Standardized Uptake Value Ratio-Independent Evaluation of Brain Amyloidosis

Andrea Chincarini[a,*], Francesco Sensi[a,b], Luca Rei[a], Irene Bossert[h], Silvia Morbelli[g], Ugo Paolo Guerra[c], Giovanni Frisoni[d,f], Alessandro Padovani[e], Flavio Nobili[h] and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]*Istituto Nazionale di Fisica Nucleare, Sezione di Genova, Genova, Italy*
[b]*Dipartimento di Fisica, Università degli Studi di Genova, Genova, Italy*
[c]*Department of Nuclear Medicine, Fondazione Poliambulanza, Brescia, Italy*
[d]*IRCCS Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy*
[e]*Department of Medical and Experimental Sciences, Unit of Neurology, Brescia University, Brescia, Italy*
[f]*University Hospitals and University of Geneva, Geneva, Switzerland*
[g]*Nuclear Medicine Unit, Department of Health Sciences (DISSAL), Genoa University and IRCCS AOU S. Martino-IST, Genova, Genova, Italy*
[h]*Clinical Neurobiology Unit, Department of Neuroscience (DINOGMI), Genoa University and IRCCS AOU S. Martino-IST, Genova, Genova, Italy*

Handling Associate Editor: Patrizia Mecocci

Accepted 12 July 2016

Abstract. The assessment of *in vivo* ^{18}F images targeting amyloid deposition is currently carried on by visual rating with an optional quantification based on standardized uptake value ratio (SUVr) measurements. We target the difficulties of image reading and possible shortcomings of the SUVr methods by validating a new semi-quantitative approach named ELBA. ELBA involves a minimal image preprocessing and does not rely on small, specific regions of interest (ROIs). It evaluates the whole brain and delivers a geometrical/intensity score to be used for ranking and dichotomic assessment. The method was applied to 504 ^{18}F-florbetapir images from the ADNI database. Five expert readers provided visual assessment in blind and open sessions. The longitudinal trend and the comparison to SUVr measurements were also evaluated. ELBA performed with area under the roc curve (AUC) = 0.997 versus the visual assessment. The score was significantly correlated to the SUVr values ($r = 0.86$, $p < 10^{-4}$). The longitudinal analysis estimated a test/retest error of $\simeq$2.3%. Cohort and longitudinal analysis suggests that the ELBA method accurately ranks the brain amyloid burden. The expert readers confirmed its relevance in aiding the visual assessment in a significant number (85) of difficult cases. Despite the good performance, poor and uneven image quality constitutes the major limitation.

Keywords: Alzheimer's disease, amyloid, image analysis, mild cognitive impairment, PET, standardized uptake value ratio

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

*Correspondence to: Andrea Chincarini, INFN, via Dodecaneso 33, I-16146 Genova, Italy. Tel.: +39 010 353 6496; Fax: +39 010 313358; E-mail: andrea.chincarini@ge.infn.it.

INTRODUCTION

The latest research criteria for diagnosing Alzheimer's disease (AD) suggest that *in vivo* quantification of brain amyloid-β (Aβ) deposition using positron emission tomography (PET) is emerging as a crucial tool in diagnosing or in excluding AD ([1] but also [2, 3]) and is likely to play a pivotal role in upcoming clinical trials of disease modifying agents [4–6].

The foreseeable increase in the use of this testing methodology, the sensitive nature of the outcome for the patient prognosis, and the availability of different radioligands pose new challenges to clinicians.

Currently, human readers who want to acquire the necessary expertise are counseled to attend ligand-specific courses, often brought in by pharmaceutical companies, where visual evaluation procedures tend to favor a dichotomic reading (positive/negative). This approach is justified by the apparent scarcity of cases where the ligand uptake distribution cannot be easily identified or where its spatial extent is limited. Indeed, while most amyloid-PET images are rather easily evaluated by a trained eye, as amyloid-PET becomes a widespread tool uncertain instances are going to be met more and more frequently.

Because of the non-trivial visual assessment in a non–negligible number of cases, a more sophisticated approach is required, which provides quantification (and rank) besides classification.

Clinicians can currently rely on commercially available quantification software, usually based on the numerical estimation of the Standardized Uptake Value ratio (SUVr) [7]. Briefly, SUVr procedure calculates the ratio of PET counts between a number of target regions of interest (ROI) versus a reference one. This way entails a significant image preprocessing to ensure that the ROIs are properly placed. In addition, ROI number, placement and size vary among implementations and they often require human feedback.

We believe therefore that there is room to improve the PET image reading by relaxing some of the constraints imposed by the SUVr approach (accurate image registration, the use of uptake and reference ROI) while providing robust ranking among subjects and proportionality to the visual assessment. With that, we do not wish to replace or belittle the established visual and SUVr-based semi-quantifications. Rather, the intent is to complement those with a novel and independent approach, with the goal of providing more robust and diversified knowledge on difficult-to-read cases.

This work proposes a method for the *E*va*L*uation of *B*rain *A*myloidosis (hereafter named "ELBA") on images of one of the new ^{18}F ligands (i.e., ^{18}F-Florbetapir). ELBA is designed to deliver the whole-brain amyloid-burden estimation and a ranking system to aid in the visual assessment. Comparison to SUVr semi-quantification, clinical evaluation at follow-up visits and cerebrospinal fluid (CSF) analysis is provided to complement the method validation.

MATERIALS AND METHODS

The ELBA method was developed on scans currently available in the ADNI database and acquired with ^{18}F-Florbetapir, which was chosen by ADNI to be the reference radioligand in the evaluation of brain amyloidosis [8].

The analysis procedure is automatic and does not need any human supervision save for an optional check after the spatial registration process, to ensure that the processed image is consistent and has acceptable characteristics.

In this study we first introduce the processing steps to characterize a PET scan using two features which are combined to give the ELBA score, next we proceed with visual assessment in blind and open sessions, finally we use the consensus visual assessment to set a cut-off value for ELBA and SUVr measures, and compare results.

PET scans and subject selection

We downloaded ^{18}F-Florbetapir scans of 244 subjects from the ADNI archive in the most fully processed format (series description in LONI Advanced Search: AV45 Coreg, AVG, Std Img. and Vox Siz, Uniform Resolution, subjects identification in Supplementary Table 1). Subjects were selected to have at least two scans (at baseline and after an approximately 2 years of follow-up), and 16 subjects came with three scans for a total of $n_i = 504$ PET images (i.e., 228 subject with two scans and 16 subjects with 3 scans). The ensemble properties of these images are shown in Table 1.

Subjects' clinical evaluation was taken to be the closest diagnosis to the baseline PET scan date. Cohorts were grouped by the ADNI core clinical criteria [9] as: normal subjects (NS, N = 70), early mild cognitively impaired (EMCI, N = 86), mild cognitively impaired (MCI, N = 26), late mild cognitively

Table 1
Subjects and scanner ensemble statistics

Maker	Model	N. of scans	N. of subj	M/F	Age
SIEMENS	1093	16	8	3/5	77.3 (8.0)
SIEMENS	1094	34	21	6/15	70.4 (6.7)
Siemens/CTI	ACCEL	12	6	5/1	72.3 (7.7)
GEMS	Advance	16	8	3/5	73.9 (4.3)
Philips Medical Systems	Allegro Body(C)	4	4	3/1	74.0 (7.3)
SIEMENS	Biograph20 mCT	1	1	1/0	72.0 (0.0)
SIEMENS	Biograph64	5	5	2/3	73.5 (7.4)
GE MEDICAL SYSTEMS	Discovery 600	4	2	1/1	71.5 (11.2)
GE MEDICAL SYSTEMS	Discovery 710	2	2	0/2	68.2 (1.3)
GE MEDICAL SYSTEMS	Discovery LS	19	11	7/4	71.2 (7.2)
GE MEDICAL SYSTEMS	Discovery RX	8	4	2/2	71.5 (4.2)
GE MEDICAL SYSTEMS	Discovery ST	32	23	14/9	75.1 (6.6)
GE MEDICAL SYSTEMS	Discovery STE	74	43	31/12	72.6 (7.3)
Philips Medical Systems	GEMINI TF Big Bore	6	5	4/1	75.2 (6.6)
Philips Medical Systems	GEMINI TF TOF 16	25	12	5/7	71.5 (6.8)
Philips Medical Systems	GEMINI TF TOF 64	14	9	5/4	72.7 (13.6)
Philips Medical Systems	Guardian Body(C)	10	5	3/2	75.6 (5.1)
Siemens/CTI	HR+	103	54	26/28	72.9 (7.7)
Siemens ECAT	HRRT	59	26	14/12	73.5 (8.6)
Philips	Ingenuity TF PET/CT	3	3	2/1	70.6 (18.1)
CPS	LSO PET/CT HI-REZ	45	24	11/13	70.2 (7.6)
Philips Medical Systems	NULL	2	1	0/1	70.7 (0.0)
SIEMENS	SOMATOM Definition AS mCT	10	10	5/5	75.1 (7.3)
–	–	504	244†	128/116	72.6 (7.6)

†The number of subjects in the total is the number of unique subject identifiers. This is not the sum of the respective column because 38 subjects were scanned on a different system at baseline and follow-up.

impaired (LMCI, N = 51), and Alzheimer's disease (AD, N = 11).

Image processing

The intent of ELBA is to capture intensity distribution patterns rather than actual counts in specific ROIs. Considering the brain as a whole, we observed that geometrical appearances of iso-intensity surfaces are rather characteristic in typical negative and positive subjects, the latter showing a sparser and more convoluted appearance than the former. In addition, whole-brain intensity histograms appear to be skewed toward higher intensities in positive subjects.

As qualitative interpretation, the PET signal clusters onto the gray matter patches with significant amyloid load, often surpassing the adjacent non-specific white matter intensity. The presence of higher intensity patches biases the counts statistics and, when thresholded, gives a more complex surface (with notches and several non-connected components).

To capture and quantify these characteristics we developed two features: one that gauges the iso-intensity surface complexity and another that assess the histogram propensity toward higher values. These features are global properties of the whole brain and do not require a reference ROI. An infographic showing these steps is provided in Supplementary Fig. 2.

Still, basic image processing consisting in a spatial registration to MNI coordinates is performed to allow the segmentation of the brain from the head. This is necessary because non-cortical regions like ventricles, cerebellum, and scalp do not carry information specific to the amyloid burden.

Reference PET

The reference PET (*RP*) is a mean image of 40 subjects acquired during the AVID-18 clinical trial at one center (Fondazione Poliambulanza Istituto Ospedaliero Brescia, Italy) with ^{18}F-Florbetapir tracer and the following acquisition parameters: injected dose = 370 MBq, acquisition time = 10 min (50 min after the injection), image reconstruction on a 256 × 256 matrix with 4 iterations, 21 subset, Gaussian filter with FWHM = 2 mm.

These subjects delivered a mix of negative and positive scans (14 and 26, respectively), whose evaluation was visually confirmed by one of the expert readers (UG). They were used to generate a spatial reference only and were not included in the ELBA score development.

The mix of positive and negative subjects was only used to provide a balanced template, with the aim to provide an average image, sampled from a typical population, which is used as registration target. In this respect, the evaluation of the single reader needs not to be confirmed, as it only serves to generate an average, smoothed reference.

