European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-019-04595-y

ORIGINAL ARTICLE

The clinical feasibility of deep learning–based classification of amyloid PET images in visually equivocal cases

Hye Joo Son[1] · Jungsu S. Oh[1] · Minyoung Oh[1] · Soo Jong Kim[1] · Jae-Hong Lee[2] · Jee Hoon Roh[2] · Jae Seung Kim[1]

Received: 16 May 2019 / Accepted: 4 November 2019

Abstract

Purpose Although most deep learning (DL) studies have reported excellent classification accuracy, these studies usually target typical Alzheimer's disease (AD) and normal cognition (NC) for which conventional visual assessment performs well. A clinically relevant issue is the selection of high-risk subjects who need active surveillance among equivocal cases. We validated the clinical feasibility of DL compared with visual rating or quantitative measurement for assessing the diagnosis and prognosis of subjects with equivocal amyloid scans.

Methods ^{18}F-florbetaben scans of 430 cases (85 NC, 233 mild cognitive impairment, and 112 AD) were assessed through visual rating–based, quantification–based, and DL–based methods. DL was trained using 280 two-dimensional PET images (80%) and tested by randomly assigning the remaining (70 cases, 20%) cases and a clinical validation set of 54 equivocal cases. In the equivocal cases, we assessed the agreement among the visual rating, quantification, and DL and compared the clinical outcome according to each modality-based amyloid status.

Results The visual reading was positive in 175 cases, equivocal in 54 cases, and negative in 201 cases. The composite SUVR cutoff value was 1.32 (AUC 0.99). The subject-level performance of DL using the test set was 100%. Among the 54 equivocal cases, 37 cases were classified as positive (Eq(deep+)) by DL, 40 cases were classified by a second-round visual assessment, and 40 cases were classified by quantification. The DL- and quantification-based classifications showed good agreement (83%, κ = 0.59). The composite SUVRs differed between Eq(deep+) (1.47 [0.13]) and Eq(deep−) (1.29 [0.10]; $P < 0.001$). DL, but not the visual rating, showed a significant difference in the Mini-Mental Status Examination score change during the follow-up between Eq(deep+) (− 4.21 [0.57]) and Eq(deep−) (− 1.74 [0.76]; $P = 0.023$) (mean duration, 1.76 years).

Conclusions In visually equivocal scans, DL was more related to quantification than to visual assessment, and the negative cases selected by DL showed no decline in cognitive outcome. DL is useful for clinical diagnosis and prognosis assessment in subjects with visually equivocal amyloid scans.

Keywords Deep learning · Alzheimer's disease · Amyloid · ^{18}F-florbetaben PET · Equivocal scan

This article is part of the Topical Collection on Neurology.

Hye Joo Son and Jungsu S.Oh contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00259-019-04595-y) contains supplementary material, which is available to authorized users.

✉ Jee Hoon Roh
alzheimer@naver.com

✉ Jae Seung Kim
jaeskim@amc.seoul.kr

Hye Joo Son
firstlady1231@gmail.com

Jungsu S. Oh
jungsu_oh@amc.seoul.kr; jungsu.oh@gmail.com

[1] Department of Nuclear Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea

[2] Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea

Published online: 06 December 2019

Introduction

Amyloid β (Aβ) positron emission tomography (PET) can affect a physician's diagnostic confidence and treatment for suspected Alzheimer's disease (AD) patients [1]. The applicability of amyloid PET as an imaging biomarker depends on the practicality and accuracy of image interpretation. Approximately 10% of AD patients present equivocal amyloid PET images [2]. The critical question regarding the functional implication of equivocal scans is whether equivocal scans are indicative of positive status, an independent entity with unique clinical characteristics, or a combination of subgroups that can be reclassified as positive or negative. As a fundamental limitation in the interpretation of equivocal scans, there is limited histopathology to use as a reference, and discrepancies exist between visual and quantitative measurements [3].

These challenges have stimulated fully automated deep learning (DL)–based analyses for AD diagnosis. Convolutional neural networks (CNNs) represent a DL architecture in which the convolutional layers and pooling layers are stacked one on top of another [4]. The intrinsic parameters of CNNs can be adjusted until the most predictive representation is found directly from the images [5]. Such systems have several benefits, including the reproducibility of interpretation, high accuracy, and rapid output of results [6].

Several DL-based studies using magnetic resonance imaging (MRI) [7, 8], functional MRI [9], and FDG PET [10] have been published. Although most studies report an excellent classification accuracy of 80%, they have generally targeted typical AD or normal cognition (NC). For typical cases, conventional visual assessment performs well, and therefore, DL has little additional value. A more clinically relevant issue is selecting high-risk subjects who are eligible for active surveillance and early treatment candidates among visually equivocal cases with a near-threshold standardized uptake value ratio (SUVR). We implemented a CNN-based DL algorithm for the automated interpretation of equivocal amyloid scans. We validated the clinical feasibility of the CNN-based DL algorithm for assessing the diagnosis and prognosis of patients with equivocal amyloid scans.

Materials and methods

Study population

Subjects were recruited from the Florbetaben Imaging in Alzheimer's and Related Neurological Conditions (FLORIAN) cohorts at Asan Medical Center between February 2015 and December 2017. A total of 430 patients—85 NC, 233 mild cognitive impairment (MCI), and 112 AD—underwent T1 volumetric MRI, ^{18}F-florbetaben PET/CT at baseline, and cognitive measures at baseline and follow-up. ^{18}F-florbetaben PET images were assessed for Aβ positivity through visual reading, quantification, and DL. The study has been approved by the Asan Medical Center institutional review board (2013-0847, 2014-0783, 2016-0588, 2016-0589, 2016-0590), and all subjects signed an informed consent form. See the Supplement for details regarding image acquisition and processing.

