ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414009

ACOUSTIC AND LINGUISTIC ANALYSES TO ASSESS EARLY-ONSET AND GENETIC ALZHEIMER'S DISEASE

P. A. Pérez-Toro[1,2,3⋆], *J. C. Vásquez-Correa*[1,2], *T. Arias-Vergara*[1,2,6], *P. Klumpp*[2], *M. Sierra-Castrillón*[5], *M. E. Roldán-López*[5], *D. Aguillón*[4], *L. Hincapié-Henao*[4], *C. A. Tobón-Quintero*[5], *T. Bocklet*[3], *M. Schuster*[6], *J. R. Orozco-Arroyave*[1,2], *E. Nöth*[2]

[1] Facultad de Ingeniería. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[2] Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Erlangen, Germany
[3] Technische Hochschule Georg Simon Ohm, Nürnberg, Germany
[4] Grupo Neurociencias de Antioquia GNA, University of Antioquia UdeA, Colombia
[5] Grupo de Neuropsicología y Conducta GRUNECO, Universitad de Antioquia UdeA, Medellín, Colombia
[6] Department of Otorhinolaryngology, Head and Neck Surgery. Ludwig-Maximilians University, Munich, Germany
⋆corresponding author: paula.andrea.perez@fau.de

ABSTRACT

The PSEN1-E280A or *Paisa mutation* is responsible for most of Early-Onset Alzheimer's (EOA) disease cases in Colombia. It affects a large kindred of over 5000 members that present the same phenotype. The most common symptoms are related to language disorders, where speech fluency is also affected due to the difficulty to access semantic information intentionally. This study proposes the use of acoustic and linguistic methods to extract features from speech recordings and their transcriptions to discriminate people with conditions related to the *Paisa mutation*. We consider state-of-the-art word-embedding methods like Word2Vec and Bidirectional Encoder Representations from Transformer to process the transcripts. The speech signals are modeled by using traditional acoustic features and speaker embeddings. To the best of our knowledge, this is the first study focused on evaluating genetic Alzheimer's and EOA using acoustics and linguistics.

Index Terms— PSEN1–E280A, Alzheimer's Disease, Acoustic Analysis, Linguistic Analysis

1. INTRODUCTION

Early-onset Alzheimer's (EOA) disease due to the PSEN1-E280A or *Paisa mutation* is commonly diagnosed at a mean age of 49 years [1]. This mutation affects 25 families with more than 5,000 members who historically lived in isolated regions of the Andes mountains in the Colombian state of Antioquia. This population is remarkable for its unusual size and for the high level of participation in longitudinal studies. The members of this kindred can be Genetic Carriers (GC) or Non-Genetic Carriers (NGC). GCs inherit this mutation but they do not show any symptom of Alzheimer's Disease (AD). However, GCs are able to pass the allele onto their offspring. Simlar to GCs, NGCs do not present any symptoms, however they are not carriers of the mutation. EOA only represents 5–10% of all AD cases worldwide. It is characterized by typical symptoms of AD such as memory deficits in the third decade of life, development of progressive cognitive impairments related to verbal disfluency, changes in personality and behavior, among others [2]. AD is highly characterized by the deterioration of the capability to produce coherent language that affects lexical, grammatical, and semantic processes. Different studies have shown abnormalities in language production, characterized by the difficulty to access semantic information intentionally, which affects the speech fluency of the patients [3]. The standard scales to evaluate the cognitive function of the patients are the Mini-Mental State Examination (MMSE) [4] and Montreal Cognitive Assessment (MoCA) [5], which are 30-point scales that contain items related to language production, immediate memory, naming, and spatial attention. Scores of over 24 and 26 indicate normal cognition for MMSE and MoCA respectively.

Nowadays, the most common linguistic methods to analyze dementia are related to syntactic, semantic, word occurrence analysis and word-embeddings [6, 7]. Additionally, some studies consider deep learning methods combined with Natural Language Processing (NLP) techniques [6, 8]. Conventional prosodic measures in dementia focus on temporal aspects, intensity, voice quality, interruptions, voice periods, and variation in fundamental frequency (F_0). Additionally, acoustic features suggest and contextualize interpretations from the acoustic information such as formant frequencies, Mel-Frequency Cepstral Coefficients (MFCCs), Energy distributed in the Bark scale (BBE), among others [9]. Recently,

speaker embeddings are considered to asses AD [10, 11]. Those embeddings aim to capture in a compact form the relevant information about the identity of the speaker. Previous work has demonstrated that automatic language understanding and linguistic analysis are suitable to detect and evaluate patients with dementia and Mild Cognitive Impairments (MCI) [7, 12]. State-of-the-art studies [6, 7, 8] are mostly focused on evaluating AD based on acoustic and linguistic information using the Dementia Bank [13] dataset. It consists of a set of English recordings and transcripts, in which the participants described the cookie theft picture [14].

