Journal of Nuclear Medicine, published on March 12, 2021 as doi:10.2967/jnumed.120.250290

Towards a Universal Readout for Fluorine-18 Labelled Amyloid Tracers: The CAPTAINs Study

Gérard N Bischof, PhD[1], Peter Bartenstein[2], Henryk Barthel[3], Bart van Berckel[4], Vincent Doré[5,6], Thilo van Eimeren[1,7,8], Norman Forster[9], Jochen Hammes[1], Adriaan A. Lammertsma[4], Satoshi Minoshima[9], Chris Rowe[5,6], Osama Sabri[3], John Seibyl[10], Koen Van Laere[11], Rik Vandenberghe[12], Victor Villemagne[5, 6], Igor Yakushev[13] & Alexander Drzezga[1,8,14]

(1) University Hospital Cologne, Multimodal Neuroimaging Group, Department of Nuclear Medicine, Cologne, Germany (2) Department of Nuclear Medicine, LMU Munich, Germany (3) University Hospital of Leipzig, Department of Nuclear Medicine, Leipzig, Germany (4) Amsterdam University Medical Centers, Location VUmc Radiology and Nuclear Medicine, Amsterdam, Netherlands (5) CSIRO Health and Biosecurity, Parkville 3052, Victoria, Australia (6) Department of Molecular Imaging & Therapy, Austin Health, Melbourne, Australia (7) Department of Neurology, University Hospital Cologne, Cologne, Germany (8) German Center of Neurodegenerative Disease (DZNE) (9) Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, USA (10) Institute for Neurodegenerative Disorders, New Haven, Connecticut, USA (11) Nuclear Medicine and Molecular Imaging, University Hospital Leuven and Department of Imaging and Pathology KU Leuven, Leuven, Belgium (12) Memory Clinic, University Hospital Leuven and Department of Neurosciences, KU Leuven, Belgium (13) Department of Nuclear Medicine, Technical University of Munich, Germany, (14) Institute of Neuroscience and Medicine (INM-2), Molecular Organization of the Brain, Forschungszentrum Jülich, Germany

Corresponding Author:
Gerard N Bischof, PhD
Department of Nuclear Medicine
Multimodal Imaging Laboratory University Hospital of Cologne
Kerpener Str. 62, 50937
Cologne, Germany
Phone: +49 221 478 86621
Fax: +49 221 478 89085
Email: gerard.bischof@uk-koeln.de

ABSTRACT:

To date, three fluorine-18 labelled PET tracers have been approved for assessing cerebral amyloid plaque pathology in diagnostic work-up of suspected Alzheimer's Disease (AD). Although scanning protocols are relatively similar across tracers, U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) approved visual rating protocols differ between the three tracers. This proof-of-concept study assessed the comparability of the three approved visual rating protocols to classify a scan as amyloid-positive or -negative, when applied by groups of experts and non-experts to all three amyloid tracers.

Methods:

In an international multicentre approach, both experts (N=4) and non-expert raters (N=3) rated scans acquired with ^{18}F-Florbetaben, ^{18}F-Florbetapir and ^{18}F-Flutemetamol. Scans obtained with each tracer were presented for reading according to all three approved visual rating protocols. In a randomized order, every single scan was rated by each reader according to all three protocols. Raters were blinded for the amyloid tracer used and asked to rate each scan as positive or negative, giving a confidence judgement after each response. Percentage of visual reader agreement, inter-rater reliability and agreement of each visual read with binary quantitative measures (fixed SUVR-threshold for positive/negative scans) were computed. These metrics were analyzed separately for expert and non-expert groups.

Results:

No significant differences in using the different approved visual rating protocols were observed across the different metrics of agreement in the group of experts. Nominal differences suggested that the Florbetaben visual rating protocol achieved the highest interrater reliability and accuracy especially under low confidence conditions. For the group of non-expert raters, significant differences between the different visual rating protocols were observed with overall moderate-to-fair accuracy and with the highest reliability for the Florbetapir visual rating protocol.

Conclusion:

We observed high interrater agreement despite applying different visual rating protocols for all ^{18}F-labelled amyloid tracers. This implies that the results of the visual interpretation of amyloid imaging can be well standardized and do not depend on the rating protocol in experts. Consequently, the creation of a universal visual assessment protocol for all amyloid imaging tracers appears feasible, which could benefit especially the less experienced readers.

Key Words: Florbetapir, Florbetaben, Flutemetamol, Amyloid PET, Visual reading standardization

INTRODUCTION

The advent of biomarkers of neuritic β-amyloid pathology (Aβ) using either cerebrospinal fluid or positron emission tomography (PET) has shifted the conceptualization of a strictly clinical diagnosis of Alzheimer's disease (AD) (*1*) to the diagnosis of the presence or absence of the underlying pathology itself (*2*). Cerebrospinal fluid biomarkers measuring the concentration levels of Aβ42 or Aβ40 peptides show substantial variability in sensitivity [sensitivity (Range) = 48.0-93.3] and specificity [specificity(Range) = 67.0-100.0] in discriminating healthy controls from AD dementia patients (*3*) . Although the ratio of Aβ 42/Aβ 40 may improve the diagnostic accuracy in advanced cases of the prodromal phase of AD (*3*), some heterogeneity using cerebrospinal fluid biomarkers of Aβ pathology exist and so far there has been no agreement on harmonizing analysis protocols or thresholds (*4*). Furthermore, cerebrospinal fluid measures are generally not suitable for assessing regional accumulation of Aβ pathology, have only a moderate test-retest reliability and hence are not ideal in evaluating disease progression. *In vivo* PET imaging with selective Aβ tracers can capture regional burden and progression and may therefore be better suited as progression marker and as a primary outcome measure in pharmaceutical clinical trials.

The use of amyloid PET biomarkers in the clinical work-up of patients with cognitive decline and its relevance for diagnosis and subsequent patient management has now been evaluated in both North America (*5*) and Europe (*6*). At present, three fluorine labelled tracers (^{18}F) Florbetapir (FBP), Flutemetamol (FLUTE) and Florbetaben (FBB) are approved by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). These tracers are commercially distributed under the following names: Amyvid (Eli Lilly; Florbetapir), Vizamyl (GE; Flutemetamol) and Neuraceq (Florbetaben; Life Molecular Imaging).