Visual grading

The visual grading was performed by two board-certified nuclear medicine physicians blinded to the clinical category. In the first round, images were classified as positive, equivocal, or negative based on the agreement between readers. Equivocal scan was defined as a case in which the evaluations by the two readers did not completely match. In the second round, equivocal scans were reclassified as positive (Eq(second visual+)) or negative (Eq(second visual−)) in the consensus read by two expert readers. Positive scans were defined as higher uptake in gray matter than in white matter in the majority of slices within at least one of four brain regions [11, 12]. Negative scans were defined as lower uptakes in gray matter than in white matter with clear gray-white matter contrast in all four brain regions [11, 12]. Equivocal scans were defined as any other findings other than typical positive or negative scans.

DL algorithm

We developed a two-dimensional (2D) deep CNN for scoring slice-level amyloid positivity (Supplemental Fig. 1). Before deciding to use 2D CNN to score the slice-level amyloid positivity for our purposes, we compared the diagnostic performances (i.e., amyloid positive/negative classification accuracy) of 2D CNN and 3D CNN in a test set and in a clinical validation set composed of 54 visually equivocal cases. After removing the peripheral pixels (the upper 6 slices and lower 6 slices from 33 slices, the left 20 pixels and right 20 pixels from 128 pixels in the left-right orientation, and the anterior 20 pixels and posterior 20 pixels from 128 pixels in anterior-posterior orientation), we could achieve 97% patch-level accuracy and 100% subject-level accuracy in the test set, which was similar to the 2D CNN–based results (95% slice-level accuracy and 100% subject-level accuracy). However, in the clinical validation set composed of equivocal cases, the accuracy of the 3D CNN–based classification with an experienced expert readers' visual assessment as the gold standard was 44%, which is much lower than that of 2D CNN–based classification (69%) (Supplemental Table 1).

The 2D CNN was trained using 2D axial slices (matrix size of 128 × 128, 33 slices per subject), and the visual assessment performed by the experts as a gold standard in each axial slice.

Since our DL system was trained using a training set that consisted of typical positive and negative cases, we could use the visual assessment performed by the experts as a ground truth. Image intensity was normalized in the range of 0–1 using whole-brain maximum uptake. The data were randomly divided into training and testing sets on the subject level. The deep CNN was trained on 80% of positive or negative cases (280 cases including 134 positive and 146 negative cases) and tested on the randomly assigned remaining 20% of cases (70 cases). We implemented an independent clinical validation set of 54 visually equivocal cases. For equivocal cases, we did not label equivocal scans by visual assessment or quantification. We defined equivocal scans as cases for which the two expert visual readers could not reach a consensus in the first-round visual assessment. Since there was no pathologic gold standard for equivocal scan, we compared the agreement of the DL-based assessment with visual- and quantification-based assessment and investigated whether the DL-based classification provides clinically useful information about the prognosis.

To increase the amount of training data multiplied by the epoch number (i.e., 300), we conducted slice-based augmentation using an axial rotation range of 15° and a scaling range of 0.9–1.1. The images passed the 16 channels of the 2D convolutional layer, producing 16 feature maps using 3 × 3 kernels and no stride was applied during the convolution. Blocks of a 2D convolutional layer and a leaky rectified linear unit (ReLU) layer were cascaded to efficiently use combinations of smaller number of filter banks of cascaded convolution. Subsequently, a max-pooling layer (2 × 2 pooling size) was used. The dropout rate was 0.25. Four convolution layers and one fully connected layer were used for assigning slice-based positivity.

Finally, the outputs of the aforementioned CNN (i.e., slice-level amyloid positivity) were fed into input layer of an additional fully connected network with one hidden layer for classifying subject-level amyloid positivity (i.e., for determining if the subject has amyloid positive scan or not).

Interpretation of feature vectors

High-dimensional feature vectors (512 nodes) were condensed to 50 dimensions using principal component analysis (PCA) and subsequently decreased to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) [13]. The t-SNE was visualized in a 2D scatter plot with each point matching an individual image in the feature space and we attempted to minimize the distances between similar features while maximizing those between different features [13].

Statistics

Categorical variables were compared using the chi-square test. Parametric data were analyzed using Student's t test or one-way ANOVA, and nonparametric data were analyzed using the Mann-Whitney U test. Intermethod agreement was assessed using Cohen's or Fleiss' κ. Analysis of covariance (ANCOVA) was performed using amyloid status as an independent variable and clinical indices as covariates. SPSS for Windows, version 18.0 (SPSS, Chicago, IL, USA) was used. $P < 0.05$ was considered statistically significant.

Results

Visual reading and quantitative measurement of PET images

Visual readings were positive for 175 patients (composite SUVR, 1.71 [0.18]), equivocal for 54 patients (1.41 [0.15]), and negative for 201 patients (1.14 [0.06]; effect size = 0.79, $P < 0.001$; Supplemental Table 2, Supplemental Fig. 2). Patients in the equivocal group (74.8 [7.3]) were older than those in the negative (69.8 [10.0]) and positive (69.5 [9.9]) groups ($P = 0.001$) (Table 1). The ROC cutoff of composite SUVR in visual negative NC and positive AD cases was 1.32 (AUC 0.99).

Performance of CNN on test set

The classification accuracy of the CNN on the test set with visual reading as a reference was 100% at the subject level and 95% at the slice level. Visual-, DL-, and quantification-based classifications showed excellent agreement (97.9%, Fleiss' $\kappa = 0.97$).