Generally speaking, the registration process between same-modality images is more robust than the cross-modality counterpart. The main benefit of a RP therefore, was to relax the need of matching PET scans directly onto the MNI-MRI template, either directly or having the subject's MRI as guide.

To generate the RP, the 40 scans underwent a recursive registration process, each step delivering a mean template. The first step consisted of an affine registration onto the ICBM-152 MRI template using a mutual-information metric. An intermediate PET template was generated by averaging over the registered images. The intermediate template was used as reference for another registration batch to generate a second reference. The subsequent steps used an affine plus a weak non-linear registration—that is a non-linear warping using a large ($12\,mm$) smooth operator on the deformation field [10]—to improve on the reference image generation. The iterative process ended after no more than 5 steps, when the generated reference did not show significant changes with respect to the previous step.

The MNI-provided lobes, ventricles and sub-cortical regions segmentations were mapped onto the RP and visually inspected for consistency.

ADNI image processing

All downloaded scans (labeled with i, $i = 1..n$) were spatially registered with an affine transformation onto the RP, delivering p_i MNI-aligned images. We then segmented the cortical surfaces (c_i) and ventricles regions (v_i) from each p_i by means of non-linear mapping of the available pre-segmented masks on the reference image RP: the RP was registered with a non-linear transformation onto the target PET and the deformation field was applied to the segmented masks.

We extracted the brain ROI b_i considering all brain lobes delimited by the cortical surface c_i, neglecting the cerebellum, the brain stem and the ventricles. A sample of the image processing result is illustrated in Supplementary Fig. 1. To reduce the processing errors, longitudinal scans from the same subject were treated as a batch from the beginning (i.e., sharing registration parameters and masks).

Images were then measured with two methods: intensity-based and geometry-based, each delivering a characteristic feature.

Geometric feature

The selected brain volume b_i is partitioned into $n_L = 48$ iso-intensity levels $0 < l_j < 1$, $j = 1..n_L$ taken at equal quantile distances of the whole intensity distribution (i.e., n_L quantiles in the interval $[1 - 1/n_L, n_L]$). The number of levels is of little consequence provided it is chosen $n_L \gtrsim 10$ to ensure adequate sampling of the distribution (we tested the feature outcome with $n_L = 16, 24, 32, 48$ levels). Partitions consist of s_j surfaces and V_j enveloped volumes defined as

$$V_j = \left\{num.\ of\ voxels \in b_i,\ voxel\ intensity \geq l_j\right\}$$

$$s_j = \sum_{\partial V_j} 1$$

where the ∂ symbol denotes the boundary, that is s_j counts the number of voxels on the perimeter. The s_j and the enveloped volumes V_j are not required to be a connected set.

Each partition is characterized by two numbers: one representing the radius r_j^s of an equivalent sphere having the same surface extent as s_j, and another is the radius r_j^v of the equivalent sphere of volume V_j, that is

$$r_j^s = \left(\frac{s_j}{4\pi}\right)^{\frac{1}{2}},\quad r_j^v = \left(\frac{3V_j}{4\pi}\right)^{\frac{1}{3}}$$

If we plot r_j^v and r_j^s for all $j = 1..n_L$ on a Cartesian plane we get a characteristic curve inferiorly bounded by the bisector line $r^v = r^s$, which is the limit for all s_j being actual spheres. The characteristic curve distances itself from the bisector line the most when the s_j is rough and notched. Because of the peculiar spatial distribution of counts in the brain, typical appearance of the characteristic curves is rather different for amyloid-positive and negative scans (Fig. 1).

When we subtract the trivial bisector line, typically positive scans show a higher surface-to-volume ratio on the higher intensity levels (low r^v) with respect to the lower intensity levels (high r^v), and vice versa for negative scans.

The characteristic curve is integrated without the bisector area on the lower and higher half of its domain D (i.e., the range of r^v) to deliver the geometric feature G_i:

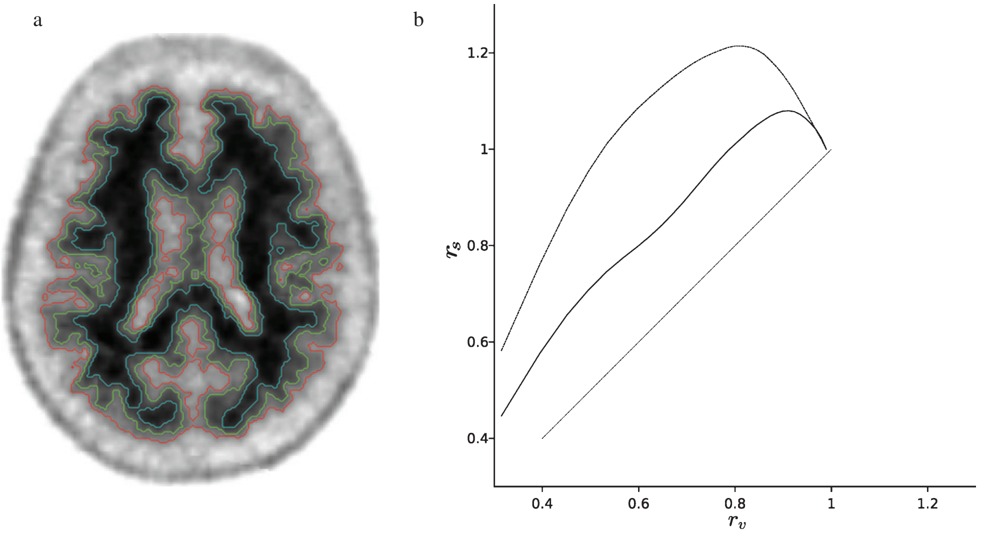

Fig. 1. Plot a: iso-intensity partition illustration on an axial projection, related to three percentile values of the intensity counts within the brain ROI (0.25, 0.50, and 0.75; red, yellow, and blue). Plot b: characteristic curve (r_v,r_s) for two typical cases: a low and a high amyloid burden scan (thick curve/dotted curve, respectively). Values are normalized to the respective brain volume and boundary. The thin line is the bisector.

$$G_i = \frac{\int_{D_1} (r^s(r) - r)\ dr}{\int_{D_2} (r^s(r) - r)\ dr},$$

$$D_1 = [min(r^v), r^v/2], D_2 = [r^v/2, max(r^v)]$$

Intensity feature

This feature gauges the intensity and contrast values in b_i and divides them into clusters. The chosen clustering method was *kmeans* [11, 12] with two classes ($High$, Low). Since *kmeans* uses an iterative algorithm starting form a random sample, to ensure reproducibility we run it for 10 repetitions, then choosing the one with minimum within-cluster sums of point-to-centroid distances.

In each class, we computed the number of elements K_{High} and K_{Low} and the class median intensity value I_{High}, I_{Low}. We conventionally linearly scale the intensity histogram so that the values corresponding to the 1% and the 99% percentiles are mapped onto the [0, 1] interval. We then defined the intensity feature C_i as

$$C_i = ln\left(\frac{K_{High}}{K_{Low}}\frac{I_{Low}}{I_{High}}\right)$$

The intensity feature modulates the relative number of elements in the classes with their contrast. As in the geometric feature, this latter formulation is expressed as a ratio too, so that both features are internally (intra-subject) normalized.

ELBA score

The two image features G_i and C_i were plotted on a Cartesian plane and used to fit a quadratic polynomial. Each point can be projected onto the curve to get two new coordinates: a curvilinear abscissa x_c (arc length) and a curvilinear ordinate y_c (see Fig. 2).

The ELBA score is simply x_c after a linear scaling and a shift to conveniently place the origin at the cut-off between negative and positive scans.

Up to this point, the construction of the ELBA score did not require any indication on the subject amyloid burden, or provide any hint about its age or clinical status. It was merely a way to combine information on the geometrical distribution of PET counts in the brain and information on the contrast between the brightest and darkest intensity components.

Alternative implementations

There can be equivalent implementations of the ELBA score which can replace the formulas listed above. For instance, any formulation which differentiates highly notched versus smoother surfaces can be used in place of the geometric feature. Similarly, intensity bias in a scan histogram can be equivalently assessed by the following ratio

$$I_b = \frac{< I > -q_1}{q_2 - < I >}$$

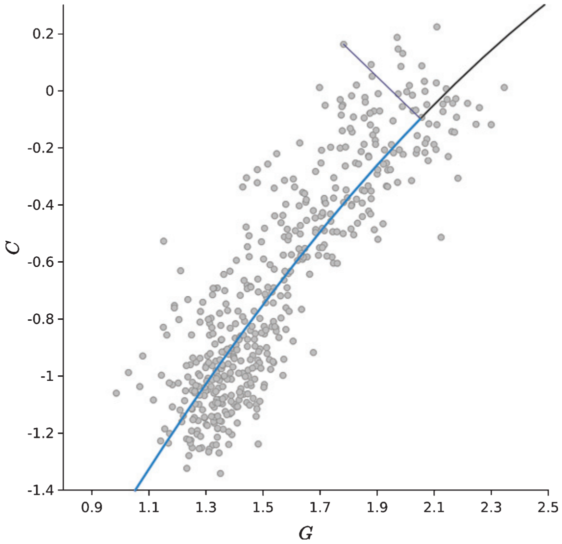

Fig. 2. ELBA scatter plot (intensity feature C versus geometric feature G). The black line is the quadratic model. For each scan (dots) the blue part of the line represents the curvilinear abscissa x_c (arc length), and the perpendicular line is the curvilinear ordinate y_c.

where $< I >$ is the mean intensity and q_1 and q_2 are the 1% and the 99% intensity percentiles.

Even the use of the curvilinear abscissa described above is not mandatory and an alternative score E' can be formulated with a simple geometric mean[2] as

$$E' = \sqrt{G \cdot C}$$

In addition, given the apparent high correlation between the two features G and C (Fig. 2), one might also be tempted to use one feature only rather than a combination of the two.

As example of possible different implementations, we show in Supplementary Fig. 7 the performance comparison among the ELBA score, the geometric and intensity features alone, and the alternative score E'.

The analysis shows the equivalence of the curvilinear abscissa versus the geometric mean approach, whereas the single feature comparison seems to favor the intensity feature C. For the sake of robustness, we kept both features in the ELBA score.

VALIDATION

All PET images were evaluated by five independent readers: two nuclear medicine (NM) physicians and one neurologist who have been trained to give teaching courses to NM physicians and read more than 200 scans with supervision (expert readers) and two moderately expert readers (NM physician) who read more than 200 images under supervision.

Upon coarse data examination, the readers noticed an apparent quality difference among scans. They agreed therefore on an operational definition of "*sub-optimal image quality*" for the purpose of keeping track of quality-induced mis-readings during the validation. They considered a subjective evaluation of low quality reconstruction, motion artefacts, or low signal-to-noise ratio according to each own clinical experience. The aim was to tag scans whose characteristics could interfere in (or impair) the visual assessment.

Regardless of the their quality, all scans were processed and visually evaluated. The quality label was used for retrospective analyses only, to keep track of possible grounds for difficult cases.

The 504 images were divided into 488 among baseline and first follow-up scans, plus 16 second follow-up scans. For the validation purpose we used the 488 baseline and follow-up images, while the additional 16 scans only were used in the longitudinal evaluation (described below).

