Toward a Universal Readout for ^{18}F-Labeled Amyloid Tracers: The CAPTAINs Study

Gérard N. Bischof[1], Peter Bartenstein[2], Henryk Barthel[3], Bart van Berckel[4], Vincent Doré[5,6], Thilo van Eimeren[1,7,8], Norman Foster[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], and 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, 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, The 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), Bonn, Germany;* [9]*Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, Utah;* [10]*Institute for Neurodegenerative Disorders, New Haven, Connecticut;* [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; and* [14]*Institute of Neuroscience and Medicine (INM-2), Molecular Organization of the Brain, Forschungszentrum Jülich, Germany*

To date, 3 ^{18}F-labeled PET tracers have been approved for assessing cerebral amyloid plaque pathology in the diagnostic workup of suspected Alzheimer disease (AD). Although scanning protocols are relatively similar across tracers, U.S. Food and Drug Administration– and the European Medicines Agency–approved visual rating protocols differ among the 3 tracers. This proof-of-concept study assessed the comparability of the 3 approved visual rating protocols to classify a scan as amyloid-positive or -negative, when applied by groups of experts and nonexperts to all 3 amyloid tracers. **Methods:** In an international multicenter approach, both expert (n = 4) and nonexpert 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 3 approved visual rating protocols. In a randomized order, every single scan was rated by each reader according to all 3 protocols. Raters were blinded for the amyloid tracer used and asked to rate each scan as positive or negative, giving a confidence judgment after each response. Percentage of visual reader agreement, interrater reliability, and agreement of each visual read with binary quantitative measures (fixed SUV ratio threshold for positive or negative scans) were computed. These metrics were analyzed separately for expert and nonexpert 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 ^{18}F-florbetaben visual rating protocol achieved the highest interrater reliability and accuracy especially under low confidence conditions. For the group of nonexpert raters, significant differences between the different visual rating protocols were observed with overall moderate-to-fair accuracy and with the highest reliability for the ^{18}F-florbetapir visual rating protocol. **Conclusion:** We observed high interrater agreement despite applying different visual rating protocols for all ^{18}F-labeled 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 rating standardization

J Nucl Med 2021; 62:999–1005
DOI: 10.2967/jnumed.120.250290

The advent of biomarkers of neuritic β-amyloid pathology (Aβ) using either cerebrospinal fluid or PET has shifted the conceptualization of a strictly clinical diagnosis of Alzheimer 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 (range = 48.0–93.3) and specificity (range = 67.0–100.0) in discriminating healthy controls (HCs) 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 thus 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 a progression marker and as a primary outcome measure in pharmaceutical clinical trials.

The use of amyloid PET biomarkers in the clinical workup of patients with cognitive decline and its relevance for diagnosis and subsequent patient management have now been evaluated in both

Revised Jun. 9, 2020; revision accepted Oct. 21, 2020.
For correspondence or reprints contact, Gerard N. Bischof (gerard.bischof@uk-koeln.de).
Published online Mar. 12, 2021.

North America (*5*) and Europe (*6*). At present, 3 fluorine-labeled tracers (^{18}F-florbetapir, ^{18}F-flutemetamol, and ^{18}F-florbetaben 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 Healthcare; flutemetamol), and Neuraceq (Life Molecular Imaging; florbetaben).

