A comparison of two Statistical Mapping Tools for Automated Brain FDG-PET analysis in predicting conversion to Alzheimer's Disease in Subjects with Mild Cognitive Impairment

Valentina Garibotto, Sara Trombella, Luigi Antelmi, Paolo Bosco, Alberto Redolfi, Claire Tabouret-Viaud, Olivier Rager, Gabriel Gold, Panteleimon Giannakopoulos, Silvia Morbelli, et al.

A comparison of two Statistical Mapping Tools for Automated Brain FDG-PET analysis in predicting conversion to Alzheimer's Disease in Subjects with Mild Cognitive Impairment

Valentina Garibotto*, MD (1,2); Sara Trombella, PhD (1,3), Luigi Antelmi, PhD (4), Paolo Bosco, PhD (5), Alberto Redolfi, PhD (6), Claire Tabouret-Viaud, MD (2), Olivier Rager, MD (2), Gabriel Gold, MD (7), Panteleimon Giannakopoulos, MD (7), Silvia Morbelli, MD, PhD (8), Flavio Nobili, MD, PhD (8), Robert Perneczky, MD (9,10,11,12), Mira Didic (13), Eric Guedj (13), Alexander Drzezga (14,15), Rik Ossenkoppele (16), Bart van Berckel (16), Osman Ratib, MD (2,7) and Giovanni B. Frisoni, MD (3,7)

1. Laboratory of Neuroimaging and innovative molecular tracer, University of Geneva, Geneva, Switzerland;
2. Division of Nuclear Medicine, University Hospitals of Geneva, Geneva, Switzerland;
3. Laboratory of neuroimaging of aging, University of Geneva, Geneva, Switzerland;
4. University of Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France;
5. IRCCS Fondazione Stella Maris, Viale del Tirreno 331, Pisa;
6. IRCCS Fatebenefratelli, Brescia, Italy;
7. University of Geneva, Geneva, Switzerland;
8. Department of Nuclear Medicine, IRCCS AOU San Martino–IST, University of Genoa, Genoa, Italy;
9. Department of Psychiatry and Psychotherapy, Ludwig- Maximilians-Universitaet Muenchen, Munich, Germany;
10. Department of Psychiatry and Psychotherapy, Technische Universitaet Muenchen, Munich, Germany;
11. Neuroepidemiology and Ageing Research Unit, School of Public Health, Imperial College London, London, UK;
12. West London Mental Health NHS Trust, Cognitive Impairment and Dementia Ser- vice, Lakeside Mental Health Unit, London, UK;
13. Aix-Marseille Université, CNRS, Ecole Centrale Marseille, UMR 7249, Institut Fresnel, Marseille, France;
14. Department of Nuclear Medicine, Technische Universitaet, Munich, Germany;
15. Department of Nuclear Medicine, University of Cologne, Cologne, Germany;
16. Department of Nuclear Medicine and PET Research, VU University Medical Center, Amsterdam, Netherlands

* corresponding author:
Valentina Garibotto
Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital
Rue Gabrielle-Perret-Gentil 4
1205 Geneva, Switzerland
Tel: +41 22 3727252
Fax: +41 22 3727169
valentina.garibotto@hcuge.ch

Running title: **Mapping tools comparison for brain FDG-PET**

Abstract: 250 words
Text: 3473
References: 25
Tables and Figures: 1 + 5

Abstract

Objective: Automated voxel-based analysis methods are used to detect cortical hypometabolism typical of Alzheimer's Disease (AD) on FDG-PET brain scans. We compared the accuracy of two clinically validated tools for their ability to identify those MCI subjects progressing to AD at follow-up, to evaluate the impact of the analysis method on FDG-PET diagnostic performance.

Methods: SPMGrid and BRASS (Hermes Medical Solutions, Stockholm, Sweden) were tested on 131 MCI and elderly healthy controls from the EADC PET dataset. The concordance between the tools was tested by correlating the quantitative parameters (z- and t-values), measured by the two software tools, which measured the topographical overlap of the abnormal regions (Dice score). Three independent expert readers blindly assigned a diagnosis based on the two map sets. We used conversion to AD dementia as the gold standard.

Results: The t-map and z-map calculated with SPMGrid and BRASS, respectively, showed a good correlation (R > .50) for the majority of individual cases (128/131) and for the majority of selected regions of interest (ROIs) (98/116 [22]). The overlap of the hypometabolic patterns from the two tools was, however, poor (Dice score .36). The diagnostic performance was comparable, with BRASS showing significantly higher sensitivity (.82 versus .59) and SPMGrid showing higher specificity (.87 versus .52).

