ORIGINAL ARTICLE
https://doi.org/10.30773/pi.2023.0052

Print ISSN 1738-3684 / On-line ISSN 1976-3026
OPEN ACCESS

A Case-Control Clinical Trial on a Deep Learning-Based Classification System for Diagnosis of Amyloid-Positive Alzheimer's Disease

Jong Bin Bae[1,2*], Subin Lee[3*], Hyunwoo Oh[4], Jinkyeong Sung[4], Dongsoo Lee[4], Ji Won Han[1,2], Jun Sung Kim[1,5], Jae Hyoung Kim[6], Sang Eun Kim[7,8], and Ki Woong Kim[1,2,3] ✉

[1]Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[2]Department of Psychiatry, Seoul National University, College of Medicine, Seoul, Republic of Korea
[3]Department of Brain and Cognitive Sciences, Seoul National University College of Natural Sciences, Seoul, Republic of Korea
[4]VUNO Inc., Seoul, Republic of Korea
[5]Institute of Human Behavioral Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea
[6]Department of Radiology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[7]Department of Nuclear Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[8]Center for Nanomolecular Imaging and Innovative Drug Development, Advanced Institutes of Convergence Technology, Suwon, Republic of Korea

Objective A deep learning-based classification system (DLCS) which uses structural brain magnetic resonance imaging (MRI) to diagnose Alzheimer's disease (AD) was developed in a previous recent study. Here, we evaluate its performance by conducting a single-center, case-control clinical trial.

Methods We retrospectively collected T1-weighted brain MRI scans of subjects who had an accompanying measure of amyloid-beta (Aβ) positivity based on a 18F-florbetaben positron emission tomography scan. The dataset included 188 Aβ-positive patients with mild cognitive impairment or dementia due to AD, and 162 Aβ-negative controls with normal cognition. We calculated the sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve (AUC) of the DLCS in the classification of Aβ-positive AD patients from Aβ-negative controls.

Results The DLCS showed excellent performance, with sensitivity, specificity, positive predictive value, negative predictive value, and AUC of 85.6% (95% confidence interval [CI], 79.8–90.0), 90.1% (95% CI, 84.5–94.2), 91.0% (95% CI, 86.3–94.1), 84.4% (95% CI, 79.2–88.5), and 0.937 (95% CI, 0.911–0.963), respectively.

Conclusion The DLCS shows promise in clinical settings where it could be routinely applied to MRI scans regardless of original scan purpose to improve the early detection of AD.

Psychiatry Investig 2023;20(12):1195-1203

Keywords Alzheimer disease; Magnetic resonance imaging; Clinical trial; Deep learning.

INTRODUCTION

The number of individuals with dementia is increasing globally, with more than 130 million people expected to live with dementia in 2050.[1] Alzheimer's disease (AD) is the most prevalent type and cause of dementia,[2] with no cure currently available yet. Early detection of AD is crucial because it allows for advanced treatment planning and improves prognosis. However, due to the insidious nature of the disease, more than 60% of people living with dementia in the community go undetected.[3] In order to improve the accuracy and advance the timing of AD diagnosis, the National Institute on Aging-Alzheimer's Association proposed new diagnostic criteria for AD that incorporates neuroimaging biomarkers such as amyloid beta (Aβ) deposition and neuronal degeneration.[4-7] Assessment of Aβ deposition (performed using positron emission tromography [PET]) is an earlier and more specific biomarker of AD than assessments of neurodegeneration (performed using magnetic structural imaging [MRI]). However, PET has practical drawbacks in clinical practice because they involve radiation and are not available in all clinical settings.

Structural brain MRI is more widely available but less ex-

Received: February 16, 2023 **Revised:** August 13, 2023
Accepted: September 12, 2023

✉ **Correspondence:** Ki Woong Kim, MD, PhD
Department of Neuropsychiatry, Seoul National University Bundang Hospital, 82 Gumi-ro 173beon-gil, Bundang-gu, Seongnam 13620, Republic of Korea
Tel: +82-31-787-7432, **Fax:** +82-31-787-4058, **E-mail:** kwkimmd@snu.ac.kr

*These authors contributed equally to this work.

pensive and invasive than PET. It can also detect structural changes related to many brain diseases other than AD. Several recent studies,[8-13] including our previous work,[14] have developed artificial intelligence (AI)-based algorithms for classifying AD using structural brain MRI, with promising results in terms of processing time and classification accuracy. In the case of our previous work, our deep learning-based classification system for AD using structural brain MRI (DLCS) demonstrated excellent accuracy in classifying probable AD patients from cognitively normal controls (area under the curve [AUC]=0.88–0.94). However, previous studies included the following limitations. First, most previous studies[9,10,12-14] are likely to have overestimated performance because the training and validation datasets were constructed by randomly splitting a single population into subsets, leaving its performance in newly seen data unknown. Second, it is unclear as to what proportion of true AD patients and normal controls were used in the previous studies, because most did not confirm the presence or absence of Aβ deposition in their AD patients and normal controls, respectively.[9-14] Aβ positivity is an important supporting evidence for presence of AD.[5] When considering that about 12% of clinically diagnosed probable AD patients are Aβ-negative[15] and 10%–40% of cognitively normal controls are Aβ-positive,[16] checking for the presence of Aβ is an important control factor in order to exclude any seemingly AD cases of different etiology from the AD group and to exclude preclinical AD cases from the normal control group.[14] Third, many studies[10-14] only included dementia patients in their AD group, which may have exaggerated performance in normal cognition (NC) vs. AD classification tasks. Including AD patients with mild cognitive impairment (MCI) is expected to offer a more comprehensive measure of the model's performance across the AD continuum.