Appropriate use criteria have been formalized for these tracers (*7*). FDA- and 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 3 tracers is required before user certification (*8–10*). The general principle underlying the visual rating schemes is similar across the 3 tracers. Specifically, a physician is trained in identifying the loss in contrast of neocortical gray 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 or negativity, and translation from regional to global positivity or negativity. This readout variability may contribute to the observed diagnostic variability in sensitivity (range = 89.0–97.0) and specificity (range = 63.0–93.0) measures among all ^{18}F-labeled amyloid tracers (*11–13*). However, thus 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 ^{18}F-labeled amyloid tracers with the gold standard ^{11}C-labeled amyloid tracers, such as the centiloid scale, have been proposed (*14,15*). However, 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 rating approaches to avoid potential dependence of diagnostic and therapeutic decisions on the type of tracer or the interpretation protocols used. Therefore, the goal of the current study was to compare amyloid PET tracer–associated interpretation strategies (CAPTAINs) of the 3 FDA- and EMA-approved visual rating protocols for the 3 approved Aβ tracers in a group of expert and nonexpert raters. A specific aim was to identify which aspects of the 3 visual rating protocols allowed the most reliable identification of Aβ-positive and -negative scans across expert and nonexpert 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 nonexpert raters was paramount.

MATERIALS AND METHODS

PET Images

The study included data from all 3 FDA- and EMA-approved ^{18}F-labeled tracers for imaging of neuritic Aβ pathology (i.e., ^{18}F-florbetapir, ^{18}F-florbetaben, ^{18}F-flutemetamol) from HCs, individuals clinically diagnosed with mild cognitive impairment (MCI), and AD dementia patients.

For each tracer, we included 10 scans (in total 30 unique scans), from 10 HCs, 10 individuals with MCI, and 10 AD patients. With 7 readers and 3 different reading systems, our approach resulted in a total of 630 responses across the sample of experts and nonexperts. The inclusion criteria for the subjects in the sample were derived from the Australian Imaging, Biomarkers and Lifestyle flagship study of aging. In brief, participants were allocated to 1 of the 3 diagnostic groups on the basis of a clinical review that used the National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association (NINCDS-ARDA) criteria for AD, the criteria of Petersen et al. for MCI, and criteria for normal cognitive function for HCs (*16*). We matched the selected images from each tracer by age ($Mean_{(age)}$ = 73.9, $SD_{(age)}$ = 6.9; $F_{(2,29)}$ = 2.65, nonsignificant); Mini–Mental State Examination (MMSE) score ($Mean_{(MMSE)}$ = 23.7, $SD_{(MMSE)}$ = 5.6; $F_{(2,29)}$ = 2.1, nonsignificant); and Education ($Mean_{(Education)}$ = 12.9, $SD_{(Education)}$ = 1.91; $F_{(2,29)}$ = 1.10, nonsignificant).

Scans of each of the 3 Aβ tracers were prepared for visual reading according to all 3 of the recommended and FDA- and EMA-approved guidelines as provided by the vendors in their respective package inserts. All scans were then presented for rating according to all 3 of the approved visual rating protocols (Supplemental Fig. 1; supplemental materials are available at http://jnm.snmjournals.org). Thus, in a randomized order, every scan was rated by each reader according to all 3 protocols (e.g., ^{18}F-florbetapir scans were rated according to florbetapir, florbetaben, and flutemetamol guidelines). Additionally, to examine intrarater reliability we added repetitions of the same image and the visual rating protocol, totaling 12 responses from each rater. Six hundred thirty responses were collected for the interrater analysis and 84 responses for the intrarater analysis, totaling 714 overall.

Raters were blinded to 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 ^{18}F-florbetapir, the cerebellar cortex as a reference region for ^{18}F-florbetaben, and the pons as a reference region for ^{18}F-flutemetamol to create SUV ratio (SUVR) images (further details are provided in the supplemental materials). 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 histopathologic Aβ-amyloid plaque burden and corresponding SUVRs for each of the tracers, ^{18}F-florbetapir (*17*), ^{18}F-florbetaben (*18*), and ^{18}F-flutemetamol (*19*). 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 were provided by the Department of Molecular Imaging & Therapy, Austin Health, Melbourne, Australia. These scans were acquired on different PET scanners, which are summarized in Table 1. Each participant underwent a 20-min PET scan with 1 of the 3 ^{18}F tracers. The scan was performed 50 min after injection of 370 MBq (±10%) of ^{18}F-florbetapir, or 90 min after injection of 185 MBq (±10%) of ^{18}F-flutemetamol or 300 MBq (±10%) of ^{18}F-florbetaben. PET scans were spatially normalized using CapAIBL (https://milxcloud.csiro.au/ (*20*)). The images were then scaled to the SUV of the cerebellum cortex to generate the SUVR.

