Received: 19 February 2021 | Revised: 4 October 2021 | Accepted: 7 October 2021 | Published online: 23 December 2021

DOI: 10.1002/dad2.12264

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Deep learning improves utility of tau PET in the study of Alzheimer's disease

James Zou[1] | David Park[1] | Aubrey Johnson[1] | Xinyang Feng[1] | Michelle Pardo[2] | Jeanelle France[1] | Zeljko Tomljanovic[1] | Adam M. Brickman[1,3] | Devangere P. Devanand[1,4] | José A. Luchsinger[2,5] | William C. Kreisl[1] | Frank A. Provenzano[1,3] | for the Alzheimer's Disease Neuroimaging Initiative[1]

[1] The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, New York, New York, USA

[2] Department of Medicine, Columbia University Medical Center, New York, New York, USA

[3] Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, New York, USA

[4] New York State Psychiatric Institute and Department of Psychiatry, Columbia University Medical Center, New York, New York, USA

[5] Department of Epidemiology, Columbia University Medical Center, New York, New York, USA

Correspondence
Frank A. Provenzano, 630 W 168th St. New York, NY 10032, USA.
E-mail: fap2005@cumc.columbia.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01AG050440, R01 AG055422, RF1AG051556, RF1AG051556-01S2, R01AG055299, K99AG065506, K24AG045334, UL1TR001873; National Institute of Aging (NIA), Grant/Award Numbers: P01AG07232, R01AG037212, RF1AG054023

Abstract

Introduction: Positron emission tomography (PET) imaging targeting neurofibrillary tau tangles is increasingly used in the study of Alzheimer's disease (AD), but its utility may be limited by conventional quantitative or qualitative evaluation techniques in earlier disease states. Convolutional neural networks (CNNs) are effective in learning spatial patterns for image classification.

Methods: 18F-MK6240 (n = 320) and AV-1451 (n = 446) PET images were pooled from multiple studies. We performed iterations with differing permutations of radioligands, heuristics, and architectures. Performance was compared to a standard region of interest (ROI)-based approach on prediction of memory impairment. We visualized attention of the network to illustrate decision making.

Results: Overall, models had high accuracy (> 80%) with good average sensitivity and specificity (75% and 82%, respectively), and had comparable or higher accuracy to the ROI standard. Visualizations of model attention highlight known characteristics of tau radioligand binding.

Discussion: CNNs could improve tau PET's role in early disease and extend the utility of tau PET across generations of radioligands.

1 | INTRODUCTION

The primary pathologic signs of Alzheimer's disease (AD) are amyloid beta plaques and neurofibrillary tau tangles.[1] Recently developed positron emission tomography (PET) radioligands that bind to tau tangles in vivo—such as 18F-AV-1451 ("flortaucipir"), 18F-MK-6240, and 18F-RO948[2]—can be used to detect regional tau pathology in vivo, which can be correlated to magnetic resonance imaging (MRI) findings, which are definitionally nonspecific to AD. Uptake of these radioligands can correspond to longitudinal biomarker change,[3,4] correlate with performance on cognitive tests throughout the disease spectrum,[4–6] and be used to identify phenotypically distinct forms of AD pathophysiology.[7]

The evaluation of tau PET images usually uses either regional standardized uptake value ratio (SUVR) quantification—which inherently does not capture off-target binding or visuospatial patterns of binding

present in disease variants;[8] and visual reads[9]—which can suffer from inter-reader reliability issues stemming from this method's subjective nature.[10] Recent work has also sought also to integrate multiple generations of tau radioligands[7,11,12] to help clarify tau deposition across centers, which could allow for the construction of models from larger and more diverse populations. This could facilitate disease detection at earlier states (at which point tau PET is currently thought to be a less effective test[13]), a point at which intervention could more effectively modify disease course.

Machine learning techniques, especially convolutional neural networks, have shown promise in a number of contexts important to neuroimaging, including with MRI[14,15] and fluorodeoxyglucose (FDG) PET.[16] Similarly, deep learning with PET radioligands has demonstrated improved research efficacy[17] and diagnostic utility.[18,19] To our knowledge, relatively few studies exist that use tau PET,[20–22] in part because of current limitations/challenges of applying deep learning in clinical research (especially the relatively small number of absolute subjects). In approaching this problem with deep learning, we hope to extend tau-PET imaging into the identification of a feature of the disease, chiefly cognitive impairment, at earlier disease states, when a deep learning model could incorporate additional features which may be undetected in tau scans of advanced disease cases. Furthermore, as more potential therapies become available, a framework that is built on one of the core A/T/[N] (amyloid/tau/neurodegeneration) criteria may be adaptable to target engagement and therapeutic response, especially when the expected effect is thought to be too subtle for existing methods of biomarker measurement.

As such, we created a neural network framework with tau PET images—from two separate sources representing two PET radioligands, from different generations, with highly similar functional binding (18F-MK-6240 and 18F-AV-1451)—trained on the classification task of predicting cognitive impairment. Using the probability of impairment status generated by these various neural networks from imaging inputs, we attempt to provide a better predictor of disease state based on tau PET imaging findings than a traditional SUVR-based method. We also implement various strategies proposed and, inspired by recent literature, aimed to improve model performance and attempt to visualize the logic of this neural network in selecting imaging features useful for this classification.

Research in Context

1. Systematic review: The authors reviewed the literature using traditional sources (e.g., PubMed), indexing services (e.g., Google Scholar), meeting abstracts, and presentations. We attempted to include the most pertinent work published for both AV-1451 and MK-6240; and relevant work done with neural networks, especially involving positron emission tomography (PET) imaging.
2. Interpretation: Our results suggest that deep learning can be successfully applied to the analysis of multi-generational tau PET images, and that this procedure may offer improvements over standard standardized uptake value ratio–based approaches. Our visualizations suggest qualitative and semi-quantitative evidence that neural networks have explainable elements.
3. Future directions: Incorporating differing modalities (especially multiple tau radioligands) may improve performance of our model and provide new insights into the correlation of tau PET with co-occurring physiology, especially as tau imaging is expected to eventually become more available for clinical use. More work is needed to investigate whether advances in frameworks provide any benefits in accuracy for early disease state.

