Brain ^{18}F-FDG PET analysis via interval-valued reconstruction: proof of concept for Alzheimer's disease diagnosis

Florentin Kucharczak, Marie Suau, Olivier Strauss, Fayçal Ben Bouallègue & Denis Mariano-Goulart

Springer

Springer

ORIGINAL ARTICLE

Brain ^{18}F-FDG PET analysis via interval-valued reconstruction: proof of concept for Alzheimer's disease diagnosis

Florentin Kucharczak[1,2] · **Marie Suau**[1] · **Olivier Strauss**[2] · **Fayçal Ben Bouallègue**[1,3] · **Denis Mariano-Goulart**[1,3]

Abstract

Objective We propose an innovative approach for ^{18}F-FDG PET analysis based on an interval-valued reconstruction of ^{18}F-FDG brain distribution. Its diagnostic performance for Alzheimer's disease (AD) diagnosis with comparison to a validated post-processing software was assessed.

Methods Brain ^{18}F-FDG PET data from 26 subjects were acquired in a clinical routine setting. Raw data were reconstructed using an interval-valued version of the ML–EM algorithm called NIBEM that stands for Non-Additive interval-based expectation maximization. Subject classification was obtained via interval-based statistical comparison (intersection ratio, IR) between cortical regions of interest (ROI) including parietal, temporal, and temporo-mesial cortices and a reference region, the sub-cortical grey nuclei, known not to be affected by AD. In parallel, PET images were post-processed using a validated automated software based on the computation of ROI normalized uptake ratios standard deviation (SUVr SD) with reference to a healthy control database (Siemens Scenium). Clinical diagnosis made during follow-up was considered as the gold-standard for patient classification (16 healthy controls and 10 AD patients).

Results Both methods provided cortical ROI indices that were significantly different between controls and AD patients. The area under the ROC curve for control/AD classification was statistically identical (0.96 for NIBEM IR and 0.95 for Scenium SUVr SD). At the optimal threshold, the sensitivity, specificity, accuracy, positive predictive value, and negative predictive value were, respectively, 100%, 88%, 92%, 83%, and 100% for both Scenium SUVr SD and NIBEM IR methods.

Conclusion This preliminary study shows that interval-valued reconstruction allows self-consistent analysis of brain ^{18}F-FDG PET data, yielding diagnostic performances that seem promising with respect to those of a commercial post-processing software based on SUVr SD analysis.

Keywords Brain ^{18}F-FDG PET · Alzheimer's disease · Interval-valued reconstruction · Assisted diagnosis

Introduction

As the world population ages, early diagnosis of neurodegenerative dementia represents a challenge for both society as a whole and the medical community. It is estimated that 46.8 million people worldwide are living with dementia in 2015. This number will almost double every 20 years, reaching 74.7 million in 2030 and 131.5 million in 2050 [1]. Alzheimer's disease (AD) is considered as the main etiology of neurodegenerative dementia (approximately 2/3 of cases have AD). AD impairs patient cognitive functions, and impacts daily life autonomy, and as such raises major public health policy issues in terms of home care and institutional placement [2]. Early diagnosis is a key feature in current patient management since treatments become less effective with worsening patient state.

Considerable effort is ongoing to identify and develop reliable biomarkers of incipient AD [3, 4] and other types of dementia [5] to target individuals who would most benefit from early therapeutic intervention [6]. Among these biomarkers, ^{18}F-FDG positron emission tomography (PET) has been recognized as a valuable

✉ Florentin Kucharczak
  florentin.kucharczak@umontpellier.fr

1 Département de médecine nucléaire, Montpellier University Hospital, CHU Lapeyronie, 191 av. du Doyen Giraud, 34295 Montpellier cedex 5, France

2 LIRMM, CNRS, Univ. Montpellier, 860 rue de Saint Priest, 34095 Montpellier, France

3 PhyMedExp, INSERM, CNRS, Univ. Montpellier, CHU Arnaud de Villeneuve, 34295 Montpellier, France

Table 1 Characteristics of the study groups

	Healthy controls ($N = 16$)	Alzheimer's disease ($N = 10$)
Male	9 (56%)	3 (33%)
Age (years)	61 ± 13 [36–77]	76 ± 6 [64–84][a]
Follow-up (days)	158 ± 162 [0–590]	320 ± 204 [101–606]
Injected activity (MBq)	147 ± 17 [121–173]	141 ± 18 [121–166]
Plasma glucose level (g/L)	1.06 ± 0.26 [0.71–1.62]	0.96 ± 0.07 [0.85–1.10]
Injection-acquisition delay (min)	30 ± 3 [20–35]	32 ± 3 [30–36]

Categorical variables are given as number (percentage). Continuous variables are given as mean ± standard deviation [range].

[a]Significantly different from healthy controls

imaging modality for the positive and differential diagnosis of neurodegenerative dementias [7–10]. Since absolute quantification of cortical glucose metabolism requires dynamic acquisitions [11, 12] that rarely fit with clinical constraints, visual interpretation is usually performed using a static scan [13]. ^{18}F-FDG uptake is a sensitive marker of synaptic dysfunction [14], and as such it has been shown to have good sensitivity in the detection of early brain dysfunction [15, 16] and follow-up of disease evolution over time [12]. Nevertheless, visual rating of relative cortical ^{18}F-FDG distribution yields undesirable inter-reader variability [11, 17] and sub-optimal specificity especially among moderately skilled readers [18, 19].

To overcome these limitations of visual interpretation, several automated semi-quantitative techniques have been developed. Although complex computational approaches based either on deep learning [20], neural networks [21], or Support Vector Machine (SVM) [22] showing promising classification capabilities have recently been reported, most semi-quantitative methods rely on the statistical mapping of voxel-based normalized cortical uptake ratio with respect to some reference region [23–27].

The aim of the present study is to propose and highlight the potential and the relevance of brain ^{18}F-FDG PET analysis based on an interval-valued reconstruction of ^{18}F-FDG distribution [28, 29]. The proposed framework allows for a direct estimation of voxel-wise confidence intervals accounting for the statistical variability of voxel values. Subject classification was obtained via interval-based statistical comparison between cortical regions of interest and a reference region (extracted from the same reconstruction), known to be spared by AD. For the analysis, only raw data from current PET acquisition were required. Diagnostic performances were compared to those of PET data post-processing using a validated automated software based on the computation of regional normalized uptake ratios SD with reference to a healthy control database [30].