SUVR 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 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

TABLE 1
Summary of Scanner and Acquisition Time by ^{18}F-Labeled Amyloid Tracer

Characteristic	^{18}F-florbetaben	^{18}F-florbetapir	^{18}F-flutemetamol
Scanner	Allegro	Biogram128/Allegro	Allegro/Geminin TF64
Acquisition time (p.i.)	90–110 min	50–70 min	90–110 min

p.i. = after injection.

imaging. Importantly, all raters had undergone the tracer-specific reading training for all 3 ^{18}F Aβ tracers, culminating in a 3-fold expert certification. Further, all expert raters had several years of experience of visual rating and were familiar with all reading approaches.

Nonexpert raters ($n = 3$) were medical doctoral students enrolled in the medical program of the University Cologne, Germany. All 3 nonexpert raters were pursuing a medical doctoral thesis at University Hospital Cologne, Germany, and had some general experience in nuclear medicine acquired during their doctoral training, but little experience with image reading. Nonexpert raters underwent a 30-min standardized introduction to the published guidelines for visual readings for all 3 tracers and completed 5 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 before distribution of the personalized links to each rater. Images were displayed in random order and suffixed with the respective rating protocol (i.e., ^{18}F-florbetaben, ^{18}F-florbetapir, ^{18}F-flutemetamol rating protocol). All images were displayed in the recommended color scale according to each visual rating protocol (i.e., gray-scale, black-and-white, and Sokoloff/Spectrum, respectively). Datasets for each rater included all images presented in all 3 visual rating scales independently of the PET tracer used and raters were asked to judge if they were positive or negative based on the corresponding visual rating protocol (Supplemental Fig. 2). Raters were able to review the guidelines of all 3 visual rating protocols on the main homepage. Images appeared on 3 windows including axial, sagittal, and coronal views, with the main window displayed on an axial plane by default. A rating form was available on 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 (additional details are provided in the supplemental materials).

Statistical Analysis

Intrarater reliability was performed on the responses related to the repetitions and was computed using the 2-way intraclass coefficient (ICC) for experts and nonexperts separately.

To evaluate the interrater agreement across experts and nonexpert raters separately, 3 statistical metrics were used: consistency, given as the percentage of scans rated identical across raters; accuracy, computed as the percentage agreement with tracer specific quantitative SUVR positivity/negativity measures; and Krippendorff's α, a metric of interrater reliability used for more than 2 raters. Krippendorff's α calculates the α coefficient of reliability by comparing the observed disagreement with the expected disagreement (*21*). As the consistency measures include only a simple percentage of agreement, Krippendorffs α reflects the individual error-corrected agreement, similar to the Fleiss κ coefficient of reliability (*22*). Whereas an $\alpha = 1$ indicates perfect reliability and an $\alpha = 0$ indicates the absence of reliability, some authors have suggested the following range of benchmarks to assist with the interpretation of Krippendorff's α: 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1, near-perfect agreement (*23*).

The generalized estimating equation (*24*) was used to assess differences in responses as a function of visual rating method (i.e., main effect method). Significance threshold was set at a P value of <0.05. 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. Responses were included only from those expert ($n = 3$) and nonexpert raters ($n = 3$) who used the entire range of confidence judgments and binned their responses into low (0–5) and high confidence (6–10) and analyzed accuracy values on the basis of the quantitative SUVR measures for all 600 ratings.

RESULTS

Intrarater Reliability

Intrarater reliability was high among the 4 experts (ICC = 0.92) and moderate among the 3 nonexperts (ICC = 0.68).

