CrossMark

ORIGINAL ARTICLE

Quantitation of PET signal as an adjunct to visual interpretation of florbetapir imaging

Michael J. Pontecorvo[1] · Anupa K. Arora[1] · Marybeth Devine[1] · Ming Lu[1] ·
Nick Galante[1] · Andrew Siderowf[1] · Catherine Devadanam[1] · Abhinay D. Joshi[1] ·
Stephen L. Heun[1] · Brian F. Teske[1] · Stephen P. Truocchio[1] · Michael Krautkramer[1] ·
Michael D. Devous Sr.[1] · Mark A. Mintun[1]

Abstract

Purpose This study examined the feasibility of using quantitation to augment interpretation of florbetapir PET amyloid imaging.

Methods A total of 80 physician readers were trained on quantitation of florbetapir PET images and the principles for using quantitation to augment a visual read. On day 1, the readers completed a visual read of 96 scans (46 autopsy-verified and 50 from patients seeking a diagnosis). On day 2, 69 of the readers reinterpreted the 96 scans augmenting their interpretation with quantitation (VisQ method) using one of three commercial software packages. A subset of 11 readers reinterpreted all scans on day 2 based on a visual read only (VisVis control). For the autopsy-verified scans, the neuropathologist's modified CERAD plaque score was used as the truth standard for interpretation accuracy. Because an autopsy truth standard was not available for scans from patients seeking a diagnosis, the majority VisQ interpretation of the three readers with the best accuracy in interpreting autopsy-verified scans was used as the reference standard.

Results Day 1 visual read accuracy was high for both the autopsy-verified scans (90%) and the scans from patients seeking a diagnosis (87.3%). Accuracy improved from the visual read to the VisQ read (from 90.1% to 93.1%, $p < 0.0001$).

Importantly, access to quantitative information did not decrease interpretation accuracy of the above-average readers (>90% on day 1). Accuracy in interpreting the autopsy-verified scans also increased from the first to the second visual read (VisVis group). However, agreement with the reference standard (best readers) for scans from patients seeking a diagnosis did not improve with a second visual read, and in this cohort the VisQ group was significantly improved relative to the VisVis group (change 5.4% vs. −1.1%, $p < 0.0001$).

Conclusion These results indicate that augmentation of visual interpretation of florbetapir PET amyloid images with quantitative information obtained using commercially available software packages did not reduce the accuracy of readers who were already performing with above average accuracy on the visual read and may improve the accuracy and confidence of some readers in clinically relevant cases.

Keywords Alzheimer's disease · Amyloid imaging · Amyloid PET · PET quantitation · Florbetapir · Amyvid

Introduction

Biomarkers have the potential to aid in the diagnosis of patients with cognitive impairment by providing information regarding the presence or absence of relevant neuropathology, when used as part of a comprehensive clinical evaluation in patients with a mild, atypical course or atypically early onset of cognitive impairment [1]. PET imaging ligands including Pittsburgh compound B (^{11}C-PIB) [2], ^{18}F-florbetaben [3], ^{18}F-flutemetamol [4] and ^{18}F-florbetapir [5] have been developed for estimation of cortical beta amyloid (Aβ) neuritic plaque deposition, a hallmark pathology, and a required element for the evaluation of neuropathological changes in patients with Alzheimer's disease (AD) [6]. As shown by their

✉ Michael J. Pontecorvo
pontecorvo@avidrp.com

1 Avid Radiopharmaceuticals (a wholly owned subsidiary of Eli Lilly and Company), 3711 Market St, Philadelphia, PA 19104, USA

Springer

respective package inserts/summaries of product characteristics, as well as the published literature [7–9], reader accuracy in the pivotal trials for the ^{18}F-labeled agents averaged close to 90% for discriminating patients found at autopsy to have no or sparse neuritic plaques (amyloid-negative, Aβ−) from those found to have moderate to frequent plaques (amyloid-positive, Aβ+). However, as might be expected, within each of these development programs, there were individual readers with sensitivity or specificity values below the average, and in some with values below 80%. While having a lower accuracy in a specific research trial might not always predict lower accuracy in the clinical setting, the range of performance suggests the potential for augmentation of the reading method to improve interpretation accuracy.

It has been suggested that image quantitation could be helpful in assisting visual interpretation of PET amyloid images [10–13]. Quantitation has been applied extensively in other realms of nuclear medicine imaging including PET [14–17], and quantitative analyses have proven useful for characterizing amyloid tracer binding and relationships to other biomarkers [18–20]. In the case of ^{18}F-florbetapir, the use of an exploratory quantitative approach resulted in an accuracy of 97% in relation to autopsy [7]. Until recently, approaches to quantitating PET amyloid images have been limited to research methods that are nonstandard and may require manual intervention and technical expertise. However, the emerging availability of commercial software packages for quantitation of PET amyloid images raises the possibility that quantitative estimates of tracer uptake/amyloid binding could be integrated into an algorithm for interpretation of scans in a clinical setting.

Although promising, the use of software programs may be vulnerable to variations in the PET image including, but not limited to, movement, atrophy or count limitations. Automatically applied, preselected target or reference regions may inadequately cover the full range of anatomical variation in the target population, and some packages may be difficult to navigate, resulting in unacceptable variations in quantitation. In addition to these software-specific issues, there may be differences in how users incorporate quantitation into the visual read decision algorithm. One approach could be to set a firm quantitative threshold beyond which images are considered positive regardless of visual appearance. Alternatively, methods could be developed for using the overall or regional quantitative values to guide reexamination of the visual interpretation. In spite of these potential issues, only one study to date has evaluated the performance of quantitative software as an adjunct to visual interpretation. Specifically, Nayate et al. [21] recently reported that the use of Siemens Scenium software to quantitate florbetapir PET scans significantly increased interreader reliability. Although an increase in interreader reliability is encouraging, it does not necessarily mean that there has been an increase in reader accuracy.

The present study was designed to examine the feasibility of an approach to incorporating quantitation into the standard visual interpretation algorithm for florbetapir PET amyloid imaging. Three representative software packages were evaluated, each by a separate cohort of physician readers. It was hypothesized that the addition of quantitation as an adjunct to visual interpretation (VisQ method) would significantly improve the total accuracy of florbetapir scan interpretation by readers whose accuracy of scan interpretation by visual read alone was less than the historical average accuracy of 90% (below-average readers), with no significant negative impact on accuracy of above-average readers (>90% accuracy).

Materials and methods

Software packages

The software packages used in this study, MIM (MIMneuro®), Siemens (Siemens syngo.PET Amyloid Plaque) and Hermes (Hermes Brain Analysis Software Suite™ BRASS, 2.0; CE 0413), are all commercially available and approved in the US and EU for visual examination and quantitation of PET images, with specific routines designed to quantitate ^{18}F-florbetapir PET images. Although the individual packages use different proprietary algorithms to perform the quantitation, the three packages share the following features:

1. They use spatial normalization to apply template-based predefined regions of interest (ROIs) on the florbetapir PET scan.
2. They employ ROIs that sample cortical regions from multiple lobes as well as cerebellum. These ROIs sample regions similar (albeit not necessarily identical) to those used by Clark et al. [7] including: frontal cortex, anterior cingulate, temporal cortex, lateral parietal cortex, medial parietal cortex (precuneus), posterior cingulate, and cerebellum.
3. They provide the ability for the reader to verify location of the ROIs on the spatially normalized florbetapir PET scan.
4. They provide cortex-to-cerebellum standardized uptake value ratios (SUVr) for each of the cortical ROIs as well as a cortical average SUVr (across the ROIs).
5. They have been shown to produce values highly correlated with the Avid research method for SUVr generation [22]. Thus, SUVr values for each program can be linked to the range of SUVr associated with none to sparse and moderate to frequent neuritic plaques found at autopsy as shown by the Avid method [7]. (Calibration for the Siemen software package has been described separately [23]. Calibration for the Hermes software package is included in the Supplementary material. Calibration for the MIM software package is planned for a separate publication.)