Appropriate use criteria have been formalized for these tracers (e.g.,(*7*)). FDA/EMA approved tracer-specific visual rating guidelines, to determine whether an Aβ scan is positive or negative have been provided, and a detailed training program for all three tracers is required before user certification (*8*–*10*). The general principle underlying the visual rating schemes is similar across the three tracers. Specifically, a physician is trained in identifying the loss in contrast of neocortical grey matter compared with adjacent white matter regions. In detail, however, there is considerable variability among the visual rating guidelines, such as color scale used, intensity scaling, definition of target regions, or number of regions, as well as spatial and signal thresholds to determine regional positivity/negativity, and translation from regional to global positivity/negativity. This readout variability may contribute to the observed diagnostic variability in sensitivity [sensitivity (Range) = 89.0-97.0] and specificity [specificity (Range) = 63.0-93.0]

measures among all flourine-18 labelled amyloid tracers (*11–13*). However, so far they have not been cross-evaluated in a head-to-head study design.

Current alternatives to visual reads for the assessment of Aβ-positivity are quantitative measures and harmonization approaches of flourine-18 labelled amyloid tracers with the gold-standard carbon-11 labelled amyloid tracers such as the Centiloid scale have been proposed (*14*,*15*). However, it is important to note that despite the development of standardized quantification approaches, the default in the clinical routine for the assessment of Aβ-status is the application of the approved visual rating approaches. Here, we aim to gather information for a possible harmonization approach for the approved visual reading approaches to avoid potential dependence of diagnostic and therapeutic decisions on the type of tracer and/or the interpretation protocols used. Therefore, the goal of the current study was to compare amyloid PET tracer-associated interpretation strategies (CAPTAINs) of the three FDA/EMA -approved visual rating protocols for the three approved Aβ-tracers in a group of experts and non-expert raters. A specific aim was to identify which aspects of the three visual rating protocols allowed the most reliable identification of Aβ positive and negative scans across experts and non-expert raters and which reading parameters could potentially be suitable for a unified visual rating scheme. Finally, to evaluate the effect of visual reader training the inclusion of non-expert raters was paramount.

MATERIALS AND METHODS

PET Images

The study included data from all three FDA/EMA approved fluorine-18-labelled tracers for imaging of neuritic Aβ pathology (i.e., FBP, FBB, FLUTE) from healthy controls, individuals clinically diagnosed with mild cognitive impairment and AD dementia patients.

For each tracer we included 10 scans in total 30 unique scans, with 10 HC, 10 MCI and 10 AD patients. With 7 readers and 3 different reading system this resulted in a total of 630 responses across the sample of experts and non-experts. The inclusion criteria for the subjects in the sample were derived from the Australian Imaging, Biomarkers and Lifestyle (ABIL) flagship study of aging. In brief, participants were allocated to one of the three diagnostic groups based on a clinical review that used the NINCDS-ARDA criteria for AD, Petersen et al., criteria for MCI and criteria for normal cognitive function for healthy controls (*16*). We matched the selected images from each tracer by age ($Mean_{(age)}$ = 73.9, $STD_{(age)}$ =6.9; F (2, 29) = 2.65, ns) MMSE ($Mean_{(MMSE)}$ = 23.7, $STD_{(MMSE)}$

=5.6; F (2,29=2.1, ns) and Education (Mean $_{(Education)}$ = 12.9; STD $_{(Education)}$ =1.91; F(2,29) =1.10, ns).

Scans of each of the three Aβ-tracers were prepared for visual reading according to all three of the recommended and FDA/EMA approved guidelines as provided by the vendors in their respective package-inserts. All scans were then presented for rating according to all three of the approved visual rating protocols (*see* Figure S1). Thus, in a randomized order, every single scan was rated by each reader according to all three protocols (e.g. Florbetapir scans were rated according to Florbetapir-, Florbetaben- and Flutemetamol-guidelines, etc.). Additionally, to examine intra-rater reliability we added repetitions of the same image and the same visual rating protocol totally 12 responses from each rater. The number responses collected were *N* = 630 for the interrater analysis and *N* = 84 responses for the intrarater analysis totaling to *N* =714 overall. Raters were blinded for the Aβ-tracer used. To assess standard of truth measures of positivity and negativity, SUV images were intensity-normalized using the whole cerebellum as reference region for Florbetapir and cerebellar cortex as a reference region for Florbetaben and the pons as a reference region for Flutemetamol to create standard uptake value ratio images (SUVR) (further details are provided in the *supplementary material*). Importantly, thresholds for positivity and negativity were not derived from the current sample but defined on the basis of previously published end-of-life studies of corresponding histopathological Aβ-amyloid plaque burden and corresponding SUVRs for each of the tracers, FBP (*17*), FBB (*18*) and FLUTE (*19*). Noteworthy, autopsy data were not available for the current sample, so that thresholds of positivity and negativity defined here, do not allow direct conclusions about the true underlying neuropathology.

Acquisition protocol for PET Images:

All scans of the study were provided by the Department of Molecular Imaging & Therapy, Austin Health, Melbourne, Australia. These scans were acquired on different PET scanners which is summarized in the Table 1 below. Each participant received a 20-minute PET scan with one of the three ^{18}F tracers. The scan was performed 50 minutes post-injection of 370MBq (+/-10%) Florbetapir or 90 minutes post-injection of 185MBq (+/-10%) Flutemetamol or 300MBq (+/-10%) Florbetaben. PET scans were spatially normalised using CapAIBL (https://milxcloud.csiro.au/,(*20*))). The images were then scaled to the SUV of the cerebellum cortex to generate a tissue ratio termed SUV ratio (SUVR).

A Global measure of Aβ burden was computed using the mean SUVR in the frontal, superior parietal, lateral temporal, occipital and anterior and posterior cingulate regions of the brain.