Interrater Reliability

Expert Raters. Among the 4 expert raters, only slight variations across the visual rating protocols were observed. Consistency measures of ^{18}F-florbetaben and ^{18}F-flutemetamol visual rating protocols produced similar values among expert raters (0.95 and 0.94, respectively). The use of the ^{18}F-florbetapir rating protocol showed overall the lowest consistency judgments across raters (0.90). When visual ratings were compared with SUVRs for positivity and negativity agreement (i.e., accuracy), slight differences were observed. Specifically, whereas reading according to ^{18}F-florbetaben and ^{18}F-flutemetamol visual rating protocols showed accuracy values of 0.86 and 0.89, respectively, the use of the ^{18}F-florbetapir reading protocol showed accuracy values of 0.90 among raters. A summary of the reading accuracy is depicted in Figure 1.

Finally, interrater agreement (Krippendorf's α) was highest for the ^{18}F-florbetaben (0.79) and the ^{18}F-flutemetamol visual rating protocol (0.75) and lowest for the ^{18}F-florbetapir visual rating method (0.68) (Fig. 1). Estimating if expert rater responses differ as a function of visual rating procedure, we used the generalized estimating equation on the consistency and accuracy measures and observed no significant main effect of method on either metric (consistency: $W_{chisquare} = 3.56$, $P = 0.17$; accuracy: $W_{chisquare} = 2.55$, $P = 0.28$). A summary of these results is displayed in Table 2. Together, we observed no significant differences between the use