Baseline and follow-up scans were read as independent images, all mixed together with random order so that evaluators were very unlikely to see the same subject twice during the reading sessions.

Blind phase

Images were presented after the preprocessing steps described above in "Image processing". The blind evaluation was carried out by each reader without interaction, without support from any automatic analysis software, blind to the clinical data, blind to the ELBA output and according to each reader's own practice and experience.

Readers were initially allowed to judge on a three classes base: *negative*, *positive*, and *uncertain*. Readers were asked to use *negative* and *positive* tags on images where they were absolutely confident of the visual assessment. Scans whose evaluation implied a more elaborated visual inspection and where the possibility of doubt existed should have been initially tagged as *uncertain*. In addition, readers were

[2] The geometric mean is often used when comparing different items to find a single "figure of merit" when each item can span different numeric ranges.

also asked to add a tag on the perceived image quality.

Besides the individual evaluations, the analysis of the blind phase delivered 4 image set: the *P* set and the *N* set, that is scans which were consistently marked as *positive* and *negative* by all readers; the *U* set, that is scans which received an *uncertain* comment from at least one reader; and the *C* set, that contains those scans which received contrasting judgment (*positive* and *negative* together from one or more readers).

More specifically, the *U* set consists of scans with at least one *uncertain* comment but otherwise no contrasting labels, whereas the *C* set consists of scans which received contrasting evaluations but which may also have had one or more *uncertain* comment.

Open phase

In this phase, all scans in the *U* and *C* set were presented again. Each scan was evaluated by all five readers in an open session, with interaction, where they were invited to reach a consensus on either *negative* or *positive* label.

This time though the ELBA output was partially used to aid in the analysis. Readers were not made aware of the ELBA score but the image to evaluate was shown side-by-side with two other images, ordered on the ELBA score scale: the nearest one from the *P* set and the nearest one from the *N* set (see Fig. 3b for an example). This visualization was meant to help in the assessment, by comparing the scan under scrutiny to the most similar, validated assets.

The original scans (i.e., not spatially normalized) were also available for consultation. They were used to cross-check the consensus evaluation during the open discussion.

Comparison with SUVr-based methods

We calculated the average cortico-cerebellar SUVr on all scans, to compare the ELBA score to this widely used semi-quantification method. We used a data-driven approach with the whole cerebellum (white and gray matter) as reference and a number of cortical ROIs as uptake regions, as displayed in Fig. 4. The SUVr information was used neither in the blind nor in open validation phase.

The cortical regions were: medial frontal gyrus, lateral frontal cortex (middle frontal gyrus), lateral temporal cortex (middle temporal gyrus), lateral parietal cortex (inferior parietal lobule), insula, caudate nucleus, and precuneus-posterior cingulate region. They were obtained by a data-driven approach similar to the one described in [13].

Briefly, we took 50 negative and 50 positive subjects from the *N* and *P* set (i.e., those subjects tagged *negative* and *positive* independently by all readers). Each image was spatially normalized into MNI space and intensity normalized by the mean counts in the whole cerebellum. Then a positive and negative mean image was generated. The negative mean was subtracted from the positive one, the result was left-right symmetrized and smoothed with a 3D-Gaussian filter ($\sigma = 3$ mm). We found an optimal threshold by maximization of the area under the ROC curve of the SUVr between the 50 positive and 50 negative scans.

SUVr measures were provided by ADNI too, but only on a fraction of the baseline PET scans (111). The ADNI-provided values are calculated according to protocols described in [8], and they are the average cortical-cerebellar SUVr computed with two methods: one with the syngo PET Amyloid Plaque (sPAP) software and another with the Avid Semi-Automated Method (AVID [14]). We checked that SUVr values computed with the data-driven approach agreed with those already provided on the subset of baseline scans using correlation and linear regression analysis.

The final SUVr cut-off was chosen to maximize accuracy using the consensus negative/positive labels after the open phase session.

Comparison with CSF Aβ_{42} quantification

CSF was acquired by lumbar puncture following procedures and criteria identified by ADNI (analysis details and quality control procedures are available at http://adni.loni.usc.edu/). The biomarker data set used in this study were taken from the file series upennbiomk4.csv to upennbiomk8.csv. We considered CSF and PET data whose measurements were closest in time, restricting to lumbar punctures and PET measurements performed within 100 days of each other, which resulted in a selection of 203 subjects.

The cut-off on CSF Aβ_{42} values was 174 ng/L (from [15]), a value Mattsson et al. found to be optimal to maximize accuracy between stable and progressive MCI.

We compared CSF with ELBA and SUVr scores and with visual assessment. We also evaluated the diagnostic performance for NS versus AD (57 versus 51 subjects, respectively), where clinical assessment was taken at the latest possible follow-up visit.

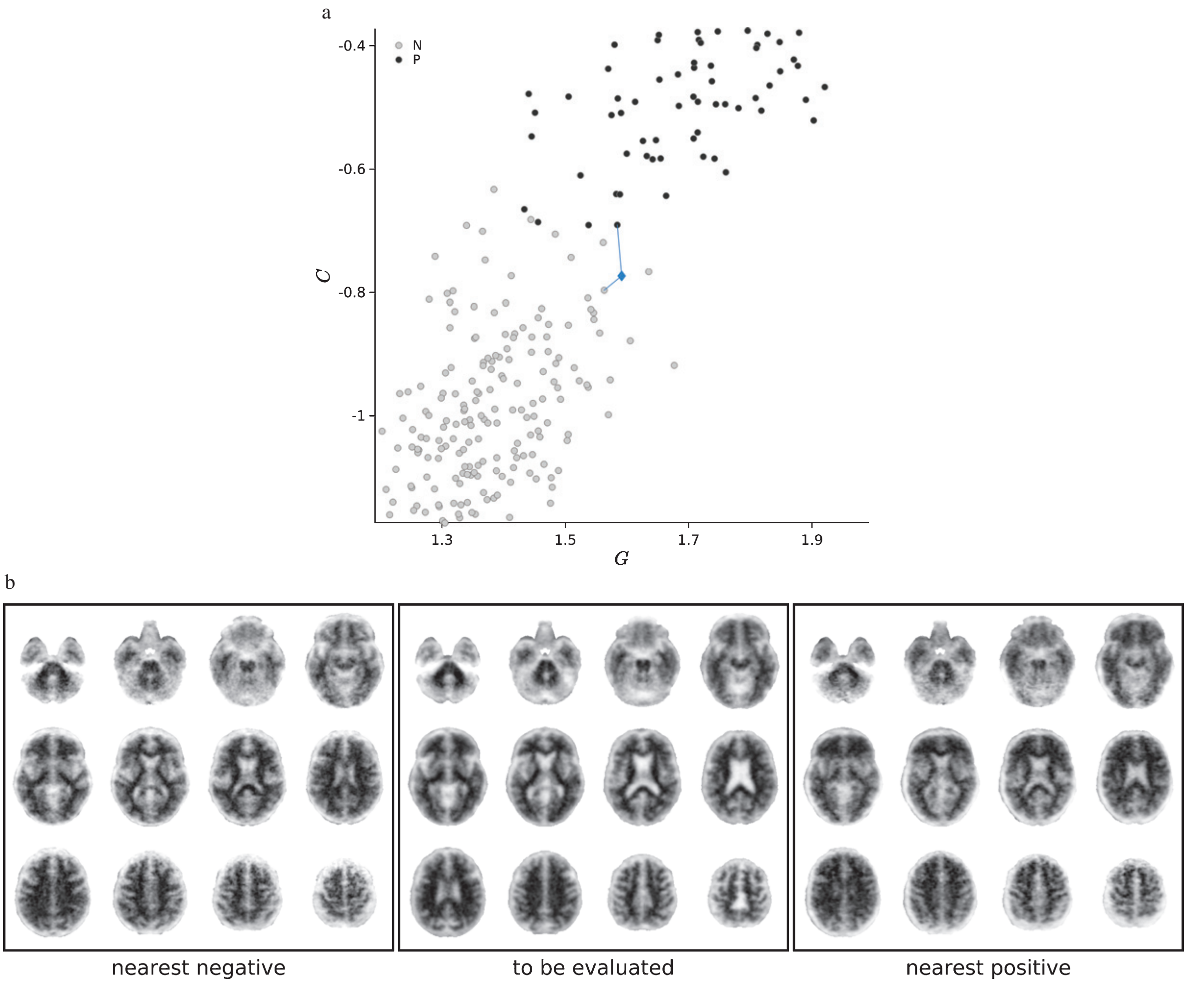

Fig. 3. Plot a: zoom on the geometric (G) and intensity feature (C) plane around a new scan to be evaluated (blue diamond). The scan is compared to the nearest tagged cases taken from the N and P set only, i.e., scans with concordant independent evaluation by all readers given in the blind session. The distance is evaluated on the ELBA plane, where the nearest positive and negative scans are indicated by the blue line. A trans-axial representation of the three scans is shown in plot b as sample of the additional information used in the open phase session.

Longitudinal evaluation

We computed the annualized score change (ASC) using follow-up scans. For the 228 subjects with a baseline and a follow-up scan we used the formula

$$\mathrm{ASC} = \frac{s_f - s_b}{t_f - t_b}$$

where f and b label the follow-up and baseline score and subject's age. For the 16 subjects with two follow-up we computed the least square linear regression

$$s + \epsilon = mt + b$$

of the score s versus time t (b is the intercept and ϵ the residuals) and the ASC is simply

$$\mathrm{ASC} = m$$

To compare the numbers with other indexes (such as those computed with SUVr quantification), we considered the normalized quantity

$$\delta = \frac{\mathrm{ASC}}{\mathrm{iqr}(s)} \times 100$$

where the interquartile range (iqr) of the score s is used as normalization factor to estimate the relative score change over the observed population variability.