Standard uptake value ratio image computation:

Neocortical retention was estimated using a composite region of frontal (dorsolateral, ventrolateral and orbitofrontal), parietal superior parietal and precuneus), lateral temporal (superior, middle and inferior), lateral occipital lobe (lateral temporal and temporo-occipital), gyrus supramarginalis, gyrus angularis and anterior and posterior cingulate. The scaling of the images generates a tissue ratio called the Standardized Uptake Value Ratio (SUVR), which is the ratio of the global composite and the tracer-specific reference region.

Raters

Expert raters (N=4) were either licensed neurologists or licensed nuclear medicine physicians with outstanding expertise in molecular imaging (HB, BvB, CR, JS). Importantly, all raters had undergone the tracer-specific reading training for all three ^{18}F Aβ-tracers, culminating in a threefold expert certification. Further, all expert raters had several years of experience of visual reading and were very familiar with all reading approaches.

Non-expert raters (N=3) were medical doctoral students (HT, MM, OR) enrolled in the medical program of the University Cologne, Germany. All three non-expert raters were pursuing a medical doctoral thesis at University Hospital Cologne, Germany with some general experience in Nuclear Medicine acquired during their doctoral training, but little experience with image reading. Non-expert raters underwent a 30-minutes standardized introduction to the published guidelines for visual readings for all three tracers and completed five examples.

Rating Procedure

An in-house online rating platform was created to ensure remote accessibility for the international group of raters from their home institution. Specific instructions on how to maneuver the online platform were made available prior to distributing personalized links to each rater. Images were displayed in random order and suffixed with the respective rating protocol (i.e., FBB, FBP, FLUTE rating protocol). All images were displayed in the recommended color scale according to each visual rating protocol (i.e., grey-scale, black-and-white and Sokoloff/Spectrum respectively). Datasets for each rater included all images presented in all three visual rating scales independently of the PET tracer utilized and raters were asked to judge if they were positive or negative based on the corresponding visual rating protocol (*see* Figure S2). Raters were able to review the guidelines of all three visual rating protocols on the main homepage. Images appeared on three windows including axial, sagittal and coronal views, with the main window displayed on

an axial plane by default. A rating form was available upon mouse click and required the rater to assess whether the scan was amyloid positive or -negative and to indicate the corresponding confidence on a scale from 1 to 10. The online platform automatically recorded the response and confidence level paralleled with a time stamp (Details see supplementary material).

Statistical Analysis

Intra-rater reliability was performed on the responses related to the repetitions and was computed using the two-way intraclass coefficient (ICC) for experts and non-experts separately. To evaluate the inter-rater agreement across experts and non-expert raters separately, three statistical metrics were used: (1) consistency given as the percentage of scans rated identical across raters, (2) accuracy computed as the percentage agreement with tracer specific quantitative SUVR positivity/negativity measures, and (3) Krippendorff's alpha, a metric of interrater-reliability used for more than two raters. Krippendorff`s alpha calculates the alpha coefficient of reliability by comparing the observed disagreement with the expected disagreement (*21*). As the consistency measures only include a simple percentage of agreement, Krippendorff`s alpha reflects the individual error-corrected agreement, similar to the Fleiss Kappa coefficient of reliability (*22*). Whereas an alpha = 1, indicate perfect reliability and an alpha = 0 indicate the absence of reliability, some authors have suggested the following range of benchmarks of .21-.40 "fair" agreement, .41 to .60 "moderate" agreement, .61-.80 "substantial" agreement and .81 to 1" near perfect" to assist with the interpretation of Krippendorff's alpha (*23*).

The Generalized Estimating Equation (*24*) was used to assess differences in responses as a function of visual reading method (i.e., main effect method). Significance threshold was set at a *p*-value of <*.05*. Finally, we examined confidence-accuracy characteristic (CAC) across all responses to evaluate if accuracy is moderated as a function of confidence and if this relationship potentially differs by tracer. Only responses were included from those experts (N=3) and non-expert raters (N=3) who utilized the entire range of confidence judgements and binned their responses into low (0-5) and high confidence (6-10) and analyzed accuracy values based on the quantitative SUVR measures for all 600 ratings.

RESULTS

Intrarater Reliability:

Intra-rater reliability was high among the four experts (ICC=.92) and moderate among the three non-experts (ICC=.68).

Interrater Reliability

Expert Raters.

Among the four expert raters only slight variations across the visual reading protocols were observed. Consistency measures of FBB and FLUTE visual rating protocols produced similar values among expert raters (.95 and .94 respectively). The use of the FBP rating protocol showed overall the lowest consistency judgements across raters (.90). Comparing visual ratings to SUVR values for positivity and negativity agreement (i.e. accuracy), slight differences were observed. Specifically, whereas reading according to FBB and FLUTE visual reading protocols showed accuracy values of .86 and .89 respectively. The use of the FBP reading protocol showed accuracy values of .90 among raters. A summary of the reading accuracy is depicted in Figure 1.

Finally, interrater-agreement (Krippendorf's alpha) was highest for the FBB (.79) and the FLUTE visual reading protocol (.75) and lowest for the FBP visual reading method (.68) *see* Figure 1. Estimating if expert rater responses differ as a function of visual rating procedure, we employed the generalized estimating equation on the consistency and accuracy measures and observed no significant main effect of method on either metric (Consistency: $W_{chisquare}$ =3.56, p=.17; Accuracy: $W_{chisquare}$=2.55 p=.28). A summary of these results is displayed in *Table 1.1*. Together, we observed no significant differences between the use of the three visual rating protocols to render a scan positive or negative and the overall rater agreement was high.

Non-experts.

Visual reading methods among non-experts were less consistent. Specifically, whereas the use of FBB (.70) and FBP (.72) visual rating protocols showed acceptable consistency values, the FLUTE protocol reached consistency at the chance level across non-expert raters (.50). When responses were compared to the SUVR thresholds, accuracy was highest for the FBP visual rating protocol (.62), followed by the FBB (.55) and lowest for the FLUTE (.51) protocols (see Figure 1). This general result pattern is reflected in measures of interrater-agreement (see Figure 1 Visual reading method; FLUTE: .35, FBB=.47, and FBP = .63). Finally, both consistency and accuracy showed a significant main effect of method (Consistency: $W_{chisquare}$ =20.62, p<.001; Accuracy: $W_{chisquare}$= 9.08 p=.001). A summary of these results is displayed in *Table 1.2*.