Unlike previous works this study considers a new dataset where the data recording is ongoing. The participants of this dataset are Colombian Spanish speakers with conditions related to the *Paisa mutation* and EOA. We propose the use of acoustic and linguistic analyses to model these conditions. The speech models include the extraction of traditional acoustic features based on articulation and prosodic analyses such as measures of duration, formant frequencies, MFCCs and BBEs. Additionally, speaker embeddings are extracted using i-vectors and x-vectors. Linguistic analysis includes the computation of word embeddings such as Word2Vec (W2V) and Bidirectional Encoder Representations from Transformers (BERT). To the best of our knowledge, this is the first study focused on automatically evaluating genetic AD using acoustics and linguistics.

2. GENETIC ALZHEIMER'S DATASET

This database is being recorded since 2018 in the University of Antioquia by Grupo de Neurociencias de Antioquia and GITA Lab. The data consist of spontaneous speech recordings and their transliterations from 114 Spanish speakers from Colombia, 28 asymptomatic subjects belonging to families with the *Paisa mutation* that are GC and 36 that are NGC, 23 MCI patients with EOA, and 27 Healthy Control (HC) subjects. The task consisted of the description of the cookie theft picture [14]. The average duration of the recordings is 84±48 seconds for the GC subjects, 83±42 seconds for the NGC subject, 53±25 for the MCI patients, and 42±19 for HC subjects. The transliterations were produced by a professional for linguistics following the verbatim protocol. The whole vocabulary of the lemmatized transcriptions without stopwords consists of 980 words. The data was labeled by expert listeners according to the MMSE and MoCA scales. Demographic information about the participants is included in Table 1.

3. METHODS

Acoustic and linguistic features are considered to process information from the speech recordings and their transliterations. Figure 1 shows the general methodology addressed in this study. Speech features include classical acoustic analysis as well as speaker-embeddings. Linguistic features are based on word-embedding methods.

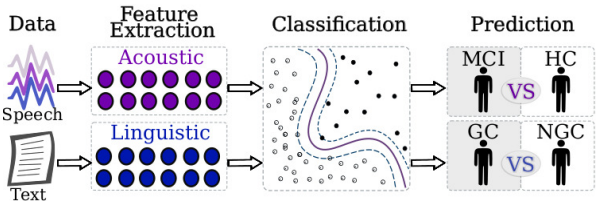

Fig. 1: Scheme of the general methodology addressed in this study

3.1. Acoustic analysis

The classical acoustic analysis in this study is based on the articulation and prosody features proposed previously in [15]. The transitions from unvoiced to voiced segments (onset) and from voiced to unvoiced segments (offset) are detected based on the presence of the fundamental frequency. Chunks of 40 ms are taken to the left and to the right of each border to form the transitions. The source code to compute the following articulation and prosody features is available online[1]

Articulation: It is based on the energy content and the formant frequencies. The energy content is modeled considering 22 BBEs, and 12 MFCCs along with their delta, and delta-delta. Both methods were computed in onset / offset transitions separately. The feature set is completed with the first two formant frequencies and their delta, and delta-delta, computed from all of the voiced segments in the signal. From these 122 descriptors, four statistical functionals are computed (mean, standard deviation, kurtosis, and skewness), forming a 488-dimensional feature vector per utterance.

Prosody: It is based on the F_0, the energy, and the speech rate. The contour of the F_0 and the energy are computed. F_0 aims to model intonation patterns, while energy aims to model intensity of the voice. The tilt and the mean square error are computed from the contours. The same four functionals are computed for these descriptors. Additionally, features based on duration measures were also considered. A total of 103 descriptors are extracted.