Concordance among visual-, quantitative-, and DL-based classifications of equivocal cases

Among 54 equivocal cases, 37 (68.5%) were classified as positive by DL (Eq(deep+)), 40 (74.1%) by a second-round visual assessment, and 40 (74.1%) by the SUVR cutoff (Eq(quantification+)). DL and the quantification showed good agreement (Supplementary Table 3a; $\kappa = 0.59$; 95% CI, 0.36–0.83) in 45 of 54 scans (83.0%) (Fig. 1; sectors I (positive concordance) and III (negative concordance)). In contrast, DL and second-round visual assessment yielded poor agreement (Supplementary Table 3c; $\kappa = 0.23$; 95% CI, − 0.04 to 0.51), showing discordance in 17 of 54 scans (31.5%) (Fig. 1; the red circle in sectors I and II ($n = 7$) and the blue circle in sectors III and IV ($n = 10$)). Negative cases (negative agreement 13%) showed greater discordance than positive cases (positive agreement 56%).

Table 1 Patient characteristics

	Negative (n = 201)	Equivocal (n = 54)	Positive (n = 175)	Total (n = 430)	P value (all groups)
Age (years)	69.8 (10.0)	74.8 (7.3)	69.5 (9.9)	70.3 (9.8)	0.001
Sex (M/F)	76/125	21/33	62/113	159/271	0.850
Education years (SD)	9.2 (5.4)	9.9 (6.1)	10.3 (6.4)	9.7 (5.9)	0.170
MMSE total score (SD)	25.8 (4.3)	24.2 (4.7)	20.9 (4.9)	23.7 (5.1)	< 0.001
CDR score (SD)	0.4 (0.4)	0.5 (0.2)	0.7 (0.4)	0.6 (0.4)	< 0.001
BAPL					
1	201	17	0	218	< 0.001
2	0	32	1	33	
3	0	5	174	179	
Clinical diagnosis					
AD	21	13	83	117	< 0.001
MCI	106	35	87	228	
NC	74	6	5	85	

MMSE, Mini-Mental State Examination; *CDR*, Clinical Dementia Rating; *BAPL*, brain amyloid plaque load

Among 17 discordant subjects, ten were Eq(deep−/second visual+); of them, two had focal and asymmetric parietal or temporo-occipital uptake, four had focal posterior cingulate uptake, and four had both findings. Seven Eq(deep+/second visual−) subjects showed globally increased cortical uptake. There was no difference in clinical indices between the discordant and concordant groups (Supplemental Table 4). Figure 2 presents four representative equivocal cases with different combinations of negative or positive categorization by DL or second-round visual assessment.

Comparison between DL-based positive and negative groups in visually equivocal cases

Composite SUVRs differed between Eq(deep+) (1.47 [0.13]) and Eq(deep−) (1.29 [0.10]; P < 0.001) (Fig. 3a), but not between Eq(second visual+) (1.47 [0.14]) and Eq(second visual−) (1.36 [0.16], P = 0.14) (Fig. 3b). Additionally, when comparing the composite SUVR between the DL-based positive and negative groups among the visual negative cases, Eq(deep+/visual−) (1.47 [0.13]) showed a higher composite SUVR than the Eq(deep−/visual−) (1.29 [0.15], P = 0.07) (Supplemental Fig. 3). Compared with Eq(deep−), Eq(deep+) showed higher uptakes in the bilateral frontal and cingulate, the left parietal and temporal cortex, and the right postcentral and superior temporal gyrus (Fig. 4a). Eq(deep−) showed higher uptakes in AD signature areas, including the bilateral frontal, temporal, parietal, and occipital cortices (Fig. 4b), while Eq(deep+) showed higher uptake in all brain regions compared with that of NC (Fig. 4c).

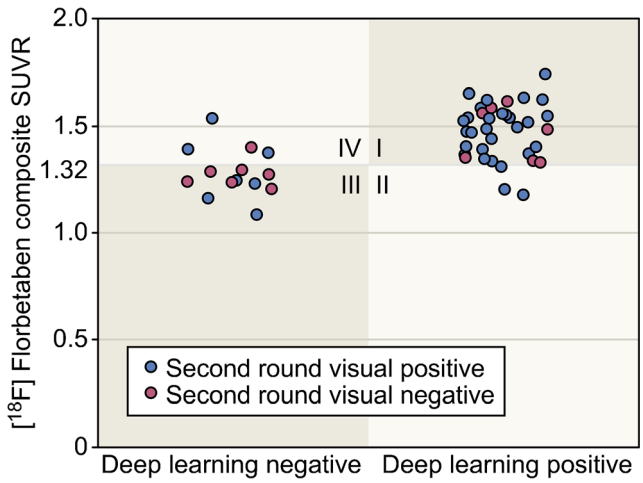

Fig. 1 Concordance between DL-based, quantitative-based, and second-round visual rating–based categorization in equivocal cases

Visualization of feature vectors

In PCA and t-SNE, positive and negative groups formed two distinct clusters connected by Eq(deep+) and Eq(deep−), demonstrating a continuum pattern (Fig. 5).