2 | METHODS

2.1 | Subject selection

2.1.1 | MK-6240 subjects

MK-6240 PET scans (n = 320) were gathered from participants in various ongoing studies at Columbia University Irving Medical Center (CUIMC). These include the Washington Heights Inwood Columbia Aging Project[23] (n = 4), the Northern Manhattan Study of Metabolism and Mind (NOMEM)[24] (n = 200), the VALAD (Valacyclovir Treatment of Alzheimer's Disease) study[25] (n = 57), and other studies recruiting participants at CUMC[26] (n = 59). These studies—as well as the current study—were reviewed by the institutional review board, and all participating subjects gave consent for all procedures.

2.1.2 | AV-1451 subjects

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI was launched in 2003 as a public–private partnership, led by principal investigator Michael W. Weiner, MD. We selected subjects with available AV-1451 PET scans (n = 446) obtained from ADNI, comprising patients from the ADNI3 cohort.[27,28] When multiple scans for the same participant were available, we used the scan performed most recent to the data download.

A CONSORT diagram detailing subject selection is in supporting information (Figure S1).

2.2 | Clinical status determination

2.2.1 | MK-6240 subjects

As a variety of studies were included, participants underwent differing cognitive testing, including the Mini-Mental State Examination

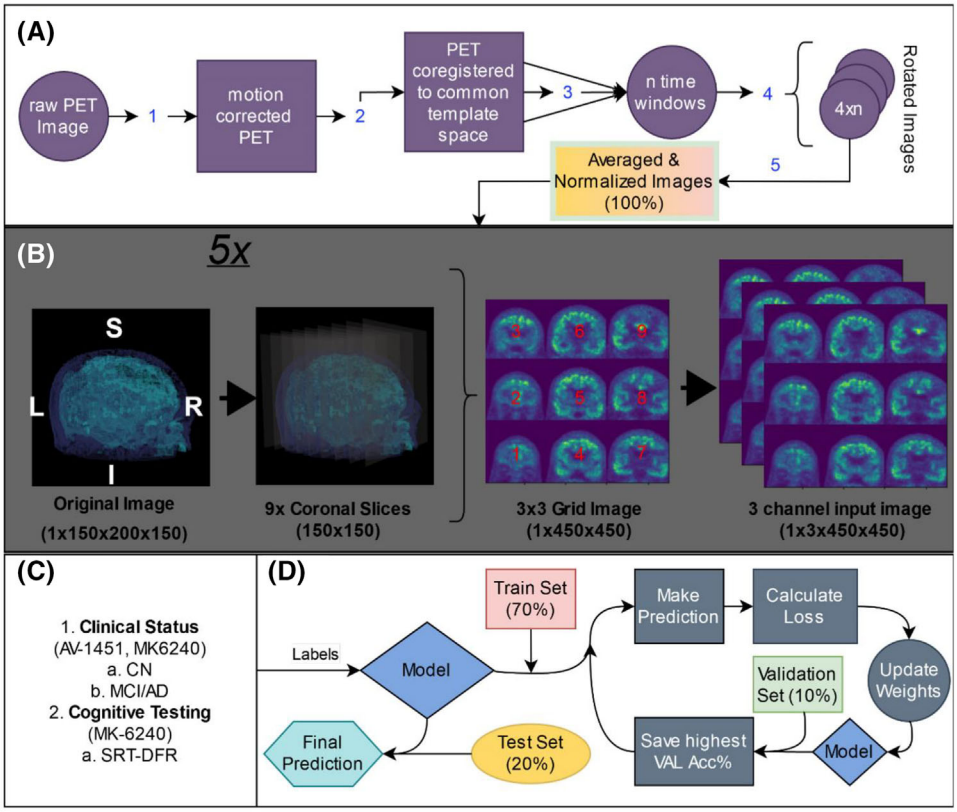

FIGURE 1 Summary overview of project pipeline. A, Image preprocessing steps, involving (1) motion correction (with FSL), (2) registration of each image to a template generated using ANTS, (3) creation of three time acquisition windows (80–100, 85–105, and 90–110 minutes post-injection) from each scan's available windows (80–110) for the MK-6240 dataset (the 80–100 minute time window was used for AV-1451), (4) rotation of each time window image by 7, 14, and 21 degrees along the sagittal plane for data augmentation (see Supplementary Methods 2.3.2 for details), (5) averaging and internal normalization of uptake values. B, 2D image generation for input into 2D inception model. The orientation of each coronal slice is shown (R = right, L = left, S = superior, I = inferior) and numbered here to show slice order from rostral to caudal. All images subsequently shown are the same orientation. We elected to generate five such images for each subject, with differing coronal slices used. C, Determination of binary label with either clinical status (MK-6240 dataset and AV-1451 dataset) or cognitive test result when formal determination unavailable (MK-6240 dataset). D, Each cycle of 5-fold cross-validation, which involves input of train set images (pink) into the model returning a scalar prediction of likelihood of binary impairment status, followed by model weight adjustment based on accuracy of the prediction (using batch gradient descent), followed by testing external validity of the model on an independent validation set (green). The model with the highest accuracy after 30 epochs is then tested on a holdout test set (yellow). AD, Alzheimer's disease; CN, cognitively normal; MCI, mild cognitive impairment; SRT-DFR, Selective Reminding Test, Delayed Free Recall

(MMSE) in some participants[29] and Selective Reminding Test, Delayed Free Recall (SRT-DFR)[30] in all (only the SRT-DFR was available from NOMEM participants). Participants were then designated as cognitively normal (CN) or with presumed mild cognitive impairment or AD (MCI/AD) if they met clinically determined criteria; or when not formally assessed, a score on the SRT-DFR lower than 1.5 standard deviations based on external norms (adjusted for age, sex, ethnicity, and education).[23]

2.2.2 | AV-1451 subjects

Procedures for cognitive testing and determination of clinical status (CN vs. MCI/AD as above) are detailed elsewhere.[27,28]

2.3 | Imaging

18F-MK-6240 PET images were acquired in 5-minute time windows, 80 to 110 minutes post injection (mean injected activity = 165.8 MBq). 18F-AV-1451 PET images were also acquired at 5-minute time windows, at 75 to 105 minutes to their site-specific protocols (mean injected activity = 370 MBq).