Materials and methods

Patient characteristics

Twenty-six patients were prospectively recruited from the outpatients of the Nuclear Medicine Department at Montpelier University Hospital from November 2016 to July 2017. All included patients were referred for brain ^{18}F-FDG PET in routine conditions for the exploration of memory or executive dysfunction. Clinical diagnosis made during follow-up was considered as the gold-standard for patient classification.

The AD group (10 patients, 3 men and 7 women, 76 ± 6 years, range 64–84 years) was composed of patients diagnosed with probable AD according to the NINCDS-ADRDA criteria [31] during a clinical follow-up of 320 ± 204 days (range 101–606).

The control group (16 patients, 9 men and 7 women, 61 ± 13 years, range 36–77 years) was composed of patients with null or low pre-test probability of neurodegenerative disease, for whom 18F-FDG PET data were within the range of normal, and/ or rated as normal during a clinical follow-up of 158 ± 162 days (range 0–590).

PET data acquisition and reconstruction

PET examinations were performed on a Siemens Biograph mCT 20 Flow PET-CT system (Siemens Medical Solutions Knoxville, USA) about 30 min after intravenous (IV) injection of 2.5 MBq/kg of ^{18}F-FDG. Data were acquired in three-dimensional (3D) time-of-flight mode for 10 min. Emission data were corrected for attenuation using the embedded computerized tomography (CT) scanner. Random coincidences (where the two photons did not arise from the same annihilation event), scatter coincidences, and dead-time were also corrected using dedicated manufacturer tools. Patient characteristics and technical data are
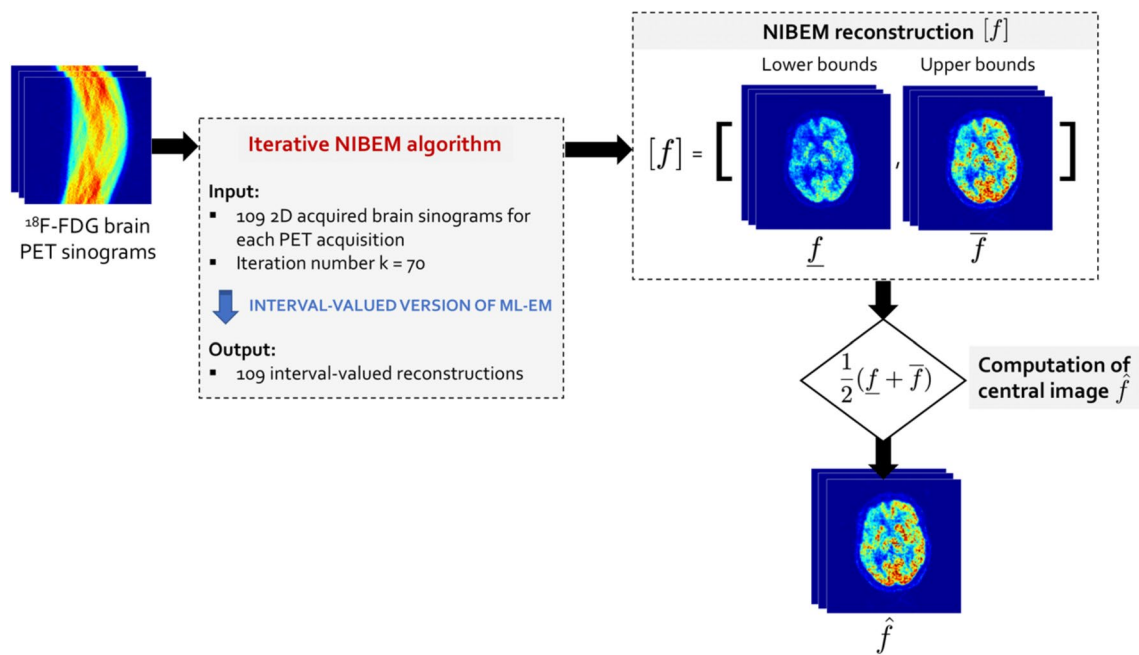
Fig. 1 PET data reconstructed with interval-valued algorithm NIBEM

summarized in Table 1. For the needs of the comparison experiments, PET data were reconstructed with two different reconstruction methods.

First, PET images were obtained using the routine workflow implemented at our institution for brain PET imaging. Iterative reconstruction was performed using 3D OSEM (21 subsets, 8 iterations) including PSF correction, followed by a 5, 5, 5 mm FWHM Gaussian post-smoothing procedure. Images were sampled on a $400 \times 400 \times 109$ grid with a voxel size of $2.04 \times 2.04 \times 2.03$ mm^3.

Second, PET data were reconstructed using an interval-valued extension of the maximum likelihood-expectation maximization (ML–EM) algorithm [32, 33] called NIBEM [28, 29] that stands for Non-Additive interval-based expectation maximization. The main motivation for using this algorithm resides in its ability to directly reconstruct voxel-wise confidence intervals. The considered confidence intervals account for the statistical variability affecting reconstructed voxel values. The confidence level associated with these intervals was shown to be about 90% (29). As the current version of the mentioned algorithm was only described in 2D, 3D emission data were rebinned into a stack of 109 two-dimensional (2D) sinograms using the Fourier rebinning (FORE) algorithm [34]. Then the 2D sinograms were reconstructed using NIBEM on a 200×200 grid with a voxel size of $4.1 \times 4.1 \times 2.03$ mm^3. Reconstructions were performed using 70 iterations, which makes it possible to reach images with a similar noise level as obtained with 3D OSEM algorithm described in the routine workflow. Time-of-flight information was not exploited in this preliminary work.

