therefore cannot explain the poor differentiation. However, a possible explanation for the low AUC could be that the supermarket SVF has a lower power in discriminating between SCD and MCI. This is in line with some recent studies in which the supermarket SVF had a relatively higher differentiating power when the diagnostic groups in the cognitive spectrum are further apart [35, 36].

A strength of this study is the automatically derived scores of both the VLT and the SVF; this is the first study to examine this within a semiautomated phone assessment. Moreover, the results of this study suggest that, particularly for the VLT, the phone assessment could be a convenient way to screen participants for clinical trials, as there would be no need to score the task manually, which is more time efficient. Furthermore, the higher sensitivity of the VLT automatically derived score would give opportunities for using it as a screening tool for clinical trials, keeping in mind that for clinical practice, a higher specificity would be important.

First, a limitation of this study is that the results cannot be generalized outside a memory clinic sample without due consideration. Future research should include healthy controls from the general population to be able to implement this semiautomated phone assessment as a screening tool for clinical trials. Second, a limitation of this study was the exclusion of people with dementia, which reduces the generalizability of the results to the memory clinic population, as this population encompasses the entire cognitive spectrum. Therefore, future studies should include people with dementia to determine whether the phone assessment is suitable for people with dementia and whether the phone assessment could differentiate between the three diagnostic groups (healthy control or SCD, MCI, and dementia).

In conclusion, the ASR-based outcomes of the semiautomated phone assessment were comparable to the manually derived scores from the VLT and SVF, although the mean difference between both scores had a clinically relevant impact as participants were more likely to be classified as MCI than SCD by using the ASR score. The phone-based VLT was able to differentiate between SCD and MCI. All in all, this semiautomated phone assessment, particularly for the VLT, could be implemented within the screening for clinical trials.

Statement of Ethics

The Local Medical Ethical Committee (METC MUMC/UM) approved the study (MEC 15-4-100, MEC 2009-3-037, MEC 2009-3-038). All participants gave written informed consent to participate in the study.

Conflict of Interest Statement

J.T., A.K., and N.L. are employed by the company ki elements, which developed the mobile application and calculated the speech and linguistic features. N.L., A.K., and J.T. own shares in the company.

Funding Sources

This work was supported by EIT Health (Grant No. 19249), as well as Janssen Pharmaceutica NV through a collaboration agreement (Grant No. is not applicable).

Author Contributions

The data collection was performed and coordinated by D.t.H., N.P., and I.R. N.L. and J.T. calculated the speech and linguistic features. D.t.H., N.P., I.R., A.K., M.d.V., and F.V. all contributed to this study’s conception of the research question. D.t.H. drafted the manuscript and performed the statistical analyses. The machine learning models were performed by D.t.H., where the scripts were developed in collaboration with N.L. and J.T. All authors (D.t.H., N.P., L.B., A.G., A.K., N.L., J.T., K.L., F.V., M.d.V., and I.R.) participated in the interpretation of the data and revised drafts of the manuscript for important intellectual content. All authors read and approved the final manuscript.

Data Availability Statement

All data generated or analyzed during this substudy are included in this article and its online supplementary material. Further inquiries can be directed to the corresponding author.

References

1 Prado CE, Watt S, Treeby MS, Crowe SF. Performance on neuropsychological assessment and progression to dementia: a meta-analysis. Psychol Aging. 2019 Nov;34(7):954–77.

2 Lezak MD, Howieson DB, Bigler ED, Tranel D. Neuropsychological assessment. 5th ed. Oxford University Press; 2016.

3 Schatz P, Browndyke J. Applications of computer-based neuropsychological assessment. J Head Trauma Rehabil. 2002 Oct;17(5):395–410.

4 Bloch A, Maril S, Kavé G. How, when, and for whom: decisions regarding remote neuropsychological assessment during the 2020 COVID-19 pandemic. Isr J Health Policy Res. 2021 May 3;10(1):31.

5 Geddes MR, O'Connell ME, Fisk JD, Gauthier S, Camicioli R, Ismail Z, et al. Remote cognitive and behavioral assessment: report of the alzheimer society of Canada task force on dementia care best practices for COVID-19. Alzheimers Dement. 2020 Sep 22;12(1):e12111.

6 Parlar ME, Spilka MJ, Wong Gonzalez D, Ballantyne EC, Dool C, Gojmerac C, et al. "You can't touch this": delivery of inpatient neuropsychological assessment in the era of COVID-19 and beyond. Clin Neuropsychol. 2020 Oct–Nov;34(7–8):1395–410. Epub 2020 Sep 10.

7 Sumpter R, Camsey E, Meldrum S, Alford M, Campbell I, Bois C, et al. Remote neuropsychological assessment: acceptability and feasibility of direct-to-home teleneuropsychology methodology during the COVID-19 pandemic. Clin Neuropsychol. 2023 Feb;37(2):432–47. Epub 2022 May 3.

8 Tröger J, Linz N, König A, Robert P, Alexandersson J. Telephone-based dementia screening I: automated semantic verbal fluency assessment. Proceedings of the 12th EAI International Conference on pervasive computing technologies for healthcare. 2018.