2.3.1 | Amyloid status

Amyloid status was determined in subjects with MK-6240 PET with 18F-Florbetaben (FBB) using a visual read from trained readers.[31] Amyloid status in subjects with AV1451 PET was determined with a threshold using either FBB or 18F-AV45, which are detailed elsewhere.[28]

2.3.2 | Data preprocessing

For MK-6240, preprocessing steps are summarized in Figure 1A. Because multiple images were generated for the same subject with these above procedures, care was taken to ensure no subject scans were in both training and/or testing/validation sets, to prevent the confound of data leakage.[32]

For AV-1451, we performed the same registration and normalization procedure. We averaged acquisition windows at 80 to 100 minutes (as has been previously done[3,33]), and we elected to use this time window when both AV-1451 and MK-6240 are used together as a dataset. All subjects were registered (with aid of MRI) to the same custom atlas space created using Advanced Normalization Tools (ANTS).[34]

2.4 | Model architecture

We trained two models, both based on the InceptionV3 architecture.[39] We implemented both a conventional 2D model as well as a 3D model using PyTorch,[35] as well as the Python packages Numpy,[36] Pandas,[37] and Sci-Kit Learn.[38] See Table S1 in supporting information for summary, and Supplementary Methods for details on configurations for training.

2.4.1 | Model architectures

The 2D model is a direct extension of the Inceptionv3 architecture.[39] The network was pretrained on images from the ImageNet dataset and these weights were preloaded onto the model prior to training. Similar to Ding et al.,[16] we used multiple coronal slices as inputs (see Figure 1B).

Additionally, we adapted the inception architecture to work with 3D images, changing applicable 2D kernels to 3D and adding a convolution operation (to represent the added third dimension) in each of the applicable InceptionV3 layers (where appropriate). See Figure S2 in supporting information for a basic diagram of the model.

2.5 | Model training

We used 5-fold cross-validation on all iterations of our training/testing (70-10-20 train/validation/test split, see Figure 1C). We evaluated the performance of the models on the MK-6240 images, the AV-1451 images, and the combination of the two datasets, and results for all three configurations are reported. Configurations for all experiments are available in Table S2 in supporting information.

2.6 | Evaluation

To evaluate the performance of the various models, we performed receiver-operator characteristic (ROC) analysis and calculated the area under curve (AUC) and F1 score for the model generated from each validation fold.

2.6.1 | Comparison method

Briefly, all regions of interest (ROIs) were based on the Braak ROIs first reported in Schöll et al.[40] All PET images were then registered to either the aforementioned custom atlas space (MK-6240 images) or to their respective MRI (AV-1451 images). We elected to use both an entorhinal cortex SUVR (i.e., Braak I) as well as a composite SUVR using early Braak regions (I–IV). The same ROIs were used for each radioligand to determine SUVR values, including an inferior cerebellar reference region. All PET images were partial volume corrected using the geometric transfer matrix (GTM) method.[41] Thresholds for tau "positivity" to derive accuracy measures for the comparison method were calculated using Youden's Index.[42]

2.7 | Statistical analysis

Group-wise statistical analyses were performed for both demographic comparisons (i.e., between-subject characteristics between controls and patients) and model results (i.e., between various models' prediction for an individual's subject tau PET image), including analysis of variance (ANOVA) for continuous variables and Chi-squared tests for categorical variables were performed with R (version 3.6.2). ROC analyses for impairment prediction for each test set were performed using the Python package Sci-Kit Learn.[38] We compared ROC curves directly with the DeLong test for correlated ROC curves.[43]

To test for potential influences of imbalanced participant characteristics (age, education, amyloid status, sex) on the classification tasks, we fit multivariate mixed effects models with cognitive impairment (represented as a z-score transformed composite cognitive score based on MMSE and SRT-DFR available for each subject) as the response variable; each tau-based measure (i.e., scaled SUVRs and neural network derived impairment probability) and amyloid status as fixed effects; and age, education, and sex as random effects. We test for relative goodness of fit of each model with tau compared to amyloid only models, and each model using a neural network–derived measure against SUVR measures with an ANOVA on model residuals. We used Satterthwaite approximation of degrees of freedom for the t-test to test the significance of fixed effects on our respective models, and used ANOVAs to derive likelihood ratios for significance of random effects.[44]

2.8 | Visualization

To visualize the activity of select 2D networks, we applied an occlusion sensitivity analysis on select participants, as well as an averaged sensitivity analysis across testing folds.[45] We also performed a 3D version of this analysis for two selected participants, and performed

t-distributed stochastic neighbor embedding (t-SNE) analysis[45] to semi-quantitatively visualize our models' feature learning.

3 | RESULTS

3.1 | Demographic characteristics

Demographic data are summarized in Table 1. Impaired patients were older and more likely to be men than unimpaired participants. The two datasets were statistically different from each other in all categories (*P*'s < .02).

This study consists of White (W), Black (B), Hispanic (H), and other (Oth). "Other" includes Asian ethnicity (n = 11), mixed race (n = 5), American Indian (n = 1), and declined to report (n = 2).

For participants with MK-6240, 49 participants met criteria for MCI and 72 for AD. For AV-1451, 84 participants met criteria for MCI and 43 for AD.

Amyloid status was unavailable for 5 participants with MK-6240 scans and 15 participants with AV-1451 scans. The average (±standard deviation) time between amyloid and tau scan for MK-6240 participants was 145 days (±151), while for AV-1451 participants was 157 days (±286).

Jitter plots for SUVR as stratified by CN/MCI/AD status are shown in Figure S3 in supporting information.

3.2 | Model performance

Overall, we were able to train every iteration of the models with average accuracy of at least 80% with a mean sensitivity 0.82 and mean specificity 0.75 on respective test sets. Accuracy of each iteration of the model and other metrics (AUC, F1 score) are summarized in Table S3a/b in supporting information. Model accuracies (by fold); and AUC, sensitivities, and specificities are summarized in Table S4a and S4b in supporting information, respectively. *P*-values for statistical comparisons between models are summarized in Table S5 in supporting information. We show t-SNE maps for all models in Figure S4 in supporting information.