Instead of computing scalar values for each voxel, the algorithm used was interval-valued (i.e. the algorithm produces interval values for each voxel), then the measure associated with each voxel i was a real interval denoted $[f_i]$, which lower and upper bounds are respectively denoted $\underline{f_i}$ and $\overline{f_i}$ Here, $[f_i]$ represents the set of real numbers lying between its respective lower et upper bounds $\underline{f_i}$ and $\overline{f_i}$. The central image $\hat{f}$ (i.e. the image that minimizes the Hausdorff distance between $\underline{f}$ and $\overline{f}$) was defined, for each voxel i, as $\hat{f_i} = \left(\underline{f_i} + \overline{f_i}\right)\big/2$. In other words, for each voxel i of the reconstructed volume, the value $\hat{f_i}$ is the center of the interval $[f_i]$ reconstructed with NIBEM. A graphical illustration of the reconstruction procedure is presented in Fig. 1.

Patient classification using SUV ratio

The first set of images (3D OSEM) was post-processed using the manufacturer's tool dedicated to statistical analyses of brain PET data (Siemens Scenium) [30]. PET images were automatically registered to the Montreal Neurological Institute (MNI) space, then segmented into cortical regions of interest (ROI) including whole brain, parietal, temporal, and temporo-mesial cortices. Mean standardized uptake values (SUV) were measured in each cortical ROI and normalized to mean whole brain SUV to produce cortical SUV ratios (SUVr). Last, these cortical SUVr were converted to standard deviation scores (SUVr SD), commonly called z score, based on reference distributions from age-matched control populations. As the ROI affected by AD appears

Fig. 2 PET data spatial registration procedure
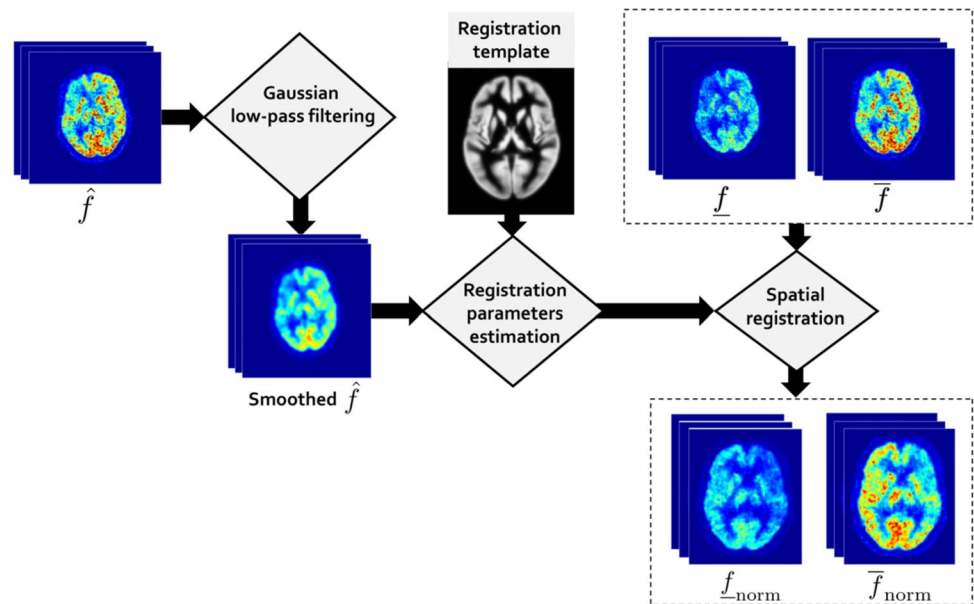
hypometabolic, ROI SUVr SD of AD patients are usually negative. For example, a SUVr SD that equals -1 means that the considered SUVr is 1 standard deviation below the mean SUVr of the control group. For a given classification threshold, a subject was considered as "not AD" if all the SUVr SD computed in 6 reference ROIs (parietal left/right, temporal left/right, and temporo-mesial left/right) were above the threshold. If at least one SUVr SD was below the threshold, the subject was classified as "AD".

Patient classification using NIBEM

As Scenium is a commercial software, registration procedure and reference database are not accessible. We then designed our own procedure to compare our method. For each subject, the NIBEM central image $\hat{f}$ was smoothed with a Gaussian kernel of 5, 5, 5 mm FWHM, then spatially normalized to the MNI space using the widely known tissue probability map template provided by SPM12 (Wellcome Trust Centre, London, UK). Reconstructed upper and lower NIBEM PET volumes were then spatially normalized using the estimated transformation parameters computed over $\hat{f}$. As upper, lower, and central volumes are defined in the same space, normalizing the data with the same transformation parameters allows minimization of the spatial registration bias between upper and lower registered volumes. Resulting PET images were sampled on a $91 \times 109 \times 91$ grid with cubic voxels. Image voxels were labeled according to the AAL2[1] anatomical atlas [35]. A graphical illustration of the proposed spatial registration procedure is shown in Fig. 2.

Spatially normalized NIBEM images were then segmented based on voxel labels to define cortical ROIs including left (L) and right (R) parietal, temporal and temporo-mesial lobes.

The clinical diagnosis of AD in ^{18}F-FDG PET is made by highlighting the hypo-metabolism of some specific ROIs that are affected by the disease with respect to reference regions that are known not to be affected by AD. Sub-cortical grey nuclei can be considered as the reference region for such a comparison [36].

Thus, as NIBEM reconstructions are interval-valued, the interval-valued distribution associated with each relevant ROI (L and R parietal, temporal, and temporo-mesial lobes) needs to be compared to the interval-valued distribution associated with the reference ROI (sub-cortical grey nuclei).

To our knowledge, the literature concerning interval-valued distribution comparison remains rather scarce. Although an interval-valued generalization of the Wilcoxon rank-sum test was proposed [37], it appears to be insufficiently specific for the statistical comparison needed to be performed in this present work. Indeed, robust tests like Wilcoxon's test are affected by the imprecision of the data: the more imprecise the data the more difficult it is for the statistical tests to reject the null hypothesis. In the presence of a considerable imprecision (which is the case in the present application), the Wilcoxon test is too weakly specific to reject such a hypothesis.