Comparison of the impact of DL- or second-round visual rating–based amyloid status on clinical outcome in equivocal amyloid PET cases

Among 54 equivocal cases, subsample analyses including MMSE and GDS tests were conducted on 34 cases (retention rate, 63%) who were eligible to return after 1.76 years of follow-up. Patients who participated in follow-up MMSE and GDS tests did not differ from those who did not participate in sex, age, clinical diagnosis, years of education, and medication status (Supplemental Table 5). Eq(deep+) was less educated and consisted of more AD and fewer NC than Eq(deep−), although there were no differences in sex, age, follow-up duration, or medication status (Supplemental

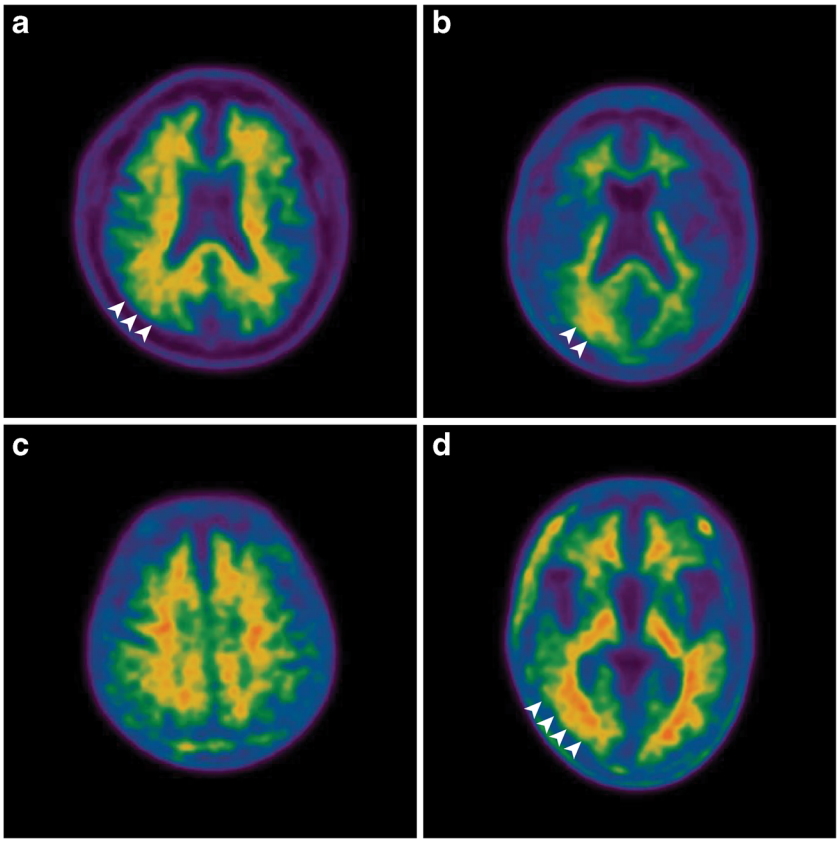

Fig. 2 Representative ^{18}F-florbetaben PET images of concordant and discordant results between DL-based and second-round visual rating–based classification. **a** A 62-year-old NC male showing well-preserved contrast between gray and white matter with mild focal uptake in the right parietal cortex (arrowheads) resulted in negativity for both visual assessment and the DL method. SUVRs were 1.19 (composite) and 1.31 (right parietal). Baseline and follow-up MMSE scores were 28 and 29, respectively. **b** A 71-year-old MCI female with focal, asymmetric increased uptake in the right parietal cortex (arrowheads) resulted in visual positivity but DL negativity. SUVRs were 1.39 (composite) and 1.68 (right posterior cingulate). Baseline and follow-up MMSE scores were 24 and 25, respectively. **c** A 77-year-old MCI female with globally increased gray matter uptake, especially in the bilateral frontal and superior parietal cortex, resulted in visual negativity but DL positivity. SUVRs were 1.60 (composite) and 1.70 (parietal). Baseline and follow-up MMSE scores were 17 and 14, respectively. **d** A 75-year-old MCI female with extensively increased uptake in the right temporo-occipital cortex (arrowheads) resulted in positivity for both visual assessment and the DL method. SUVRs were 1.38 (composite) and 1.33 (right temporal). Baseline and follow-up MMSE scores were 27 and 21, respectively

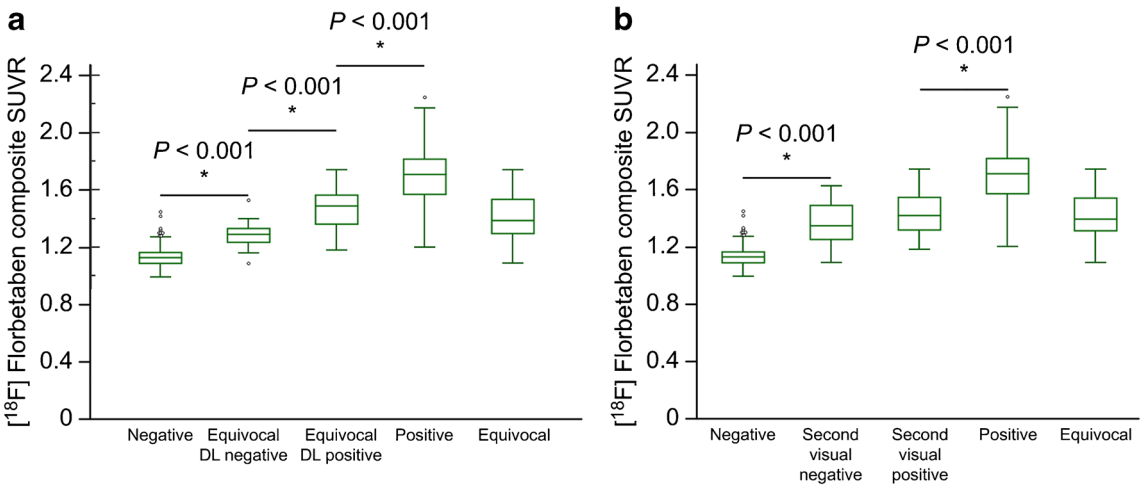

Fig. 3 Composite SUVR distribution according to **a** DL-based or **b** second-round visual rating–based amyloid status in visually equivocal cases. Boxes with median, 25%, and 75% quartiles. Whiskers extended to 1.5 × interquartile range