Participating physicians

A total of 80 physicians participated as scan readers in this study. The study was conducted in three separate replications in different cohorts of readers using the three different software packages (MIM, Siemens, Hermes). The MIM and Siemens replications (NCT 01946243) were performed with US physicians at ACR Image Metrix, Philadelphia, PA. The Hermes replication (NCT 02107599) was performed with readers from Spain and the UK at Bioclinica, Inc. (Leiden, The Netherlands). For each replication (MIM, Siemens, Hermes), imaging physicians who had completed a florbetapir PET reader training course were contacted at random and invited to participate. Physician readers were excluded from the study if they had more than minimal experience with or had previously been trained personally to perform quantitation of amyloid PET.

For each replication, readers who met the above qualifying criteria were invited to the testing facility in cohorts of three to ten readers to complete day 1 (visual read) and day 2 (quantitative read). The testing continued in each replication until a minimum of seven readers with visual read accuracy ≤90% (below-average readers; accuracy less than the mean accuracy expected based on previous studies) and a minimum of five readers with visual read accuracy >90% (above-average readers) were recruited.

Study flow

Upon arrival at the core laboratory read facility, all readers underwent a brief refresher training utilizing portions of the online (US) reader training program, highlighting the steps for visual interpretation and criteria for determining a scan as positive or negative for amyloid plaques. The core laboratory provided training on the respective software to facilitate visual reads, and readers practiced with nine image sets under supervision. The readers then independently visually interpreted a test group of 20 florbetapir PET scans (without supervision). These interpretations served as a practice exercise and were not used in the primary or secondary analyses, nor were these results used to disqualify readers from the study.

All readers then underwent training related to the use of quantitation with florbetapir PET images. Training consisted of teaching the operation of the quantitative software, and the method for generation of SUVr values. The readers were shown the validation of the research quantitation method in autopsy-verified cases [7] and the relationship between the quantitation results from the research method and the results from the respective commercial quantitation package [23] (see also Online Resource 1), which allowed them to estimate the approximate SUVr values associated with a positive scan. Readers were then taught the principles for applying quantitation as an adjunct to visual interpretation, including algorithms for comparing the quantitative results to their initial visual interpretation. The training included supervised practice of the visual with adjunct quantitation (VisQ) interpretation approach on the same nine sample cases used for the initial practice of visual interpretation.

On day 1 of the study, the readers visually interpreted 96 florbetapir scans comprising the 46 autopsy-verified scans [7], and 50 randomly selected scans from a trial with patients seeking a diagnosis for cognitive impairment [24]. The readers did not have access to quantitation tools during this reading session.
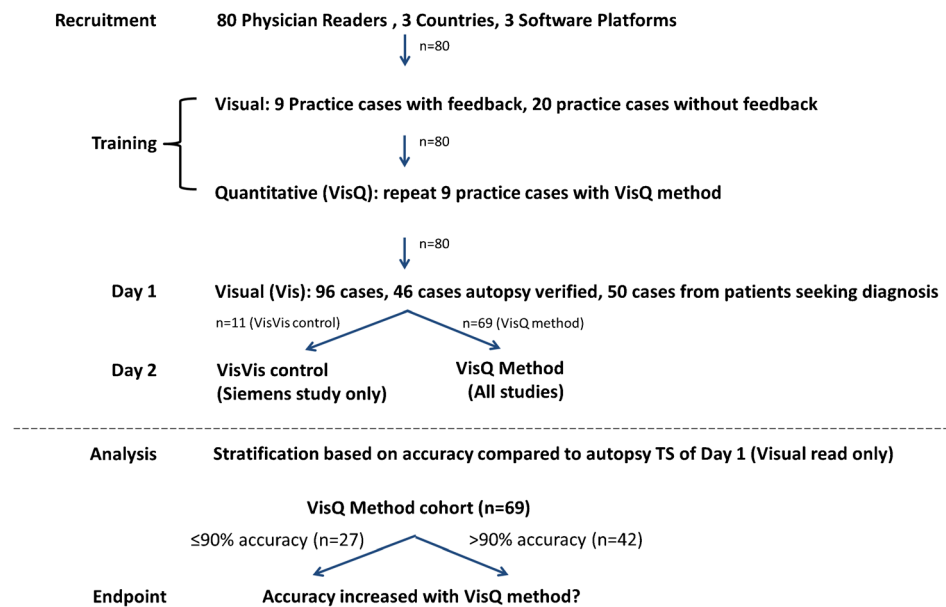
On the following day (day 2), readers in the MIM and Hermes replications were presented these same 96 florbetapir PET scans for interpretation using the VisQ approach. The readers obtained SUVr values for the predefined ROIs, as well as an overall cortical average SUVr using the respective quantitation software in accordance with the software manufacturer's instructions. For each scan, the reader had the opportunity to review their previous interpretation based on visual assessment alone and was then asked to make a final read interpretation using the VisQ interpretation principles. In addition to the final interpretation, the SUVr values for the individual regions and the average SUVr value were recorded.

For the Siemens replication (on day 2), readers were randomized to either an experimental arm (VisQ) or a control arm (VisVis). Procedures in the experimental arm were identical to those described for the MIM and Hermes replications above. For the readers randomized to the control arm, the only difference was that they were not allowed to use the quantitative software or the VisQ approach during the second review of the 96 florbetapir PET cases; these readers had the opportunity to review their previous interpretation (Aβ+ or Aβ−) based on visual assessment alone and were then asked to make a final read interpretation using only the visual interpretation method (hence VisVis). This condition was intended to control for any learning or other benefit derived from reviewing the scans a second time. A diagram of the study design is shown in Fig. 1.

Florbetapir PET images

The images used in this study included florbetapir PET scans from 46 end-of-life patients recruited from hospice, long-term care facilities and community healthcare facilities who came to autopsy within 1 year of their scan in the florbetapir pivotal trial [7] and 50 scans randomly selected from a previous study of florbetapir use in patients with diagnostic uncertainty [24] (Table 1). In general, the patients seeking a diagnosis for cognitive impairment were younger, included more mildly impaired patients, and a lower proportion of patients with AD and other non-AD dementia than the end-of-life patients. Both

Fig. 1 Schematic representation of study design. *TS* truth standard, *Vis* visual read, *VisQ* visual read with quantitation, *VisVis* visual read with second visual read

previous studies were approved by the relevant institutional review boards and subjects or other family members of subjects contributing PET scans used in these studies gave written informed consent. All florbetapir PET scans used in these studies were acquired under standard methods described previously [7, 24]. A 10-min PET acquisition was performed approximately 50 min after administration of approximately 370 MBq (10 mCi) of ^{18}F-florbetapir. Images were acquired and reconstructed with iterative or maximum likelihood algorithms with a postreconstruction gaussian filter. Images were displayed for visual interpretation, and quantitation was performed using the MIM, Siemens, or Hermes software in accordance with the respective replication.