Therefore, we propose to compare interval-valued distributions using a consistency measure between the two interval-valued distributions: $\left[f_{\text{TEST}}^{1}\right], \ldots, \left[f_{\text{TEST}}^{N}\right]$ for the tested ROI and $\left[f_{\text{REF}}^{1}\right], \ldots, \left[f_{\text{REF}}^{M}\right]$ for the reference region. These two distributions can be considered as being two random empirical distributions of intervals. Note that the probability

---

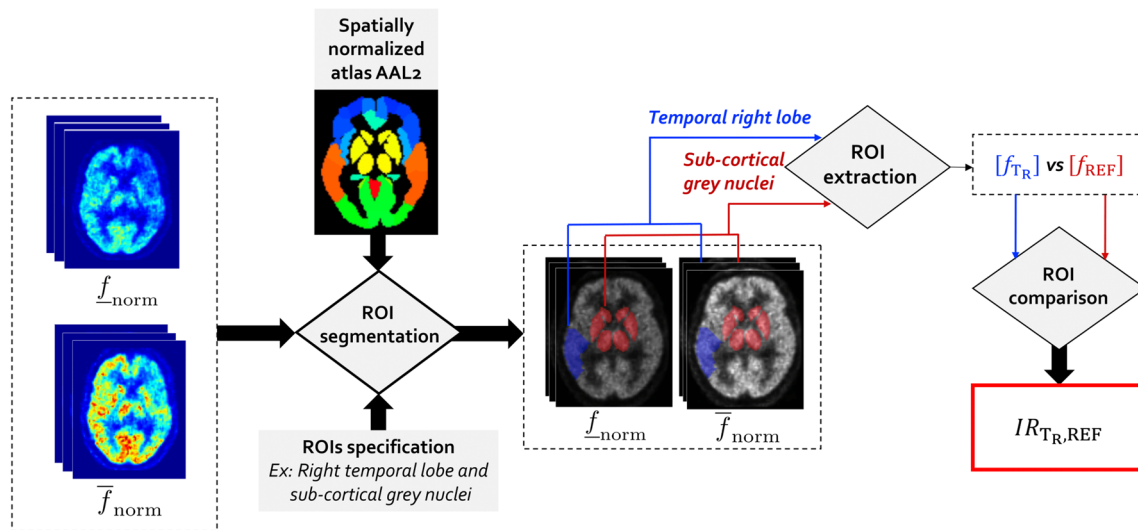[1] Available at: https://www.gin.cnrs.fr/AAL2.

Fig. 3 PET data segmentation and ROI comparison step

density associated with each interval-valued reconstruction is unknown.

Two intervals $\left[f_{\text{TEST}}^i\right]$ and $\left[f_{\text{REF}}^j\right]$ are said to be inconsistent if $\left[f_{\text{TEST}}^i\right] \cap \left[f_{\text{REF}}^j\right] = \emptyset$ and consistent if $\left[f_{\text{TEST}}^i\right] \cap \left[f_{\text{REF}}^j\right] \neq \emptyset$ The consistency function $g\left(\left[f_{\text{TEST}}^i\right], \left[f_{\text{REF}}^j\right]\right)$ is equal to 1 if both intervals are consistent, 0 if they are inconsistent. If we consider the pair $\left(\left[f_{\text{TEST}}^i\right], \left[f_{\text{REF}}^j\right]\right)$, assume these two intervals independent and since there are $N.M$ possible comparisons, the weight $1/(N.M)$ can be associated with the pair. Under this assumption, in the theory of belief functions [38], the more common way to calculate a consistency measure between the two interval distributions is to consider the cumulative mass on the pairs that are consistent. The consistency measurement denoted IR between the region to test and the reference region is thus defined by the equation:

$$\text{IR}_{\text{TEST,REF}} = \frac{1}{N.M} \sum_{i=1}^{N} \sum_{j=1}^{M} g\left(\left[f_{\text{TEST}}^i\right], \left[f_{\text{REF}}^j\right]\right), \quad (1)$$

where $\text{IR}_{\text{TEST,REF}} = 1$ means that all intervals of the tested region intersect the intervals of the reference region. On the opposite, $\text{IR}_{\text{TEST,REF}} = 0$ means that none of the intervals intersect between the two considered regions. Varying a given classification threshold from 1 to 0, a subject was considered as "not AD" if all the concordance measures IR computed independently between the 6 ROIs (left/ right temporal, left/right temporo-mesial and left/ right parietal lobes) and the reference (sub-cortical grey nuclei) regions were above the threshold. If at least one intersection ratio was below the threshold, the subject was classified as "AD". ROI segmentation and comparison processes are graphically

detailed in Fig. 3. Illustration of both post-processing and proposed methods outputs for an AD patient extracted from the database can be found in Fig. 4. For each method, one score by ROI is extracted from the analysis, albeit with different meanings: deviation scores for SUVr SD analysis and consistency measure between the analyzed ROI and a reference ROI for NIBEM IR analysis. An overview of the reconstructions for both compared methods is presented in Fig. 5.

Statistical analysis

Differences between the control group and AD group in the distribution of SUVr SD and NIBEM IR computed on an ROI basis were assessed using Mann–Whitney's test.

The ability of the two methods to discriminate between controls and AD subjects was evaluated using a receiver operating characteristic (ROC) analysis by varying the classification thresholds (from − 10 to 10 with a step. ROC analysis is performed with pre-defined criteria that, respectively, state false-positive (FP), false-negative (FN), true-positive (TP), and true-negative (TN) definitions:

- FP: a patient classified as "AD" whilst labeled "control"
- FN: a patient classified as "not AD" whilst labeled "AD"
- TP: a patient classified as "AD" whilst labeled "AD"
- TN: a patient classified as "not AD" whilst labeled "control"

For each threshold, FP, FN, TP, and TN were computed. Then always for each threshold, (sensitivity) vs (1− specificity) was computed and plotted to obtain the corresponding ROC curve. Areas under the ROC curve (ROC AUC) were