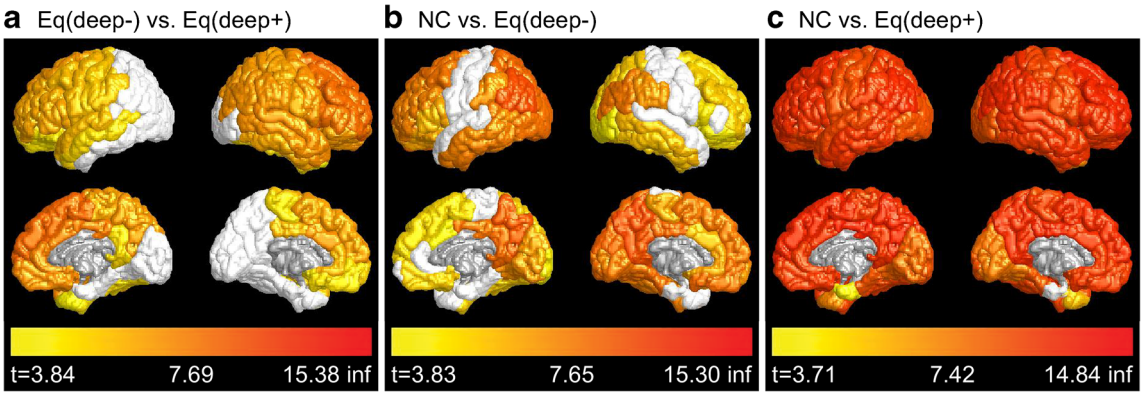

Fig. 4 Visual comparison between **a** Eq(deep−) and Eq(deep+), **b** NC and Eq(deep−), and **c** NC and Eq(deep+). Two-sample *t* tests were computed at each FreeSurfer parcellation VOI ($P = 0.05$) corrected for family-wise error of multiple comparisons. Inf indicates infinity

Table 6). The effect of visual- or DL-based amyloid status on cognitive decline was estimated using clinical indices as covariates (Supplemental Table 7, Fig. 6). The follow-up MMSE score was significantly lower than the baseline score in 34 patients (baseline, 24.09 [4.11]; follow-up, 20.82 [5.72]; $P < 0.001$). There was a significant difference in the MMSE score change during the follow-up between Eq(deep+) (− 4.21 [0.57]) and Eq(deep−) (− 1.74 [0.76]) in DL-based classification ($P = 0.02$). There was a significant difference in the MMSE score change during the follow-up between Eq(quantification+) (− 3.88 [2.94]) and Eq(quantification−) (− 1.25 [2.12]) ($P = 0.03$). Intriguingly, there was no difference in the MMSE score change during the follow-up between Eq(second visual+) (− 3.00 [0.51]) and Eq(second visual−) (− 4.28 [1.05]) ($P = 0.30$) (Supplemental Table 7). However, in the MCI group, there was no difference in the 18-month AD conversion ratio between Eq(deep+) (0.3 (7/23)) and Eq(deep−) (0.2 (2/9)).

Discussion

The incorporation of DL can automate the interpretation of equivocal amyloid scans. We investigated agreement among DL, visual, and quantification methods as well as clinical implications for predicting future cognitive outcomes by DL-based Aβ positivity. DL was more related to the quantification than to visual assessment, and negative cases selected by DL showed no decline in cognitive outcomes.

Computer-based image analysis has been applied to AD/NC classification in neuroimaging field. A support vector machine study reported a sensitivity of 85.2% and a specificity of