Using the 16 subjects with three scans each, we attempted to estimate the analysis stability and robustness. With the present data we could not evaluate a test/retest paradigm on the same subject (i.e., two repeated scans with subject reposition) so we

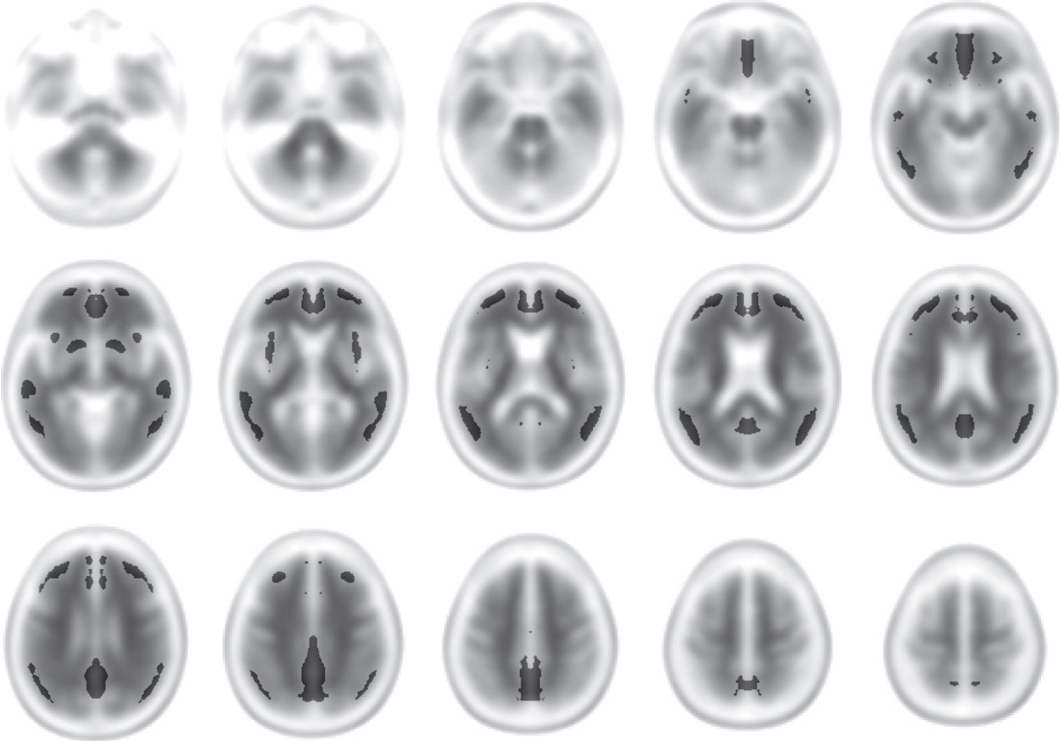

Fig. 4. Uptake regions for the computation of SUVr values superimposed on the reference PET. The counts normalization region (not shown here) is the whole cerebellum.

used the residuals ϵ on the linear regression of the score versus time as a proxy. This works under the hypothesis that the amyloid burden piles up slowly and linearly at least within the follow-up time (approx. 4 years)—an assumption that agrees with the accepted neuropathological models—and assuming that the technical errors on repeated scans are independent from the subject's amyloid burden.

Deviation from the linear behavior is then used as a surrogate to the test/retest error and treated as analysis uncertainty (due to protocol, image acquisition, reconstruction and processing) and used to estimate the error on the single examination. It can be thus compared to literature works on the various Aβ ligands [16–19], which show an average test/retest relative error using the global SUVr measurement in the range $3\% - 7\%$.

Obviously, the measured uncertainty would not be due to the feature processing only, but also to the different acquisition conditions, scanners, and reconstruction parameters. For instance, we remark that among the 16 subjects with three scans, 11 were acquired on a single scanner on baseline and follow-ups, and 5 subjects were acquired with two different scanners at some follow-up.

Finally, we selected a subset of subjects with either negative or borderline ELBA score (i.e., ELBA <0.5) which were also evaluated by CSF analysis. These subjects were divided into three groups, of progressively lower average Aβ_{42} concentration (ng/L): group A, A$\beta_{42} > 230$; group B, $174 <$ A$\beta_{42} < 230$; group C, A$\beta_{42} < 174$.

These three groups contained an approximately equal number of subjects (52, 48, and 34, respectively) where the above-cutoff ensemble was divided into A and B to better reflect the possible trend in longitudinal behavior with respect to CSF outcome. An analysis of variance was applied to check for significant differences among groups.

Clinical follow-up

The latest clinical evaluation was checked and compared to the initial assessment. We found 78 subjects who had their assessment reviewed from baseline (latest clinical data sheet downloaded on May 10, 2016). The average time to conversion

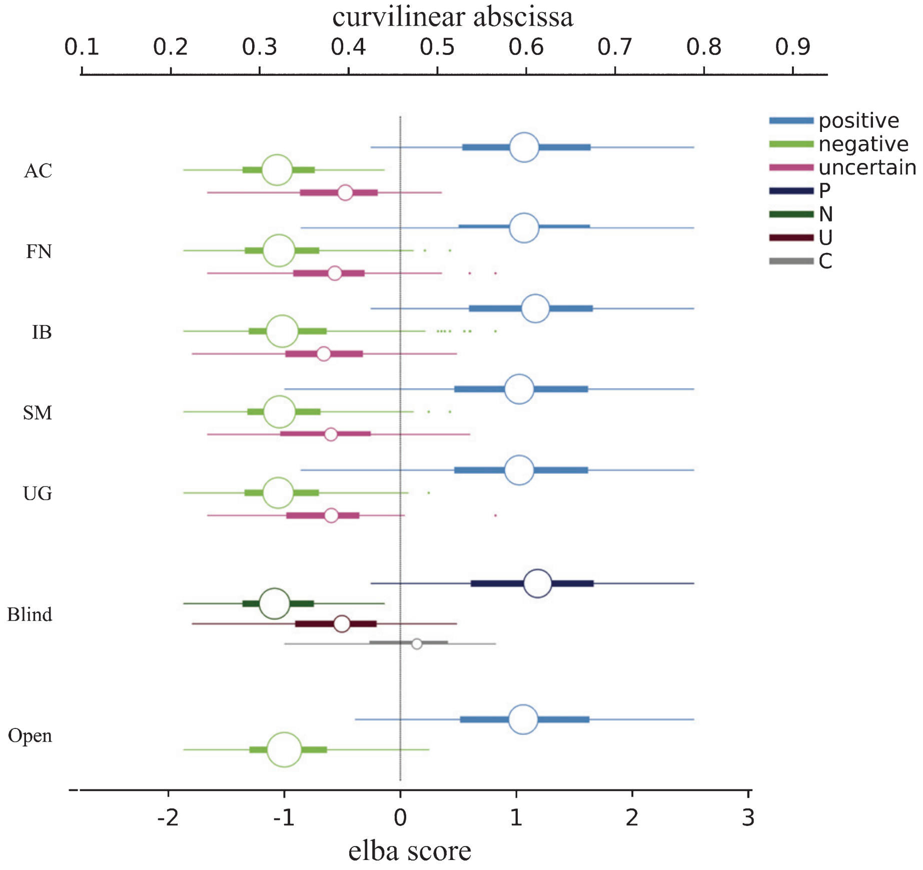

Fig. 5. Visual evaluation versus the ELBA score (or equivalently the curvilinear abscissa) before and after the blind phase session; *negative*, *positive*, and *uncertain* labels are given by each reader (AC, FN, IB, SM, and UP). In the blind session summary ('Blind'), P and N refers to scans consistently (i.e., by all readers) tagged *positive* and *negative*, respectively. U refers to scans which received at least one *uncertain* tag but no contrasting assessment; C refers to scans which received contrasting assessment (even when together with *uncertain* tags). The open session summary ('Open') shows the dichotomic consensus. Circles are centered on the median value of the respective cohort and their areas are proportional to the sample size. The vertical line marks the cut-off. Thick lines mark the 25% and 75% percentile, thin lines extends up to 1.5 times the interquartile range.

was 35 ± 10 months (mean and standard deviation). They were divided into 4 classes according to the baseline/follow-up assessment: MCI→AD (46), MCI→NS (16), NS→MCI (10), and NS→AD (6).

Their ELBA and SUVr scores at baseline were used to measure consistency between semi-quantification methods and the agreement with the diagnosis at baseline and follow-up. When present, CSF Aβ_{42} level was used to help discussing borderline cases and evaluate diagnosis agreement with the biomarkers.

Further methodological considerations

We checked for dependency on tracer doses administered, scan start time after the injection, and white matter integrity as possible factors which might influence the intensity mapping. Data set containing scan information were downloaded from ADNI (files av45meta.csv, ucd_adni1_wmh.csv and ucd_adni2_wmh.csv, additional analysis details are available at http://adni.loni.usc.edu/). These additional data were available for 242 scans.

All parameters were linearly regressed against the ELBA score, showing no significant trend (95% CL on the line slope include the zero). A visual representation of this analysis is found in Supplementary Fig. 6.

In addition, we estimated the generalized accuracy, sensitivity, and specificity of ELBA and SUVr scores versus the visual assessment (open phase) by means of an iterative procedure. Briefly, we randomly selected 50% of the PET datasets as the training group (for the coefficient estimation for SUVr and

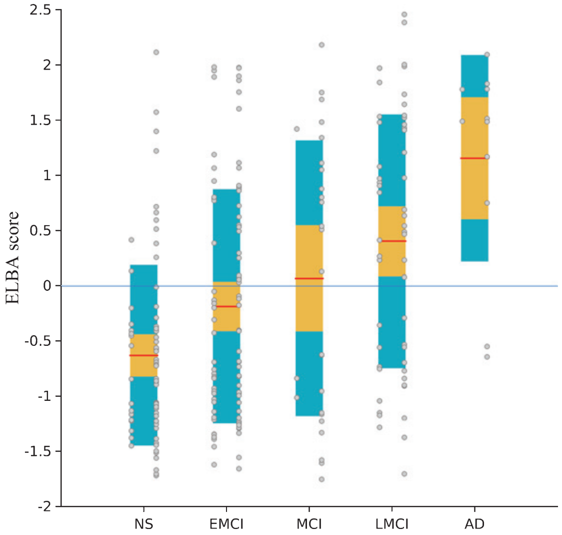

Fig. 6. Mean (baseline and follow-up) ELBA score distribution versus clinical cohort. Dots (subjects) are shown with the median (red line), the 95% conf. level on the median (yellow band) and the interquartile range (azure band). In each cohort, the leftmost [rightmost] dots represents age class <70 [>70] years old. NS, cognitively normal subjects; EMCI/LMCI/MCI, early- / late- / mild cognitive impairment; AD, Alzheimer's disease.

ELBA) and 50% of the PET datasets as the testing group (for the cross-validation of the estimated coefficients). Each group contained approximately an equal number of positive and negative scans, based on the concordant judgment of the experts (i.e., after the open session). Cut-off values for SUVr and ELBA scores were computed on the training group (maximizing accuracy) and applied to the testing group. The procedure was iterated 500 times.

RESULTS

Blind and open phase performance

The score distribution was grouped by visual assessment and results are displayed in Fig. 5, where we show the single reader evaluation and the combined set after the blind and the open sessions. An equivalent plot using the curvilinear ordinate y_c is provided in Supplementary Fig. 4.

Of the 488 scans, 186 were consistently marked as *positive* by all readers (*P* set), 217 marked as *negative* (*N* set), 63 received an *uncertain* comment from at least one reader (*U* set), and 22 received contrasting judgment (*positive* and *negative* together, *C* set). The agreement among the readers after the blind session was measured by the intraclass correlation coefficient (ICC) and was found to be ICC=0.94, ($p < 10^{-4}$).

Scans labeled *uncertain* were rather consistent both in number—59, 45, 49, 42, and 48 for AC, IB, UG, SM, and FN, respectively—and ELBA score. Moreover, all readers consistently tagged 34 scans as *uncertain*. Not surprisingly, 45 out of 63 (71%) of all *uncertain* scans were also flagged as "sub-optimal quality" by at least one reader.

To define the cut-off value on the score, we used the open phase results. The cut-off was chosen to maximize accuracy and the original curvilinear abscissa (x_c) was scaled and shifted to have cut-off = 0 and the mean score on the negative scans = –1 (lower x-axis in Fig. 5). The linear scaling is not a necessary step *per se*; it is applied only to facilitate the score interpretation.