----------Figure 1 about here----------

----------Table 2 about here----------

Confidence-accuracy characteristic (CAC) Analysis

In both groups, experts and non-expert, low confidence judgements were associated with lower accuracy values independent of the actual visual rating scheme used (see Figure 2). Furthermore, in the expert group, even in low confidence conditions, experts showed the highest accuracy values for the FBB visual rating protocol, whereas for the FBP and FLUTE protocols, accuracy values dropped to chance level when experts indicated low confidence in rating a scan as either positive or negative.

For non-expert raters, the FBP visual rating protocol ´showed the highest accuracy (.58) for low confidence judgements, whereas FBB and FLUTE protocols either approached (.56) or fell even below chance level (.41) for responses accompanied with low confidence.

----------Figure 2 about here----------

DISCUSSION

The main purpose of the present study was to determine the comparability and potential interchangeability of the three FDA/EMA-approved visual rating protocols on the three amyloid-tracers both in experts and non-experts. To this end, experts and non-experts together rated over 700 scans as positive or negative accompanied with a confidence judgement. All FBB, FBP and FLUTE images were presented in all three visual interpretation modes.

We observed that different metrics of interrater agreement did not significantly differ by visual rating protocols in the group of experts. Qualitatively, nominal differences were observed in favor of the FBB visual rating protocol, as interrater reliability was highest and confidence-accuracy analysis suggests that even in low confidence conditions visual rating mostly agreed with quantitative SUVR measures across experts.

For non-expert raters, accuracy and interrater-reliability was dependent on the visual rating protocol and was highest when using the FBP visual rating protocol. Overall non-expert raters' responses showed only moderate and fair agreement confirming that specific training is required in order to accurately evaluate Aβ images. The results also suggest that particularly inexperienced readers may additionally benefit from a universal visual rating protocol for all three FDA/EMA approved Aβ-tracers. In the following we will discuss in more detail the implications of our study findings.

Standardization of visual rating protocols for fluorine-18 labelled amyloid tracers

As Aβ-tracers evidenced improved utility in the differential diagnosis, patient care and management in both North America and Europe (*5*,*6*), it is expected that *in vivo* imaging of Aβ-amyloid pathology will be increasingly used in the routine clinical work up in patients with suspected neurodegenerative disease, as well as inclusion for therapeutic trials. Our data in the group of experts showed that sufficient levels of agreement on rendering a scan as positive or negative can be reached independently of the visual rating protocol used. Consequently, these results indeed suggest that the available rating protocols in combination with suitable reader training ensure adequate levels of standardization of the visual assessment of Aβ-amyloid pathology across the AD spectrum. Additional standardization efforts to simplify and standardize the visual reading may be feasible and particularly meaningful for less experienced readers, as significant heterogeneity among the three visual rating protocols was detected in the group of non-expert raters. From a practical point of view the development of a universal readout for ^{18}F-Aβ-tracers may indeed be a straight forward solution to ensure comparability across differently trained specialist in regions where not all three FDA/EMA approved Aβ-tracers are available (e.g., Europe: FBB and FLUTE but not FBP), as well as in multicenter international therapeutic trials where the three tracers are used. The universal readout includes a consistent starting point and the demarcation of standardized landmarks where the reader would examine significant loss of white/gray matter contrast, a clear definition of the size of a region and a recommendation for the type of reading scale.

Optimally, a universal readout could possibly be validated against neuropathological Aβ-amyloid plaque burden in the previous conducted end-of-life studies. Standardization approaches for quantitative purposes to reduce heterogeneity when measuring SUVRs have been suggested to achieve comparability between fluorine-18 labelled amyloid tracers and 11C-PiB, the gold-standard tracer for beta-amyloid pathology (*25*). For this purpose, the centiloid scale has been introduced, which linearly scales the measurement of the tracer from zero to 100, with zero representing the average uptake of young amyloid-negative individuals and 100 the retention of a typical Alzheimer's disease patient. When the centiloid scale is used, thresholds of 20 to 25 centiloids correspond to positive visual assessment (*15*). Although quantitative retention measures may aid in the visual assessment of Aβ-amyloid scans, they are currently not part of clinical routine work-up. Also, centiloids are based on SUVR measures, which have been discussed to be susceptible to asymmetric perfusion changes over time in reference and target regions, potentially affecting longitudinal evaluation e.g. of therapy effects (*26*). Nevertheless, it would be of great interest in future research to include centiloid values across ^{18}F Aβ-tracers to

assist in the visual readings and systematically examine if interrater reliability improves significantly among experts and non-expert raters. A combination of data-driven and/or artificial intelligence driven approaches for amyloid imaging with different fluorine-18 labelled tracer may an additional future direction that could potentially assist in clinical read outs.

Limitations

The present study has some limitations. Although, experts and non-expert rated over 700 images in total, a differential analysis by tracer or disease category was not possible due to the limited number of scans available per category. Further, this convenience sample may not have captured the wider range of potential cases present in the general population. Adding more scans to the existing sample would certainly allow additional analyses, but inadvertently increase the amount of rating time. Such an effort may, however, improve the design of a universal readout, and may reveal some nuances in advancing the validity of a universal readout. As we intend, in a planned follow-up study, to increase the set of images beyond the convenience sample of images presented here, we aim to encompass the entire range of cases that may be present within a clinical context. In this first step of the CAPTAINs Project we intended to focus on matching the images carefully by several characteristics including, age, gender, demographic information, SUVR threshold and by diagnostic category.

The chosen standard of truth method for positivity were SUVR measures which were informed by previous end-of life studies and inferred from histopathological correlation. However, pathological confirmation was not available for the rated scans, which would have been the ideal standard of truth confirmation for positive and negative scans.