Image interpretation

The initial visual interpretation was performed in accordance with the instructions in the ^{18}F-florbetapir package insert. Briefly, images were reviewed using a black-and-white palette (gray scale) with the maximum intensity of the scale set to the maximum intensity brain pixel. Starting at the bottom of the brain, primarily in transaxial orientation, the cerebellum (presumed amyloid-free normal tissue) was examined followed in succession by the temporal lobes and occipital cortex, the prefrontal cortex and parietal lobes. A scan was defined as positive (Aβ+) if at least two regions contained areas with reduced gray–white matter contrast, or

Table 1 Characteristics of patients who contributed PET images

	Autopsied patients	Patients seeking a diagnosis
Number (%) of patients	46 (47.9)	50 (52.1)
Age (years), mean (SD)	79.0 (12.38)	75.0 (7.351)
Clinical diagnosis, *n* (%)		
Alzheimer's disease	20 (43.5)	13 (26.0)
Mild cognitive impairment	4 (8.70)	34 (68.0)
Other or non-Alzheimer's dementia	11 (23.9)	3 (6.00)
Cognitively intact normal control	11 (23.9)	0 (0.00)
Mini-Mental State Examination score, *n* (%)[a]		
28–30	NA	20 (40.0)
25–27	NA	12 (24.0)
20–24	NA	9 (18.0)
<20	NA	9 (18.0)
Patients Aβ+ by truth/reference standard, *n* (%)	28 (60.9)	29 (58.0)

NA not applicable

[a] Mini-Mental State Examination not applicable as cognitive testing in end-of-life patients is not a reliable indicator of neurological disease status

if at least one region had an area of gray matter uptake more intense than the adjacent white matter uptake. Readers recorded an interpretation as either Aβ− (indicative of no or sparse neuritic plaques) or Aβ+ (moderate to frequent plaques). For positive scans the regions of positivity were also recorded.

Image quantitation involved spatial normalization of the ^{18}F-florbetapir PET scan into a standard coordinate system, application of predefined ROIs, checking the quality of the normalization and application of the ROI, and refitting of the image if applicable for a software package. A series of cortical target region to whole cerebellum count ratios (SUVr) and a cortical average SUVr were then generated.

Readers were instructed to use quantitation as an adjunct to the visual read, not as an alternative. Thus, they compared the calculated cortical average SUVr to the expected range for Aβ+/Aβ− scans for the particular software package. If the quantitative result was consistent with the initial visual read, the readers were expected not to change their initial interpretation. In the event of apparent disagreement between visual interpretation and quantitation, readers were instructed to perform the following actions. First, the readers checked the spatial normalization and fit of the scan to the template. They confirmed the accuracy of the placement of the ROIs, checking for cerebrospinal fluid or bone within the ROI, and evaluated the potential impact of atrophy or ventriculomegaly on quantitation. Next they reviewed the basis for making a visual Aβ+ or Aβ− determination. They looked for loss of gray–white contrast in at least two regions or intense uptake in one region. In the case of an Aβ+ initial visual read and an apparent Aβ− quantitation, readers were instructed to consider whether the positive visual interpretation might be based on tracer retention in regions other than the six ROIs that contribute to the composite SUVr (e.g., intense tracer retention in the occipital lobe could support an Aβ+ visual determination but would not contribute to the SUVr). In the case of an Aβ− initial visual read and an Aβ+ quantitation, readers visually examined the regions corresponding to the ROIs with elevated SUVr to confirm whether there was a loss of gray–white contrast in these areas. Finally, the readers visually examined the cerebellum region, confirming the fit of the ROI (which can affect the denominator of the SUVr) and the level of gray–white contrast (which provides a standard for comparison to the cortex), and looking for possible structural anomalies (e.g., stroke) that could influence quantitation of the cerebellar region. The final interpretation was then based on a visual read augmented by quantitative information.

Statistical analysis

The prespecified primary efficacy hypothesis in all three replications was that addition of quantitative information (VisQ method) would significantly improve overall accuracy of florbetapir PET scan interpretation. The primary analysis utilized the 46 images from patients who received a florbetapir scan within 1 year of autopsy. The neuropathologist's diagnosis, that was based on the modified Consortium to Establish a Registry for Alzheimer's Disease (CERAD) plaque score, was used as the truth standard such that an image was considered correctly interpreted as Aβ+ when there were moderate or frequent plaques and Aβ− when there were no or sparse plaques as previously described [7]. In the MIM and Siemens replications, the primary analysis population comprised those readers with a visual read (day 1) average accuracy of 90% or less based on historical studies. Paired t tests were used to determine whether accuracy (percent agreement with the truth standard) increased between the day 1 visual interpretation and the day 2 interpretation incorporating quantitative information (VisQ). In the Hermes replication, all readers were included in the primary analysis and the net reclassification index (NRI) [25] (see Online Resource 3 for statistical description) was used to evaluate differences in accuracy between the day 1 visual interpretation and the day 2 VisQ interpretation.

Since the three replications were designed similarly, the results were also integrated to assess the possible benefit from the addition of quantitative information, across all readers. The integrated analyses were done in two ways. In the first analysis, for scans from the autopsied patients [7], the neuropathologist's diagnosis was used as the truth standard, as specified above, and the changes in reader accuracy in the VisQ group between the day 1 visual interpretation and the day 2 interpretation incorporating quantitative information were compared with the changes between the day 1 and day 2 interpretations in the VisVis group (control group from Siemens software) that did not involve the use of quantitative information and performed on both day 1 and day 2. An analysis of covariance (ANCOVA) model was used for this comparison, adjusting for readers' day 1 accuracy, and replication. Secondary analyses were also performed looking at sensitivity and specificity relative to the truth standard.

In the second analysis, because an autopsy-based truth standard was not available for the scans from patients seeking a diagnosis for cognitive impairment [24], the majority VisQ interpretation of the three readers with the best interpretation accuracy on the autopsy-verified scans was used as the reference standard. The majority interpretation from these readers (coincidentally all from the Siemens replication) was 100% accurate relative to the neuropathologist's diagnosis of the autopsied patients. All three readers agreed on 45 of 50 scans from patients lacking autopsy truth standard. A sensitivity analysis was performed excluding these five cases to ensure that they did not influence the results. Thus, the best three readers' majority interpretation was

considered a reasonable reference standard for the scans from the non-autopsy patients.

Analysis of the scans from the non-autopsy patients was similar to that described above. The primary analysis evaluated the impact of the addition of quantitative information by comparing the VisQ condition with a second visual read (VisVis condition) in terms of their agreement with the reference standard (accuracy). Positive agreement (sensitivity) and negative agreement (specificity) were also calculated. To control for the multiplicity, Bonferroni's correction was applied to adjust the p values from these analyses. The three readers serving as the reference standard were included in the VisQ group for the primary analysis because they were among the best readers on the visual read alone and had some of the smallest improvements when quantitation (VisQ) was added (hence the most conservative analysis). However, a sensitivity analysis was performed excluding these readers and yielded similar conclusions.