3.2.1 | Model performance versus ROI-based standard

The highest accuracy of the entorhinal cortex SUVR for MK-6240 was 69.4% at a cutoff of 1.28 (sensitivity = 0.53, specificity = 0.86), while for our composite SUVR the highest accuracy was 69.8% at a cutoff of 1.51 (sensitivity = 0.51, specificity = 0.90). These two SUVR measures did not have significantly different AUC (0.72 vs. 0.73, Z = -0.6, *P* = .54). For AV-1451, highest accuracy for entorhinal cortex SUVR was 78.4% at a cutoff of 3.15 (sensitivity = 0.61, specificity = 0.77), while for our composite SUVR, the highest accuracy was 79.4% at a cutoff of 2.49 (sensitivity = 0.22, specificity = 0.98). These two SUVR measures similarly did not differ (0.71 vs. 0.74, Z = −1.0, *P* = .31).

We report comparisons with our neural network models to the composite SUVR in the main text below. With AV-1451, the 2D model measure trained solely on this radioligand had a better AUC compared to AUC derived from SUVR measurements, though not significantly (0.78 vs. 0.73, Z = −1.4, *P* = .17), whereas the measure derived from the model trained using both radioligands had a significantly greater AUC (0.89 vs. 0.72, Z = 5.2, *P* < .0001). There was no significant difference between the 3D model trained solely on this radioligand (0.73 vs. 0.72, Z = 0.22, *P* = .83), whereas the 3D model trained on both radioligands was significantly better than SUVR measurement (0.80 vs. 0.73, Z = 2.00, *P* = .04).

With MK-6240, both the 2D models trained on the single radioligand (0.83 vs. 0.74, Z = −3.5, *P* = .001) and on both radioligands (0.85 vs. 0.72, Z = 2.7, *P* = .006) were significantly better than SUVR measurement. Similar results were found for our 3D models trained on the sole radioligand (0.83 vs. 0.74, Z = 3.2, *P* = .001) and on both radioligands (0.82 vs. 0.74, Z = 2.5, *P* = .01).

Comparisons with entorhinal cortex SUVR were similar. These, and other outcomes, are summarized in Table S5 in supporting information, and in Figure 2.

3.2.2 | Mixed effect models incorporating tau derived measures

We wished to test the relative goodness of fit of tau measures in modeling predicted scores on cognitive testing, when accounting for amyloid status and adjusted for the random effects of sex, age, and education. A multivariate model using only amyloid status as a fixed effect was significant for both radioligands (estimates ← 0.4, *P*'s < .0001). As expected, education level was generally a strong significant random effect (likelihood ratio tests [LRTs] > 3.9, *P*'s < .05), although age tended to only have a trend level effect (LRTs > 2.0, *P*'s ≈.14).

Every tau-based measure had a significant, negative main effect in their respective models (estimates < −0.15, *P*'s < .01). Interestingly, amyloid status was not a significant fixed effect for MK subjects with values derived from 2D MK images, or 2D and 3D modeled data using both sets of radioligands.

Unsurprisingly, we found that each model incorporating a tau-derived measure had superior fit compared to models using only amyloid status and random effects (χ^2's > 4.1, *P*'s < .05). Generally, multivariate models using tau measures from neural networks trained solely on AV-1451 images did not have superior fit compared to AV-1451 SUVRs, whereas multivariate models using tau values derived from models trained on MK-6240 images or the combined dataset did tend to have strongly superior fit (χ^2's > 37.6, *P*'s < .0001). Interestingly, sex was a significant random effect for AV-1451 subject models that did not use neural network derived measures (LRTs > 5.8, *P*'s < .02). Age had a significant effect in the AV-1451 subject model using both radioligands (LRT = 4.6, *P* = .03), whereas education level had a significant

TABLE 1 Basic demographics for each cohort, separated by disease status

	MK-6240 participants (n = 320)			AV-1451 Participants (n = 446)			Combined participants (n = 766)			AV-1451 vs. MK-6240 participants
	Controls (n = 199)	MCI/AD* (n = 121)	test-stat (*P*-value)	Controls (n = 319)	MCI/AD* (n = 127)	test-stat (*P*-value)	Controls (n = 518)	MCI/AD* (n = 248)	test-stat (*P*-value)	test-stat (*P*-value)
Age (y)	66.0 (5.5)	70.2 (8.8)	−4.71 (< .001)	73.5 (8.2)	75.3 (8.3)	3.5 (< .001)	70.2 (8.6)	72.1 (8.7)	−4.4 (< 0.0001)	−9.7 (< .0001)
Sex (F/M)	140/59	58/63	15.1 (.0001)	183/136	55/72	6.2 (.01)	323/195	113/135	19.3 (< .0001)	5.16 (.02)
Ethnicity (W/B/H/Oth)†	28/29/142/0	68/5/48/0	65.0 (< .001)	282/15/18/4	111/6/5/5	3.76 (.28)	310/44/160/4	179/11/53/5	15.5 (.001)	311 (< .0001)
MMSE score	28.9 (1.6)	23.6 (3.6)	10.2 (< .0001)	29.3 (.7)	25.1(2.2)	18.0 (< .0001)	29.1 (1.2)	25.7 (3.4)	21.64 (< .0001)	−6.9 (< .0001)
Education (y)	12.0 (4.2)	13.9 (4.9)	−3.1 (< .002)	16.7 (2.4)	15.3 (2.4)	4.6 (< .0001)	14.7 (4.0)	15.1 (3.7)	0.69 (0.40)	−12.6 (< .0001)
Amyloid (+/-)‡	14/181	75/45	70.3 (< .0001)	96/216	41/78	0.66 (.42)	110/397	116/123	55.4 (< .0001)	8.58 (.003)
SRT-DFR (z-score)	−0.23 (0.85)	−2.46 (0.74)	20.1 (< .0001)	n/a	n/a	n/a	-0.23 (0.85)	-2.46 (0.74)	20.1 (< .0001)	n/a
EC SUVR§	1.12 (0.32)	1.76 (0.82)	−8.0 (< .0001)	2.0 (0.5)	2.8 (1.0)	−7.3 (< .0001)	n/a	n/a	n/a	n/a
Composite SUVR§	1.23 (0.45)	2.10 (1.13)	90.9 (< .0001)	1.7 (0.3)	2.1 (0.5)	76.9 (< .0001)	n/a	n/a	n/a	n/a

Abbreviations: AD, Alzheimer's disease; B, Black; EC, entorhinal cortex; H, Hispanic; MCI, mild cognitive impairment; Oth, other ethnicity; SRT-DFR, Selective Reminding Test, Delayed Free Recall; SUVR, standardized uptake value ratio; W, White.