Additionally, all scans were provided from the same research center, but scans were acquired from different scanners, so this study design does not account for potential differences or similarity that are scanner- and/or site-dependent. Potentially, different scanner types may have impacted visual rating results. However, potential differences based on the scanner type would have affected all three rating protocols equally and differences were minimized by ensuring that preprocessing was done using the same analysis pipeline (details see supplementary material). Finally, the visual rating protocols recommend the use of co-registered CT/MRI scans particularly in cases of low image quality to discern possible anatomical boundaries that may have been influenced by atrophy. In the current study we refrained from providing additional CT information to focus on the standard visual rating procedure.

CONCLUSION

Our study indicates that the results of the visual interpretation of amyloid imaging can be well standardized and do not depend relevantly on the visual rating protocol in expert readers. At the same time, these results suggest that the creation of a universal visual readout protocol for all amyloid-imaging tracers may be feasible. Especially less experienced readers could benefit from such a universal readout protocol.

ACKNOWLEDGEMENT

The authors are very grateful for the contribution of Hendrik Theis (HT), Michelle Meier (MM) und Omer Rainer (OR) for their time and assistance in the study design.

Disclosure:

GNB reports the following conflicts of interest: Speaker Honorary: Life Molecular Imaging.

AD reports the following conflicts of interest: Research support: Siemens Healthineers, Life Molecular Imaging, GE Healthcare, AVID Radiopharmaceuticals, Speaker Honorary/Advisory Boards: Siemens Healthineers, Sanofi, GE Healthcare, Stock: Siemens Healthineers
Patents: Patent pending for 18F-PSMA7 (PSMA PET imaging tracer for prostate cancer).

JS reports the following disclosures: Consultancy: Biogen, Roche, AbVie, Life Molecular Imaging, LikeMinds, Invicro. Equity stake: Invicro

No other potential conflicts of interest relevant to this article exist.

KEY POINTS:

QUESTION: Are the FDA-approved visual rating protocols for the three currently available 18F-labeled tracers for amyloid-imaging considerably different in evaluating an amyloid scan as positive or negative?

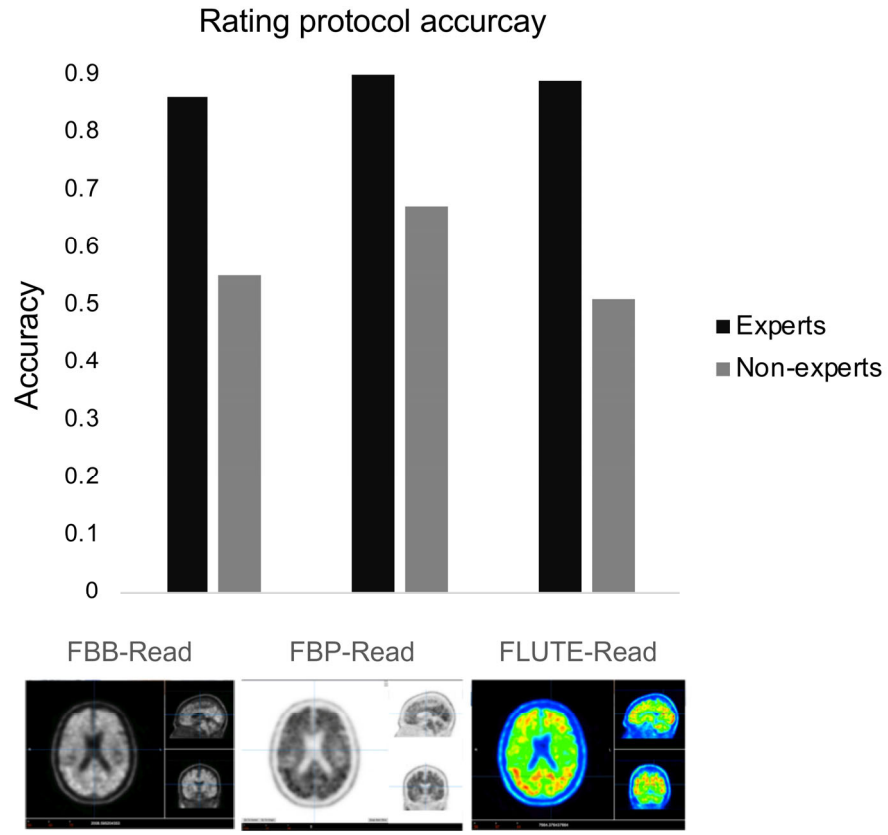
FINDINGS: We demonstrate that overall accuracy was high and that experts did not significantly differ in their accuracy or interrater agreement as a function of the visual rating procedure utilized. In non-experts' significant differences arose suggesting that reader training is necessary to examine beta-amyloid scans.

IMPLICATIONS FOR PATIENT CARE: These results support the notion that rating of amyloid-imaging achieves high levels of standardization which may serve as an important argument to justify the application of a modern Nuclear Medicine procedure for clinical and scientific purposes and to prefer it over other available options.