In this study, we performed a clinical trial with a well-defined case-control population that can address the aforementioned limitations in our previous work. We investigated the performance of the DLCS in discriminating Aβ-positive patients with MCI or AD dementia (MCI/dementia due to AD) from Aβ-negative cognitively normal controls, all of whom were from a sample independent of the population used for the development of the DLCS.

METHODS

Study participants

A single-center, case-controlled clinical trial was conducted and registered in the Korean Clinical Trials Registry (KCT0004758, 21/02/2020). Data of subjects over 50 years of age who visited Seoul National University Bundang Hospital (SNUBH) and underwent a T1-weighted MRI scan between January 2010 and September 2019 were retrospectively collected. Our data include brain MRI scans with clinical assessment and 18F-florbetaben PET scans from visitors to dementia clinic at the Department of Neuropsychiatry in SNUBH as well as from participants of the Korean Longitudinal Study on Cognitive Aging and Dementia (KLOSCAD).[17]

A group of patients with AD and a group with NC matched for age and sex were screened and enrolled using the following inclusion criteria. The AD group included patients diagnosed as MCI due to AD or dementia due to AD,[5] according to the following criteria: 1) a diagnosis of probable or possible AD according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria, or MCI according to the International Working Group on MCI, and 2) amyloid deposition as determined by a positive 18F-florbetaben PET scan. The NC group included those who 1) had no subjective cognitive complaints, 2) had no objective cognitive decline in the Korean version of the Consortium to Establish a Registry for AD (CERAD-K) neuropsychological assessment battery, 3) were functioning independently in the community, and 4) had no amyloid deposition as determined by a negative 18F-florbetaben PET scan. Subjects who had any of the following conditions were excluded: 1) diagnosis of dementia with a cause other than or in addition to AD, i.e., mixed dementia, 2) brain pathologies on T1-weighted MRI that may cause cognitive deficits, 3) more than 1 year between the date of clinical assessment and date of MRI scan (NC and MCI participants only), and 4) white matter hyperintensities with a Fazeka's rating of 3 or higher on fluid-attenuated inversion recovery images.

The data of the participants were retrospectively screened and collected starting from April 27, 2020 to June 5, 2020 (6 weeks). The employment of the DLCS on the data were conducted between June 8, 2020 to June 19, 2020 (2 weeks).

Sample size calculation

We employed both the sensitivity and specificity of DLCS to AD as primary outcome measures. We calculated the sample size needed to evaluate whether DLCS performed better than a reference, based on a one-sided α of 2.5% (Z_{α}=1.96), statistical power of 80% ($Z_{1-\beta}$=0.842), and the results of a pilot study. The pilot study tested the performance of DLCS using a dataset consisting of 367 AD patients and 316 controls with NC: 130 AD and 130 NC from SNUBH and 237 AD and 186 NC from the Alzheimer's Disease Neuroimaging Initiative database. At a threshold value of 0.38, the DLCS yielded a sensitivity of 82.0% (95% confidence interval [CI], 77.7%–85.8%) and specificity of 83.2% (95% CI, 78.6%–87.2%). To calculate the sample size n, we used the following formula[18]:

$$n = \frac{(Z_\alpha\sqrt{p_0(1-p_0)} + Z_{1-\beta}\sqrt{p_1(1-p_1)})^2}{(p_1-p_0)^2},$$

where p_0 is the assumed sensitivity/specificity under the null hypothesis H_0, and p_1 is the targeted sensitivity/specificity under alternative hypothesis H_1. The p_0 and p_1 values were defined as the lower and higher bounds of the 95% CI of the sensitivity and specificity from the pilot study (p_0=0.777 and p_1=0.858 for sensitivity; p_0=0.786 and p_1=0.872 for specificity). The null hypothesis was that the sensitivity/specificity of the DLCS is less than or equal to the lower boundary of the assumed sensitivity/specificity. The alternative hypothesis was that it is higher. Based on this, the necessary number of subjects with the disease was 188, and the number of subjects without the disease was 162. Therefore, the final estimated sample size was 350 subjects, consisting of 188 patients with AD and 162 normal controls that were matched for age (<5 years apart) and sex to the AD group.

Image acquisition

We acquired three-dimensional (3D) T1-weighted MR images in Digital Imaging and Communications in Medicine format using Philips Achieva and Ingenia scanners (Philips Medical Systems, Eindhoven, The Netherlands). The parameters were as follows: voxel dimensions=1.0×0.5×0.5 mm^3, slice thickness=1.0 mm, echo time=8.15 or 8.20 ms (for Achieva and Ingenia, respectively), repetition time=4.61 ms, flip angle=8°, and field of view=240×240 mm.

We acquired 18F-florbetaben PET scans in 3D using a Discovery VCT scanner (General Electric Medical Systems, Milwaukee, WI, USA). The subjects were injected with 8.1 mCi (300 MBq) 18F-florbetaben (Neuraceq; Life Molecular Imaging Ltd., Berlin, Germany) through a slow single intravenous bolus (6 MBq) in a total volume of 10 mL. After a 90-min uptake period, 20-min PET images comprising four 5-min dynamic frames were obtained. Images of each time frame were reframed into one summed frame. Board-certified nuclear medicine physicians then determined Aβ-positivity based on visual interpretation of tracer uptake in the gray matter compared to neighboring subcortical white matter in the following four brain regions: the temporal lobes, frontal lobes, posterior cingulate cortex/precuneus, and parietal lobes.