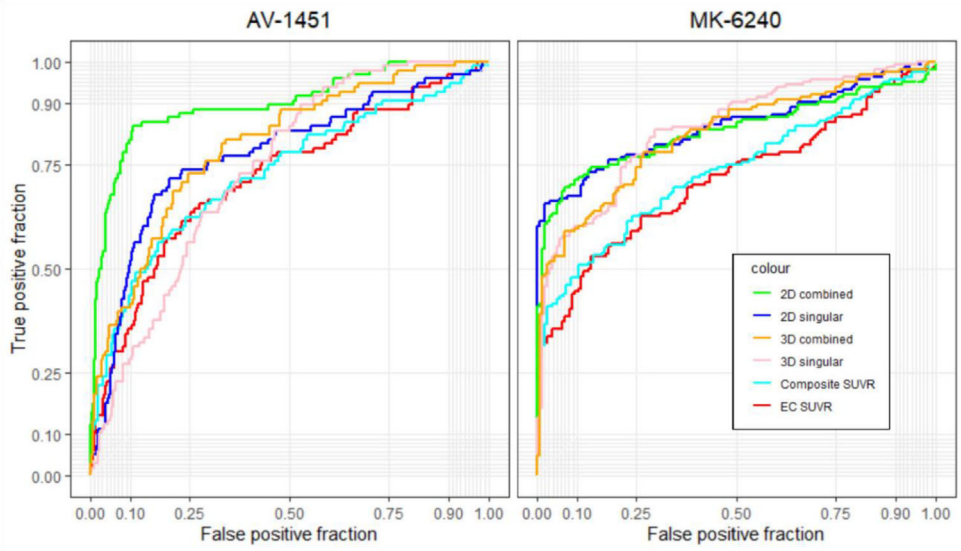

FIGURE 2 ROC curves for 2D and 3D models versus composite/entorhinal cortex (EC) SUVR. Shown by radioligand. Training our neural network models involved using either a single ("singular") radioligand or pooling together both radioligands ("combined"). Predictions from our "combined" 2D/3D model configurations are assigned to the appropriate comparison group. 2D, two-dimensional input model; 3D, three-dimensional input model; EC, entorhinal cortex; ROC, receiver operating characteristic; SUVR, standardized uptake value ratio

effect in all models (LRTs > 3.9, *P*'s < .05). Results for all models are summarized in Table S6 in supporting information.

3.3 | Occlusion sensitivity analyses

3.3.1 | 2D models

We show select example occlusion sensitivity maps for our MK-6240 derived model (Figure 3). We see in these examples that this model tended to assign importance to cortical areas (especially medial temporal areas, see Figure 3 Image 1) for prediction of impairment, and against white matter areas for prediction of lack of impairment. Similar general patterns are exhibited for the AV-1451 models (see Figure S5 in supporting information). See the description in Figure 3 for further information, and Figure S6 in supporting information for an averaged activation map for both models.

3.3.2 | 3D models

Given that the processing time for our occlusion sensitivity image creation for our 3D images was computationally expensive, we created images for two selected participants (one with a highly probable true positive prediction and one with an intermediate but true negative prediction) from each dataset model. As with results of our 2D models, we found in general that cortical—especially temporal—areas tended to positively activate the network, whereas white matter regions tended to suppress activation. Visualizations for these are shown in Figure S7 in supporting information.

4 | DISCUSSION

We demonstrate that neural networks can be feasibly applied to tau PET images through either a 2D or 3D analytical framework, with comparable or superior performance to standard SUVR-based methods. We additionally show that multiple generations of tau radioligands can be integrated into a single framework.[7,11] Earlier deep learning efforts with tau PET have sought to simplify[20] and augment[17] preprocessing steps; and have provided initial proof of feasibility and model interpretation.[22] The main metrics of the models compare similarly with prior deep learning classification tasks that used different sets of PET radioligands,[16,22,46] though comparisons are of course difficult given differences in available clinical outcomes. Furthermore, while population differences within and between control and patient groups do somewhat limit empirical conclusions on the data itself, results of our goodness-of-fit analyses point to the promising utility of neural network–derived measures for accurately estimating the contribution of tau burden to cognitive performance because of AD pathophysiology.

As both these radioligands have been used primarily for research purposes (and with AV-1451 only recently gaining Food and Drug Administration approval) with limited work on systems of visual interpretation,[9] we did not test the models against clinicians. No standard method exists to reliably aid diagnostic and management guiding decisions with these scans, though there is evidence that they—in conjunction with other biomarkers—can provide reproducible quantitative biomarkers in routine clinical decision making. Our successful application of deep learning to a framework using both radioligands suggests that our deep learning–derived measure could be standardized/harmonized across multiple radioligands, and provide