REFERENCES

1. McKhann GM, Knopman DS, Chertkow H, 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.
2. Jack CR, Bennett DA, Blennow K, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14:535-562.
3. Ritchie C, Smailagic N, Noel-Storr AH, et al. Plasma and cerebrospinal fluid amyloid beta for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*. 2014:CD008782.
4. Hansson O, Lehmann S, Otto M, Zetterberg H, Lewczuk P. Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease. *Alzheimer's Research & Therapy*. 2019;11:34.
5. Rabinovici GD, Gatsonis C, Apgar C, et al. Association of Amyloid Positron Emission Tomography With Subsequent Change in Clinical Management Among Medicare Beneficiaries With Mild Cognitive Impairment or Dementia. *JAMA*. 2019;321:1286-1294.
6. de Wilde A, van der Flier WM, Pelkmans W, et al. Association of Amyloid Positron Emission Tomography With Changes in Diagnosis and Patient Treatment in an Unselected Memory Clinic Cohort. *JAMA Neurol*. 2018;75:1062-1070.
7. Johnson KA, Minoshima S, Bohnen NI, et al. Update on appropriate use criteria for amyloid PET imaging: dementia experts, mild cognitive impairment, and education. Amyloid Imaging Task Force of the Alzheimer's Association and Society for Nuclear Medicine and Molecular Imaging. *Alzheimers Dement*. 2013;9:e106-109.
8. Buckley CJ, Sherwin PF, Smith APL, Wolber J, Weick SM, Brooks DJ. Validation of an electronic image reader training programme for interpretation of [18F]flutemetamol β-amyloid PET brain images. *Nucl Med Commun*. 2017;38:234-241.
9. Seibyl J, Catafau AM, Barthel H, et al. Impact of Training Method on the Robustness of the Visual Assessment of 18F-Florbetaben PET Scans: Results from a Phase-3 Study. *J Nucl Med*. 2016;57:900-906.
10. Pontecorvo MJ, Arora AK, Devine M, et al. Quantitation of PET signal as an adjunct to visual interpretation of florbetapir imaging. *Eur J Nucl Med Mol Imaging*. 2017;44:825-837.
11. Martínez G, Vernooij RW, Fuentes Padilla P, Zamora J, Bonfill Cosp X, Flicker L. 18F PET with florbetapir for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*. 2017;11:CD012216.
12. Martínez G, Vernooij RW, Fuentes Padilla P, Zamora J, Flicker L, Bonfill Cosp X. 18F PET with florbetaben for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*. 2017;11:CD012883.
13. Martínez G, Vernooij RW, Fuentes Padilla P, Zamora J, Flicker L, Bonfill Cosp X. 18F PET with flutemetamol for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*. 2017;11:CD012884.
14. La Joie R, Ayakta N, Seeley WW, et al. Multisite study of the relationships between antemortem [11C]PIB-PET Centiloid values and postmortem measures of Alzheimer's disease neuropathology. *Alzheimers Dement*. 2019;15:205-216.
15. Amadoru S, Doré V, McLean CA, et al. Comparison of amyloid PET measured in Centiloid units with neuropathological findings in Alzheimer's disease. *Alzheimers Res Ther*. 2020;12:22.
16. Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle

(AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr*. 2009;21:672-687.
17. Clark CM, Schneider JA, Bedell BJ, et al. Use of florbetapir-PET for imaging beta-amyloid pathology. *JAMA*. 2011;305:275-283.
18. Sabri O, Sabbagh MN, Seibyl J, et al. Florbetaben PET imaging to detect amyloid beta plaques in Alzheimer's disease: Phase 3 study. *Alzheimer's & Dementia*. 2015;11:964-974.
19. Ikonomovic MD, Buckley CJ, Heurling K, et al. Post-mortem histopathology underlying β-amyloid PET imaging following flutemetamol F 18 injection. *Acta Neuropathol Commun*. 2016;4:130.
20. Bourgeat P, Villemagne VL, Dore V, et al. Comparison of MR-less PiB SUVR quantification methods. *Neurobiol Aging*. 2015;36 Suppl 1:S159-166.
21. Krippendorff K. Content Analysis: An Introduction to Its Methodology. SAGE Publications; 2018.
22. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971;76:378-382.
23. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33:159-174.
24. Hardin JW (James W. Generalized estimating equations. Boca Raton, Fla. : Chapman & Hall/CRC; 2003.
25. Klunk WE, Koeppe RA, Price JC, et al. The Centiloid Project: Standardizing Quantitative Amyloid Plaque Estimation by PET. *Alzheimers Dement*. 2015;11:1-15.e4.
26. van Berckel BNM, Ossenkoppele R, Tolboom N, et al. Longitudinal amyloid imaging using 11C-PiB: methodologic considerations. *J Nucl Med*. 2013;54:1570-1576.


Rating protocol accurcay
Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Experts
Non-experts
FBB-Read
FBP-Read
FLUTE-Read

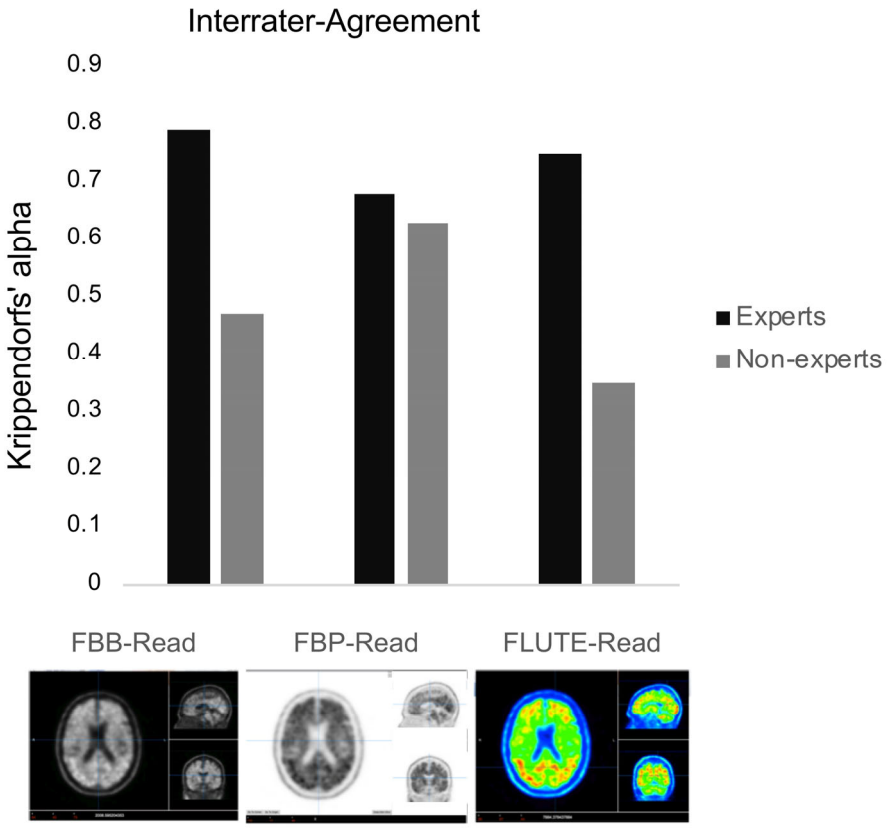

Interrater-Agreement
Krippendorfs' alpha
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Experts
Non-experts
FBB-Read
FBP-Read
FLUTE-Read

Figure 1: Upper panel: Reading accuracy (determined by SUVR measurement) displayed as a function of visual reading method for experts (black bars) and non-experts (gray bars). Below an image presented in the CAPTAINs Tool in the three different visual reading approaches. Lower panel: Interrater agreement assessed with Krippendorfs' alpha as a function of visual reading method for both groups.