The discriminating power was measured by the area under the receiver operating characteristic curve (AUC), giving AUC = 1.000 for *N* versus *P* in blind condition (i.e., on the scans on which all readers independently concurred), and AUC = 0.997 [0.993 – 0.999] for *negative* versus *positive* (accuracy = 0.97) after the open phase discussion (CL = 0.95 within brackets).

We also show the distribution of the score grouped by clinical cohorts in Fig. 6; 244 subjects are plotted, grouped by their clinical classification at baseline. The values on the y axis are the average between the baseline and the follow-up ELBA scores. To enhance the reading, subjects in Fig. 6 are also grouped by age: for each cohort, the leftmost and rightmost dots represents subjects of age <70 and >70 years, respectively.

Comparison with SUVr-based methods

Fig. 7 and Table 3 present semi-quantification values and binary summary separately for the different diagnosis groups, using the visual assessment after the blind session. Fig. 8 shows the comprehensive ELBA score versus SUVr semi-quantification on all scans labeled according to the visual evaluation after the open phase session; the confusion matrix is provided in Table 2a and b. In these figures and tables, the optimal SUVr cut-off is 1.23 and it was computed maximizing the accuracy using the open session results, in the same way as with the ELBA score.

The SUVr AUC = 0.978 [0.964 – 0.985] (accuracy = 0.94). The Pearson correlation between ELBA and SUVr scores is $r = 0.86$ ($p < 10^{-4}$).

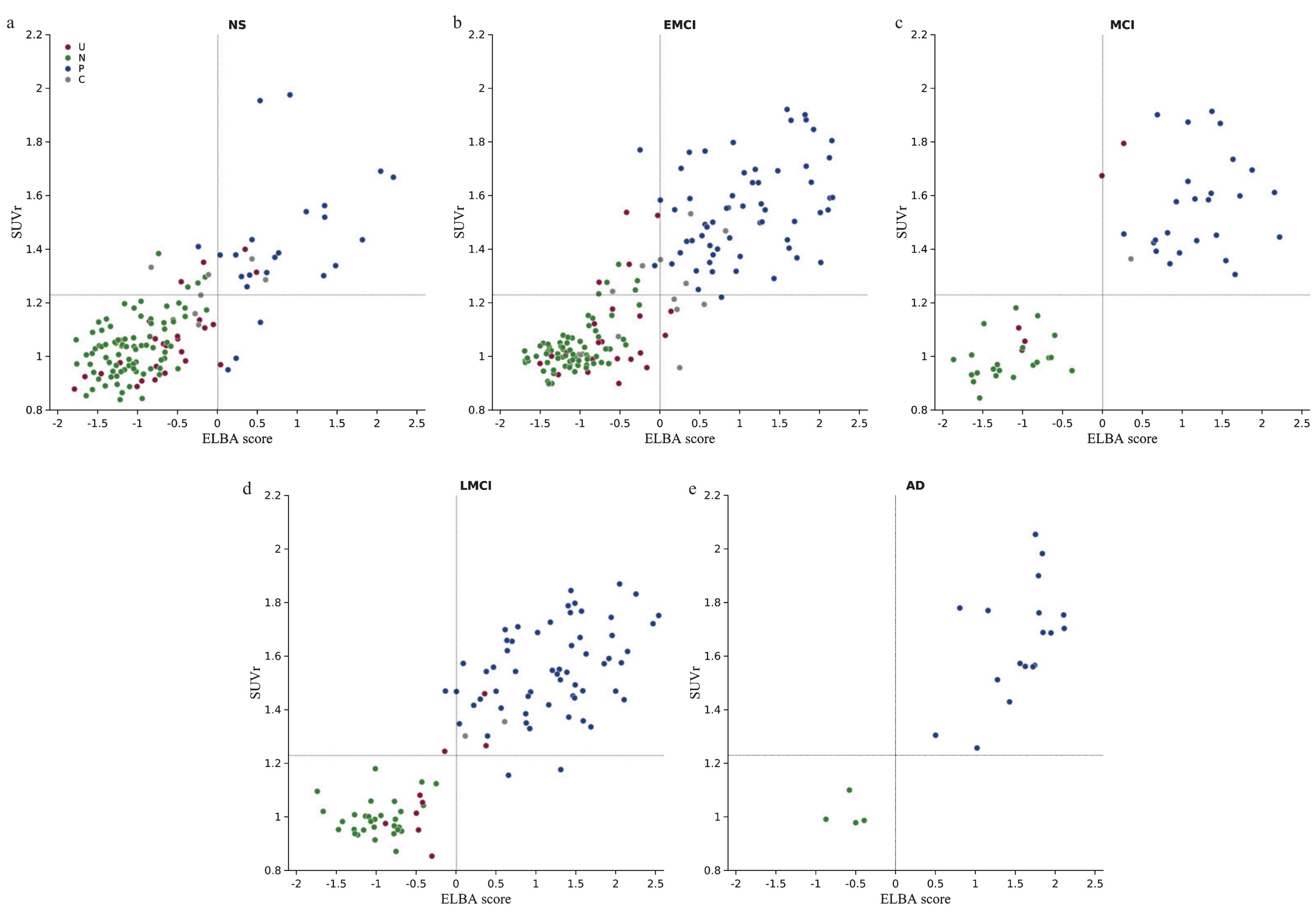

Fig. 7. ELBA score / SUVr values versus visual assessment after the blind session, grouped by baseline clinical evaluation. P and N refers to scans consistently tagged *positive* and *negative*, respectively. U refers to scans which received at least one *uncertain* tag but no contrasting assessment; C refers to scans which received contrasting assessment. NS, cognitively normal subjects; EMCI/LMCI/MCI, early- / late- / mild cognitive impairment; AD, Alzheimer's disease.

Table 2
Confusion matrices for ELBA and SUVr scores alone (a), and compared to the visual assessment (b)

a) Confusion matrix (ELBA score/SUVr)

N = 488		ELBA	
		positive	*negative*
SUVr	*positive*	190	25
	negative	13	260

b) Confusion matrix versus visual assessment (open session)

SUVr	ELBA	Visual assessment	
		positive	*negative*
positive	*positive*	189	1
positive	*negative*	7	18
negative	*positive*	7	6
negative	*negative*	1	259

We matched our SUVr computation on 111 baseline scans which were provided with independent SUVr values by ADNI (sPAP and AVID methods). Pearson correlation is: $r = 0.98$ ($p < 10^{-4}$) this work versus sPAP; $r = 0.99$ ($p < 10^{-4}$) this work versus AVID. The least square line $y = ax + b$ between our SUVr and the AVID one has a slope [CL=0.95] $a = 0.90$ [0.87, 0.92] and intercept $b = 0.07$ [0.03, 0.10] which translates in an equivalent optimized cut-off = 1.18 on the AVID values.

Using the cross-validation procedure to establish cut-off values and estimate the generalized performance, we found that the combined accuracy, sensitivity, and specificity for ELBA are 0.96, 0.97, and 0.95, respectively. Similarly, SUVr results are: 0.95, 0.96, and 0.93. The ELBA cut-off range was found within the interval [–0.14 – 0.18] (95% CL).

Comparison with CSF Aβ_{42}

The scatterplot between baseline ELBA score and CSF Aβ_{42} concentration is provided in Fig. 9, where the open-phase visual assessment was used to group data.

Concordance between ELBA score and Aβ_{42} score classification was achieved in 184 out of 203 (90.6%) instances; in 7.5% of patients an altered Aβ_{42} score was found with normal ELBA score and, on opposite, 1.5% of patients had a (slightly) increased ELBA score with normal Aβ_{42} levels.

The related SUVr representation is in Supplementary Fig. 9. The confusion matrix for CSF versus visual assessment (open phase), ELBA and SUVr scores is provided in Table 4, where the accuracy is found to be 0.90, 0.91, and 0.89, respectively.

The number of subjects with confirmed NS and AD clinical status at follow-up and with CSF analysis is 108. The area under the ROC curve (auc, CL = 0.95) for NS (57) versus AD (51) is found to be auc = 0.88 [0.78 – 0.95] (CSF), auc = 0.96 [0.90 – 0.98] (ELBA) and auc= 0.92 [0.84 – 0.97] (SUVr). A graphical representation is in Supplementary Fig. 10.

Clinical follow-up

Fig. 10a shows the scores for those subjects whose clinical evaluation changed over time. The agreement between ELBA and SUVr dichotomized scores is 97.4%. Most MCI→AD (89%) do fall into the SUVr positive/ELBA positive quadrant, as well as most MCI→NS (87.5%) fall into the SUVr negative/ELBA negative quadrant.

When considering CSF Aβ_{42} values, which were available for 63 out of 78 converters, the agreement with ELBA was slightly better than with SUVr (93.7% versus 90.5%, Figs. 10b and c).

Very limited discrepancies with the clinical evaluation are apparent. For instance, there are two subjects (1 MCI→AD and 1 NS→AD) who exhibit consistently negative markers (ELBA, SUVr, and CSF) despite their final clinical assessment, while another NS→AD is borderline negative for CSF and ELBA and borderline positive for SUVr.

Longitudinal evaluation

Longitudinal analysis on the subjects with a baseline and one follow-up scans showed rather scattered values, although a pattern could be clearly discerned. Fig. 11 shows the distribution of δ versus the average ELBA score together with a 2^{nd}-order polynomial

Table 3
Binary semi-quantification versus visual assessment and clinical evaluation

ELBA/SUVr	Visual assess.	Baseline clinical eval.					
quadrant	(blind)	NS	EMCI	MCI	LMCI	AD	*Scan tot.*
–/–	N	81	71	21	31	4	208
	P	0	0	0	0	0	0
	U	20	18	3	6	0	47
	C	3	2	0	0	0	5
–/+	N	4	5	0	0	0	9
	P	1	2	0	1	0	4
	U	2	4	1	1	0	8
	C	2	2	0	0	0	4
+/–	N	0	0	0	0	0	0
	P	3	1	0	2	0	6
	U	1	2	0	0	0	3
	C	0	4	0	0	0	4
+/+	N	0	0	0	0	0	0
	P	19	57	25	57	18	176
	U	2	0	1	2	0	5
	C	2	4	1	2	0	9
Scan tot.		140	172	52	102	22	488

Classification summary of all scans using the visual assessment after the blind session (see Fig. 7). P and N refers to scans consistently (i.e., by all readers) tagged *positive* and *negative* respectively. U refers to scans which received at least one *uncertain* tag but no contrasting assessment; C refers to scans which received contrasting assessment (even when together with *uncertain* tags). NS, cognitively normal subjects; EMCI/LMCI/MCI, early- / late- / mild cognitive impairment; AD, Alzheimer's disease. Value represent the number of scans in each class.

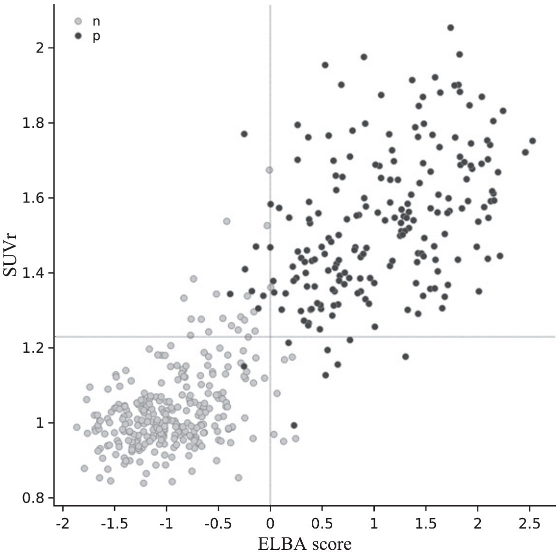

Fig. 8. ELBA and SUVr scores scatter plot for all scans. 'n' and 'p' tag the consensus evaluation after the open phase session. Dotted lines mark the cut-off values.