The interreader reliability of scan interpretation by the VisQ and VisVis methods was also assessed using Fleiss' kappa statistics for both day 1 and day 2. In addition, for each read method, a change in interreader reliability was calculated as the kappa value based on day 2 interpretations minus the kappa value for the day 1 interpretations. The 95% confidence interval around this difference was calculated using a bootstrap method [26]. If the lower bound of this 95% confidence interval was greater than 0, then a statistically significant improvement on interreader consistency for day 2 interpretations over day 1 interpretations was demonstrated. Percent agreement is also provided to assess the interreader reliability, calculated as the number of reader pairs who agreed when interpreting the same scan, divided by all possible pairs of readers for the same scan. A logistic regression model with robust variance estimation by a generalized estimated equation was used to compare the change in percent agreement between day 1 and day 2 for the VisQ and VisVis methods. Finally, reader confidence (low, medium, high) was recorded on day 1 and day 2 and the changes from day 1 to day 2 in the VisQ and VisVis groups were compared using the Wald chi-squared test from a proportional odds model.

Results

Table 2 summarizes the characteristics of the 80 readers participating in this study. The readers were similar across regions with respect to their experience with PET scans, brain scan interpretation and amyloid scan interpretation, and their experience with quantitation. In addition, there were no statistically significant differences in any of the recorded characteristics among the readers with above average accuracy (>90%) and those with below average

accuracy (≤90%) on the visual reads of the autopsy-verified scans, nor were there any clear differences among the readers in the VisVis control arm and the remaining readers (Note: the 90% threshold for reader accuracy was based on the historical average from previous studies). Most readers read no more than 20 brain scans per week, had interpreted ten or fewer clinical amyloid PET scans and all readers had no previous experience quantitating amyloid PET. All readers completed the study.

Table 3 shows the primary results for the individual replications. In all three replications, the mean visual read accuracy in the autopsy-verified scans on day 1 was close to 90% (88.7% Hermes, 89.5% MIM, 91.6% Siemens, 90.1% overall). In all three replications, the use of quantitative information in the visual read on day 2 (VisQ condition) resulted in increased accuracy, and all three replications showed significantly improved results in terms of the prespecified primary endpoints. When the results of the three replications were pooled, accuracy compared to the autopsy truth standard across all 69 readers increased from 90.1% with the visual read method to 93.1% using the VisQ method. This increase was statistically significant whether judged by the paired t test or the NRI. Table 3 also shows the sensitivity (positive agreement) and specificity (negative agreement) for the reader cohorts. In the cohort of all 69 readers, specificity significantly increased with the addition of the VisQ method from 86.7% to 92.8% ($p < 0.0001$). Sensitivity remained above 90% with a slight numerical improvement (92.2% to 93.3%, $p = 0.1259$) with the VisQ method. In each of the three individual replications, specificity increased significantly, and sensitivity either improved (MIM replication) or was not significantly changed with the addition of the VisQ method (further details are provided in the Online Resource 4 and 5).

Figure 2a shows an example subject with Parkinson's disease in life and confirmed to be Aβ− (no neuritic plaques) at autopsy, where quantitation may have aided in image interpretation. Although the majority of readers in both the VisQ (51 of 80) and VisVis (11 of 11) cohorts interpreted this scan as positive on the initial visual read, a net 23 VisQ readers (in contrast to only 1 of 11 VisVis readers) changed to a negative interpretation on the second read (i.e., after quantitation). Figure 2a, b give a clue as to the the readers' possible thought process during the study. After obtaining a negative quantitation result (mean SUVr 0.94) readers should have checked the fit of the ROI to the PET scan, and in doing so might have noticed that the areas of greatest tracer retention were medial to the temporal lobe ROI and likely reflected retention in the white matter rather than the gray matter.

Figure 2c, d shows images from another example patient, a 71-year-old man with a 15-month history of cognitive impairment and an Mini-Mental State Examination score

Table 2 Characteristics of participating physician readers

Characteristics	Below-average readers (≤90%) using visual alone (N = 32)	Above-average readers (>90%) using visual alone (N = 48)	VisQ (N = 69)	VisVis (N = 11)
Country				
US	20 (62.5%)	39 (81.3%)	48 (69.6%)	11 (100%)
Spain	9 (28.1%)	8 (16.7%)	17 (24.6%)	0 (0.00%)
UK	3 (9.38%)	1 (2.08%)	4 (5.80%)	0 (0.00%)
Number of PET scans read per week				
20 or fewer	14 (43.8%)	18 (37.5%)	27 (39.1%)	5 (45.5%)
21–50	13 (40.6%)	21 (43.8%)	29 (42.0%)	5 (45.5%)
51–100	5 (15.6%)	8 (16.7%)	12 (17.4%)	1 (9.09%)
101 or more	0 (0.00%)	1 (2.08%)	1 (1.45%)	0 (0.00%)
Number of brain PET scans				
20 or fewer	32 (100%)	47 (97.9%)	68 (98.6%)	11 (100%)
21–50	0 (0.00%)	1 (2.08%)	1 (1.45%)	0 (0.00%)
Total number of amyloid PET scans read in the past				
0	12 (37.5%)	21 (43.8%)	29 (42.0%)	4 (36.4%)
1–10	10 (31.3%)	17 (35.4%)	22 (31.9%)	5 (45.5%)
11–20	3 (9.38%)	6 (12.5%)	7 (10.1%)	2 (18.2%)
21 or more	7 (21.9%)	4 (8.33%)	11 (15.9%)	0 (0.00%)
Experience with quantitating amyloid PET scans				
No	32 (100%)	48 (100%)	69 (100%)	11 (100%)

Below/above-average readers are defined as those whose accuracy of scan interpretation by visual read alone (day 1) was ≤90%/>90% of the historical average based on previous studies

Vis qualitative visual read, *VisQ* visual read with quantitation

of 25, who was undergoing evaluation for mild cognitive impairment of uncertain origin at the time of the florbetapir PET scan. The majority of readers in both the VisQ and VisVis cohorts interpreted this scan as Aβ+ on the initial visual read, but eight VisQ and one VisVis reader returned Aβ− interpretations on day 1. The quantitation result was positive (mean SUVr 1.39, with regional SUVr approximately 1.55 in both the precuneus and posterior cingulate), and all eight VisQ readers revised their interpretation to Aβ+, whereas the only change among the VisVis readers was an additional reader who recorded an Aβ− interpretation on day 2. According to the VisQ interpretation algorithm, after obtaining a positive quantitation result, readers should have checked the fit of the ROI to the PET scan (Fig. 2d), and then reviewed the gray–white contrast in regions that overlapped the quantitative ROI. In doing so might have noticed the high level of signal in the precuneus/posterior cingulate regions (Fig. 2c, top row, second and third slices). The positive quantitative values may also have reminded readers that the gray–white contrast in

the cortex should be evaluated with respect to the presumed normal level of gray–white contrast seen in the cerebellum. In this case, even where the gray matter signal did not exceed that of the white matter (e.g., temporal lobe) the gray–white contrast was reduced relative to the cerebellum.