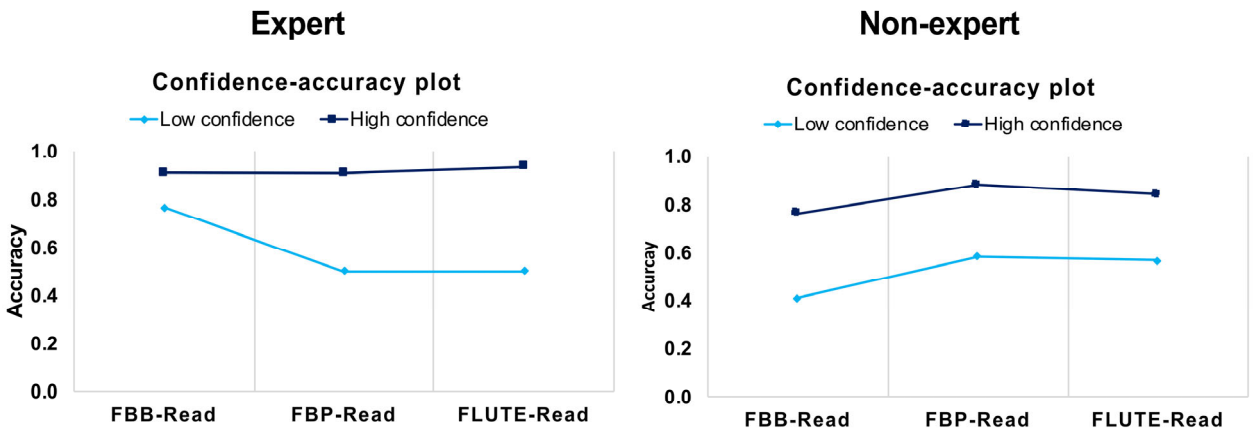


Figure 2: Confidence-accuracy analysis (CAC) separately by experts (left) and non-experts (right). Light blue represents low confidence judgements by accuracy values, and dark blue represents high confidence judgments by accuracy. CAC are shown by visual reading method. FBB= Florbetaben, FBP= Florbetapir, FLUTE= Flutemetamol.

Tracer	**Florbetaben**	**Florbetapir**	**Flutemetamol**
Scanner	Allegro	Biogram128/Allegro	Allegro/Geminin TF64
Acquisition time (p.i.)	90-110 min	90-110 min	50-70 min

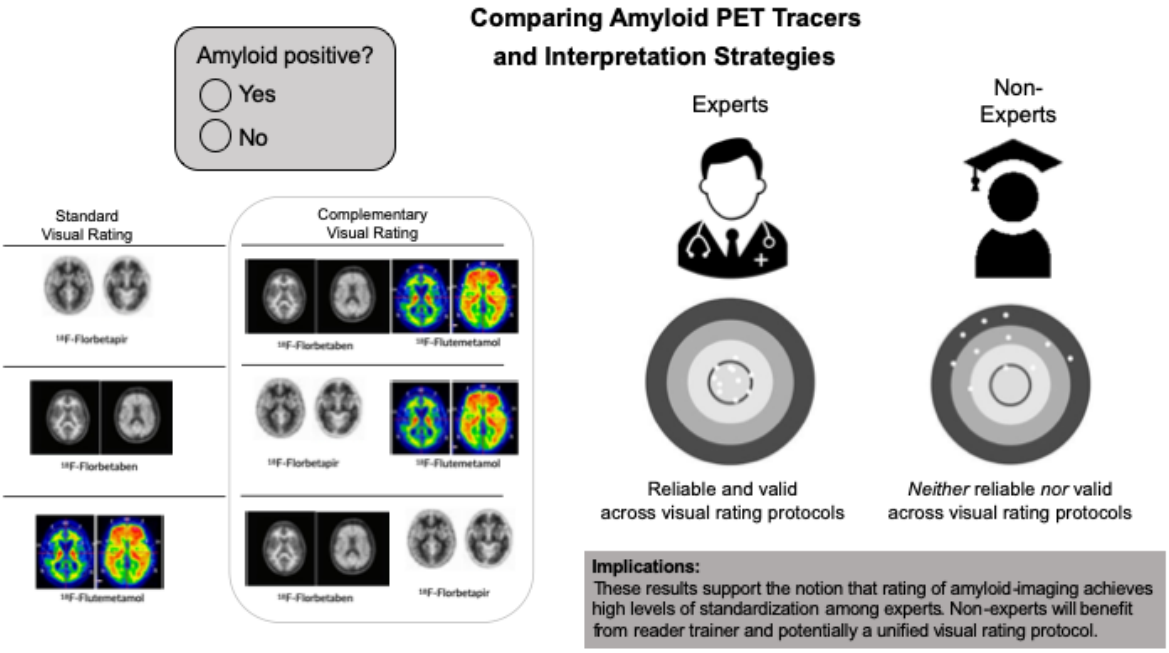
Table 1 Summary of scanner and acquisition time by fluorine-18 labelled amyloid Tracer.