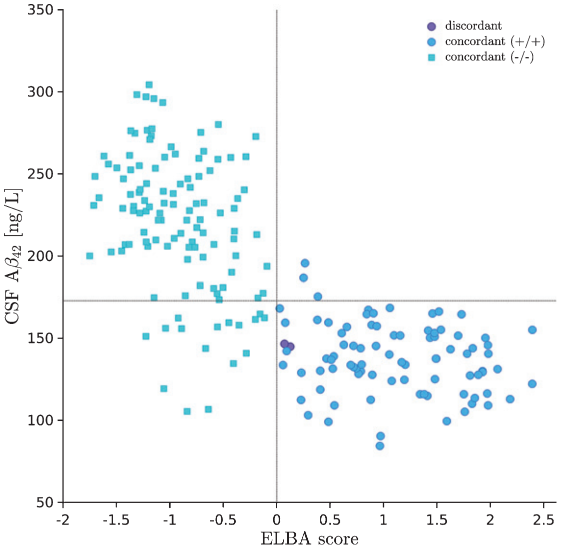

Fig. 9. CSF Aβ_{42} level versus baseline ELBA score. Markers are grouped by binarized ELBA score being either concordant or discordant with the consensual visual evaluation (open phase). Cut-offs are marked with thin dotted lines.

model, used to fit the data. The interquartile value used for normalization of the ASC is $\text{iqr}(s) = 1.89$. An equivalent graph using SUVr values is reported in Supplementary Fig. 5.

Subjects with two follow-up scans are also plotted on the same graph. Cohortwise, these subjects belonged to NS (5), EMCI (2), and MCI (9). Using these latter 16 subjects, we computed the residuals ϵ from the linear fit, to be used as proxy of the test/retest error. The standard deviation of the residuals is approximately $\sigma = 0.084$, which amounts to an estimated relative error $\sigma_r = \sigma/\text{iqr}(s) = 4.4\%$ ($\sigma_r = 2.3\%$ if we normalize on the 95% percentile of the score range). The equivalent

Table 4
Confusion matrix for CSF Aβ_{42} concentration versus the consensus visual assessment (a), ELBA (b) and SUVr scores (c)

	N = 203		Visual evaluation	
a)			*negative*	*positive*
	Aβ_{42} [ng/L]	>174	100	3
		<174	18	82
	N = 203		ELBA score	
b)			<0	>0
	Aβ_{42} [ng/L]	>174	100	3
		<174	16	84
	N = 203		SUVr score	
c)			<1.23	>1.23
	Aβ_{42} [ng/L]	>174	96	7
		<174	15	85

SUVr values are $\sigma_{SUVr} = 0.023$ and $\sigma_r = \sigma_{SUVr}/\text{iqr}(SUVr) = 4.6\%$

We computed also the relative annualized score change δ grouped by cohorts. Reported values (expressed in $\%/year$) are the median and its confidence interval (at CL = 95%): $\delta = 4.6$ [3.1, 6.1] (NS), $\delta = 8.1$ [6.6, 9.6] (EMCI), $\delta = 3.4$ [0.6, 6.2] (MCI), $\delta = 6.0$ [3.9, 8.2] (LMCI), $\delta = 5.6$ [−0.7, 11.9] (AD). A t-test found the values to be significantly different between EMCI and NS cohorts ($p = 0.02$) and between EMCI and MCI cohorts ($p = 0.02$).

Finally, we analyzed the ELBA ASC against the CSF Aβ_{42} concentration (Fig. 12 and Supplementary Fig. 8 for SUVr ASC) for 134 subjects with negative or borderline ELBA score (<0.5). The Pearson correlation between the two quantities is small but significant ($r = -0.30$, $p = 0.0004$). The grouping by CSF intervals allows to see a trend beyond the noisy ASC data. From the highest mean Aβ_{42} concentrations to the lowest, the ASC ranges from group A = 0.07 [0.03 – 0.10] to group B = 0.14 [0.11 – 0.17], to group C = 0.20 [0.14 – 0.27] (mean and 95% CL on the mean). The analysis of variance indicates that group A and C, as well as group B and C are significantly different ($p < 10^{-4}$ and $p = 0.02$, respectively). The corresponding analysis on SUVr ASC gives: group A = 0.005 [–0.00 – 0.01], group B = 0.01 [0.00 – 0.02], and group C = 0.03 [0.01 – 0.04]. The between-groups comparison is $p = 0.001$ (A versus C), and $p = 0.03$ (B versus C).

DISCUSSION

The proposed analysis shows that it is feasible to construct a semi-quantification method on amyloid-PET images without relying on counts-ratio approaches. The ELBA method shows good performance versus the dichotomic visual assessment and has ranking characteristics, proven both on cohort-based and longitudinal analyses.

From a methodological point of view, both SUVr-based and ELBA approaches require image registration techniques (spatial normalization), although the lack of small cortical ROIs in ELBA renders the registration process and the template choice less demanding. ELBA tries to mimic the human visual process, in that it captures information on global contrast and intensity distribution rather than weighing intensity in predefined regions. While further tests are necessary, particularly on all other major PET tracers and with histologically validated scans, it is worth noting that this process delivers comparable (if not slightly better) information in terms of semi-quantification, classification, and ranking, with respect of the widely used SUVr methods.

Blind and open phase performance

Expert readers were concordant and confident in reaching a diagnosis (*positive* or *negative*) on 403 scans (82.6%), without the influence of clinical information and other imaging data. This finding confirms that a trained reader can safely rely on his/her experience on over 80% of images, when evaluating by means of visual analysis.

As a second point, the five expert readers were not able to give a concordant diagnosis (i.e., amyloidosis present/absent) in 22 (4.5%) images and in 63 (12.9%) more images at least some of them declared that the initial assessment was unclear (i.e., doubtful). Interestingly, *uncertain*-tagged scans were mostly located in the negative domain whereas conflicting scans were evenly placed around the cut-off. As a consequence, looking at the fraction of uncertain and contrasting labels versus clinical cohort (blind session, Table 3), we find that for cognitively normal subjects and EMCI it is 23% and 21%, respectively, versus a 11% found in MCI and LMCI (AD has 0%).

Overall, these 85 (17.4%) scans may represent the borderline scans where a quantification approach can help in reaching a definite diagnosis. This fraction of scans is polled from 61 subjects and could derive both from healthy subjects or mainly early MCI patients in a stage when Aβ is accumulating in the brain but still in a limited amount [20, 21], thus leading to difficult reading. In facts, they belonged to NS (22), EMCI (27), MCI (4), and LMCI (8).

When the experts were aided by the closest negative and positive images, as yielded by the ELBA output, the 22 conflicting cases and 63 uncertain were solved, which is a non-trivial aid this automatic system can provide even to expert readers. We have not tested the potential aid ELBA can give to less experienced readers but it should be intuitively higher.

A substantial number (about 19%) of healthy subjects had a positive score, although the positivity fraction is mainly ascribed to the elderly (>70 years) ones, on a par with literature findings. Also, a few patients with AD had a normal score, which is in keeping with the literature [22] and raising the possibility of wrong clinical diagnosis (ADNI subjects do not have pathological confirmation), patients with discrepant normal amyloid-PET scan but abnormal Aβ_{1-42} levels in CSF [23] or, alternatively, of patients with suspected non-Alzheimer pathology (SNAP [24]).

In the middle, MCI subjects in progressive stages of cognitive impairment were roughly halved between positive and negative scans, which highlights the presence of causes of MCI (such as frontotemporal lobar degeneration or vascular cognitive impairment) other than AD, and possibly of SNAP. Ranking based on the overall amyloid burden is also apparent and it is a benefit of the method, as the average ELBA score progressively increases from healthy subjects to patients with AD.

Comparison with SUVr-based methods

The performances of both approaches are very similar, with a modest but statistically significant edge in favor of ELBA. The optimized SUVr cut-off (=1.23 and based on the visual assessment) is higher than the values typically considered in literature for this ligand (1.10 - 1.14, [14, 22, 25]) but it agrees with other works (e.g., [15]), where their cut-off value (1.24) was obtained by AUC optimization.

Tables 2b and 3 summarizes the binary classification (ELBA and SUVr) versus the visual assessment in both blind and open session. The comparison to cortico-cerebellar SUVr in the cases where both methods agree shows that SUVr values are not alone in providing good classification relative to visual assessment, and that and alternative and independent approaches can enrich the information obtainable from the PET scan.

Although the ensemble on which the two methods disagree is limited to a small number of subjects (quadrants for which ELBA/SUVr are pos/neg and neg/pos), the apparent trend is that subjects negative to the visual assessment are likely to be considered negative by ELBA (although not far from its cut-off) but are considered positive for SUVr. This suggests that whole-brain amyloid burden evaluation is more concordant with respect to a visual read than (small) ROI-based quantification on borderline cases. When keeping the blind visual assessment as reference (that is using only the *N* and *P* set, Table 3) this distinction is even more pronounced: 15 out of 19 cases agree with ELBA versus 4 out of 19 for SUVr.

Comparison with CSF Aβ_{42}

The comparison between ELBA score and CSF Aβ_{42} assay yielded a satisfactory agreement with 90.6% concordance, in the same range [26] or even a bit higher [23, 27] than those achieved in previous works using the SUVr approach. Discrepancies (7.5%) were mostly found in patients with abnormal Aβ_{42} levels and a normal ELBA score. Such a limited discrepancy may be explained with the notion that decreased Aβ_{42} reduction in CSF can precede brain amyloid deposition [28].

Clinical follow-up

Results shown in Fig. 10 suggest a cautionary attitude when using clinical evaluation (even at follow-up visits) as gold standard. For instance, the lack of amyloid markers, as evidenced by both CSF and PET analysis, is virtually incompatible with the diagnosis/conversion to AD. It has recently been shown that a small but non-trivial part of AD patients of the ADNI population shares a negative Florbetapir scan [26]. These Florbetapir negative patients would have a variety of clinical and neurodegeneration biomarker features distinct from Florbetapir positive patients, suggesting that one or more non-AD etiologies—such as cerebrovascular disease and SNAP—may be the main cause of their cognitive deficit, mimicking AD.

On the other hand, MCI who reverted to NS condition may still have a negative or borderline amyloid PET burden; among them an occasional patient may show abnormally low levels of CSF Aβ_{42} and might theoretically become Florbetapir positive in the future due to the earlier CSF positivity [28].