DLCS

We used VUNO Med-DeepBrain AD (version 1.0.0; VUNO Inc., Seoul, South Korea), which is the DLCS for AD. The convolutional neural network model used in VUNO Med-DeepBrain AD has been previously described.[14] Briefly, the DLCS uses as its backbone the Inception-V4 architecture, which is a 2D image classification convolutional neural network that achieved very good performance with low computational cost.[19] The network uses pretrained weights (https://github.com/Cadene/pretrained-models.pytorch#inception) obtained from a subset of ImageNet,[20] which is a training dataset of 1.28 million nautral images. The DLCS receives a subject's T1-weighted image, extracts 30 coronal slices from areas that span the medial temporal lobe, and feeds each coronal slice as a separate input into the pretrained Inception-V4 architecture. From this, the network extracts various features that include structural and textural information of the brain from the coronal slice. The feature vector is then concatenated with the subject's age and sex information (which is input to the system at the beginning with the MRI scan) and the location information (slice number) of the coronal slice. The concatenated feature vector is entered into a fully connected network that calculates the probability of the slice belonging to that of a patient with AD. The probabilities of each slice are averaged to calculate a final score that represents the subject's probability of having AD (score ranges from 0 to 1).

In this clinical trial, we processed the MRI data of subjects anonymously, omitting information that could identify the individual (name, sex, birth date, and hospital number). A researcher (K.J.S.), who was blinded to the subjects' clinical diagnoses and did not participate in the construction of the study dataset, performed the processing of the subjects' data with DLCS. The DLCS was installed on a desktop PC with the following specifications: Intel hexa-core 2.90 GHz CPU with 16 GB RAM running on Ubuntu 18.04.4 LTS.

Statistical analysis

We evaluated the accuracy of the DLCS in the diagnosis of AD by comparing its output (a continuous probability ranging from 0 to 1) with the subject's clinical diagnosis. We defined sensitivity and specificity as the primary outcomes, and the area under the receiver operating characteristic curve (AUC) as the secondary outcome. We calculated the outcomes on the whole dataset, as well as in sex, age, and Mini-Mental State Examination (MMSE) subgroups. The age groups were divided into ≥75 years and <75 years of age, based on the median age of 75 years. The MMSE subgroups were divided into ≥26 and <26 scores. In addition, in a subgroup of subjects with MMSE scores, we compared the performance of DLCS with that of MMSE score for the diagnosis of all AD (MCI and dementia due to AD as well as of just MCI due to AD), on the basis of the following evaluation metrics: AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

For demographics, continuous variables were compared using independent samples t-test, and categorical variables were compared using the chi-square test between groups. We estimated the 95% CI of sensitivity and specificity using the Clopper-Pearson method[21] and the AUC using the DeLong test.[18] For comparison of evaluation metrics between sex-, age-, and MMSE-subgroups, we used chi-square test. In the comparison of evaluation metrics between the DLCS and MMSE, we used McNemar test[22] for comparing accuracy, sensitivity, and specificity, and two-sample z-test for comparing PPV and NPV. All statistical analyses were performed using IBM SPSS, version 20 (IBM Corp., Armonk, NY, USA) and MedCalc (version 16.4.3; MedCalc Software, Mariakerke, Belgium).

Standard protocol approvals, registrations, and patient consents

This clinical trial (Korean Clinical Trials Registry identifier: KCT0004758) was approved by the Ministry of Food and Drug Safety in South Korea and the Institutional Review Board of SNUBH (E-2001/588-001). The design and conduct of this study were in accordance with the principles outlined in the Declaration of Helsinki.[23] Because this clinical trial was conducted retrospectively, participation consent forms from subjects or legal guardians of the subjects were waived. All methods were carried out in accordance with relevant guidelines and regulations.

RESULTS

We enrolled a total of 350 subjects who met the eligibility criteria, with 162 (46.3%) in the NC group and 188 (53.7%) in the AD group. The demographic and clinical characteristics of the participants are summarized in Table 1. The mean age of the whole dataset was 73.3±7.23 (range, 55 to 92) years. Age and sex were comparable between the NC and AD groups, while years of education were higher in the NC group. In the patient group, 76 (40.4%) had MCI due to AD, the rest had dementia due to AD (12 [6.4%] possible AD and 99 [52.7%] probable AD). All participants with MCI due to AD had a clinical dementia rating (CDR) score of 0.5. Among the 112 participants with AD dementia, 68 (60.7%) had a CDR score of 0.5, 35 (31.3%) had a CDR score of 1, and the rest (8.0%) had a CDR score of 2 or 3. The models of MR scanners were comparable, while the type of head coil was different between