In the replication using data from the Siemens software, an increase in accuracy was also observed between the day 1 visual reads and the day 2 visual reads (VisVis condition). However, the study was not powered to make a statistical comparison between the VisVis and VisQ conditions. In order to facilitate a statistical comparison and to better characterize performance of readers in interpreting PET amyloid images, the data were combined across the three replications as shown in Table 4. Consistent with the results from the individual replications, the average visual (day 1) image interpretation accuracy across all readers was 90% in the autopsy-verified scans with the CERAD neuritic plaque score as the truth standard. A similar average accuracy (87.3%) was obtained in the scans from patients seeking a diagnosis, with the majority score of the best readers used as the reference standard. Only

Table 3 Results of the individual replications for the autopsy-verified scans in terms of accuracy, sensitivity and specificity for the day 1 qualitative visual read (Vis) in comparison with the day 2 visual read with quantitation (VisQ)

	Software	Below-average readers (≤90%)				Above-average readers (>90%)				All readers				
		No. of readers	Day 1 (Vis)	Day 2 (VisQ)	p value[a]	No. of readers	Day 1 (Vis)	Day 2 (VisQ)	p value[a]	No. of readers	Day 1 (Vis)	Day 2 (VisQ)	p value[a]	NRI (p value[b])
Accuracy (%)	Hermes	12	84.6	89.5	0.0043	9	94.2	93.7	0.5943	21	88.7	91.3	0.0212	0.07 (0.0001)[c]
	MIM	7	81.7	88.8	0.0029*	15	93.2	96.1	0.0001	22	89.5	93.8	<0.0001	0.09 (<0.0001)
	Siemens	8	87.0	91.8	0.0025*	18	93.7	94.8	0.1767	26	91.6	93.9	0.0038	0.05 (0.0013)
	All	27	84.5	90.0	<0.0001	42	93.6	95.0	.0043	69	90.1	93.1	<0.0001	0.07 (<0.0001)
Sensitivity (%)	Hermes	12	91.4	90.2	0.6662	9	96.8	94.8	0.0955	21	93.7	92.2	0.3412	
	MIM	7	80.1	85.2	0.0157	15	94.0	96.9	0.0053	22	89.6	93.2	0.0002	
	Siemens	8	89.3	91.5	0.3506	18	95.0	95.4	0.6309	26	93.3	94.2	0.2829	
	All	27	87.8	89.3	0.3306	42	95.1	95.8	0.1927	69	92.2	93.3	0.1259	
Specificity (%)	Hermes	12	74.1	88.4	0.0033	9	90.1	92.0	0.5447	21	81.0	89.9	0.0047	
	MIM	7	84.1	94.4	0.1017	15	91.9	94.8	0.0878	22	89.4	94.7	0.0188	
	Siemens	8	83.3	92.4	0.0417	18	91.7	93.8	0.2020	26	89.1	93.4	0.0168	
	All	27	79.4	91.2	<0.0001	42	91.4	93.8	0.0321	69	86.7	92.8	<0.0001	

NRI net reclassification index

Below/above-average readers are defined as those readers whose accuracy of scan interpretation by visual read alone (day 1) was ≤90%/>90% of the historical average based on previous studies

Asterisk (*) indicates p values for those comparisons considered primary to the prespecified analyses has been lost from this table

[a] Paired *t* test.

[b] *p* values from an asymptotic *Z* test (see Online Resource)

[c] *p* values representing the study prespecified primary analyses

four of the 80 readers achieved <80% accuracy in the autopsy-verified scans, and two of these plus two other readers achieved <80% accuracy on the scans from patients seeking a diagnosis.

The addition of quantitative information (VisQ) improved the day 2 accuracy relative to the accuracy of the day 1 visual read in interpreting both autopsy-verified scans and scans from patients seeking a diagnosis. However, this improvement from day 1 to day 2 in the VisQ group was significantly greater than that seen for a repeat visual read on day 2 (VisVis group) only for the scans from patients seeking a diagnosis. Similar results were obtained when the five scans with imperfect agreement among the reference standard readers were excluded from the analysis.

The interreader reliability, as assessed by both Fleiss' kappa and by the percentage of scans with agreement between pairs of readers, increased from day 1 to day 2 (Table 5). Considering all scans, the change from day 1 to day 2 was not different between the VisQ and the VisVis readers, although there was a trend toward a greater difference in the interpretation of scans from patients seeking a diagnosis. Consistent with the changes in accuracy (Table 4), the improvement in interreader agreement from day 1 to day 2 in the VisQ condition was greater for below-average readers (≤90%) than for above-average readers (>90%), and was greater for the scans from patients seeking a diagnosis than for the autopsy-verified scans. Finally, confidence increased

by a significantly greater amount from day 1 to day 2 in the VisQ than in the VisVis group. As shown in Table 6, in the VisQ group there was an 18% increase in the proportion of images interpreted with high confidence, in contrast to only a 7% increase in the VisVis group.

Discussion

The present study was designed to test the feasibility of an approach to incorporating quantitation into the standard interpretation algorithm for florbetapir PET amyloid imaging. The key study findings were:

1. Day 1 visual read accuracy was high for both the autopsy-verified end-of-life scans (90% accuracy compared to the autopsy truth standard) and scans from patients seeking a diagnosis (87.3% agreement with the reference standard, the majority interpretation of the three best readers).
2. For all three software packages, accuracy improved from the day 1 visual read to the day 2 read incorporating quantitative information, whether judged by the paired *t* test or the NRI. As expected this effect was largest in readers with below average accuracy (≤90%) on the day 1 qualitative visual read. Importantly, access to quantitative information did not result in a decrease in accuracy of the above-average readers.
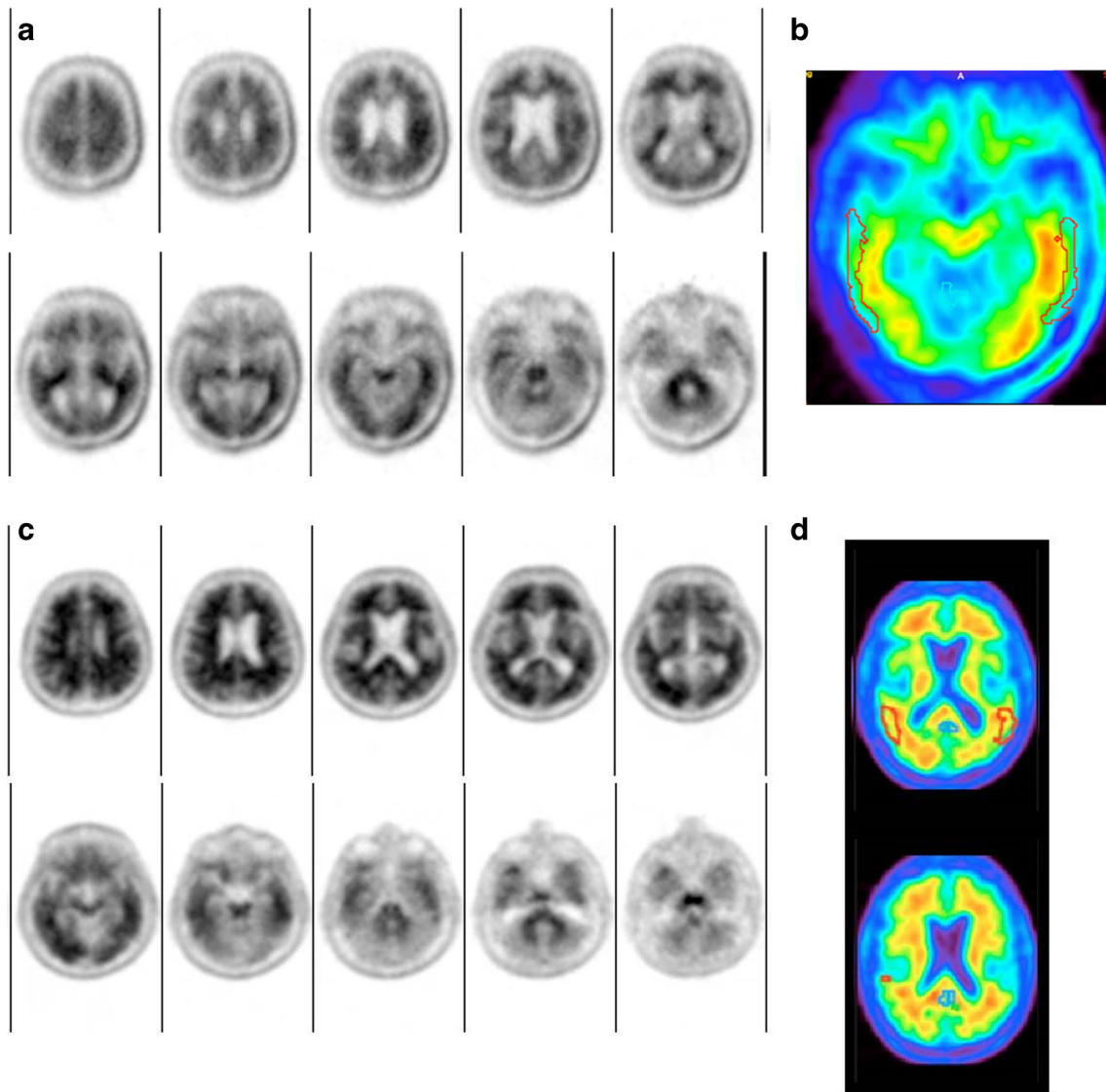
Fig. 2 Florbetapir PET quantitation as an adjunct to visual read. **a**, **b** Florbetapir PET images from a subject diagnosed with Parkinson's disease and confirmed to be Aβ− at autopsy. This scan was frequently interpreted by readers as positive (51 of 80) in the visual interpretation (day 1) but more than 50% of those incorrect interpretations were changed to negative (23 of 51) with quantitation as an adjunct. **a** Axial slices of a florbetapir PET scan from the top (*upper left*) to the bottom (*lower right*) of the brain in native space. **b** Slice from the same scan normalized to the template space using one of the commercial packages. Although the study did not record the thought processes of the readers, it is possible that they reviewed the quantitative result (normal SUVr) and the placement of the temporal lobe region of interest (*red*) and revisited their impression of whether the temporal cortex had loss of gray–white contrast. **c**, **d** Images from a 71-year-old man who was undergoing evaluation for mild cognitive impairment (no-autopsy group). Eight VisQ and one VisVis reader returned Aβ− interpretations on day 1. The quantitation result was positive (mean SUVr 1.39, with regional SUVr approximately 1.55 in both the precuneus and posterior cingulate), and all eight VisQ readers revised their interpretation to Aβ+. Possibly the readers reviewed the gray–white contrast in regions that overlapped the quantitative ROI and noticed the high level of signal in the precuneus/posterior cingulate regions (**c** *top row*, second and third slices)

3. Accuracy as compared to the autopsy truth standard also increased from the first to the second qualitative visual read (VisVis group). There was no significant difference in accuracy change between the VisQ and VisVis in the cohort of images from autopsy patients. However, in a cohort of cases from patients seeking a diagnosis, accuracy did not improve with a second visual read (VisVis), while in in this cohort the accuracy in the VisQ group was significantly improved relative to the VisVis group.

4. Across all scans interreader reliability improved from day 1 to day 2 among both the VisQ and VisVis readers but the increase in readers' confidence in their interpretation was significantly greater in the VisQ than in the VisVis group.

Although not the primary objective of this study, the results of the day 1 visual read are particularly noteworthy. The mean visual read accuracy of 90.0% (±5.4%, median 91.3% observed in relation to the autopsy truth standard in

Table 4 Impact of quantitation as an adjunct to visual read (VisQ group, combined across studies) on accuracy, sensitivity, and specificity in comparison with a second qualitative visual read (VisVis group) in interpreting autopsy-verified scans and scans from patients seeking a diagnosis

	Accuracy (%)			Sensitivity (%)			Specificity (%)		
	Day 1	Day 2	Change	Day 1	Day 2	Change	Day 1	Day 2	Change
Autopsy-verified scans[a]									
VisQ (n = 69)	90.1 (5.5)	93.1 (4.7)	3.0 (4.0)	92.2 (8.4)	93.3 (7.1)	1.0 (5.5)	86.7 (12.6)	92.8 (8.9)	6.0 (10.5)
VisVis (n = 11)	89.3 (4.7)	92.3 (6.0)	3.0 (3.1)	94.5 (4.6)	97.1 (3.5)	2.6 (4.5)	81.3 (13.0)	84.8 (16.3)	3.5 (7.1)
p value (VisQ vs. VisVis)		NS			NS			NS	
All readers (N = 80)	90.0 (5.4)			92.5 (8.0)			86.0 (12.7)		
Median	91.3			92.9			88.9		
Scans from patients seeking a diagnosis[b]									
VisQ (n = 69)	87.0 (5.0)	92.4 (4.3)	5.4 (4.8)	87.1 (9.4)	95.2 (6.3)	8.1 (7.8)	86.9 (11.8)	88.5 (9.3)	1.7 (7.4)
VisVis (n = 11)	89.3 (2.7)	88.2 (2.6)	−1.1 (3.1)	89.3 (5.7)	87.8 (5.6)	−1.6 (5.0)	89.2 (8.0)	88.7 (7.8)	−0.4 (4.0)
p value (VisQ vs. VisVis)		<0.0001			<0.0001			0.1919	
All readers (N = 80)	87.3 (4.8)			87.4 (9.0)			87.2 (11.3)		
Median	88.0			89.7			90.5		

NS not significant

The data presented are means (SD), except where indicated

[a] Truth standard based on the modified Consortium to Establish a Registry for Alzheimer's Disease (CERAD) plaque score

[b] Reference standard was the majority Interpretation of the three best readers

this study of 80 physicians from three different countries (US, UK, Spain), reading on three different software platforms robustly confirms the effectiveness of the florbetapir reader training. Additionally, although readers were split for analysis purposes into above-average and below-average readers, based on an expected visual read average accuracy of 90%, this threshold still reflects a high level of accuracy for diagnostic image interpretation. An accuracy of <80% might be a more useful threshold for identifying undesirable performance; only four of the 80 readers (5%) scored less than 80% accuracy relative to the autopsy truth standard. Similarly high agreement with the reference standard was observed in the scans from patients seeking a diagnosis, thus extending the findings, within the limits of the study design (below), to interpretation of scans from a clinically-relevant population.