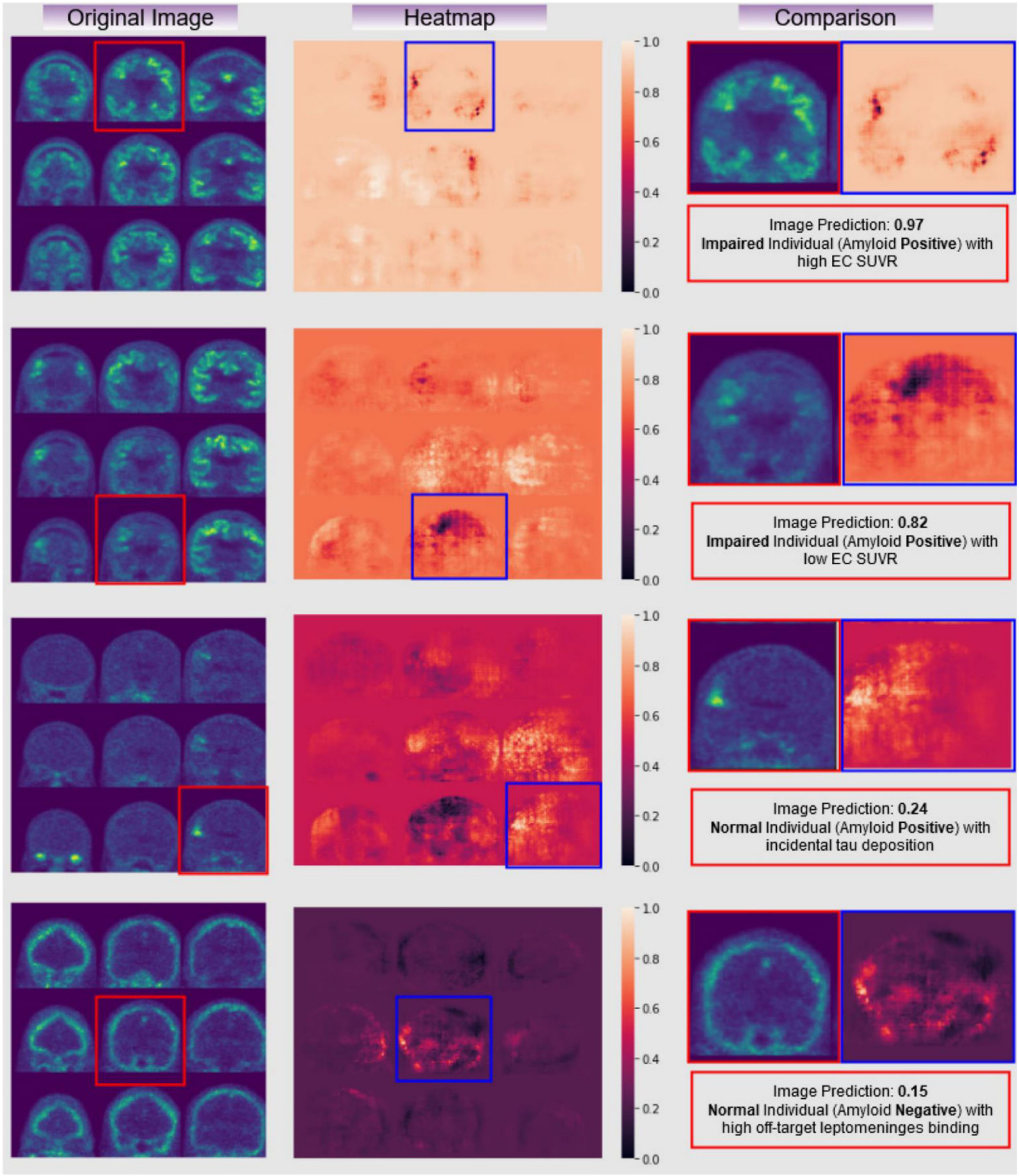

FIGURE 3 Select heat maps for our MK-6240 model with selected subjects. Negative values represent regions of positive predictive importance for impairment, whereas positive areas represent areas of negative predictive importance. The scale to the right of images represents proportional change from baseline prediction due to occlusion of specified region (i.e., more "important" areas have a larger absolute value and are brighter/darker on the maps). We highlight a specific slice for each example subject to the right. The probability of impairment as predicted by the model interpretation of the image is shown along with amyloid and actual impairment status to the right. The orientation of these heat map images (representing the 3×3 coronal slice images fed into the model) is the same as explicated in Figure 1. The leftmost column of images represents the input image, the middle column represents the generated sensitivity maps, and the paired images in the rightmost column are a representative comparison slice for each image and heatmap for each subject. The first image represents a true positive prediction, with relevance conferred by the sensitivity analysis to cortical binding in medial temporal and interestingly a contralateral parietal area. For these models, differential

an intuitive (i.e., probabilistic) and unitary measure in a way not directly possible with current methods.[11] Continued work on a multi-radioligand approach may allow the construction of higher confidence research/clinical models and ease of tau PET interpretation, perhaps as diagnostic aid in challenging clinical contexts, as has been reported for work with amyloid PET.[18,19]

Our work also adds to growing evidence that imaging tau[5,13] has utility as a high-specificity biomarker in the diagnosis of AD, with good—albeit lower—performance in less impaired individuals where tau deposition in early Braak regions may be the first sign.[47,48] A future deep-learning–based system to detect early disease and predict conversion may indeed rely on MRI,[15] if only because of greater availability of data (especially longitudinal). Indeed, our prior work has demonstrated the value of deep learning in pure MRI, both in AD and stable and progressing MCI,[14] as a standalone method and supplement to existing biomarkers. But although accuracies resulting from our study (and others) have shown exceptional promise, they are reflective of changes only observable on T1-weighted brain scans. Tau PET reflects one component of the A/T/[N] criteria and it is important to consider all core biomarkers irrespective of accuracy of the other. Some even argue that tau PET can be a "one-stop-shop" for learning about each component of this criteria.[49] It still may be important to consider a tau PET framework, as evidence emerges of the association between differential tau deposition patterns and clinically distinct trajectories of AD,[7] which could impact clinical management.

Much prior deep-learning–based classification work has focused on training models that learn features prominent in AD subjects compared to controls, and attempts to map this onto the classification of MCI.[15,22] While an effective approach, we explicitly trained our model using a more heterogeneous (but larger size) sample of impaired individuals (spanning from early impairment to highly probably AD), achieving similar "absolute" performance. Using this "noisier" sample has potential benefits for our model's generalizability and robustness to real-world tasks, where many subjects with putative "MCI" do not end up developing AD in a reasonable timeframe.[50] As such, incorporating more MCI data, with greater detail about the prodromal status (by confirming stability or progression to AD) would extend potential diagnostic utility of this model.

Our occlusion sensitivity analysis suggests that our networks can learn expected patterns of specific cortical binding in regions known to be affected during the pathogenesis of AD, which is in agreement with prior work[22] that used similar visualization techniques on AV-1451. While no specific brain region is consistently and specifically implicated through this analysis, these findings are consistent with prior work that suggests that deep learning algorithms may make classification decisions using non-linear interpretations of imaging data,[16] and seems to agree with the observation in AV-1451 that conventional Braak-staged regions may not be sufficient for accurate diagnosis or staging.[5] These visual findings also suggest that neural-network–based analysis of tau PET for diagnostic/research purposes would be mathematically and heuristically distinct from ROI-based quantification.

As for constructing a 2D versus 3D model for this data, it is important to consider that 2D data greatly benefits from pre-trained models whereas 3D data may have advantages in identifying features in higher dimensional space. Continued experimentation with some of the augmentation strategies used here may help increase the performance of these models, especially our 3D implementations. While we chose to use a more established model framework to take advantage of pretraining, newer architectures may offer improvements on computational cost or performance. Given these future improvements, deep learning could become an essential utility for the discovery of these complex spatial binding relationships, especially as we work as a field to standardize the interpretation of tau radioligands for both research and eventual clinical adaptation.