Longitudinal evaluation

The fitted model is qualitatively comparable to that shown by [20] albeit using a much shorter observation

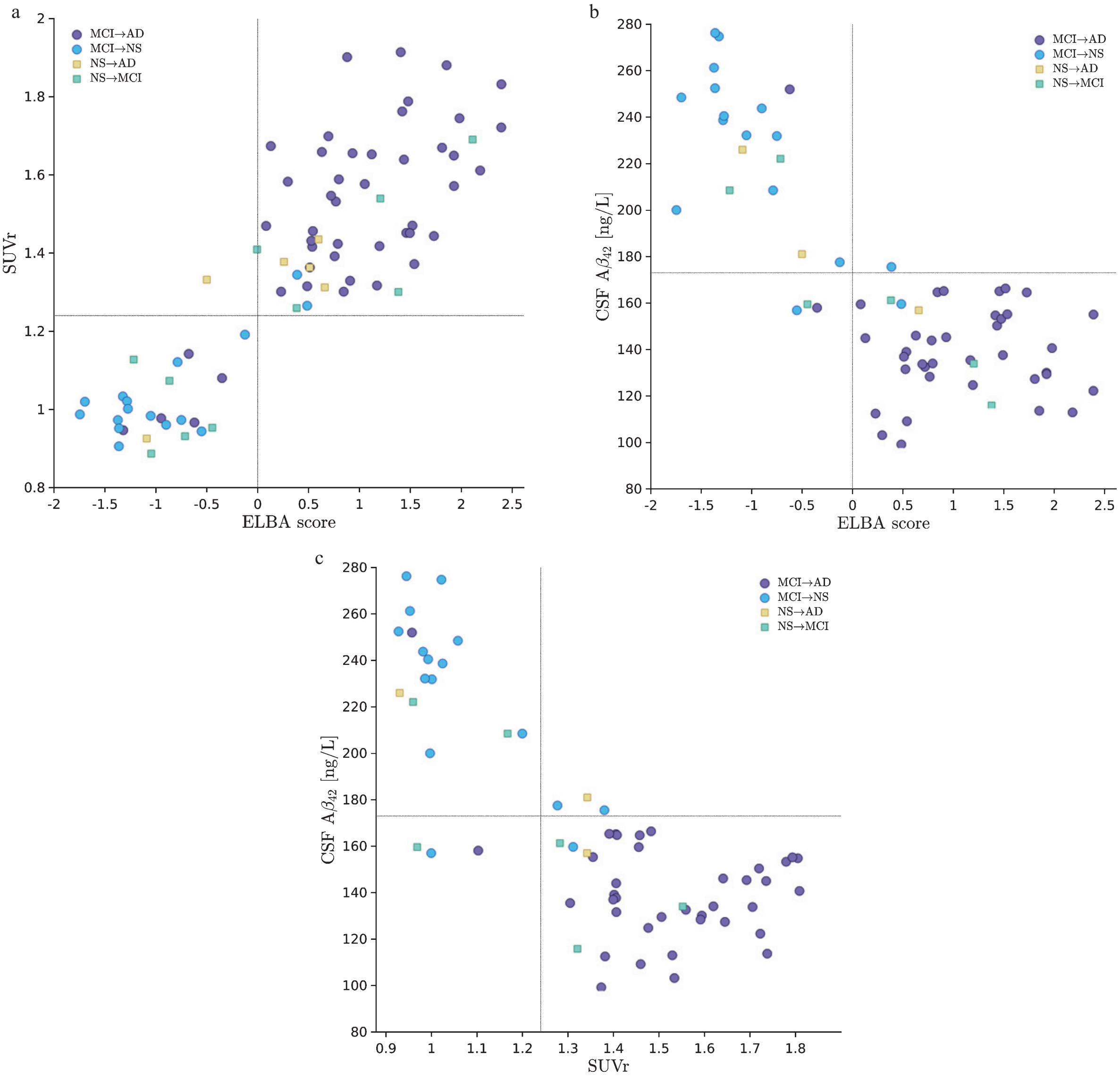

Fig. 10. Baseline CSF Aβ_{42}, ELBA, and SUVr scores grouped by clinical evaluation (baseline→follow-up) for subjects whose initial assessment changed at some later visit. Cut-offs are marked with dotted lines. In plot (a) the number of subjects is 78. In plot (b) and (c), the number of subjects is 63 out of 78, that are those for which CSF data were available too. NS, cognitively normal subjects; MCI, mild cognitive impairment; AD, Alzheimer's disease.

time (24 months), with a different ligand and on a larger pool of PET scanners. The shorter follow-up time could also explain why our findings on the SUVr longitudinal analysis exhibit a larger variability than that shown by [20]. The ASC shows a rather sparse distribution when computed on two scans only (with a wide range of positive and negative values), a behavior which is reduced in the 3 scans analysis. In all the 16 subjects with at least three repeated scans the ELBA score increased, showing the sensitivity of the method to amyloid deposition even in a relatively short time span (48 months) and opening potential applications to pharmacological studies with anti-amyloid compounds.

Even taking into account the relatively strong uncertainty due to the use of only two scans, ASC values peaked in the EMCI cohort, are mildly positive within normal subjects and are substantially compatible with zero in AD, a behavior which is expected according to the latest models of amyloid load [21].

The comparison between ELBA ASC and CSF Aβ_{42} points to a discrete, inverse relation between

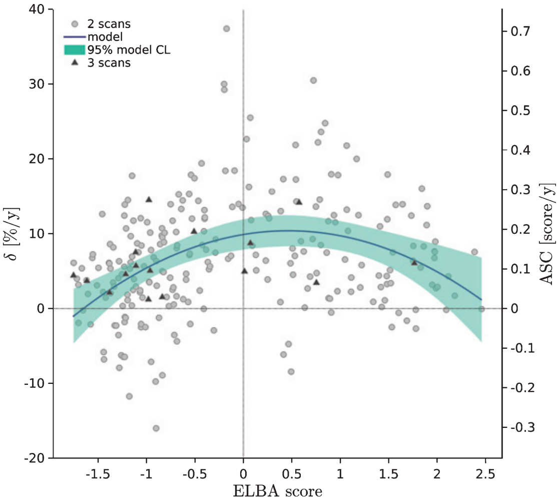

Fig. 11. Relative (δ, interquartile-range normalized) and absolute annualized score change (ASC) for ELBA versus baseline score. A quadratic model and the CL band on the model are superimposed. Subjects with three scans (triangles) are marked separately from subjects with only two scans (dots).

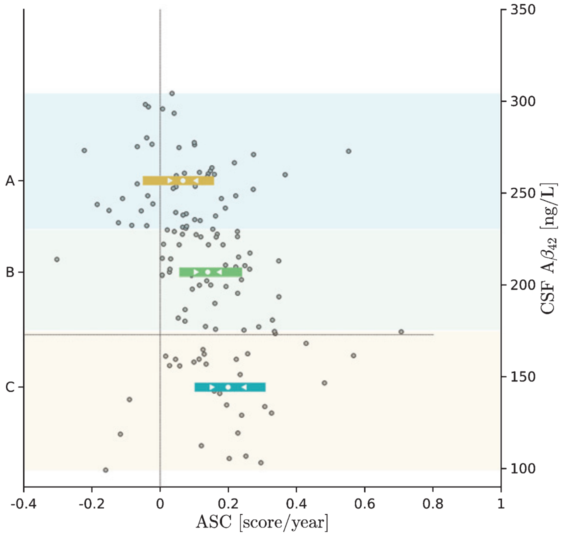

Fig. 12. CSF Aβ_{42} versus ELBA annualized score change (ASC) for 134 subjects whose baseline ELBA score was <0.5 (i.e., negative or borderline according to ELBA). Subjects are divided by CSF range (ng/L) into three groups: A ($230 < A\beta_{42}$), B ($174 < A\beta_{42} < 230$), and C ($A\beta_{42} < 174$). The corresponding ensemble statistics is summarized in the boxplot, where the box length spans the 25% to the 75% ASC percentile, the white dot and the white triangles are the mean and the 95% CL on the mean, respectively. Groups A/C and B/C are significantly different.

CSF Aβ_{42} level and variations in brain amyloid load, taking into account that they reflect indirect measures of a biologically complex phenomenon.

According to our results which focused on patients with a negative or borderline positive ELBA score at baseline, brain amyloid accumulation appears to be faster in those with a pathological low CSF Aβ_{42} levels, in line with current knowledge.

The relatively high scattering of ASC value of both ELBA and SUVr scores tells us that we are still rather far from being able to use differential measures at the single subject level. This limitation alone should sponsor new methodological approaches to semi-quantification on amyloid-PET images.

Compared to longitudinal SUVr values though, ELBA shows lesser variability. This is likely due to the fact that reference (and uptake) region is not needed, the selection of which has recently been shown to impair the reproducibility and accuracy of longitudinal SUVr measurements [29].

In addition, the average ASC values found in our work are comparable to those proposed by [21] using SUVr measurements, suggesting that the important clinical and pharmacological implications of an accurate longitudinal evaluation at the cohort level are within our reach, particularly if the protocol involves three or more PET scans.

Image quality issues

A non-negligible fraction of scans (60 out of 504) were flagged as “sub-optimal quality” by at least one reader, and 37 were flagged so by all readers (a sample image is provided in Supplementary Fig. 3). This flag did not imply the impossibility of visual assessment; it meant though that—within the boundary of the readers clinical experience—they deemed that their evaluation was made more difficult by the image quality.

Indeed, the apparent higher performance of the intensity feature C versus the geometric feature G is actually incidental for this study, as scans provenance and quality are so heterogeneous. In another study (unpublished) where data came from a single center, we observed no significant performance difference between the intensity and geometric feature.

The number of flagged scans is rather relevant, given the cost of the ligand and radiation exposure to the patient. The peculiar relevance of image quality and resolution in these investigations is most notable when dealing with difficult cases, as it is with those subjects affected by significant gray matter atrophy. In addition, the relationship between flagged scans and the blind phase *uncertain* label suggest that diagnostic errors are more likely to be made on amyloid-negative subjects.

This highlights the need of procedural guidelines which underline the importance of spatial resolution, partial volume effect correction, and the general minimization of acquisition nuisances; all to be addressed in the official nuclear medicine societies guidelines on acquisition protocols.

The general call to quality and consistency in nuclear medicine imaging peaks in longitudinal studies. Considering the absolute value of the residuals $|\epsilon|$ on the 16 subjects with three scans and dividing them into two batches: the first one consisting of those subjects acquired on the same site and scanner ($n_{same} = 11$) and a second one consisting of those who had at least one scan taken with a different system ($n_{diff} = 5$) – we find $|\epsilon_{same}| = 0.04 \pm 0.02$ and $|\epsilon_{diff}| = 0.14 \pm 0.07$. Although the modest number of samples excludes a definite statement, the influence of scan consistency (both within and among different sites) should be at least considered as a nuisance in the rather large spread found in the evaluation of the ASC. For the same reason, we suggest a minimum of three scans to attempt a reasonable estimation of the annualized score change, unless image quality and consistency can be guaranteed (and possibly quantified).

Study limitations

An unavoidable limitation of this study is the lack of a true gold standard (i.e., neuritic plaques at autopsy) that can be used to evaluate the accuracy of the imaging quantitation, and to set an absolute threshold for positivity. Unfortunately, there are relatively few examples of such cases in literature (see for instance [30] or [31]). As practical solution, also shown by [3], we used the consensus visual read as the reference standard, together with the CSF Aβ_{42} and SUVr measurements. The cross-comparison of these should provide the ground for a reasonable method validation targeted at the clinical practice.