Table 1. Subject characteristics

Characteristics	NC (N=162)*	AD (N=188)†	t or χ^2‡	p‡
Age (yr)	73.3±6.9	73.9±7.4	-0.8	0.42
Age band			13.650	0.009
50–59 yr	0 (0.0)	12 (6.4)		
60–69 yr	46 (28.4)	36 (19.1)		
70–79 yr	84 (51.9)	96 (51.1)		
80–89 yr	32 (19.7)	43 (22.9)		
≥90 yr	0 (0.0)	1 (0.5)		
Female	108 (66.6)	125 (66.5)	0.001	0.97
Education (yr)	12.4±4.5	11.1±4.9	2.57	0.01
MMSE (score)	27.5±2.2	20.9±4.9	16.21	<0.001
MRI				
Scanner			1.64	0.44
Philips Achieva	137 (84.6)	167 (88.8)		
Philips Ingenia	20 (12.3)	18 (9.6)		
Philips Ingenia CX	5 (3.1)	3 (1.6)		
Head coil			63.79	<0.001
SENSE-Head-8	73 (45.1)	39 (20.7)		
SENSE-NV-16	15 (9.3)	89 (47.4)		
Dual coil	42 (25.9)	39 (20.7)		
Multi coil	32 (19.7)	21 (11.2)		
WMH (cc)	14.3±29.6	16.0±17.5	-0.64	0.52

Values are presented as mean±standard deviation or number (%). *cognitively normal having clinical dementia rating of 0 and amyloid-β-negative; †amyloid-β-positive dementia due to AD or amyloid-β-positive mild cognitive impairment due to AD; ‡Student's t-test for continuous variables and chi-square test for categorical variables. NC, normal control; AD, Alzheimer's disease; MMSE, Mini-Mental State Examination; WMH, white matter hyperintensity; MRI, magnetic resonance imaging the two groups. The volume of white matter hyperintensity was comparable between the NC and AD groups.

As summarized in Figure 1, the DLCS demonstrated a good diagnostic performance in the classification of AD. The DLCS had a sensitivity for AD of 85.6% (95% CI, 79.8%−90.3%), and the lower bound of 95% CIs for its sensitivity was higher than the assumed value of 77.7%. Its specificity for AD was 90.1% (95% CI, 84.5%−94.2%), and the lower bound of 95% CIs for its specificity was higher than the assumed value of 78.6%. Its accuracy, PPV, and NPV for AD were 87.7%, 91.0%, and 84.4%, respectively. The AUC of DLCS for AD classification was 0.937 (95% CI, 0.911−0.963).

The distribution of the DLCS probability scores for each cognitive group, NC, MCI, and dementia, are shown as histograms and boxplots in Figure 2.

Further analyses were conducted in various subgroups, summarized in Table 2. When analyzed separately in the male and female subgroup (n=117 and 233, respectively), there were no significant differences in sensitivity, specificity, and AUC. When analyzed separately in each age group, sensitivity was higher in the ≥75 years group (n=186), while specificity was higher in the <75 years group (n=164). There were no significant age-wise differences in AUC. In the comparison between participants stratified by MMSE score (≥26 vs. <26), the AUC was comparable between the two MMSE subgroups. However, specificity was higher in the subgroup with MMSE of ≥26 (n=166), whereas sensitivity was higher in the subgroup with MMSE of <26 (n=179).

Table 3 summarizes results of the comparative analysis between DLCS and MMSE in a subgroup who have MMSE scores. In the diagnosis of AD (n=345), the AUC of DLCS (AUC, 0.936; 95% CI, 0.905−0.960) was larger than the AUC of MMSE score (AUC, 0.907; 95% CI, 0.871−0.935) by a marginally larger value (p=0.0718). The sensitivity and specificity of DLCS were 85.8% (95% CI, 79.9%−90.5%) and 90.1% (95% CI, 84.5%−94.2%). At a cutoff value of 25 indicated by the Youden's index, the MMSE had a sensitivity of 78.1% (95% CI, 71.4%−83.9%) and specificity of 91.4% (95% CI, 85.9%−

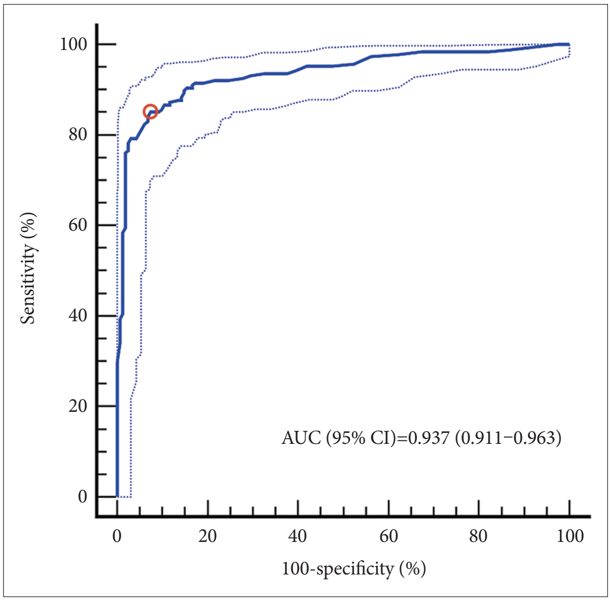

Figure 1. Receiver operating characteristic curve of Alzheimer's disease discrimination. The ROC curve and AUC (95% CI) for the DLCS in the discrimination between amyloid-negative normal controls and amyloid-positive patients with Alzheimer's disease are shown. The threshold value of 0.38 is shown as the red circle. Results for other evaluation metrics (with 95% CI in parentheses) are: ACC=87.7 (83.8–91.0), SEN=85.6 (79.8–90.3), SPE=90.1 (84.5–94.2), PPV=91.0 (86.3–94.1), and NPV=84.4 (79.2–88.5). AUC, area under the curve; CI, confidence intervals; ROC, receiver operator characteristic; DLCS, deep learning-based classification system; ACC, accuracy; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value.