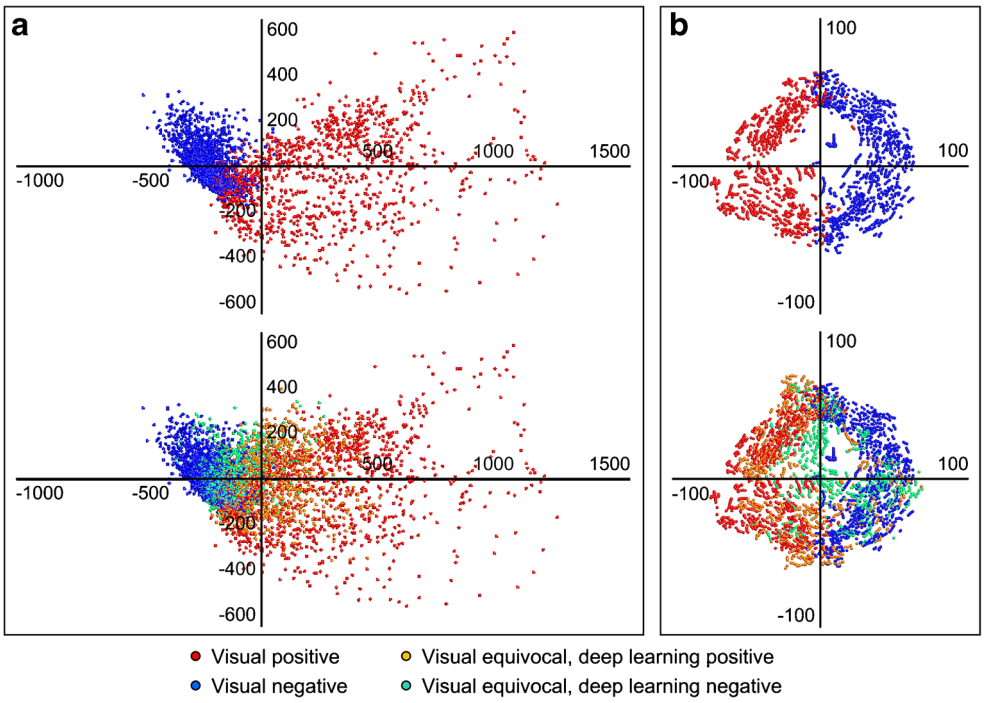

Fig. 5 Visualization of feature vectors with **a** PCA and **b** t-SNE

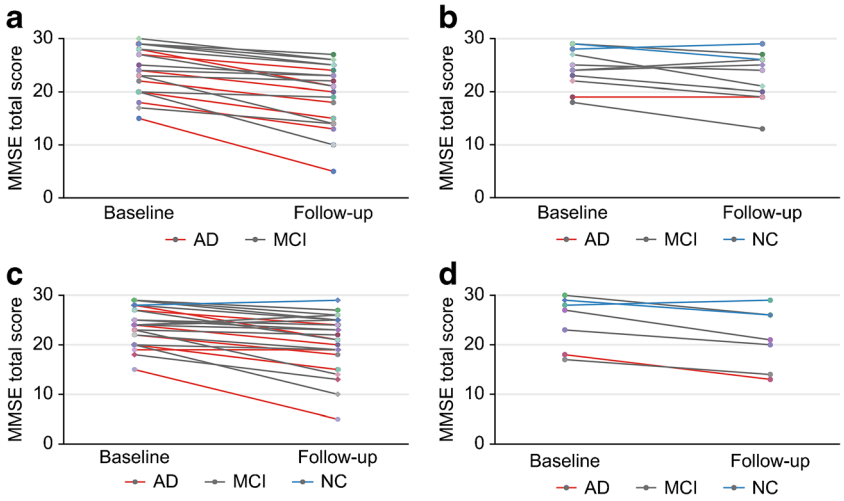

Fig. 6 Comparison of the impact of DL-based or second-round visual rating–based amyloid status on clinical outcome in patients with visual equivocal amyloid PET images. **a** Visual equivocal, DL-positive. **b** Visual equivocal, DL-negative. **c** Second-round visual rating–based positive. **d** Second-round visual rating–based negative groups

92.0% for AD/NC classification [14]. However, conventional machine learning did not perform well on raw data and required a complicated feature selection process [5]. These problems are avoided in a CNN-based method where features can be learned automatically. FDG PET–based CNN has a reported sensitivity of 100% and a specificity of 82% for AD/NC classification, outperforming a visual reader's performance (57% sensitivity, 91% specificity) [15]. However, to the best of our knowledge, no prior study has used equivocal scans as a validation set in the field of DL.

Regarding the definition of equivocal scans, Hosokawa et al. [2] focused on uptake intensity and defined it as slightly higher gray matter uptake than that in white matter. Payoux et al. [16] focused on the interrater agreement. To broadly incorporate definitions suggested in prior studies, we defined equivocal scans as images that did not fulfill the definitions of typical positive or negative scans [17]. In a study, nine of 11 equivocal cases showed cognitive impairments and FDG distribution compatible with AD, suggesting that an equivocal scan was indicative of a positive scan with mild uptake [2]. An AV45-PET study presented a contradictory viewpoint in that an equivocal scan was not a clinically separate but a quantitatively independent entity [16]. We presented a new perspective on equivocal scans, which comprised of two subgroups with differences in quantification and prognosis, and DL could distinguish between them. Model visualization with t-SNE revealed that Eq(deep+) and Eq(deep−) formed two discrete clusters that connect positive and negative clusters.

An integrated application of visual and quantitative approaches may be the optimal method for identifying true positive or negative cases; therefore, DL adds little value. However, in 10% of all cases with near-threshold SUVR, visual and quantitative results are discordant [3]. Herein, Aβ positivity between DL and visual analysis was discordant in 31.5% of equivocal cases, in contrast to the 100% agreement in the test set comprising typical cases. A total of 71% of discordant cases were multidomain amnestic MCI. DL makes diagnostic predictions in a different way from how humans interpret images. During visual inspection, clinicians determine amyloid positivity based on geographic gray-white matter differentiation and are sensitive to focal uptake, while focal uptake may be averaged by VOI-based quantification [18]. Herein, ten cases showing focal and asymmetric uptake led to visual positivity but DL negativity. However, DL placed a greater emphasis on the absolute pixel intensity related to quantitative information. Seven cases showing globally increased uptake without specific local uptake led to DL positivity but visual negativity. In a FDG PET–based CNN study, the saliency map suggested that DL considers the whole brain as a pixel-by-pixel volume [15]. Yuan et al. [19] estimated 3D CNN–based SUVR from florbetapir PET images directly without using conventional target and reference regions and reported a high correlation (0.97) between the original SUVR and 3D CNN–estimated SUVR.

Another original finding with clinical implications is that DL-based but not visual rating–based classification provided valuable information about future cognitive outcome. In eauivocal cases, there was no difference in clinical indices between follow-up and no follow-up groups. Eq(deep+) was less educated and comprised more AD and fewer NC cases than Eq(deep−). Most cases—100% of Eq(deep+) and 84.6% of Eq(deep−)—were taking a cholinesterase inhibitor or an NMDA receptor antagonist at some time points during the study. After covariate adjustment for clinical indices, cognitive outcomes differed between DL-based but not visual rating–based classification of the positive and negative groups. Although the evidence is still limited to a few studies, the clinical and biological significance of a non-negligible borderline Aβ elevation within the visual negative range has been recently investigated. In a longitudinal study using ^{18}F-florbetapir PET imaging in an Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (mean age 74.7 ±