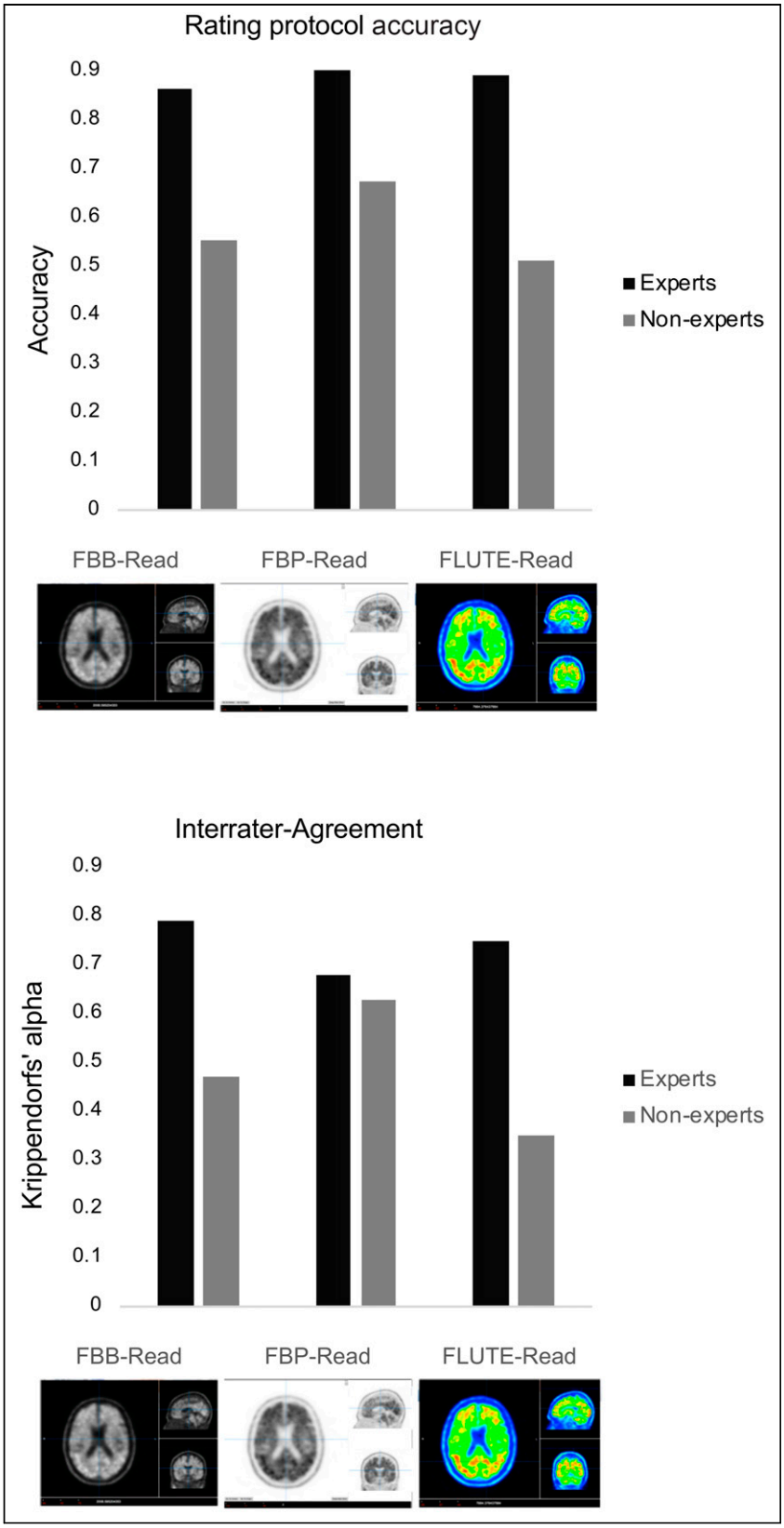

FIGURE 1. (Top) Reading accuracy (determined by SUVR measurement) displayed as a function of visual rating method for experts (black bars) and nonexperts (gray bars). Below that is an image presented in the CAPTAINs Tool in 3 different visual rating approaches. (Bottom) Interrater agreement assessed with Krippendorf's α as a function of visual rating method for both groups.

of the 3 visual rating protocols to render a scan positive or negative, and the overall rater agreement was high.

Nonexperts. Visual rating methods among nonexperts were less consistent. Specifically, whereas the use of ^{18}F-florbetaben (0.70) and ^{18}F-florbetapir (0.72) visual rating protocols showed acceptable consistency values, the ^{18}F-flutemetamol protocol reached consistency at the chance level across nonexpert raters (0.50). When responses were compared with the SUVR thresholds, accuracy was highest for the ^{18}F-florbetapir visual rating protocol (0.62), followed by the ^{18}F-florbetaben visual rating protocol (0.55), and lowest for the ^{18}F-flutemetamol (0.51) protocols (Fig. 1). This general result pattern is reflected in measures of interrater agreement (Fig. 1, visual rating method; ^{18}F-flutemetamol = 0.35, ^{18}F-florbetaben = 0.47, and ^{18}F-florbetapir = 0.63). Finally, both consistency and accuracy showed a significant main effect of method (consistency: $W_{chis-quare} = 20.62$, $P < 0.001$; accuracy: $W_{chis-quare} = 9.08$, $P = 0.001$). A summary of these results is displayed in Table 2.

Confidence-Accuracy Characteristic (CAC) Analysis

In both expert and nonexpert groups, low confidence judgments were associated with lower accuracy values, independent of the actual visual rating scheme used (Fig. 2). Furthermore, in the expert group, even in low confidence conditions, experts showed the highest accuracy values for the ^{18}F-florbetaben visual rating protocol, whereas for the ^{18}F-florbetapir and ^{18}F-flutemetamol protocols, accuracy values dropped to chance level when experts indicated low confidence in rating a scan as either positive or negative.

For nonexpert raters, the ^{18}F-florbetapir visual rating protocol showed the highest accuracy (0.58) for low-confidence judgments, whereas ^{18}F-florbetaben and ^{18}F-flutemetamol protocols either approached (0.56) or fell even below chance level (0.41) for responses accompanied with low confidence.