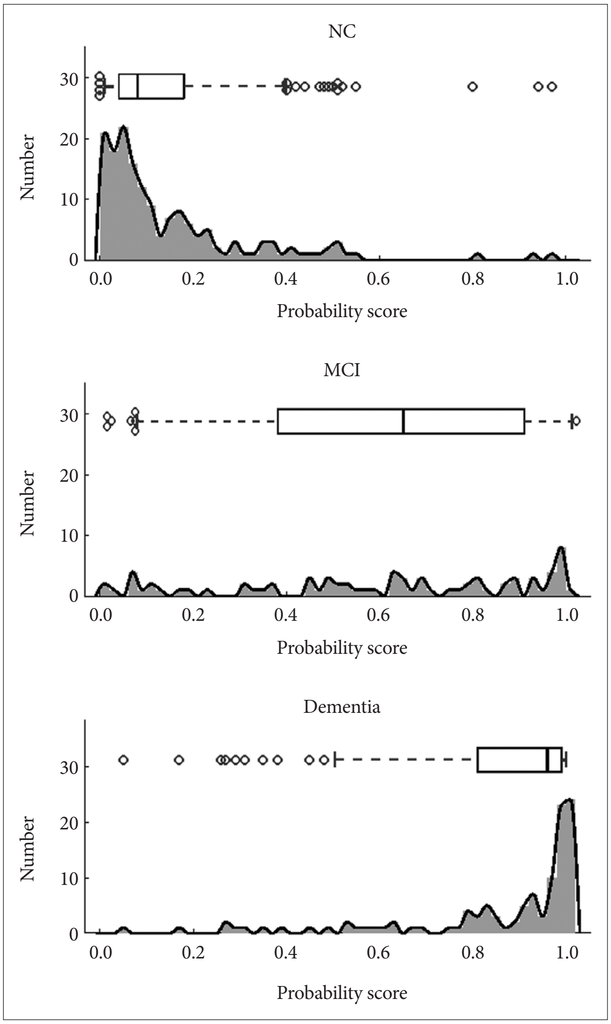

Figure 2. AD probability scores for each clinical diagnostic group. Histograms and boxplots of the generated probability scores, ranging from 0 to 1, for NC and patients with MCI and dementia. In the boxplots, the line indicate the median, the box indicates 25% and 75% quartiles, and the whiskers bound the 9% and 91%. AD, Alzheimer's disease; NC, normal control; MCI, mild cognitive impairment.

Table 2. Performance metrics according to sex, age, and Mini-Mental Status Examination score

Variable	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Sex			
Male	0.931 (0.882–0.980)	82.5 (70.9–90.9)	92.6 (82.1–97.9)
Female	0.941 (0.911–0.972)	87.2 (80.0–92.5)	88.9 (81.4–94.1)
p	0.730	0.523	0.642
Age			
≥75 yr	0.952 (0.922–0.982)	93.1 (86.2–97.1)	83.5 (73.9–90.7)
<75 yr	0.936 (0.896–0.976)	77.0 (66.8–85.4)	97.4 (90.9–99.7)
p	0.541	0.003	0.007
MMSE			
≥26 points	0.853 (0.790–0.903)	61.3 (42.2–78.2)	92.6 (86.8–96.4)
<26 points	0.907 (0.854–0.945)	90.8 (85.0–94.9)	77.8 (57.8–91.4)
p	0.346	<0.001	0.019

AUC, area under the curve; CI, confidence intervals; MMSE, Mini-Mental Status Examination

Table 3. Comparison of diagnostic performance metrics between DLCS and Mini-Mental Status Examination

Variable	DLCS	MMSE	p
For all AD			
AUC	0.936 (0.905–0.960)	0.907 (0.871–0.935)	0.0718
Accuracy	0.878 (0.839–0.911)	0.844 (0.801–0.880)	0.0440
Sensitivity (%)	85.8 (79.9–90.5)	78.1 (71.4–83.9)	0.0336
Specificity (%)	90.1 (84.5–94.2)	91.4 (85.9–95.2)	0.8145
PPV	90.8 (86.0–94.0)	91.1 (86.0–94.4)	0.9243
NPV	84.9 (79.7–88.9)	78.8 (73.7–83.0)	0.1313
For MCI due to AD			
AUC	0.877 (0.829–0.916)	0.823 (0.768–0.869)	0.0922
Accuracy	0.849 (0.797–0.892)	0.811 (0.755–0.859)	0.0725
Sensitivity (%)	73.7 (62.3–83.1)	59.2 (47.3–70.4)	0.0522
Specificity (%)	90.1 (84.5–94.2)	91.4 (85.9–95.2)	0.8145
PPV	77.8 (68.3–85.0)	76.3 (65.3–84.6)	0.8392
NPV	88.0 (83.3–91.4)	82.7 (78.4–86.3)	0.1618

Values in the parentheses are 95% confidence intervals. DLCS, deep learning-based classification system; MMSE, Mini-Mental Status Examination; AD, amyoid-β-positive Alzheimer's disease; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value

95.2%). The sensitivity of DLCS was stastistically significantly higher than that of MMSE (p=0.034). In the diagnosis of patients with MCI due to AD (n=238), the AUC of DLCS (AUC, 0.877; 95% CI, 0.829–0.916) was marginally higher than that of MMSE (AUC, 0.823; 95% CI, 0.768–0.869). The sensitivity and specificity of DLCS were 73.7% (95% CI, 62.3%–83.1%) and 90.1% (95% CI, 84.5%–94.2%), and that of MMSE were 59.2% (95% CI, 47.3%–70.4%) and 91.4 (95% CI, 85.9%–95.2%), respectively.