Fig. 4 Illustration of brain ^{18}F-FDG PET analysis for a patient diagnosed with Alzheimer's disease with both compared methods
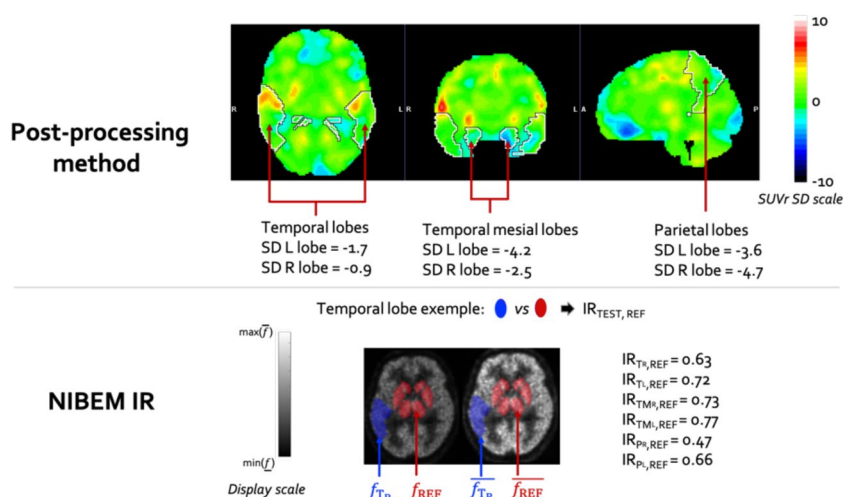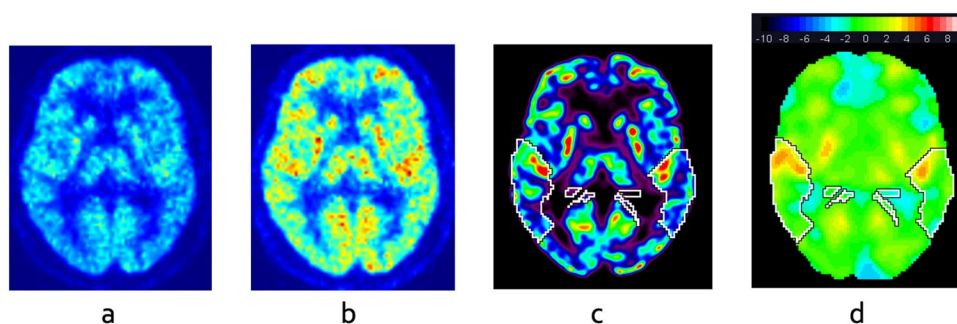

Fig. 5: ^{18}F-FDG PET brain axial slice of an AD patient. **a** NIBEM lower bound, **b** NIBEM upper bound, **c** OSEM 3D reconstruction used in SUVr SD method, **d** SUVr SD map with its color bar. Same color bar was used for (**a**) and (**b**). All volumes were registered to template space

compared [39] and the optimal threshold was defined as that maximizing Youden's index (sensitivity + specificity− 1).

Diagnostic performances of the two methods were compared using Fisher's exact test. A *p* value ≤ 0.05 was considered as significant. All statistical computations were performed using Matlab (The MathWorks, Inc).

Results

Distributions of SUVr SD by cortical ROI in controls and AD subjects are presented in Fig. 6a. It shows that distributions were significantly different in all tested cortical ROIs. It thus highlights that SUVr SD analysis is effective in classifying "AD" and "not AD" groups of the study according to the analyzed ROIs.

The same experiment carried out with cortical ROI NIBEM IR distributions in controls and AD subjects is presented in Fig. 6b. Group distributions were shown to also be significantly different between controls and AD patients for all the comparisons ROI vs reference except for the left temporo-mesial lobe. However, the classification (see Fig. 7; Table 2) is not impacted by the lack of

IR statistical difference between the "AD" and "controls" groups in this specific ROI. It thus suggests that considering these groups and this comparison framework, left temporo-mesial lobe ROI does not play a big role in the classification framework in this specific study.

Figure 7 shows the ROC curves obtained by varying the classification threshold of each method. There was no significant difference between the ROC AUCs of the two methods (0.95 vs 0.96) for SUVr SD and NIBEM IR (respectively, *p* = 0.86). Both AUC values are excellent, and highlight the good classification performance of both frameworks on the considered dataset.

In the present experiment, the optimal classification thresholds that maximize the Youden's index were 0.614 for NIBEM IR and − 2.1 for Scenium SUVr SD. These threshold values are those used for analyzing and comparing the two methods. In Table 2, where the diagnostic performance of both methods in differentiating healthy controls from AD patients is summarized, we can observe that statistical outcomes (sensitivity, specificity, accuracy, positive and negative predictive value) are equal for both methods considering the formerly specified thresholds.