ACKNOWLEDGMENTS

AV-1451: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; ElanPharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the

preference for different sides of an image (despite bilateral radioligand deposition) seems to be an important criterion. The second image represents another true positive prediction in an individual with low EC SUVR but high SUVR binding elsewhere, placing high diagnostic importance on a right midline parietal region. This suggests that the proposed method may have additional uses in non-conventional AD subtypes. The third image represents an amyloid-positive participant who had incidental tau deposition in a temporoparietal region without complaints of memory impairment and normal performance on cognitive testing. Notably, this area of incidental tau signal seems to have been identified by the model as having negative predictive value. The fourth image represents an amyloid negative without impairment, who exhibits high off-target binding load. Interestingly, the neural network is able to identify a large portion (though not all) of this binding as non-relevant to the classification task. AD, Alzheimer's disease; EC, entorhinal cortex; SUVR, standardized uptake value ratio

Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

MK-6240: Data collection was supported by the National Institutes of Health (NIH): grants R01AG050440, R01 AG055422, RF1AG051556, RF1AG051556-01S2, R01AG055299, K99AG065506, and K24AG045334. Partial support for data collection was provided by NIH grant UL1TR001873. Data collection and sharing for this project was additionally supported by the National Institute of Aging (NIA, P01AG07232, R01AG037212, RF1AG054023). This manuscript has been reviewed by WHICAP investigators for scientific content and consistency of data interpretation with previous WHICAP Study publications. We acknowledge the WHICAP study participants and the WHICAP research and support staff for their contributions to this study.

CONFLICTS OF INTEREST

Dr. Kreisl is a consultant for Cerveau Technologies. However, Cerveau was not involved in the design or execution of this study or in the interpretation of results. Dr. Provenzano is a consultant for and has equity in Imij Technologies, an unrelated company, which was not involved in the design or execution of this study or the interpretation of results. Dr. Provenzano also holds several unrelated neuroimaging MRI patents not pertaining to the current study. Dr. Brickman has provided consultative services to Cognition Therapeutics and Regeneron. Dr. Devanand received research support from the NIA and the Alzheimer's Association. He is a scientific advisor for Acadia, BioExcel, Biogen, Eisai, Genentech, GW Pharmaceuticals, Novo Nordisk. James Zou, David Park, Aubrey Johnson, Xinyang Feng, Michelle Pardo, Jeanelle France, Zeljko Tomljanovica, and Jose A. Luchsinger have no disclosures to report.

REFERENCES

1. Alzheimer A, Stelzmann R, Schnitzlein H, Murtagh F. An English translation of Alzheimer's 1907 paper,". Uber Eine Eigenartige Erkankung Der Hirnrinde". *Clin Anat*. 1995;8:429-431.
2. Villemagne VL, Lopresti BJ, Doré V, et al. What is T+? A Gordian Knot of tracers, thresholds, and topographies. *J Nucl Med*. 2021;62(5):614-619. https://doi.org/10.2967/jnumed.120.245423
3. Jack CR Jr, Wiste HJ, Schwarz CG, et al. Longitudinal tau PET in ageing and Alzheimer's disease. *Brain*. 2018;141:1517-1528.
4. Betthauser TJ, Koscik RL, Jonaitis EM, et al. Amyloid and tau imaging biomarkers explain cognitive decline from late middle-age. *Brain*. 2020;143:320-335.
5. Pascoal TA, Therriault J, Benedet AL, et al. 18F-MK-6240 PET for early and late detection of neurofibrillary tangles. *Brain*. 2020.
6. Tanner JA, Rabinovici GD. Relationship between Tau and cognition in the evolution of Alzheimer's disease: new insights from Tau PET. *J Nucl Med*. 2020. jnumed.120.257824.
7. Vogel JW, Young AL, Oxtoby NP, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. 2021: 1-11.
8. Dronse J, Fliessbach K, Bischof GN, et al. n vivo patterns of Tau pathology, amyloid-β burden, and neuronal dysfunction in clinical variants of Alzheimer's disease. *I*. 2017;55:465-471.
9. Sonni I, Segev OHL, Baker SL, et al. Evaluation of a visual interpretation method for tau-PET with 18F-flortaucipir. *Alzheimer's Dement Diagn Assess Dis Monit*. 2020;12:e12133.
10. Harn NR, Hunt SL, Hill J, Vidoni E, Perry M, Burns JM. Augmenting amyloid PET interpretations with quantitative information improves consistency of early amyloid detection. *Clin Nucl Med*. 2017;42:577-581.
11. Leuzy A, Pascoal TA, Strandberg O, et al. A multicenter comparison of [18F]flortaucipir, [18F]RO948, and [18F]MK6240 tau PET tracers to detect a common target ROI for differential diagnosis. *European Journal of Nuclear Medicine and Molecular Imaging*. 2021;48(7):2295-2305. https://doi.org/10.1007/s00259-021-05401-4
12. Smith R, Strandberg O, Leuzy A, et al. Sex differences in off-target binding using tau positron emission tomography. *NeuroImage: Clin*. 2021: 102708.
13. Ossenkoppele R, Hansson O. Towards clinical application of tau PET tracers for diagnosing dementia due to Alzheimer's disease. *Alzheimer Dement*. 2021.
14. Feng X, Lipton ZC, Yang J, Small SA, Provenzano FA. Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging. *Neurobiol Aging*. 2020;91:15-25.
15. Ocasio E, Duong TQ. Deep learning prediction of mild cognitive impairment conversion to Alzheimer's disease at 3 years after diagnosis using longitudinal and whole-brain 3D MRI. *PeerJ Comput Sci*. 2021;7:e560.
16. Ding Y, Sohn JH, Kawczynski MG, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using ^{18} F-FDG PET of the brain. *Radiology*. 2019;290:456-464.
17. Liu H, Nai Y-H, Saridin F, et al. Improved amyloid burden quantification with nonspecific estimates using deep learning. *Eur J Nucl Med Mol Imag*. 2021;48(6):1842-1853. https://doi.org/10.1007/s00259-020-05131-z
18. Kim J-Y, Oh D, Sung K, et al. Visual interpretation of [18F]Florbetaben PET supported by deep learning–based estimation of amyloid burden. *Eur J Nucl Med Mol Imag*. 2021;48(4):1116-1123. https://doi.org/10.1007/s00259-020-05044-x
19. Son HJ, Oh JS, Oh M, et al. The clinical feasibility of deep learning-based classification of amyloid PET images in visually equivocal cases. *Eur J Nucl Med Mol Imaging*. 2020;47:332-341.
20. Alven J, Heurling K & Smith R, et al. A deep learning approach to MR-less spatial normalization for Tau PET images n.d.:9. https://doi.org/10.1007/978-3-030-32245-8_40
21. Macdonald T, Chen K, Koran ME, Moseley M, Zaharchuk G. DualNet: a deep neural network to predict individual tau and amyloid PET images from a combined dose image using the disambiguation of dual dose amyloid-tau PET scans using the ADNI dataset. *J Nucl Med*. 2020;61:3009-3009.
22. Jo T, Nho K, Risacher SL, Saykin AJ. for the Alzheimer's Neuroimaging Initiative. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. *BMC Bioinformatics*. 2020;21:496.
23. Manly JJ, Bell-McGinty S, Tang M-X, Schupf N, Stern Y, Mayeux R. Implementing diagnostic criteria and estimating frequency of mild cognitive impairment in an urban community. *Arch Neurol*. 2005;62:1739-1746.
24. Luchsinger JA, Palta P, Rippon B, et al. Sex Differences in in vivo Alzheimer's Disease Neuropathology in Late Middle-Aged Hispanics. *J Alzheimer's Dis*. 2020: 1-10.
25. Devanand DP, Andrews H, Kreisl WC, et al. Antiviral therapy: valacyclovir Treatment of Alzheimer's Disease (VALAD) Trial: protocol for a randomised, double-blind,placebo-controlled, treatment trial. *BMJ Open*. 2020;10:e032112.
26. Zou J, Tao S, Johnson A, et al. Microglial activation, but not tau pathology, is independently associated with amyloid positivity and memory impairment. *Neurobiol Aging*. 2020;85:11-21. https://doi.org/10.1016/j.neurobiolaging.2019.09.019