DISCUSSION

In this clinical trial, we evaluated the use and performance of a DLCS in the identification of an individual's risk of AD. To the best of our knowledge, this is the first clinical trial in the field of AI-based AD diagnosis to obtain regulatory approval as a class III medical device from the Ministry of Food and Drug Safety.

The results of this clinical trial demonstrate that the DLCS has excellent diagnostic performance for AD, according to the criteria of excellent biomarkers proposed by the Ronald and Nancy Reagan Research Institute of the Alzheimer's Association and the National Institute on Aging Working Group on "Molecular and Biochemical Markers of Alzheimer's Disease."[24] The working group suggested that an excellent evaluating biomarker should have a sensitivity approaching or exceeding 85%, a specificity of approximately 75%–85% or greater, and a PPV of approximately 80% or more. The sensitivity, specificity, and PPV of the DLCS were 85.8%, 90.1%, and 90.8%, respectively, which met the requirements proposed by the working group. This is also comparable to or better than that of existing methods. The clinical diagnosis of probable AD according to the NINCDS-ADRDA criteria has shown a sensitivity of 70.9%–75.3% and a PPV of 59.5%–70.8% for autopsy-proven AD.[25] Fluorodeoxyglucose PET has sensitivities of 84% and 75.8% and specificites of 74% and 74.3% for autopsy-proven AD[26] and amyloid PET-proven AD,[27] respectively. Cerebral blood flow single-photon emission computed tomography (SPECT) showed sensitivies of 63% and 42.9% and specificities of 82% and 82.9% for autopsy-proven AD[28] and amyloid PET-proven AD,[27] respectively.

The DLCS was evaluated in relation to convetional measures such as MMSE score. As seen in Table 2, the notable specificity of the DLCS for AD in a population with high MMSE score is pivotal because it offers the potential to be used in conjunction with highly sensitive quick cognitive tests in AD diagnosis. Also as seen in Table 3, the DLCS exhibited a slightly higher AUC and significantly higher sensitivity in detecting dementia or MCI due to AD compared to MMSE. This has clinical implications, given that the preclinical and prodromal phases of AD last for more than a decade. The DLCS shows potential to be used as a means to increase the sensitivity of the diagnosis of AD, especially when combined with cognitive tests such as the MMSE.

The rate of undetected dementia in community-dwelling elderly is pooled to be 61.7% (95% CI=55.0%−68.0%), according to a meta-analysis.[3] The rate of undetected cases of dementia due to AD, which is the largest cause of dementia, is even higher because of its slow progressive onset.[29] This warrants a method that can effectively and accurately diagnose AD in its early stages. While amyloid PET[30] has the potential to detect AD in its preclinical stages, their use in clinical settings is restricted to cases of diagnostic uncertainty in patients with cognitive impairment.[31] In contrast, structural brain MRI is also capable of detecting early changes[32-35] and have the additional benefit of being widely administered for a variety of purposes such as diagnosing various types of dementia and neurologic disorders and even for health checkups. Thus, utilizing structural MRI allows for an opportunity to screen a broader range of individuals for possible dementia.

The DLCS offers a structural brain MRI-based diagnosis of AD, with strengths that allow ease of use in clinical practice. The DLCS takes 3 simple inputs−patient age, patient gender, and a T1-weighted MRI scan of the patient. Once uploaded, it outputs the result (probability of the subject having AD), within 23 seconds. The DLCS uses a pre-learned neural network, which eliminates the need for other preprocessing steps,[14] and extracts information that is expected to comprehensively reflect volumetric, shape, and textural information. The short processing time and simplicity of use make it feasible to use in clinical settings, in contrast to other previously developed structural MRI-based diagnostic markers[32-35] that have good diagnostic performance but require heavy data processing and longer processing time. Furthermore, the DLCS can be embedded to clinical routine workflow such that MRI scans taken for whichever purpose can be quickly and efficiently screened for possible AD. This can extend AD screening to patients who had visited the hospital for reasons other than cognitive complaints, and may help clinicians to catch potential AD cases that may otherwise go unnoticed and direct them for a fuller battery of tests that can lead to a timely diagnosis of AD.

To accurately assess the performance of DLCS for AD, our study population consisted of amyloid PET-confirmed cases of AD and non-AD. This contrasts with most previous studies which used clinical diagnosis as the criteria for defining the AD group and control group.[9-13] Clinically determined AD dementia patients may have AD-like symptoms but not actually have AD pathology, and cognitively normal-seeming individuals may have underlying amyloid pathology in which case they would be recognized as preclinical AD.[16] Thus we used amyloid PET results to prevent enrollment of non-AD dementia cases to the AD patient group and preclinical AD cases to the normal control group, thereby minimizing misclassification bias.

Second, we included patients with MCI due to AD in the patient group so that we can see the performance of DLCS in detecting AD from varying degrees of cognitive deficits along the AD continuum.[5] This allows for a more accurate assessment of the software's ability to discriminate AD across the cognitive severity spectrum.