3.2. Speaker-embeddings

Speaker recognition/verification models compress relevant information about the identity of the speaker in a low-dimensional representation. Chunks of 25 ms, mean normalized over a sliding window of up to 3 seconds were considered to compute the following embeddings.

i-vectors: These embeddings are an extension of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) combined with the concept of Joint Factor Analysis (JFA) model for speaker verification proposed in [16]. This approach assumes that the speaker and the channel share the same subspace, i.e., each subspace is not decoupled covering the total variability in the utterance. The GMM-UBM models were trained using the Vox-Celeb [17] dataset

[1] https://github.com/jcvasquezc/DisVoice

Table 1: General information of the subjects in the Genetic Alzheimer's Dataset

	MCI	HC	GC	NGC
Gender [F / M]	12 / 11	15 / 12	16 / 12	22 / 14
Age [F / M]	48.1 (5.5) / 51.0 (7.9)	49.5 (7.7) / 53.2 (7.7)	31.5 (5.6) / 31.8 (5.0)	32.1 (6.4) / 33.5 (4.8)
Education [F / M]	5.8 (3.4) / 7.3 (6.2)	7.3 (3.6) / 8.8 (4.6)	10.6 (2.8) / 11.0 (3.3)	13.5 (2.6) / 12.4 (2.7)
MMSE [F / M]	25.0 (4.3) / 25.7 (2.3)	28.5 (1.4) / 28.8 (1.2)	29.4 (0.8) / 29.5 (0.9)	29.5 (0.9) / 29.6 (0.8)
MoCA [F / M]	14.9 (5.2) / 15.7 (4.0)	20.7 (4.2) / 22.3 (5.4)	24.5 (2.3) / 24.3 (2.7)	25.5 (2.3) / 25.9 (2.7)

MCI patients: AD patients with mild cognitive impairment. **GC** subjects: asymptomatic genetic carriers. **NGC** subjects: asymptomatic non-genetic carriers. **HC** subjects: healthy control subjects. Values are expressed as mean (standard deviation). F: female. M: male. Age and education are given in years.

by extracting 23 MFCCs together with the log energy, and considering the respective delta and delta-delta. The i-vectors model the variability present in the GMM-UBM super-vector by using JFA to project them into a low-dimensionality space, forming a 400-dimensional feature vector.

x-vectors: This method consists of a Deep Neural Network (DNN) approach proposed in [18] as an alternative to i-vectors. x-vectors aim to model characteristics to discriminate between speakers, unlike i-vectors, which represents the speaker and channel variability. The model works on 30 MFCCs, and on a pre-trained five-layer Time Delay Neural Network (TDNN) that models the temporal context in variable length speech segments. The last layer (512 units) is taken as the speaker-embedding representation. The pre-trained model uses augmented Vox-Celeb [17].

3.3. Linguistic embeddings

NLP approaches are considered to perform linguistic analysis to process information from the transcripts. Those methods are based on the "word-embeddings" paradigm.

Word2Vec: This method consists of a Neural Network (NN) with two layers to reconstruct the linguistic contexts. The inputs of the NN are "one-hot encodings" representation, i.e., binary vectors of the term in the vocabulary. The activations of the hidden layer are stored as "word vectors" during the training step. The NN is trained using neighbor words to predict a target word, which is known as the continuous bag of words algorithm. The number of neighbor words for each context is selected depending on the "window size" which was set to 5. For this study, we considered the Spanish Billion Word Corpus (SBWC), which contains 1.5 billion words [19]. The length of the word-embedding vector was set to 300. Finally, the four statistical functionals are computed, resulting in 1200-dimensional feature vectors.

Bidirectional Encoder Representations from Transformers: It is an unsupervised and deep bidirectional pre-trained model proposed in [20]. Unidirectional models consider previous words to predict a target word, unlike bidirectional models that use the previous and the following words. This allows the words in the corpus to be represented into lower dimensional feature vectors based on the encoder part from "Transformers" method [21]. We considered two different pre-trained BERT models: (1) Bert-Base trained with the Multi-Genre Natural Language Inference (MultiNLI) corpus, that was translated from English to Spanish, and (2) Bert-Base trained with the Spanish Unannotated Corpora [22]. This last model is mostly known as BETO. The last layer (768 units) is taken as the word-embedding representation. The four functionals are computed over all word-embeddings to form a 3072-dimensional static vectors. Instead of sentence embeddings, word-embeddings are considered to compare this method with W2V. The source code to compute the BERT and BETO embeddings is also available online[2] [23].

3.4. Optimization and classification

The classification was performed using a Radial Basis Function-Support Vector Machine (RBF-SVM). The validation process is a modification of the regular Leave-One-Speaker-Out (LOSO) strategy. During the regular LOSO, the meta-parameters of the classifier have to be decided on the best parameters in the development for all the N speakers. However, these results are optimistic, since up to N parameter sets are found, i.e., N different classifiers. The proposed validation in this study used the regular strategy with an internal 6-fold cross-validation to optimize the hyper-paramaters of the SVM. N different optimal hyper-parameters were obtained and stored. The found settings were sorted and the median of all of these values was obtained in order to have only one C and one γ. Finally, the LOSO strategy was performed again with the fixed parameters. The optimal parameters of the RBF-SVM were found through a grid search where $C \in \{10^{-6}, 10^{-3}, ..., 10^{6}\}$ and $\gamma \in \{10^{-6}, 10^{-3}, ..., 10^{6}\}$. The optimization criterion was the F-score obtained in development, and as a tiebreaker method the Area Under the Curve (AUC). The classification using early fusion strategy was performed. It consisted of merging by concatenating linguistic and acoustic features before performing the classification and making the final decision.