Because of the peculiar image treatment which evaluates the intensity distribution patterns, the main drawback for the ELBA analysis is the image quality and consistency. Although convincing evidence can be drawn even from multi-centric and blind analysis such as this one, the weight of the acquisition-related variables can be significant. To correctly estimate this effect we are planning a more detailed analysis on images coming from a single center but with different acquisition protocols and image reconstruction parameters.

Another open area of investigation is the brain ROI specialization. ELBA was developed in order to overcome the weaknesses inherent to the SUVr computation, and for this reason the method is not suitable to be applied on small ROIs such as those used in SUVr analysis. Still, ELBA could be specialized on brain macro regions such as the frontal or parietal lobes, with potential benefits for the human reader in the process of clinical assessment.

Finally, the proposed method mimics the visual process in that it captures global geometrical and intensity features, and it is therefore reasonably correlated to the reader's assessment once image quality issues are solved. Nevertheless its usefulness in a clinical setting is likely to be most informative when combined with independent measures such as SUVr and CSF analyses.

ACKNOWLEDGMENTS

This research was supported by Istituto Nazionale di Fisica Nucleare (INFN), Italy. This research was also directly supported by grants to FS and to LR from INFN.

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

The proposed method (ELBA) is protected under patent n. WO2016016833 (A1), WO2015IB55758 20150730.

Authors' disclosures available online (http://j-alz.com/manuscript-disclosures/16-0232r1).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-160232.

REFERENCES

[1] Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R, Cappa S, Crutch S, Engelborghs S, Frisoni GB, Fox NC, Galasko D, Habert MO, Jicha GA, Nordberg A, Pasquier F, Rabinovici G, Robert P, Rowe C, Salloway S, Sarazin M, Epelbaum S, de Souza LC, Vellas B, Visser PJ, Schneider L, Stern Y, Scheltens P, Cummings JL (2014) Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *Lancet Neurol* **13**, 614-629. Erratum in: *Lancet Neurol* **13**, 757, 2014.

[2] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) 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* **7**, 263-269.

[3] Johnson Ka, Minoshima S, Bohnen NI, Donohoe KJ, Foster NL, Herscovitch P, Karlawish JH, Rowe CC, Carrillo MC, Hartley DM, Hedrick S, Pappas V, Thies WH (2013) 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. *Alzheimers Dement* **9**, e–1-16.

[4] Fagan AM, Mintun MA, Mach RH, Lee SY, Dence CS, Shah AR, LaRossa GN, Spinner ML, Klunk WE, Mathis CA, DeKosky ST, Morris JC, Holtzman DM (2006) Inverse relation between *in vivo* amyloid imaging load and cerebrospinal fluid Abeta42 in humans. *Ann Neurol* **59**, 512-519.

[5] Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, Bergström M, Savitcheva I, Huang Gf, Estrada S, Ausén B, Debnath ML, Barletta J, Price JC, Sandell J, Lopresti BJ, Wall A, Koivisto P, Antoni G, Mathis CA, Långström B (2004) Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Ann Neurol* **55**, 306-319.

[6] Rowe CC, Ng S, Ackermann U, Gong SJ, Pike K, Savage G, Cowie TF, Dickinson KL, Maruff P, Darby D, Smith C, Woodward M, Merory J, Tochon-Danguy H, O'Keefe G, Klunk WE, Mathis CA, Price JC, Masters CL, Villemagne VL (2007) Imaging beta-amyloid burden in aging and dementia. *Neurology* **68**, 1718-1725.

[7] Kinahan PE, Fletcher JW (2010) Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. *Semin Ultrasound CT MR* **31**, 496-505.

[8] Jagust WJ, Bandy D, Chen K, Foster NL, Landau SM, Mathis CA, Price JC, Reiman EM, Skovronsky D, Koeppe RA (2010) The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. *Alzheimers Dement* **6**, 221-229.

[9] Aisen PS, Petersen RC, Donohue MC, Gamst A, Raman R, Thomas RG, Walter S, Trojanowski JQ, Shaw LM, Beckett LA, Jack CR, Jagust W, Toga AW, Saykin AJ, Morris JC, Green RC, Weiner MW (2010) Clinical core of the Alzheimer's Disease Neuroimaging Initiative: Progress and plans. *Alzheimers Dement* **6**, 239-246.

[10] Thirion JP (1998) Image matching as a diffusion process: An analogy with Maxwell's demons. *Med Image Anal* **2**, 243-260.

[11] MacQueen J (1967) Some methods for classification and analysis. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; vol. 233. The Regents of the University of California, pp. 281-297.

[12] Seber GAF (1984) *Multivariate Observations*. Wiley Series in Probability and Statistics; John Wiley & Sons, Inc., Hoboken, NJ, USA.

[13] Klunk WE, Koeppe RA, Price JC, Benzinger TL, Devous MD, Jagust WJ, Johnson KA, Mathis CA, Minhas D, Pontecorvo MJ, Rowe CC, Skovronsky DM, Mintun MA (2015) The Centiloid Project: Standardizing quantitative amyloid plaque estimation by PET. *Alzheimers Dement* **11**, 1-15.e4.

[14] Hutton C, Declerck J, Mintun MA, Michael J, Joshi A (2013) SPAP and Avid florbetapir Analysis Methods. http://adni.bitbucket.org/docs/SPAP_AVID_FLORBETAPIR/sPAP_Avid_Florbetapir_Analysis_Methods.pdf

[15] Mattsson N, Insel PS, Landau S, Jagust W, Donohue M, Shaw LM, Trojanowski JQ, Zetterberg H, Blennow K, Weiner M (2014) Diagnostic accuracy of CSF Aβ_{42} and florbetapir PET for Alzheimer's disease. *Ann Clin Transl Neurol* **1**, 534-543.

[16] Joshi AD, Pontecorvo MJ, Clark CM, Carpenter AP, Jennings DL, Sadowsky CH, Adler LP, Kovnat KD, Seibyl JP, Arora A, Saha K, Burns JD, Lowrey MJ, Mintun Ma, Skovronsky DM (2012) Performance characteristics of amyloid PET with florbetapir F 18 in patients with Alzheimer's disease and cognitively normal subjects. *J Nucl Med* **53**, 378-384.

[17] Vandenberghe R, Adamczuk K, Dupont P, Laere KV, Chételat G (2013) Amyloid PET in clinical practice: Its place in the multidimensional space of Alzheimer's disease. *Neuroimage Clin* **2**, 497-511.

[18] Klinger RY, James OG, Wong TZ, Newman MF, Doraiswamy PM, Mathew JP (2013) Cortical β-amyloid

levels and neurocognitive performance after cardiac surgery. *BMJ Open* **3**, e003669.

[19] Lopresti BJ, Klunk WE, Mathis CA, Hoge JA, Ziolko SK, Lu X, Meltzer CC, Schimmel K, Tsopelas ND, DeKosky ST, Price JC (2005) Simplified quantification of Pittsburgh Compound B amyloid imaging PET studies: A comparative analysis. *J Nucl Med* **46**, 1959-1972.

[20] Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, Szoeke C, Macaulay SL, Martins R, Maruff P, Ames D, Rowe CC, Masters CL (2013) Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: A prospective cohort study. *Lancet Neurol* **12**, 357-367.

[21] Jack CR, Wiste HJ, Lesnick TG, Weigand SD, Knopman DS, Vemuri P, Pankratz VS, Senjem ML, Gunter JL, Mielke MM, Lowe VJ, Boeve BF, Petersen RC (2013) Brain β-amyloid load approaches a plateau. *Neurology* **80**, 890-896.

[22] Johnson KA, Sperling RA, Gidicsin CM, Carmasin JS, Maye JE, Coleman RE, Reiman EM, Sabbagh MN, Sadowsky CH, Fleisher AS, Murali Doraiswamy P, Carpenter AP, Clark CM, Joshi AD, Lu M, Grundman M, Mintun Ma, Pontecorvo MJ, Skovronsky DM (2013) Florbetapir (F18-AV-45) PET to assess amyloid burden in Alzheimer's disease dementia, mild cognitive impairment, and normal aging. *Alzheimers Dement* **9**, S72-S83.

[23] Landau SM, Lu M, Joshi AD, Pontecorvo M, Mintun MA, Trojanowski JQ, Shaw LM, Jagust WJ (2013) Comparing positron emission tomography imaging and cerebrospinal fluid measurements of β-amyloid. *Ann Neurol* **74**, 826-836.

[24] Jack CR (2014) PART and SNAP. *Acta Neuropathol* **128**, 773-776.

[25] Camus V, Payoux P, Barré L, Desgranges B, Voisin T, Tauber C, La Joie R, Tafani M, Hommet C, Chételat G, Mondon K, De La Sayette V, Cottier JP, Beaufils E, Ribeiro MJ, Gissot V, Vierron E, Vercouillie J, Vellas B, Eustache F, Guilloteau D (2012) Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *Eur J Nucl Med Mol Imaging* **39**, 621-631.

[26] Landau SM, Horng A, Fero A, Jagust WJ (2016) Amyloid negativity in patients with clinically diagnosed Alzheimer disease and MCI. *Neurology* **86**, 1377-1385.

[27] Zwan MD, Rinne JO, Hasselbalch SG, Nordberg A, Lleó A, Herukka SK, Soininen H, Law I, Bahl JM, Carter SF, Fortea J, Blesa R, Teunissen CE, Bouwman FH, van Berckel BN, Visser PJ (2016) Use of amyloid-PET to determine cutpoints for CSF markers. *Neurology* **86**, 50-58.

[28] Palmqvist S, Mattsson N, Hansson O (2016) Cerebrospinal fluid analysis detects cerebral amyloid-β accumulation earlier than positron emission tomography. *Brain* **139**, 1226-1236.

[29] Landau SM, Fero A, Baker SL, Koeppe R, Mintun M, Chen K, Reiman EM, Jagust WJ (2015) Measurement of longitudinal beta-amyloid change with 18F-Florbetapir PET and standardized uptake value ratios. *J Nucl Med* **56**, 567-574.

[30] Clark CM, Pontecorvo MJ, Beach TG, Bedell BJ, Coleman RE, Doraiswamy PM, Fleisher AS, Reiman EM, Sabbagh MN, Sadowsky CH, Schneider Ja, Arora A, Carpenter AP, Flitter ML, Joshi AD, Krautkramer MJ, Lu M, Mintun Ma, Skovronsky DM (2012) Cerebral PET with florbetapir compared with neuropathology at autopsy for detection of neuritic amyloid-β plaques: A prospective cohort study. *Lancet Neurol* **11**, 669-678.

[31] Sabri O, Sabbagh MN, Seibyl J, Barthel H, Akatsu H, Ouchi Y, Senda K, Murayama S, Ishii K, Takao M, Beach TG, Rowe CC, Leverenz JB, Ghetti B, Ironside JW, Catafau AM, Stephens AW, Mueller A, Koglin N, Hoffmann A, Roth K, Reininger C, Schulz-Schaeffer WJ; Florbetaben Phase 3 Study Group (2015) Florbetaben PET imaging to detect amyloid beta plaques in Alzheimer's disease: Phase 3 study. *Alzheimers Dement* **11**, 964-974.