Identifying AD in MCI patients due to AD proved to be a challenge, as shown in the wide distribution of probability scores for MCI patients in Figure 2. This may be attributed to MCI being a heterogenous group with structural brain changes that are not as different from that of NC. It was also observed in the age and sex subgroup analysis that the proportion of MCI patients in the AD group may affect the DLCS performance. The relatively lower sensitivity values in the male and younger groups may be associated with the higher proportion of MCI patients in the AD group in the male subgroup (47.6%) compared to the female subgroup (37.6%), and in the younger group (48.3%) compared to the older group (34.7%). However, it is noteworthy that our proposed method was trained only on NC and possible/probable AD patients, and that the MCI patients are completely newly seen data to the DLCS. This supports the potential of MRI-based deep learning methods for AD classification, and further studies involving training with MCI patients included should lead to more promising results in the future.

It should be noted that composition of AD patients can be different across studies, depending on their definition of the AD group. In general, AD diagnosis is used to refer to detection of AD patients as defined by clinical diagnostic criteria. In this paper, because our interest was in detecting patients with amyloid PET-proven AD, AD diagnosis refers to detection of MCI due to AD and dementia due to AD,[5] which requires amyloid positivity in addition to clinically diagnosed

MCI and clinically diagnosed AD. This diagnostic criteria of AD[5] was proposed for research purposes but it is also widely used in clinical practice,[36] which supports our use of it for this clinical study. It should also be noted that there are also cases in clinical settings where the clinical trajectory seems like that of AD but are amyloid-negative (suspected non-Alzheimer's pathophysiology, SNAP). Future studies focusing on discriminating between SNAP and AD will be necessary to ensure accurate detection of AD.

There are several technical considerations to be addressed. First, all MRI scans used in this study were acquired from a single scanner (Philips) using the same protocol. Therefore, the performance of the DLCS on scans from other vendors or protocols remains to be determined. Second, the DLCS currently only takes 3D T1-weighted images as input data because 3D scans contain higher anatomical detail and resolution than conventional 2D scans. However, 3D scans are not available in all clinical settings, which may restrict the use of DLCS to fewer settings. Third, it is not clear which features contribute to the predictions made by the DLCS, which can undermine the explainability of the results. Increasing the explainability and interpretability of deep learning algorithms will be crucial in increasing the trustworthiness of the technology for use in the medical domain. This is an unresolved issue that is currently the topic of many recent research.[37]

In conclusion, DLCS, a software as a medical device using structural brain MRI, demonstrated excellent diagnostic performance for MCI or dementia due to AD. When used together during screening of MRI, taken for whichever purpose, DLCS may help improve the early detection of AD.

Availability of Data and Material

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

Ki Woong Kim, a contributing editor of the *Psychiatry Investigation*, was not involved in the editorial evaluation or decision to publish this article. J.B.B., S.L., J.S.K., J.W.H., and K.W.K. received royalty income from VUNO Inc.

H.O, J.S., and D.L. are employees at VUNO Inc. All remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Jong Bin Bae, Subin Lee, Hyunwoo Oh, Jinkyeong Sung. Data curation: Jong Bin Bae, Subin Lee. Formal analysis: Jong Bin Bae, Subin Lee. Funding acquisition: Ki Woong Kim. Investigation: Jong Bin Bae, Subin Lee. Methodology: Jong Bin Bae, Subin Lee, Hyunwoo Oh. Project administration: Jun Sung Kim, Ji Won Han. Resources: Ji Won Han, Jae Hyoung Kim, Sang Eun Kim. Software: Hyunwoo Oh, Dongsoo Lee. Supervision: Ki Woong Kim. Validation: Jong Bin Bae, Subin Lee. Visualization: Subin Lee, Dongsoo Lee. Writing—original draft: Jong Bin Bae, Subin Lee. Writing—review & editing: all authors.

ORCID iDs

Jong Bin Bae	https://orcid.org/0000-0002-3913-1011
Subin Lee	https://orcid.org/0000-0001-7583-6468
Hyunwoo Oh	https://orcid.org/0000-0002-0952-2046
Jinkyeong Sung	https://orcid.org/0000-0003-3546-6081
Dongsoo Lee	https://orcid.org/0000-0002-7057-745X
Ji Won Han	https://orcid.org/0000-0003-2418-4257
Jun Sung Kim	https://orcid.org/0000-0002-4579-8218
Jae Hyoung Kim	https://orcid.org/0000-0002-0545-4138
Sang Eun Kim	https://orcid.org/0000-0003-1434-8369
Ki Woong Kim	https://orcid.org/0000-0002-1103-3858

Funding Statement

This study was supported by a grant from the Korean Health Technology R&D Project, Ministry for Health, Welfare, and Family Affairs, Republic of Korea [grant number HI09C1379 (A092077)] and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (2018-0-00861, Intelligent SW Technology Development for Medical Data Analysis).