Fig. 6 Box plots of (**a**) SUVr SD distributions and (**b**) NIBEM IR distributions both for healthy controls (blue) and AD patients (orange) according to cortical ROI. Box: median and inter-quartile range. Whiskers: mean $\pm$ 1.5 standard deviations. Round markers stand for extreme outliers beyond the whisker's limits.
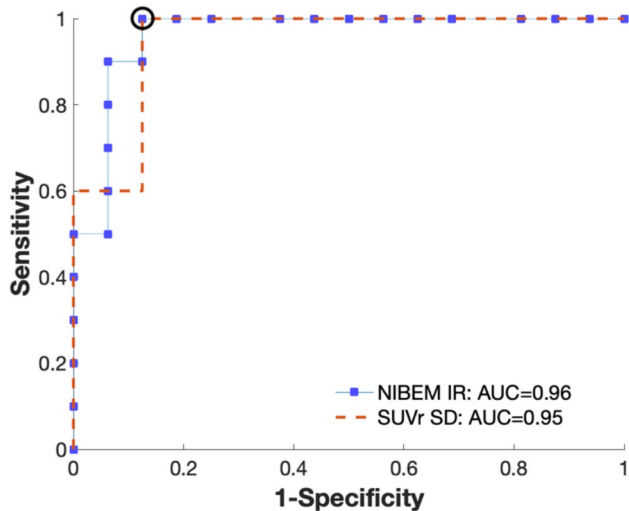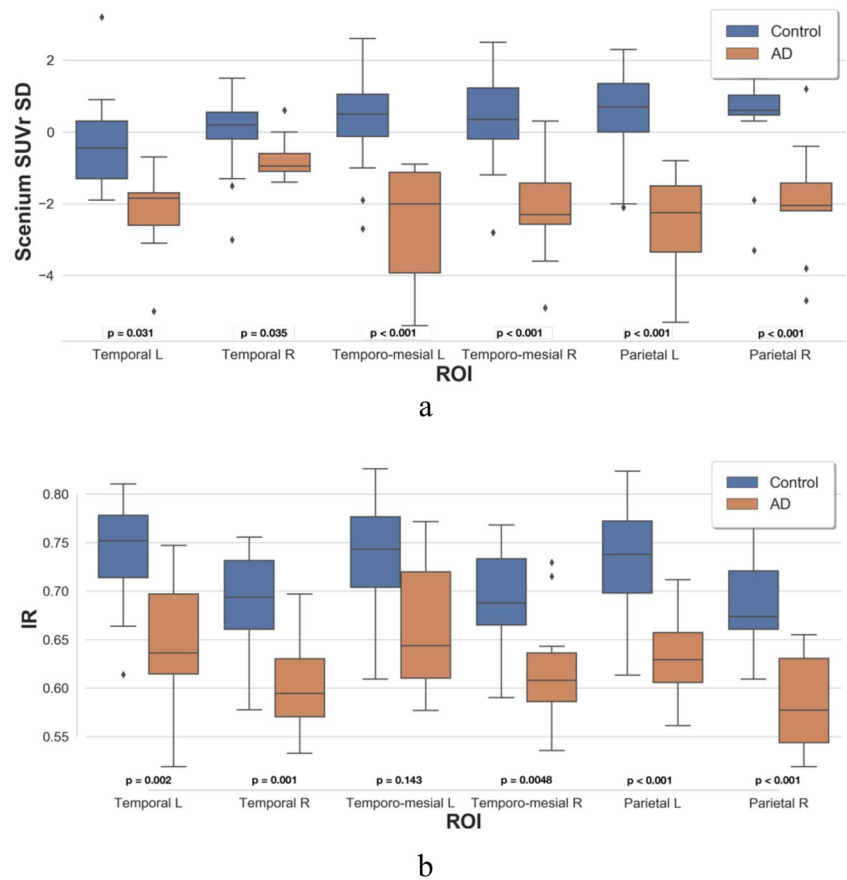

a

b

Fig. 7 Receiver operating curves (ROC) of SUVr SD (red dotted line) and NIBEM IR (blue line and boxes) in differentiating healthy controls from AD patients. The black round marker indicates the optimal threshold maximizing Youden's index that is common to both methods.

Discussion

It is well documented that in the first stages of AD, the temporal and parietal lobes are usually affected by amyloid deposition [18]. This coincides with neurofibril deposition initially confined to the medial temporal lobe and limbic structures [11, 18]. These manifestations tend to be reflected as areas of hypo-metabolism on ^{18}F-FDG PET reconstructed images. Visual rating of cortical distribution yields undesirable inter-reader variability [11, 17] and sub-optimal specificity especially among moderately skilled readers [18, 19]. To overcome these limitations, automated semi-quantitative techniques have been developed to help the physician judge whether the ROI can be considered as hypo-metabolic or not. In the first stages of the disease, hypo-metabolism is not usually observed symmetrically. It thus would seem intuitive to make a relative comparison: can the considered ROI be considered as identical in terms of radio-tracer concentration with respect to its symmetrical counterpart or a reference region?

The key limitation, for performing reliable direct ROI comparison in PET, is that no information about the statistical variability of the reconstructed data is directly available with the reconstruction algorithms used in clinical routine [28, 29].

Table 2 Diagnostic performances of SUVr SD and NIBEM IR in differentiating healthy controls from AD patients

	ROC AUC	Se	Sp	Acc	PPV	NPV
SUVr SD	0.95	100% (10/10)	88% (14/16)	92% (24/26)	83% (10/12)	100% (14/14)
NIBEM IR	0.96	100% (10/10)	88% (14/16)	92% (24/26)	83% (10/12)	100% (14/14)
p value	0.86	1	1	1	1	1

Values computed using the optimal threshold maximizing Youden's index.

ROC AUC area under the ROC curve, *Se* sensitivity, *Sp* specificity, *Acc* accuracy, *PPV* positive predictive value, *NPV* negative predictive value

To overcome this problem, most of the diagnostic assistance techniques recently proposed in the literature rely on the use of databases. These techniques can be separated into two groups. The first one is artificial intelligence (AI) based: for a very specific task, a convolutional neural network can be trained using a previously annotated database to classify whether the proposed ^{18}F-FDG PET scan belongs to a patient suffering from the disease or not. Such approaches can lead to very promising classification performance [20]. The second group of approaches relies on the statistical mapping of voxel-based normalized cortical uptake ratio with respect to some reference region [23–27]. Each investigated voxel or ROI is characterized by a score that reflects its distance to a reference score. The latter is computed based on a database of ethnic-, age- and sex-matched healthy controls. As previously mentioned, these approaches give promising results, but have certain limitations. The major drawback is that they inescapably depend on the database on which they rely or from which they have been trained. This raises the question of performance variations of these methods considering changes in acquisition equipment, in reconstruction parameters, acquisition conditions, radio-tracer dose, etc. As institution-wide, country-specific, or equipment manufacturer-wide recommendations are and will continue to different, it appears very difficult to trust a diagnosis relying only on AI or database-aided techniques. Nevertheless, comparisons with these methods that are now fashionable would be of great interest. Apart from the need to have access to raw data to perform NIBEM reconstruction, the comparison between the proposed approach and these state-of-the-art methods would be straightforward as their output is of a comparable nature and the metrics used to assess their performance are similar.

The framework proposed in this article is different from the kind of approaches described above as it apprehends the problem of ROI comparison in PET in an alternative way. Through the reconstruction of confidence intervals, what is proposed is to directly compare regions of interest for dementia with a reference region extracted from the same reconstruction thanks to concordance measures. The performed comparison is not based on any external information (like a database for instance) for the analysis. Indeed, the reference used to perform the hypo-metabolism analysis is specific to each patient. It does not require a normalization step either. Only raw data of PET acquisition are needed. The preliminary results presented in this paper seem to highlight that such a processing framework makes possible diagnostic performances that are promising as they seem similar to a validated post-processing software for AD on the tested data.