9 Carlew AR, Fatima H, Livingstone JR, Reese C, Lacritz L, Pendergrass C, et al. Cognitive assessment via telephone: a scoping review of instruments. Arch Clin Neuropsychol. 2020 Nov 19;35(8):1215–33.

10 Castanho TC, Amorim L, Zihl J, Palha JA, Sousa N, Santos NC. Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: a review of validated instruments. Front Aging Neurosci. 2014 Feb 25;6:16.

11 Ter Huurne D, Ramakers I, Possemis N, Banning L, Gruters A, Van Asbroeck S, et al. The accuracy of speech and linguistic analysis in early diagnostics of neurocognitive disorders in a memory clinic setting. Arch Clin Neuropsychol. 2023 Jan 27;38(5):667–76.

12 König A, Linz N, Tröger J, Wolters M, Alexandersson J, Robert P. Fully automatic speech-based analysis of the semantic verbal fluency task. Dement Geriatr Cogn Disord. 2018;45(3–4):198–209. Epub 2018 Jun 8.

13 American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Text rev.; 2000.

14 American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. 2013.

15 Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011 May;7(3):270–9. Epub 2011 Apr 21.

16 Jessen F, Wolfsgruber S, Wiese B, Bickel H, Mösch E, Kaduszkiewicz H, et al. AD dementia risk in late MCI, in early MCI, and in subjective memory impairment. Alzheimers Dement. 2014 Jan;10(1):76–83. Epub 2013 Jan 30.

17 O'Bryant SE, Lacritz LH, Hall J, Waring SC, Chan W, Khodr ZG, et al. Validation of the new interpretive guidelines for the clinical dementia rating scale sum of boxes score in the national Alzheimer's coordinating center database. Arch Neurol. 2010 Jun;67(6):746–9.

18 Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993 Nov;43(11):2412–4.

19 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 Nov;12(3):189–98.

20 Kok R, Verhey F. Gestandaardiseerde MMSE. Zeist: Altrecht GGZ; 2002. p. 1–2.

21 Van der Elst W, van Boxtel MP, van Breukelen GJ, Jolles J. Rey's verbal learning test: normative data for 1,855 healthy participants aged 24–81 years and the influence of age, sex, education, and mode of presentation. J Int Neuropsychol Soc. 2005 May;11(3):290–302.

22 Aschenbrenner S, Tucha O, Lange KW. Regensburger wortflüssigkeits-test: RWT. Göttingen: Hogrefe, Verlag für Psychologie; 2000.

23 Ki elements. Available from: https://ki-elements.de.

24 Linz N, Lundholm Fors K, Lindsay H, Eckerström M, Alexandersson J, Kokkinakis D. Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. 2019.

25 Tröger J, Linz N, König A, Robert P, Alexandersson J, Peter J, et al. Exploitation vs. exploration-computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. Neuropsychologia. 2019 Aug;131:53–61. Epub 2019 May 20.

26 Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.R-project.org/.

27 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. Epub 2016 Mar 31. Erratum in: J Chiropr Med. 2017 Dec;16(4):346.

28 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

29 Hosmer D, Lemeshow S. Applied logistic regression. New York: Wiley; 2000. p. 375.

30 Weitzner DS, Pugh EA, Calamia M, Roye S. Examining the factor structure of the Rey auditory verbal learning test in individuals across the life span. J Clin Exp Neuropsychol. 2020 May;42(4):406–14. Epub 2020 Mar 23.

31 Estévez-González A, Kulisevsky J, Boltes A, Otermín P, García-Sánchez C. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. Int J Geriatr Psychiatry. 2003 Nov;18(11):1021–8.

32 Ribeiro F, Guerreiro M, De Mendonça A. Verbal learning and memory deficits in mild cognitive impairment. J Clin Exp Neuropsychol. 2007 Feb;29(2):187–97.

33 Perri R, Carlesimo GA, Serra L, Caltagirone C; Early Diagnosis Group of the Italian Interdisciplinary Network on Alzheimer's Disease. Characterization of memory profile in subjects with amnestic mild cognitive impairment. J Clin Exp Neuropsychol. 2005 Nov;27(8):1033–55.

34 Campos-Magdaleno M, Facal D, Lojo-Seoane C, Pereiro AX, Juncos-Rabadán O. Longitudinal assessment of verbal learning and memory in amnestic mild cognitive impairment: practice effects and meaningful changes. Front Psychol. 2017 Jul 20; 8:1231.

35 Neves TRF, Araújo NBD, Silva FDO, Ferreira JVA, Nielsen TR, Engedal K, et al. Accuracy of the semantic fluency test to separate healthy old people from patients with Alzheimer's disease in a low education population. J Bras Psiquiatr. 2020;69(2):82–7.

36 Price SE, Kinsella GJ, Ong B, Storey E, Mullaly E, Phillips M, et al. Semantic verbal fluency strategies in amnestic mild cognitive impairment. Neuropsychology. 2012 Jul; 26(4):490–7.