REFERENCES

1. Alzheimer's Disease International. World Alzheimer Report 2015 - The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends. London: Alzheimer's Disease International; 2015.
2. Kim KW, Park JH, Kim MH, Kim MD, Kim BJ, Kim SK, et al. A nationwide survey on the prevalence of dementia and mild cognitive impairment in South Korea. J Alzheimers Dis 2011;23:281-291.
3. Lang L, Clifford A, Wei L, Zhang D, Leung D, Augustine G, et al. Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. BMJ Open 2017;7:e011146.
4. Jack CR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:257-262.
5. 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;7:270-279.
6. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280-292.
7. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia 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;7:263-269.
8. Li H, Habes M, Wolk DA, Fan Y; Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Study of Aging. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimers Dement 2019;15:1059-1070.
9. Li F, Liu M; Alzheimer's Disease Neuroimaging Initiative. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. Comput Med Imaging Graph 2018;70:101-110.
10. Luo S, Li X, Li J. Automatic Alzheimer's disease recognition from MRI data using deep learning method. J Appl Math Phys 2017;5:1892-1898.
11. Liu M, Zhang J, Adeli E, Shen D. Landmark-based deep multi-instance learning for brain disease diagnosis. Med Image Anal 2018;43:157-

168.

12. Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. Neuroimage Clin 2019;21:101645.
13. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF; Alzheimer's Disease Neuroimaging Initiative. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. Sci Rep 2018;8:5697.
14. Bae JB, Lee S, Jung W, Park S, Kim W, Oh H, et al. Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging. Sci Rep 2020;10:22252.
15. Ossenkoppele R, Jansen WJ, Rabinovici GD, Knol DL, van der Flier WM, van Berckel BN, et al. Prevalence of amyloid PET positivity in dementia syndromes: a meta-analysis. JAMA 2015;313:1939-1949.
16. Jansen WJ, Ossenkoppele R, Knol DL, Tijms BM, Scheltens P, Verhey FR, et al. Prevalence of cerebral amyloid pathology in persons without dementia: a meta-analysis. JAMA 2015;313:1924-1938.
17. Han JW, Kim TH, Kwak KP, Kim K, Kim BJ, Kim SG, et al. Overview of the Korean Longitudinal Study on Cognitive Aging and Dementia. Psychiatry Investig 2018;15:767-774.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-845.
19. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Association for the Advancement of Artificial Intelligence, editor. Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17); 2017 Feb 4-9; San Francisco, CA, USA. Washington, DC: AAAI Press; 2017. p.4278-4284.
20. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: IEEE, editor. 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20-25; Miami, FL, USA. New York, NY: IEEE; 2009. p.248-255.
21. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26:404-413.
22. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153-157.
23. General Assembly of the World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. J Am Coll Dent 2014;81:14-18.
24. The Ronald and Nancy Reagan Research Institute of the Alzheimer's Association; National Institute on Aging Working Group. Consensus report of the Working Group on: "Molecular and biochemical markers of Alzheimer's disease". Neurobiol Aging 1998;19:109-116. Erratum in: Neurobiol Aging 1998;19:285
25. Beach TG, Monsell SE, Phillips LE, Kukull W. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. J Neuropathol Exp Neurol 2012;71:266-273.
26. Jagust W, Reed B, Mungas D, Ellis W, Decarli C. What does fluorodeoxyglucose PET imaging add to a clinical diagnosis of dementia? Neurology 2007;69:871-877.
27. Nadebaum DP, Krishnadas N, Poon AMT, Kalff V, Lichtenstein M, Villemagne VL, et al. Head-to-head comparison of cerebral blood flow single-photon emission computed tomography and (18) F-fluoro-2-deoxyglucose positron emission tomography in the diagnosis of Alzheimer disease. Intern Med J 2021;51:1243-1250.
28. Jagust W, Thisted R, Devous MD Sr, Van Heertum R, Mayberg H, Jobst K, et al. SPECT perfusion imaging in the diagnosis of Alzheimer's disease: a clinical-pathologic study. Neurology 2001;56:950-956.
29. Sternberg SA, Wolfson C, Baumgarten M. Undetected dementia in community-dwelling older people: the Canadian Study of Health and Aging. J Am Geriatr Soc 2000;48:1430-1434.
30. La Joie R, Ayakta N, Seeley WW, Borys E, Boxer AL, DeCarli C, et al. Multisite study of the relationships between antemortem [(11)C]PIB-PET Centiloid values and postmortem measures of Alzheimer's disease neuropathology. Alzheimers Dement 2019;15:205-216.
31. Laforce R, Rosa-Neto P, Soucy JP, Rabinovici GD, Dubois B, Gauthier S. Canadian consensus guidelines on use of amyloid imaging in Canada: update and future directions from the specialized task force on amyloid imaging in Canada. Can J Neurol Sci 2016;43:503-512.
32. Dickerson BC, Bakkour A, Salat DH, Feczko E, Pacheco J, Greve DN, et al. The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb Cortex 2009;19:497-510.
33. Fox NC, Warrington EK, Freeborough PA, Hartikainen P, Kennedy AM, Stevens JM, et al. Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study. Brain 1996;119:2001-2007.
34. Gerardin E, Chételat G, Chupin M, Cuingnet R, Desgranges B, Kim HS, et al. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 2009;47:1476-1486.
35. Lee S, Lee H, Kim KW; Alzheimer's Disease Neuroimaging Initiative. Magnetic resonance imaging texture predicts progression to dementia due to Alzheimer disease earlier than hippocampal volume. J Psychiatry Neurosci 2020;45:7-14.
36. Bocchetta M, Galluzzi S, Kehoe PG, Aguera E, Bernabei R, Bullock R, et al. The use of biomarkers for the etiologic diagnosis of MCI in Europe: an EADC survey. Alzheimers Dement 2015;11:195-206.e1.
37. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2020;32:4793-4813.