7.0), among 142 baseline florbetapir-negative individuals, 13/142 (9.2%) individuals converted to a florbetapir-positive status over 3.9 ± 1.4 years (3.3% per year), although most individuals (130/142, 91.5%) maintained in florbetapir-negative status [20]. Another study investigated the relationships between postmortem pathology and centiloid (CL) quantitative measure of antemortem ^{11}C-PIB PET. Among the 27 visually negative cases, 26 cases had CL values < 12.2, and the other case was slightly above the threshold (13.8), but the postmortem neuropathology showed positive amyloid pathology [21]. DL was more related to quantification than to visual assessment and may detect borderline levels of amyloid deposition, which may be missed by current binary visual criteria. Also, the prediction of prognosis by the DL-based classification is similar to that by the quantification-based classification. A difference in quantitative burden may influence the difference in prognosis between Eq(deep+) and Eq(deep−), suggesting DL may carry clinically meaningful information regarding future cognitive decline. Choi et al. [22] reported a CNN-based quantitative biomarker, ConvScore, which was significantly correlated ($R = -0.61$) with 3-year cognitive outcome. Using unbiased multilayer clustering, Gamberger et al. [23] identified two distinct clusters with different 5-year prognoses of 562 late MCI patients. Rapid decliners progressed to dementia at five times the rate of slow decliners [23]. Herein, the nonsignificant difference in the AD conversion rate between Eq(deep+) and Eq(deep−) may be due to the small sample size of MCI and a short follow-up duration (1.76 years).

DL complements the limitations of conventional visual and quantitative analyses. Approach to explore the continuous Aβ burden offers more sensitive information about borderline Aβ cases that might be missed by dichotomous visual grading. However, quantitative assessment implies the use of a cutoff, which is cohort-specific and influenced by various methodological factors, such as preprocessing method, the choice of the target and reference region, and the inclusion criteria for the selected sample. However, in clinical setting, DL has practical advantages over traditional quantification because it determines positivity based on raw images and is less affected by analytical factors. In conclusion, DL provided quantitative information to identify individuals with low but meaningful amounts of Aβ, which in turn potentially reflected prognosis prediction.

Promising clinical trials aiming to develop drugs for secondary AD prevention are currently underway [24]. The study design commonly used is a randomized, placebo-controlled trial, implying that more than half of the participants receive a placebo for at least 18 months or longer [25]. This design raises the ethical dilemma of exposing high-risk subjects to a meaningless placebo and low-risk subjects to unnecessary treatment [26]. We believe that our DL-based approach will not only improve the efficiency and increase the statistical power of trials with smaller sample sizes but also help identify clinically relevant subpopulations at risk of AD and provide reassurance to patients at a low risk of aggravation.

Regarding the interpretation of amyloid PET, we believe that a 2D-based approach is not inferior to a 3D-based approach for our specific purpose of amyloid positivity scoring considering that human readers score amyloid positivity by slice-by-slice visual inspection and determine the subject-level positivity by integrating the slice-level positivity. A comparison study supported our hypothesis that the 2D CNN analysis was not inferior to the 3D CNN for test sets and showed better performance for the validation sets with equivocal cases. In 3D CNNs, due to the small amount of input/label samples, large data should be generated by either augmentation or patch-based learning. Unlike 2D CNNs, which benefit from various slicing directions, 3D augmentation is merely achieved by rotation or scaling of a single subject. Yan et al. [27] raised the drawback of 3D patch–based learning regarding tremendously increased labeling efforts. Although Choi et al. [22] increased the stride number to reduce parameters to train with a small amount of 3D data, this approach leads to omission of useful information from continuous voxels. Without conducting the compromised approach (3D augmentation, patch-based learning, or increased strides), 3D CNNs may suffer from not only the shortage of input/label samples but also high computer memory demands. To obtain benefits while overcoming the limitations of 2D CNNs, we incorporated an additional fully connected network using the 2D slice–based CNN output as input for the final subject-based decision-making network, as conducted by previous studies [27, 28]. By performing 2D-based data augmentation, we successfully increased the number of input/label samples by the factor of the slice number ($N = 33$). Thereby, we increased the recognition accuracy of subject-level decisions (i.e., the 2D CNN with a fully connected network for integrating 3D amyloid positivity information), which was 100% higher than that of slice-level decisions(94%).

Our study had several limitations. First, our sample was obtained from a single-center cohort. How technical factors, such as camera systems and reconstruction methods, affect the generalizability of the DL algorithm is unknown. Second, the age of the equivocal group (74.8 [7.29]) was higher than that of the positive group (69.53 [9.95]), which may be attributed to the mixture of late-onset AD (LOAD) and early-onset AD (EOAD) in the positive group; 49.4% of AD patients had EOAD in the positive group, whereas 15.4% of patients had EOAD in the equivocal group. Age-dependent slowing of Aβ turnover may influence the equivocal finding, such as poor differentiation between gray and white matter [29]. Third, investigating features that DL uses to make predictions remains a research frontier in this field. The image classification method, training with a single label per image, has limitations,

causing DL to associate irrelevant information with the diagnosis or to use features ignored by humans. Coupling with specific masks using the semantic segmentation method may enhance the utilization of DL for segmentation of diagnostic clues.

Conclusion

In equivocal scans, DL was more related to quantification than visual assessment, and negative cases selected by DL showed no decline in cognitive outcome. DL is useful for assessing the clinical diagnosis and prognosis in patients with visually equivocal amyloid scans and could be used to effectively filter out and provide reassurance to patients at low risk of AD progression.

Funding information This study was supported by a grant from the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute funded by the Ministry of Health and Welfare, Republic of Korea (HI14C2768, HI14C3319); the Industrial Core Technology Development Program (10060305) funded by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Evaluation Institute of Industrial Technology (KEIT); the Asan Institute for Life Sciences (2014-0783, 2016-0588); and the Ministry of Science and ICT (MIST), Republic of Korea (2017M2A2A6A02020353).