However, the current study has also some limitations that raise interesting perspectives and necessary improvements. Concerning the clinical validation, the number of patients included remains limited. Indeed, as this method requires access to raw PET data, it is impossible to use databases such as ADNI [40] because associated raw data are usually not available. It thus could be interesting to validate the results on a more consistent database. Speaking of databases, it could also be interesting to adapt such an analysis to a broader panel of neurodegenerative diseases like mild cognitive impairment, fronto-temporal dementia, and Lewy body dementia. Another important step in the clinical follow-up of patients with dementia is the quantification of the evolution of brain function disorders. A method like the one proposed in this paper could have a role to play for this task. In further studies, one could imagine testing the ability to quantify the evolution of a disease over time by analyzing the analogous IR scores of PET scans acquired at two different times. It may thus be interesting to see if the IR score evolution correlates with the clinical and expected evolution of the disease and thus allow a better follow-up of disease evolution over time. On the side of theoretical developments, to be in line with current PET standards it may be interesting to perform NIBEM reconstructions making use of time-of-flight information. However, this would require prior theoretical developments to manage time-of-flight in the NIBEM algorithm. Generalizing the NIBEM algorithm to perform 3D reconstruction of CI estimates without making use of re-binning procedures would also be a plus. Comparability of the tested methods would be greatly improved. In short, the objective of further studies is to propose a reconstruction pipeline that meets all the current PET reconstruction standards. The diagnosis-aid performance will also benefit further theoretical developments in statistical testing of interval-valued distributions that could be more specific to the statistical population characteristics we are trying to compare.

More generally, the diagnosis of neurodegenerative diseases in PET will surely continue to benefit from image quality improvements allowed by hardware and software progress of PET technology. However, chances are pretty good that tools like the one presented that facilitate better understand and characterize the acquired and processed data will also play an important role for early diagnosis in the future.

One last important point concerns the use of a database for comparative analysis. Even if the proposed analysis framework does not require external information like metadata or database references, it still requires to set beforehand a classification IR threshold that allows to state if the analyzed patient is affected by the disease or not. This information is required for all the diagnosis assistance methods mentioned above. Obviously, this preliminary work does not have the level of proof to establish a definitive threshold that could be used to differentiate AD from healthy controls in routine clinical work. As this threshold is the only dependence of the proposed method to a database, it would be interesting to carry out a study to see if the classification performance is independent of the group characteristics, acquisition conditions, radio-tracer dose, etc. In the case of a positive answer, it could be conceivable to propose a decision-making process based on the proposed methodology that is freed from the need of using any database which would be an important step for aided-diagnosis methods.

Conclusion

The method presented in this paper is the first method of the literature that exploits the uncertainty quantization associated with the data (in particular the statistical variability here) to make an assisting diagnosis tool in PET. The preliminary and original results presented in this paper highlighted the fact that clinical routine would directly benefit from theoretical progress in the field of confidence interval estimation and comparison. Moreover, the limited computational cost of the proposed method and its relatively easy implementation in the clinical routine framework should allow its use as a complementary tool for diagnosis aid and thus to improve the decision algorithm of AD diagnosis.

To conclude, this paper has shown that interval-valued reconstruction can allow promising analysis of brain ^{18}F-FDG PET data, yielding diagnostic performances similar to that of a SUVr SD-based commercial post-processing software.

Compliance with ethical standards

References

1. Prince M, Wimo AGM, Ali GC, Wu YT, Prina M. World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends. London: Alzheimer's Disease International; 2015.
2. Castro DM, Dillon C, Machnicki G, Allegri RF. The economic cost of Alzheimer's disease: family or public health burden? Dement Neuropsychol. 2010;4(4):262–7.
3. Bloudek LM, Spackman DE, Blankenburg M, Sullivan SD. Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. J Alzheimers Dis. 2011;26(4):627–45.
4. Perani D, Cerami C, Caminiti SP, Santangelo R, Coppi E, Ferrari L, et al. Cross-validation of biomarkers for the early differential diagnosis and prognosis of dementia in a clinical setting. Eur J Nucl Med Mol Imaging. 2016;43(3):499–508.
5. Siderowf A, Aarsland D, Mollenhauer B, Goldman JG, Ravina B. Biomarkers for cognitive impairment in Lewy body disorders: btatus and relevance for clinical trials: biomarkers of cognitive impairment. Mov Disord. 2018;33(4):528–36.
6. Petrella JR. Neuroimaging and the search for a cure for Alzheimer disease. Radiology. 2013;269(3):671–91.
7. Nasrallah IM, Wolk DA. Multimodality imaging of Alzheimer disease and other neurodegenerative dementias. J Nucl Med. 2014;55(12):2003–111.
8. Bohnen NI, Djang DS, Herholz K, Anzai Y, Minoshima S. Effectiveness and safety of ^{18}F-FDG PET in the evaluation of dementia: a review of the recent literature. J Nucl Med. 2012;53(1):59–71.
9. Kato T, Inui Y, Nakamura A, Ito K. Brain fluorodeoxyglucose (FDG) PET in dementia. Ageing Res Rev. 2016;30:73–84.
10. Nestor PJ, Altomare D, Festari C, Drzezga A, Rivolta J, Walker Z, Bouwman F, Orini S, Law I, Agosta F, Arbizu J, Boccardi M, Nobili F, Frisoni GB. Clinical utility of FDG-PET for the differential diagnosis among the main forms of dementia. Eur J Nucl Med Mol Imaging. 2018;45(9):1509–25.
11. Ng S, Villemagne VL, Berlangieri S, Lee ST, Cherk M, Gong SJ, et al. Visual assessment versus quantitative assessment of 11C-PIB PET and 18F-FDG PET for detection of Alzheimer's disease. J Nucl Med. 2007;48(4):547–52.
12. Mosconi L, Mistur R, Switalski R, Tsui WH, Glodzik L, Li Y, et al. FDG-PET changes in brain glucose metabolism from normal cognition to pathologically verified Alzheimer's disease. Eur J Nucl Med Mol Imaging. 2009;36(5):811–22.
13. Varrone A, Asenbaum S, Vander Borght T, Booij J, Nobili F, Någren K, et al. EANM procedure guidelines for PET brain imaging using ^{18}F-FDG, version 2. Eur J Nucl Med Mol Imaging. 2009;12:2103–10.