4. EXPERIMENTS AND RESULTS

The experiments consider two classification tasks: (1) MCI vs. HC, and (2) GC vs. NGC. Other classification tasks are not considered to avoid the influence of aging. Kruskal-Wallis test with Bonferroni correction was performed to evaluate

[2] https://github.com/PauPerezT/WEBERT

whether there was a significant difference between groups. The null hypothesis of the medians coming from the same distribution was rejected ($p \ll 0.05$) for all features sets and both classification problems. For the experiments, we use the standard python library Scikit-learn [24]. Table 2 shows the

Table 2: Results of each feature set separately

Features	# of Features	F-score	UAR	Sens	Spec	AUC
MCI-AD vs HC						
Articulation	488	0.66	66.0	69.6	63.0	0.70
Prosody	103	0.66	66.0	56.5	74.1	0.70
i-vectors	400	0.66	66.0	60.9	70.4	0.71
x-vectors	512	0.56	56.0	52.2	59.3	0.59
W2V	1200	0.48	48.0	39.1	55.6	0.46
BERT	3072	0.50	50.0	43.5	55.6	0.45
BETO	3072	0.50	50.0	39.1	59.3	0.55
GC vs NGC						
Articulation	488	0.47	46.9	39.3	52.8	0.47
Prosody	103	0.67	67.2	60.7	72.2	0.70
i-vectors	400	0.52	54.7	28.6	75.0	0.47
x-vectors	512	0.63	62.5	57.1	66.7	0.58
W2V	1200	0.53	53.1	39.3	63.9	0.52
BERT	3072	0.68	68.8	50.0	83.3	0.74
BETO	3072	0.65	65.6	53.6	75.0	0.72

UAR: unweighted average recall. **Sens**: sensitivity w.r.t. MCI/GC. **Spec**:specificity w.r.t. HC/NGC. **AUC**: Area under the ROC curve. UAR, sensitivity, and specificity are given in [%].

classification results considering each feature set individually. Lines highlighted in dark gray represent the acoustic features, in light gray the speaker embedding features, and in white the word-embedding features. The most accurate results for MCI vs. HC are obtained using acoustic features and i-vectors. Prosody and BERT produce the highest results for GC vs. NGC. Despite the fact that BERT obtained a higher F-score than prosody, sensibility and specificity are more balanced for prosody. The comparison with the regular LOSO strategy was performed, where the difference between the unweighted average recall between both validation approaches was 3.2% on average, which confirms that the regular LOSO is more optimistic. Table 3 shows the results using the early fusion

Table 3: Top five classification results using early fusion of the different feature sets

Features	# of Features	F-score	UAR	Sens	Spec	AUC
MCI-AD vs HC						
Art + Pro	591	0.74	74.0	65.2	81.5	0.77
Art + Pro + i-vec	631	0.74	74.0	69.6	77.8	0.75
Art + i-vec	888	0.72	72.0	69.6	74.1	0.71
Art + Pro + x-vec	1103	0.71	70.0	69.6	70.4	0.71
Pro + i-vec	503	0.70	70.0	60.9	77.8	0.68
GC vs NGC						
i-vec + x-vec + W2V	2112	0.70	70.3	53.6	83.3	0.80
i-vec + BERT	3472	0.67	67.2	50.0	80.6	0.68
Art+ + Pro + i-vec + BERT	4063	0.67	67.2	53.6	77.8	0.70
Art + i-vec + W2V + BERT	5160	0.66	67.2	48.9	86.1	0.63
x-vec + BERT	3584	0.65	65.6	50.0	77.8	0.67

UAR: unweighted average recall. **Sens**: sensitivity w.r.t. MCI/GC. **Spec**: specificity w.r.t. HC/NGC. **AUC**: area under the ROC curve. Art: articulation. Pro: prosody. i-vec: i-vectors. x-vec: x-vectors. UAR, sensitivity, and specificity are given in [%].

strategy. In general, when articulation and prosody features are combined, the performance of the classifier improves in comparison with each feature set separately. The most accurate results on discriminating GC subjects are obtained with the combination of i-vectors, x-vectors, and W2V. The effectiveness in the word-embeddings methods are directly linked to the number of known words by the predefined vocabulary in the corpus on which they were trained. Note that for this classification task the linguistic embeddings do not achieve satisfactory results. This may occur due to some unknown words by the algorithms related to characteristic lexicon from the region, or mispronunciations of some words. The unknown words on average are 21.4 % of the total of words per utterance using the word-embedding methods. Regarding BERT and BETO models, the results with BERT were slightly higher, which concludes that for our approach the translation from English to Spanish in BERT did not show a strong impact on the results.