Compliance with ethical standard

Conflict of interest The authors declare that they have no conflicts of interest.

Informed consent Informed written consent was obtained from all individual participants included in the study.

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and the principles of the 1964 Declaration of Helsinki and its subsequent amendments or comparable ethical standards.

References

1. Grundman M, Pontecorvo MJ, Salloway SP, Doraiswamy PM, Fleisher AS, Sadowsky CH, et al. Potential impact of amyloid imaging on diagnosis and intended management in patients with progressive cognitive decline. Alzheimer Dis Assoc Disord. 2013;27:4–15.
2. Hosokawa C, Ishii K, Hyodo T, Sakaguchi K, Usami K, Shimamoto K, et al. Investigation of (11)C-PiB equivocal PET findings. Ann Nucl Med. 2015;29:164–9.
3. Schreiber S, Landau SM, Fero A, Schreiber F, Jagust WJ. Alzheimer's Disease Neuroimaging Initiative. Comparison of visual and quantitative florbetapir F 18 positron emission tomography analysis in predicting mild cognitive impairment outcomes. JAMA Neurol. 2015;72:1183–90.
4. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278–324.
5. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
6. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
7. Liu M, Zhang D, Shen D. Alzheimer's Disease Neuroimaging Initiative. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Hum Brain Mapp. 2014;35:1305–19.
8. Sarraf S, Tofighi G. Classification of Alzheimer's disease structural MRI data by deep learning convolutional neural networks. 2017. https://arxiv.org/abs/1607.06583. Accessed 5 Apr 2019.
9. Wen D, Wei ZH, Zhou YH, Li GL, Zhang X, Han W. Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion. Front Neuroinform. 2018;12:23.
10. Singh S, Srivastava A, Mi L, Caselli RJ, Chen K, Goradia D, Reiman EM, Wang Y. Deep-learning-based classification of FDG-PET data for Alzheimer's disease categories. In: 13th international conference on medical information processing and analysis. San Andres Islands, Colombia: International Society for Optics and Photonics; 2017. pp. 105720J.
11. Minoshima S, Drzezga AE, Barthel H, Bohnen N, Djekidel M, Lewis DH, et al. SNMMI procedure standard/EANM practice guideline for amyloid PET imaging of the brain 1.0. J Nucl Med. 2016;57:1316–22.
12. NeuraCeq. NEURACEQ (florbetaben F 18 injection), highlights of prescribing information. 2017. http://piramal.com/neuraceq/images/Neuraceq_PI.pdf. Accessed 5 Apr 2019.
13. Van Maaten LD, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
14. Vandenberghe R, Nelissen N, Salmon E, Ivanoiu A, Hasselbalch S, Andersen A, et al. Binary classification of ^{18}F-flutemetamol PET using machine learning: comparison with visual reads and structural MRI. Neuroimage. 2013;64:517–25.
15. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using (18)F-FDG PET of the brain. Radiology. 2019;290:456–64.
16. Payoux P, Delrieu J, Gallini A, Adel D, Salabert AS, Hitzel A, et al. Cognitive and functional patterns of nondemented subjects with equivocal visual amyloid PET findings. Eur J Nucl Med Mol Imaging. 2015;42:1459–68.
17. Cattell L, Platsch G, Pfeiffer R, Declerck J, Schnabel JA, Hutton C. Alzheimer's Disease Neuroimaging Initiative. Classification of amyloid status using machine learning with histograms of oriented 3D gradients. Neuroimage Clin. 2016;12:990–1003.
18. Cohen AD, Mowrey W, Weissfeld LA, Aizenstein HJ, McDade E, Mountz JM, et al. Classification of amyloid-positivity in controls: comparison of visual read and quantitative approaches. Neuroimage. 2013;71:207–15.
19. Yuan Y, Wang Z, Lee W, VanGilder P, Chen Y, Reiman EM, et al. Quantification of amyloid burden from florbetapir pet images without using target and reference regions: preliminary findings based on the deep learning 3D convolutional neural network approach. Alzheimers Dement. 2018;14:P315–6.
20. Landau SM, Horng A, Jagust WJ. Initiative AsDN. Memory decline accompanies subthreshold amyloid accumulation. Neurology. 2018;90(17):e1452–e60.
21. 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(2):205–16.

22. Choi H, Jin KH. Alzheimer's Disease Neuroimaging Initiative. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav Brain Res. 2018;344:103–9.
23. Gamberger D, Lavrač N, Srivatsa S, Tanzi RE, Doraiswamy PM. Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease. Sci Rep. 2017;7:6763.
24. Mangialasche F, Solomon A, Winblad B, Mecocci P, Kivipelto M. Alzheimer's disease: clinical trials and drug development. Lancet Neurol. 2010;9:702–16.
25. Spiegel R, Berres M, Miserez AR, Monsch AU, Alzheimer's Disease Neuroimaging Initiative. For debate: substituting placebo controls in long-term Alzheimer's prevention trials. Alzheimers Res Ther. 2011;3:9.
26. Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. 2015. https://arxiv.org/abs/1502.02506. Accessed 5 Apr 2019.
27. Yan K, Bagheri M, Summers RM. 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. In: 21th international conference on medical image computing and computer assisted intervention. Granada, Spain: Springer; 2018. pp. 511–9.
28. Zhao G, Liu F, Oler JA, Meyerand ME, Kalin NH, Birn RM. Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. Neuroimage. 2018;175:32–44.
29. Patterson BW, Elbert DL, Mawuenyega KG, Kasten T, Ovod V, Ma S, et al. Age and amyloid effects on human central nervous system amyloid-beta kinetics. Ann Neurol. 2015;78:439–53.