14. Giovacchini G, Squitieri F, Esmaeilzadeh M, Milano A, Mansi L, Ciarmiello A. PET translates neurophysiology into images: a review to stimulate a network between neuroimaging and basic research. J Cell Physiol. 2011;226(4):948–61.

15. Anchisi D, Borroni B, Franceschi M, Kerrouche N, Kalbe E, Beuthien-Beumann B, et al. Heterogeneity of brain glucose metabolism in mild cognitive impairment and clinical progression to Alzheimer disease. Arch Neurol. 2005;62(11):1728–33.

16. Silverman DH, Small GW, Chang CY, Lu CS, De Kung AMA, Chen W, et al. Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome. JAMA. 2001;286(17):2120–7.

17. Grimmer T, Wutz C, Alexopoulos P, Drzezga A, Förster S, Förstl H, et al. Visual versus fully automated analyses of ^{18}F-FDG and amyloid PET for prediction of dementia due to Alzheimer disease in mild cognitive impairment. J Nucl Med. 2016;57(2):204–7.

18. Lehman VT, Carter RE, Claassen DO, Murphy RC, Lowe V, Petersen RC, et al. Visual assessment versus quantitative three-dimensional stereotactic surface projection fluorodeoxyglucose positron emission tomography for detection of mild cognitive impairment and Alzheimer disease. Clin Nucl Med. 2012;37(8):721–6.

19. Morbelli S, Brugnolo A, Bossert I, Buschiazzo A, Frisoni GB, Galluzzi S, et al. Visual versus semi-quantitative analysis of ^{18}F-FDG-PET in amnestic MCI: an European Alzheimer's Disease Consortium (EADC) project. J Alzheimers Dis. 2015;44(3):815–26.

20. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using ^{18}F-FDG PET of the brain. Radiology. 2018;290(2):456–64.

21. Liu M, Cheng D, Wang K, Wang Y. Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics. 2018;16:295–308.

22. De Carli F, Nobili F, Pagani M, Bauckneht M, Massa F, Grazzini M, et al. Accuracy and generalization capability of an automatic method for the detection of typical brain hypometabolism in prodromal Alzheimer disease. Eur J Nucl Med Mol Imaging. 2019;46(2):334–47.

23. Cerami C, Della Rosa PA, Magnani G, Santangelo R, Marcone A, Cappa SF, et al. Brain metabolic maps in mild cognitive impairment predict heterogeneity of progression to dementia. Neuroimage Clin. 2014;5(7):187–94.

24. Perani D, Della Rosa PA, Cerami C, Gallivanone F, Fallanca F, Vanoli EG, Panzacchi A, Nobili F, Pappata S, Marcone A, Garibotto V, Castiglioni I, Magnani G, Cappa SF, Gianolli L. Validation of an optimized SPM procedure for FDG-PET in dementia diagnosis in a clinical setting. Neuroimage Clin. 2014;6:445–54.

25. Yamane T, Ikari Y, Nishio T, Ishii K, Ishii K, Kato T, et al. Visual-statistical interpretation of (18)F-FDG-PET images for characteristic Alzheimer patterns in a multicenter study: inter-rater concordance and relationship to automated quantitative evaluation. AJNR Am J Neuroradiol. 2014;35(2):244–9.

26. Caminiti SP, Ballarini T, Sala A, Cerami C, Presotto L, Santangelo R, et al. FDG-PET and CSF biomarker accuracy in prediction of conversion to different dementias in a large multicentre MCI cohort. Neuroimage Clin. 2018;28(18):167–77.

27. Brugnolo A, De Carli F, Pagani M, Morbelli S, Jonsson C, Chincarini A, et al. Head-to-Head comparison among semi-quantification tools of brain FDG-PET to aid the diagnosis of prodromal Alzheimer's disease. J Alzheimers Dis. 2019;68(1):383–94.

28. Kucharczak F, Loquin K, Buvat I, Strauss O, Mariano-Goulart D. Interval-based reconstruction for uncertainty quantification in PET. Phys Med Biol. 2018;63(3):035014.

29. Kucharczak F, Ben BF, Strauss O, Mariano-Goulart D. Confidence interval constraint-based regularization framework for PET quantization. IEEE Trans Med Imaging. 2019;38(6):1513–23.

30. Jena A, Taneja S, Goel R, Renjen P, Negi P. Reliability of semiquantitative ^{18}F-FDG PET parameters derived from simultaneous brain PET/MRI: a feasibility study. Eur J Radiol. 2014;83(7):1269–74.

31. Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 2007;6:734–46.

32. Shepp L, Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE Trans Medical Imaging. 1982;1(2):113–22.

33. Lange K, Carson R. EM reconstruction algorithms for emission and transmission tomography. J Comput Assist Tomogr. 1984;8(2):306–16.

34. Defrise M, Kinahan PE, Michel DT, Sibomana C, Newport MD. Exact and approximate rebinning algorithms for 3-D PET data. IEEE Trans Med Imaging. 1997;16(2):194–204.

35. Rolls ET, Joliot M, Tzourio-Mazoyer N. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. NeuroImage. 2015;122:1–5.

36. Mosconi L. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. FDG-PET studies in MCI and AD. Eur J Nucl Med Mol Imaging. 2005;32:486–510.

37. Perolat J, Couso I, Loquin K, Strauss O. Generalizing the Wilcoxon rank-sum test for interval data. J Approx Reason. 2015;56:108–21.

38. Smets P. Analyzing the combination of conflicting belief functions. Inf Fus. 2007;8(4):387–412.

39. Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology. 1982;143:29–36.

40. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology. 2010;74(3):201–9.