Table 5 Impact of quantitation as an adjunct to visual read (VisQ group, combined across studies) on interrater agreement in comparison with a second qualitative visual read (VisVis group)

		Day 1		Day 2		Change			p value (VisQ vs. VisVis)
		Fleiss' kappa	Percent agreement	Fleiss' kappa	Percent agreement	Fleiss' kappa	95% confidence interval	Percent agreement	
All scans									
VisQ	Accuracy >90% (n = 42)	0.79	89.8	0.81	90.9	0.02	−0.02, 0.07	1.1	
	Accuracy ≤90% (n = 27)	0.60	80.9	0.73	86.8%	0.12	0.07, 0.18	5.9	
	All readers (n = 69)	0.71	85.9	0.77	89.2	0.07	0.02, 0.11	3.2	0.4737
VisVis	All readers (n = 11)	0.73	87.3	0.78	89.4	0.04	0.00, 0.09	2.1	
Autopsy-verified scans									
VisQ	Accuracy >90% (n = 42)	0.83	92.0	0.83	92.1	0.00	−0.05, 0.06	0.1	
	Accuracy ≤90% (n = 27)	0.57	79.6	0.69	85.0	0.12	0.05, 0.20	5.5	
	All readers (n = 69)	0.72	86.6	0.77	89.2	0.06	0.00, 0.12	2.6	0.5656
VisVis	All readers (n = 11)	0.72	87.2	0.79	90.5	0.07	0.01, 0.14	3.3	
Scans from patients seeking a diagnosis									
VisQ	Accuracy >90% (n = 42)	0.75	87.6	0.79	89.8	0.04	−0.03, 0.10	2.2	
	Accuracy ≤90% (n = 27)	0.63	82.1	0.76	88.4	0.13	0.06, 0.19	6.3	
	All readers (n = 69)	0.70	85.3	0.77	89.2	0.07	0.01, 0.13	3.8	0.1923
VisVis	All readers (n = 11)	0.74	87.3	0.76	88.3	0.02	−0.04, 0.08	1	

Table 6 Impact of visual read with quantitation (combined across replications) on the confidence of image interpretation

Confidence	No. (%) of images				p value (VisQ vs. VisVis)
	VisQ		VisVis		
	First read (visual)	Second read (with quantitation)	First read (visual)	Second read (visual)	
Low	788 (11.9%)	471 (7.11%)	76 (7.20%)	44 (4.17%)	<0.0001
Medium	1,948 (29.4%)	1,066 (16.1%)	247 (23.4%)	205 (19.4%)	
High	3,888 (58.7%)	5,087 (76.8%)	733 (69.4%)	807 (76.4%)	

Interpretation accuracy further improved from the day 1 visual read to the day 2 read incorporating quantitative information. As expected this effect was largest in the readers with below average day 1 accuracy (≤90%). These readers often exhibited a bias toward a positive or a negative response. This bias was attenuated on the quantitative read, resulting in higher overall accuracy. Importantly, access to quantitative information did not result in a decrease in accuracy of the above-average readers. This could have been a concern, particularly for the scans from end-of-life patients in whom atrophy and other end-of-life brain changes could have affected the accuracy of quantitation. These findings suggest that the improvement in interpretation accuracy appears to be a result of the application of the VisQ algorithm by the readers, and not a result of blind reliance on a numerical result provided by the software to determine the final scan interpretation.

The increase in accuracy from day 1 to day 2 in interpretation of the autopsy-verified scans in the VisVis group is challenging to explain. This increase was unexpected since previous studies have shown 95% agreement between sequential blinded reads [14], but in retrospect, the small (3%) improvement in accuracy was within the limits of the previous result. This increase in accuracy is consistent with the hypothesis that readers may improve their interpretation skills with experience (e.g., may improve after reading the 96 scans on day 1), or alternatively with the hypothesis that, regardless of experience, interpretation may be improved by reviewing a scan a second time. However, in contrast to the result in the autopsy-verified scans, there was no significant improvement in agreement with the reference standard between day 1 and day 2 in interpretation of the scans from patients seeking a diagnosis. This result suggests that a second visual read may not always result in improved accuracy and further suggests that the improvement seen in both the VisVis and VisQ groups on day 2 for the autopsy-verified scans may have resulted from the readers learning to deal with image features such as patient movement artifacts or atrophy, which would be expected to be more common in end-of-life patients than in patients seeking a diagnosis.

On the other hand, the finding that agreement with the reference standard for the scans from patients seeking a diagnosis improved from day 1 to day 2 by a significantly greater amount in the VisQ group than in the VisVis group suggests that quantitation could offer some benefit in this clinically relevant population. In contrast to the end-of-life patients, the patients seeking a diagnosis were younger (75 vs. 79 years) and at an earlier disease stage (68% vs. 9% mild cognitive impairment), and thus less likely to show atrophy and end-of-life brain changes that may result in poorer fitting of some ROIs, with resultant underestimation of the SUVr in some end-of-life cases. Thus, in the younger, milder patients seeking a diagnosis, quantitation may accurately help identify borderline cases with abnormal amyloid burden, thus increasing sensitivity, as shown in Tables 3 and 4.

Interreader reliability (kappa and percent agreement) also improved from day 1 to day 2. This improvement was most likely driven by the observed changes in accuracy. Finally, confidence increased significantly in the VisQ condition relative to the VisVis condition. This increase in confidence may be important in a clinical setting because it may increase the likelihood that a scan result will lead to management change.

All of these findings must be considered in light of several significant design limitations, particularly the choice to have all readers perform the visual read on all scans prior to beginning the VisQ read. As noted above, this makes it difficult to separate the impact of interpretation experience from the impact of quantitative information. However, alternative designs are potentially more problematic. It would have been possible, for example, to counterbalance across readers with some performing the VisQ read first and some the visual read first, or even counterbalance reading approaches within readers. However, in both of those designs readers obtain feedback (quantitation) during the VisQ reads that may alter their approach to the visual read. Another alternative might have been a between-group design with one set of readers performing visual reads and the other VisQ reads. A between-group design would have been adequate for an overall analysis such as that described in this paper, and, based on the visual read results of the present study, might have required more than 50–60 subjects per group to have 80–90% power to detect a 3% difference in accuracy. However, this design would not have been useful for evaluating the individual software packages (e.g., Table 2), and there would have been no way to determine the impact on readers with a low accuracy.

Another significant limitation of the current design was the absence of an autopsy-based truth standard for the clinically relevant scans from patients with cognitive impairment of uncertain origin who were seeking a diagnosis. Obviously this is a limitation that is nearly impossible to overcome, since patients seeking a diagnosis are usually relatively healthy and unlikely to come to autopsy in a reasonable amount of time. The reference standard chosen for the present study was the majority interpretation of the three readers who had the best VisQ accuracy on the autopsy-verified scans. This majority score was 100% accurate in relation to the autopsy truth standard. These readers were in unanimous agreement in the interpretation of 45 of 50 scans from patients seeking a diagnosis. A sensitivity analysis excluding the five scans yielded results similar to the primary analysis (the improvement in accuracy, sensitivity and specificity from day 1 to day 2 was significantly greater among the VisQ readers than the VisVis readers). Thus, we believe the majority rating as used here was a good reference standard for evaluating the scans without autopsy verification.