27. Aisen PS, Petersen RC, Donohue MC, et al. Clinical core of the Alzheimer's disease neuroimaging initiative: progress and plans. *Alzheimers & Dement*. 2010;6:239-246.
28. Weiner MW, Veitch DP, Aisen PS, et al. The Alzheimer's Disease Neuroimaging Initiative 3: continued innovation for clinical trial improvement. *Alzheimers Dement*. 2017;13:561-571.
29. Folstein MF, Robins LN, Helzer JE. The Mini-Mental State Examination. *Arch Gen Psychiatry*. 1983;40:812-812.
30. Grober E, Sanders AE, Hall C, Lipton RB. Free and Cued Selective Reminding Identifies Very Mild Dementia in Primary Care. *Alzheimer Dis Assoc Disord*. 2010;24:284-290.
31. Bullich S, Seibyl J, Catafau AM, et al. Optimized classification of 18F-Florbetaben PET scans as positive and negative using an SUVR quantitative approach and comparison to visual assessment. *NeuroImage: Clinical*. 2017;15:325-332.
32. Liu S, Yadav C, Fernandez-Granda C, & Razavian N. On the design of convolutional neural networks for automatic detection of Alzheimer's disease. Proceedings of the Machine Learning for Health NeurIPS Workshop, PMLR 116:184-201, 2020.
33. Palmqvist S, Janelidze S, Quiroz YT, et al. Discriminative Accuracy of Plasma Phospho-tau217 for Alzheimer Disease vs Other Neurodegenerative Disorders. *JAMA*. 2020.
34. Klein A, Ghosh SS, Avants B, et al. Evaluation of volume-based and surface-based brain image registration methods. *Neuroimage*. 2010;51:214-220.
35. Paszke A, Gross S, Chintala S, Chanan G, Yang E & DeVito Z et al. Automatic differentiation in PyTorch. in NIPS-W (Long Beach, CA). 2017.
36. Oliphant TE. *A guide to NumPy*. Trelgol Publishing USA; 2006.
37. McKinney W. *Data Structures for Statistical Computing in Python*. Austin, Texas. 2010: 56-61.
38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research*. 2011;12:2825-2830.
39. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. *Rethinking the Inception Architecture for Computer Vision*. 2016: 2818-2826.
40. Schöll M, Lockhart SN, Schonhaut DR, et al. PET imaging of tau deposition in the aging human brain. *Neuron*. 2016;89:971-982.
41. Schwarz CG, Gunter JL, Lowe VJ, et al. A Comparison of Partial Volume Correction Techniques for Measuring Change in Serial Amyloid PET SUVR. *J Alzheimers Dis*. 2019;67:181-195.
42. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biomet J*. 2005;47:458-472.
43. 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.
44. Halekoh U, Højsgaard SA. Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *J Stat Softw*. 2014;59:1-32.
45. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. *Computer Vision – ECCV 2014*. Cham: Springer International Publishing; 2014: 818-833. editors.
46. Choi H, Jin KH. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behav Brain Res*. 2018;344:103-109.
47. Braak H, Del Tredici K. The pathological process underlying Alzheimer's disease in individuals under thirty. *Acta Neuropathol*. 2011;121:171-181.
48. Braak H, Zetterberg H, Del Tredici K, Blennow K. Intraneuronal tau aggregation precedes diffuse plaque deposition, but amyloid-β changes occur before increases of tau in cerebrospinal fluid. *Acta Neuropathol*. 2013;126(5):631-641. https://doi.org/10.1007/s00401-013-1139-0
49. Hammes J, Bischof GN, Bohn KP, et al. One-Stop Shop: 18F-Flortaucipir PET Differentiates Amyloid-Positive and-Negative Forms of Neurodegenerative Diseases. *J Nucl Med*. 2021;62:240-246.
50. Stephan BCM, Minett T, Pagett E, Siervo M, Brayne C, McKeith IG. Diagnosing Mild Cognitive Impairment (MCI) in clinical trials: a systematic review. *BMJ Open*. 2013;3.

SUPPORTING INFORMATION

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

How to cite this article: Zou J, Park D, Johnson A, et al., for the Alzheimer's Disease Neuroimaging Initiative. Deep learning improves utility of tau PET in the study of Alzheimer's disease. *Alzheimer's Dement*. 2021;13:e12264.
https://doi.org/10.1002/dad2.12264