DISCUSSION

The main purpose of the present study was to determine the comparability and potential interchangeability of the 3 FDA- and EMA-approved visual rating protocols on the 3 amyloid tracers both in experts and in nonexperts. To this end, experts and nonexperts together rated more than 700 scans as positive or negative, accompanied with a

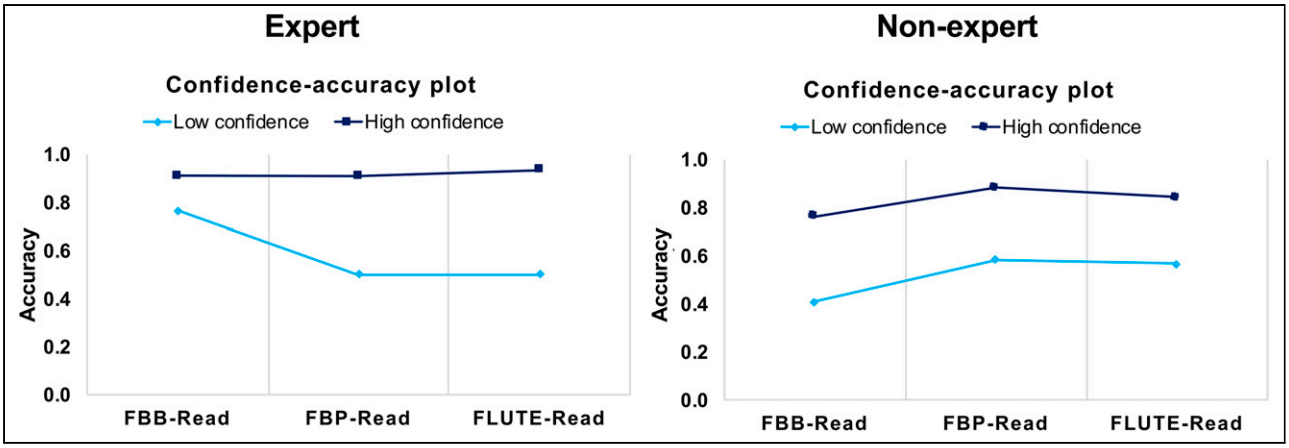

FIGURE 2. CAC separately by experts (left) and nonexperts (right). Light blue represents low confidence judgments by accuracy values, and dark blue represents high-confidence judgments by accuracy. CAC are shown by visual rating method. FBB = ^{18}F-florbetaben; FBP = ^{18}F-florbetapir; FLUTE = ^{18}F-flutemetamol.

confidence judgment. All ^{18}F-florbetaben, ^{18}F-florbetapir, and ^{18}F-flutemetamol images were presented in all 3 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 ^{18}F-florbetaben 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 nonexpert raters, accuracy and interrater reliability were dependent on the visual rating protocol and was highest when the ^{18}F-florbetapir visual rating protocol was used. Overall, nonexpert raters' responses showed only moderate and fair agreement, confirming that specific training is required to accurately evaluate Aβ images. The results also suggest that particularly inexperienced readers may additionally benefit from a universal visual rating protocol for all 3 FDA- and EMA-approved Aβ tracers.