Finally, it must be recognized that the improvement in accuracy obtained by the addition of quantitative information (VisQ) relative to a purely visual scan interpretation was small; some readers benefitted more than others and some readers did not benefit at all. The mean net increase in accuracy from day 1 (Vis) to day 2 (VisQ) was equivalent to approximately 1 in 46 (3.0%) or 2 in 50 (5.4%) additional scans correctly classified per reader for the autopsy-verified scans and for the scans from patients seeking a diagnosis, respectively. This relatively small effect should be considered in the context of the finding that readers typically misclassified only a handful of cases on day 1 (mean accuracy 90% and 87% for the autopsy-verified scans and the scans from the patients seeking a diagnosis, respectively), thus creating a potential ceiling for improvement in this study. The magnitude of effect was larger in the below-average (day 1 accuracy ≤90%) than above-average readers (Table 3), but even among above-average readers with a day 1 accuracy >90% there was no mean decrease in accuracy as a result of the addition of quantitative information. Although not the most dramatic finding of this study, this latter finding is also important. As noted above, multiple software programs have now been approved for quantitation of PET amyloid images in the US and EU. The packages may be vulnerable to various technical limitations and when used uncritically could potentially lead to image misinterpretation. However, the current results suggest that software packages that share the core features described above can be employed as adjuncts in the reading of florbetapir PET scans, according to the methods and interpretation algorithms described above, with minimal risk of increasing interpretation errors, and may possibly improve the interpretation accuracy of some imaging physicians.

In conclusion, the present study in 80 readers from three countries, using three different software platforms, demonstrated a mean visual reading accuracy of approximately 90% in relation to the truth/reference standard for both autopsy-verified scans and clinically relevant scans from patients seeking a diagnosis. The results further suggest that access to quantitative information may provide clinically improvement in performance and confidence of some readers in the interpretation of scans when used as an adjunct to a visual read, and importantly did not reduce the accuracy of readers with already above average accuracy on the visual read.

References

1. Johnson KA, Minoshima S, Bohnen NI, Donohoe KJ, Foster NL, Herscovitch P, et al. Appropriate use criteria for amyloid PET: a report of the Amyloid Imaging Task Force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association. J Nucl Med. 2013;4:476–490.
2. Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Ann Neurol. 2004;55:306–319.
3. Rowe CC, Ackerman U, Browne W, Mulligan R, Pike KL, O'Keefe G, et al. Imaging of amyloid beta in Alzheimer's disease with 18F-BAY94-9172, a novel PET tracer: proof of mechanism. Lancet Neurol. 2008;7:129–135.
4. Nelissen N, Van Laere K, Thurfjell L, Owenius R, Vandenbulcke M, Koole M, et al. Phase 1 study of the Pittsburgh compound B derivative 18F-flutemetamol in healthy volunteers and patients with probable Alzheimer disease. J Nucl Med. 2009;50:1251–1259.
5. Wong DF, Rosenberg PB, Zhou Y, Kumar A, Raymont V, Ravert HT, et al. In vivo imaging of amyloid deposition in Alzheimer disease using the radioligand 18F-AV-45 (florbetapir F 18). J Nucl Med. 2010;51:913–920.
6. Hyman BT, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Carrillo MC, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. Alzheimers Dement. 2012;8:1–13.

7. Clark CM, Pontecorvo MJ, Beach TG, Bedell BJ, Coleman RE, Doraiswamy PM, et al. Cerebral PET with florbetapir compared with neuropathology at autopsy for detection of neuritic amyloid-beta plaques: a prospective cohort study. Lancet Neurol. 2012;11: 669–678.

8. Curtis C, Gamez JE, Singh U, Sadowsky CH, Villena T, Sabbagh MN, et al. Phase 3 trial of flutemetamol labeled with radioactive fluorine 18 imaging and neuritic plaque density. JAMA Neurol. 2015;72:287–294.

9. Sabri O, Sabbagh MN, Seibyl J, Barthel H, Akatsu H, Ouchi Y, et al. Florbetaben PET imaging to detect amyloid plaques in Alzheimer disease: phase 3 study. Alzheimers Dement. 2015;11: 964–974.

10. Camus V, Payoux P, Barre L, Desgranges B, Voisin T, Tauber C, et al. Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. Eur J Nucl Med Mol Imaging. 2012;39:621–631.

11. Guerra UP, Nobili FM, Padovani A, Perani D, Pupi A, Sorbi S, et al. Recommendations from the Italian Interdisciplinary Working Group (AIMN, AIP, SINDEM) for the utilization of amyloid imaging in clinical practice. Neurol Sci. 2015;36:1075–1081.

12. Kobylecki C, Langheinrich T, Hinz R, Vardy ER, Brown G, Martino ME, et al. [18F]-florbetapir positron emission tomography in patients with frontotemporal dementia and Alzheimer's disease. J Nucl Med. 2015;56:386–391.

13. Perani D, Schillaci O, Padovani A, Nobili FM, Iaccarino L, Della Rosa PA, et al. A survey of FDG and amyloid-PET imaging in dementia and GRADE analysis. Biomed Res Int. 2014;2014, 785039. doi:10.1155/2014/785039.

14. International Atomic Energy Agency. Quantitative nuclear medicine imaging: concepts, requirements and methods; IAEA Human Health Reports No. 9. Vienna: International Atomic Energy Agency; 2014.

15. Avril N, Bense S, Ziegler SI, Dose J, Weber W, Laubenbacher C, et al. Breast imaging with fluorine-18-FDG PET: quantitative image analysis. J Nucl Med. 1997;38:1186–1191.

16. Lin C, Itti E, Haioun C, Petegnief Y, Luciani A, Dupuis J, et al. Early 18F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis. J Nucl Med. 2007;48:1626–1632.

17. Foster NL, Heidebrink JL, Clark CM, Jagust WJ, Arnold SE, Barbas NR, et al. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain. 2007;130:2616–2635.

18. Joshi AD, Pontecorvo MJ, Clark CM, Carpenter AP, Jennings DL, Sadowsky CH, et al. Performance characteristics of amyloid PET with florbetapir F 18 in patients with Alzheimer's disease and cognitively normal subjects. J Nucl Med. 2012;53:378–384.

19. Landau SM, Lu M, Joshi AD, Pontecorvo M, Mintun MA, Trojanowski JQ, et al. Comparing positron emission tomography imaging and cerebrospinal fluid measurements of β-amyloid. Ann Neurol. 2013;74:826–836.

20. Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. Lancet Neurol. 2013;4:357–367.

21. Nayate AP, Dubroff JG, Schmitt JE, Nasrallah I, Kishore R, Mankoff D, et al. Use of standardized uptake value ratios decreases interreader variability of [18F] florbetapir PET brain scan interpretation. AJNR Am J Neuroradiol. 2015;36:1237–1244.

22. Joshi AD, Pontecorvo MJ, Lu M, Skovronsky DM, Mintun MA, Devous MD Sr. A semi-automated method for quantification of florbetapir F 18 PET images. J Nucl Med. 2015;56:1736–1741.

23. Hutton C, Declerck J, Mintun MA, Pontecorvo MJ, Devous Sr MD, Joshi AD; Alzheimer's Disease Neuroimaging Initiative. Quantification of 18F florbetapir PET: comparison of two analysis methods. Eur J Nucl Med Mol Imaging. 2015;42:725–732.

24. Grundman M, Pontecorvo MJ, Salloway SP, Doraiswamy PM, Fleisher AS, Sadowsky CH, et al. Potential impact of amyloid imaging on diagnosis and intended management in patients with progressive cognitive decline. Alzheimer Dis Assoc Disord. 2013;27:4–15.

25. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157–172.

26. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.