Table 2:

1.1 Expert-Raters	Florbetaben Rating Protocol	Florbetapir Rating Protocol	Flutemetamol Rating Protocol
Consistency	.95	.90	.94
Accuracy	.86	.90	.89
Interrater-Agreement	.79	.68	.75

1.2 Non-expert raters	Florbetaben Rating Protocol	Florbetapir Rating Protocol	Flutemetamol Rating Protocol
Consistency	.70	.72	.50
Accuracy	.55	.67	.51
Interrater-Agreement	.47	.63	.35

Graphical Abstract:



Representation of online tool of CAPTAINs platform:

Displayed in Figure S1 is an example scan presented in three visual reading approaches with the recommended examples provided by the distributing companies. Written instructions were provided on the online-platform for each rater to review during the reading procedure.

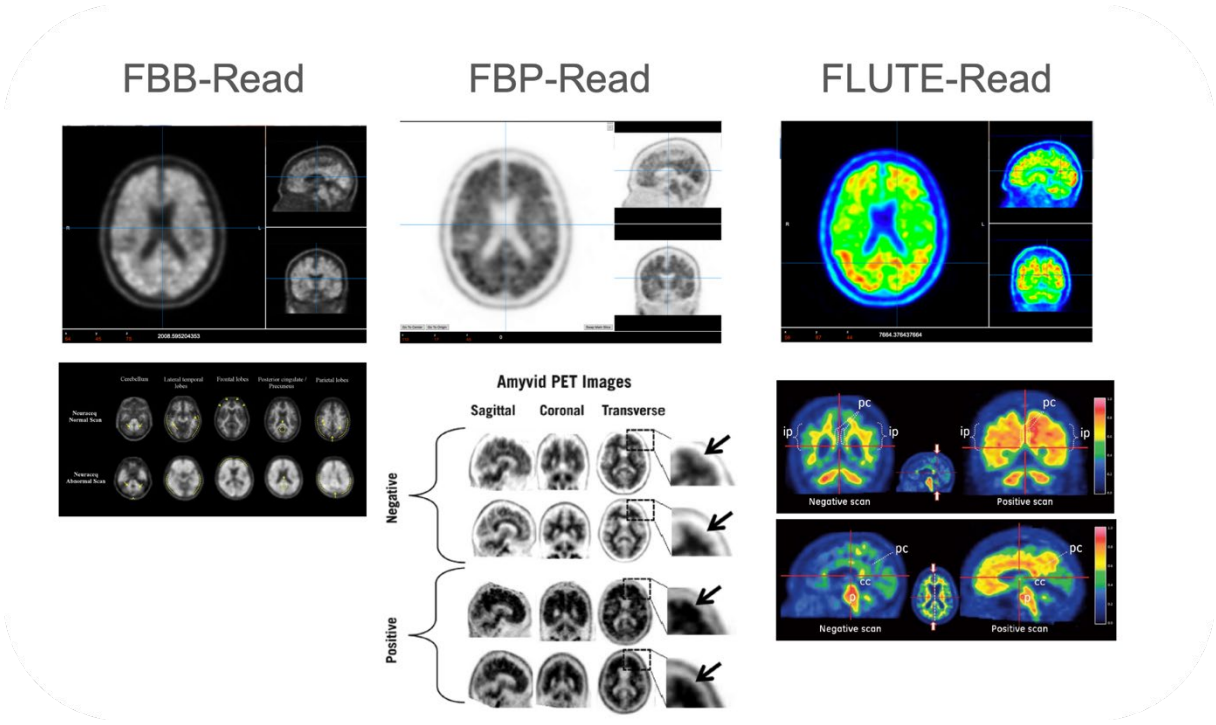


Figure S1:
Example of a scan presented in three approved visual reading methods.

Supplementary Material 1

Instructions on the CAPTAINs platform and depiction for rating tool:

The following instruction were displayed on the main page of the rating tool:

How to rate?

FBB-rating approach:
Starting point: cerebellum
Positive if: a) Small areas in the majority of slices of one region OR b) One large confluent area in one region

FBP-rating approach:
Starting point: cerebellum
Positive if: a) two or more brain regions =/> white matter OR b) one or more region >> white matter

FLUTE-rating approach:
Starting point: pons
Positive if: At least one cortical region above 50-60 % of peak or loss of grey/white contrast

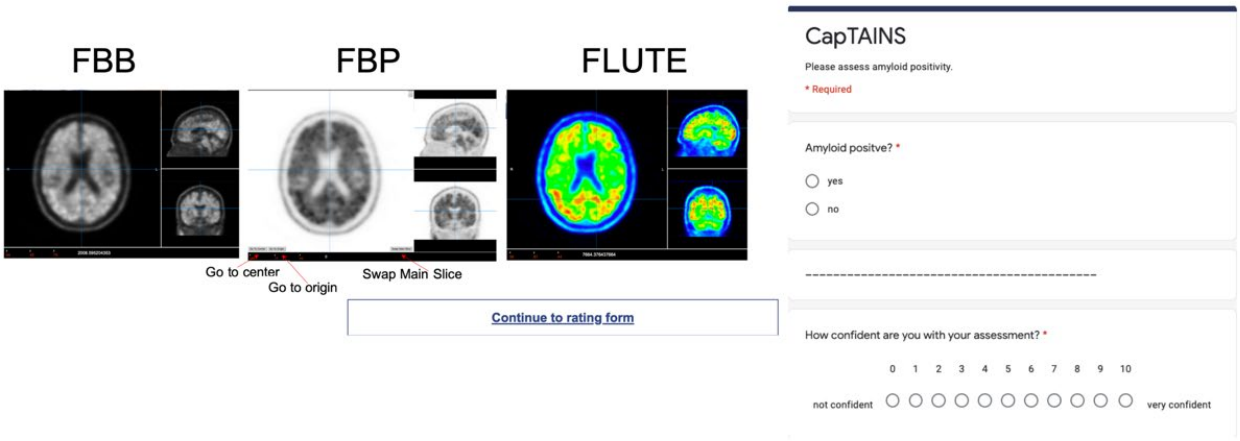


Figure S2:
Depicted are the three visual rating displays for the fluorine-labelled tracer. Raters were able to maneuver through the slices and swap to different views (i.e., axial, coronal, sagittal) to be displayed in the main slice. Clicking on the rating form promoted to the response sheet, where raters were asked to indicate whether the scan was positive or not, accompanied with a confidence assessment.