Standardization of Visual Rating Protocols for ^{18}F-Labeled 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 workup of patients with suspected neurodegenerative disease, as well as for inclusion in 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 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 efforts to simplify and standardize the visual rating may be feasible and particularly meaningful for less-experienced readers, as significant heterogeneity among the 3 visual rating protocols was detected in the group of nonexpert raters. From a practical point of view, the development of a universal readout for ^{18}F-Aβ tracers may indeed be a straightforward solution to ensure comparability across differently trained specialists in regions in which not all 3 FDA- and EMA-approved Aβ tracers are available (e.g., Europe: ^{18}F-florbetaben and ^{18}F-flutemetamol but not ^{18}F-florbetapir), as well as in multicenter international therapeutic trials in which the 3 tracers are used. The universal readout includes a consistent starting point and the demarcation of standardized landmarks in which the reader would examine significant loss of white or 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 neuropathologic Aβ-amyloid plaque burden in the previously conducted end-of-life studies. Standardization approaches for quantitative purposes to reduce heterogeneity when measuring SUVRs have been suggested to achieve comparability between ^{18}F-labeled amyloid tracers and ^{11}C-Pittsburgh compound B, the gold standard tracer for β-amyloid pathology (*25*). For this purpose, the centiloid scale has been introduced, which linearly scales the measurement of the tracer from 0 to 100, with 0 representing the average uptake of young amyloid-negative individuals and 100 the retention of a typical AD patient. When the centiloid scale is used, thresholds of 20–25 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 the clinical routine workup. 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 of, for example, 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 expert and nonexpert raters. A combination of data-driven or artificial intelligence–driven approaches for amyloid imaging with different ^{18}F-labeled tracers may provide an additional future direction that could potentially assist in clinical readouts.

Limitations

The present study has some limitations. Although, experts and nonexperts rated more than 700 images in total, a differential analysis by tracer or disease category was not possible because of the limited number of scans available per category. Further, this

TABLE 2
Summary of Interrater Reliability Statistics

Rater	^{18}F-florbetaben rating protocol	^{18}F-florbetapir rating protocol	^{18}F-flutemetamol rating protocol
Expert			
Consistency	0.95	0.90	0.94
Accuracy	0.86	0.90	0.89
Interrater agreement	0.79	0.68	0.75
Nonexpert			
Consistency	0.70	0.72	0.50
Accuracy	0.55	0.67	0.51
Interrater agreement	0.47	0.63	0.35

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. In a planned follow-up study, we intend to increase the set of images beyond the convenience sample of images presented here and 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, sex, 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 histopathologic correlation. However, pathologic 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 similarities that are scanner- or site-dependent. Potentially, different scanner types may have affected visual rating results. However, potential differences based on the scanner type would have affected all 3 rating protocols equally, and differences were minimized by ensuring that preprocessing was done using the same analysis pipeline (supplemental materials). Finally, the visual rating protocols recommend the use of coregistered CT/MRI scans, particularly in the cases of low image quality, to discern possible anatomic 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.

DISCLOSURE

Gérard N. Bischof reports receiving speaker honoraria from Life Molecular Imaging. Alexander Drzezga reports receiving research support from Siemens Healthineers, Life Molecular Imaging, GE Healthcare, AVID Radiopharmaceuticals; speaker honoraria from and being on the advisory boards of Siemens Healthineers, Sanofi, and GE Healthcare; and stock from Siemens Healthineers. There is a patent pending for ^{18}F-PSMA7 (PSMA PET imaging tracer for prostate cancer). John Seibyl reports being a consultant for Biogen, Roche, AbVie, Life Molecular Imaging, LikeMinds, and Invicro and equity stake in Invicro. No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

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

KEY POINTS

QUESTION: Are the FDA-approved visual rating protocols for the 3 currently available ^{18}F-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 used. In nonexperts, significant differences arose, suggesting that reader training is necessary to examine β-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. *Alzheimers Res Ther*. 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–e109.
8. Buckley CJ, Sherwin PF, Smith APL, Wolber J, Weick SM, Brooks DJ. Validation of an electronic image reader training programme for interpretation of [^{18}F]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 ^{18}F-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. ^{18}F 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. ^{18}F 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. ^{18}F 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 [^{11}C]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. *Alzheimers Dement*. 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–S166.
21. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications; 2018:289–315.
22. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 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. *Generalized Estimating Equations*. Chapman & Hall/CRC; 2003:85–89.
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.e1.
26. van Berckel BNM, Ossenkoppele R, Tolboom N, et al. Longitudinal amyloid imaging using ^{11}C-PiB: methodologic considerations. *J Nucl Med*. 2013;54:1570–1576.