5. CONCLUSIONS

This study proposed to discriminate genetic AD as well as EOA related to the *Paisa mutation* by using acoustic and linguistic analyses. The classification between MCI patients with early-onset Alzheimer and HC subjects was performed. Early fusion of articulation and prosody features exhibited the highest performance. The linguistic-based analysis did not show satisfactory results, which may occur due to the proportion of unknown words that could affect the performance of models. The same features were used to classify GC vs. NGC. Good results were obtained considering the difficult task that involves the classification of two healthy groups without any AD symptom. The influence of depression in the GC subjects was discarded by performing a Man-Whitney U-test (p=0.89) between GC and NGC regarding the geriatric depression scale of Yesavage [25]. We are aware of the limitations of this study, the amount of data needs to be increased and we do not have enough demographic information yet. According to Table 1 and to the results, there is no bias at cognitive (MMSE, MoCA) and depression level between the groups, even though the machine learning algorithm was able to find significant differences above chance between GC vs. NGC. Therefore we need to investigate other possible causes that influenced the results. Other late fusion methods will be explored in future research to merge the different feature sets.

6. ACKNOWLEDGMENTS

This work was mainly funded by the Colombian Ministry of Science grant # 2018-18889, call 777. Additionally, the work was also funded by the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287, the CODI PRG2017-15530 project from University of Antioquia, and the Bayerisches Hochschulzentrum für Lateinamerika (BAYLAT) grant. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS.

7. REFERENCES

[1] M.A. Lalli et al., "Origin of the PSEN1 E280A mutation causing early-onset Alzheimer's disease," *Alzheimer's & Dementia*, vol. 10, pp. S277–S283, 2014.

[2] N. Acosta-Baena et al., "Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: a retrospective cohort study," *The Lancet Neurology*, vol. 10, no. 3, pp. 213–220, 2011.

[3] A. König et al., "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[4] M. F. Folstein et al., "The mini-mental state examination," *Archives of general psychiatry*, vol. 40, no. 7, pp. 812–812, 1983.

[5] Z. S. Nasreddine et al., "The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[6] P. Klumpp et al., "ANN-based Alzheimer's disease classification from bag of words," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–4.

[7] S. Wankerl et al., "An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language.," in *INTERSPEECH*, 2017, pp. 3162–3166.

[8] J. Fritsch and others., "Automatic diagnosis of Alzheimer's disease using neural network language models," in *ICASSP 2019-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5841–5845.

[9] M. L. B. Pulido et al., "Alzheimer's disease and automatic speech analysis: a review," *Expert Systems with Applications*, pp. 1–19, 2020.

[10] E. L. Campbell et al., "Alzheimer's Dementia Detection from Audio and Text Modalities," *arXiv preprint arXiv:2008.04617*, 2020.

[11] R. Pappagari et al., "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," in *INTERSPEECH*, 2020.

[12] G. Gosztolya et al., "Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[13] J. T. Becker et al., "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[14] H. Goodglass et al., "Cookie Theft picture," *Boston diagnostic aphasia examination. Philadelphia, PA: Lea & Febiger*, 1983.

[15] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, and E. Nöth, "Current methods and new trends in signal processing and pattern recognition for the automatic assessment of motor impairments: the case of Parkinson's disease," *Neurological Disorders and Imaging Physics, Volume 5*, pp. 1–57.

[16] N. Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[17] A. Nagrani et al., "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[18] D. Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[19] C. Cardellino, "Spanish Billion Words Corpus and Embeddings," `https://crscardellino.github.io/SBWCE/`, Aug. 2019.

[20] J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] A. Vaswani et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[22] J. Cañete et al., "Spanish Pre-Trained BERT Model and Evaluation Data," in *to appear in PML4DC at ICLR 2020*, 2020.

[23] P. A. Perez-Toro, "PauPerezT/WEBERT: Word Embeddings using BERT," `https://doi.org/10.5281/zenodo.3964244`, July 2020.

[24] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[25] J. A. Yesavage, "Geriatric depression scale," *Psychopharmacol Bull*, vol. 24, no. 4, pp. 709–711, 1988.