Conclusion: Despite similar diagnostic performance in predicting conversion to AD in MCI subjects, the two tools showed significant differences, and the maps provided by the tools showed limited overlap. These results underline the urgency for standardization across FDG-PET analysis methods for their use in clinical practice.

Keywords: FDG-PET, Alzheimer's disease, MCI, Automated analysis, tatistical parametric mapping, Hypometabolic pattern

1 Introduction

The availability of biological markers of neurodegeneration and synaptic dysfunction, such as FDG-PET, provides clinicians with the opportunity to improve diagnostic accuracy in clinical practice [1]. There is now evidence that pathophysiological events leading to neurodegeneration in Alzheimer's Disease (AD) initiate several years before the clinical manifestation of dementia [2]. The identification of pathological subjects before clinical diagnosis of dementia constitutes one of the challenges of modern research in neurodegenerative diseases.

Mild cognitive impairment (MCI) is the diagnostic term that describes the prodromal/predementia stage of neurodegeneration. The discrimination between progressive and non-progressive MCI and the identification of different diagnostic outcomes within the class of progressive MCI patients currently has great practical relevance and stands as one of the key topics in dementia research. The more relevant clinical outcome in the MCI patient population is conversion to dementia, i.e., loss of functional autonomy, within the 2 to 3 years following the diagnosis. For this purpose, the biomarker providing the best diagnostic performance for the prediction of conversion in MCI has been identified as the typical AD-pattern of hypometabolism in FDG PET images and therefore, has been included in the clinical guidelines [3-7]. However, the analysis of FDG-PET images is still flawed by a large variability in the literature, and in clinical practice, and the method for assessment as well as the threshold for abnormality required to accurately predict conversion in MCI individuals is still a matter of debate [5]. The lack of a homogeneous and objective reference for defining a normal test is one of the main limitations preventing stronger recommendations for the use of FDG PET in MCI [8]. The statistical parametric mapping methods allow for a more objective investigation of the relationship between metabolic changes in the brain and their anatomical location, and there is an overall agreement in the community that they should support visual reading [3, 4, 9, 10]. The use of an automated method has the advantage of providing an objective threshold of normality, required for a systematic validation of the utility of FDG PET as a disease biomarker. A binary readout of FDG PET images, as well as of other biomarkers, is also encouraged by the new "ATN system," classifying subjects on the basis of the presence of A: amyloid, T: tau, and N: neurodegeneration, as adopted in the latest research criteria of the National Institute of Aging [11]. While there is an agreement on the need for a standardized approach, there is no consensus on which tool should be used for this purpose, and various solutions, commercial and freeware, exist and have been tested. Previous studies have observed a highly variable diagnostic performance with the support of automated analyses, which could be strongly

influenced by the use of different tools in different studies [10, 12, 13].

The aim of this study was to compare the parametric maps calculated by two largely used automated voxel-wise methods (SPMGrid and BRASS) in evaluating FDG PET images of MCI and healthy controls (HC) and to estimate and identify the associated diagnostic performance on baseline MCI subjects converting to AD dementia over a follow-up of three years.

2 Material and methods

2.1 Subject selection

The scans analyzed in this study were selected from the FDG-PET project of the European Alzheimer's Disease Consortium (EADC) (www.eadc.info), which pools FDG-PET scans with related clinical and neuropsychological information from the contributing centers.
We adopted the following inclusion criteria:
1. A diagnosis of normal cognition or MCI at baseline;
2. A conversion to probable AD during follow-up or a minimum follow-up duration of 36 months for non-converters.

We selected a total number of 131 cases, in particular from the five EADC centers of Brescia (17 subjects), Genoa (61 subjects), Munich (30 subjects), Amsterdam (17 subjects), and Marseille (6 subjects). The used diagnostic criteria have been described in detail elsewhere [12].

The selected dataset included 90 MCI and 41 HC. During the follow-up time, 56 out of the 90 MCI patients progressed to AD dementia (MCI to AD - M: F = 24:32; age: 72±1). Out of the remaining 34 MCI patients: 32 remained in the MCI stage (MCI to MCI - M: F = 20:12; age: 67±2), and 2 were reverted to normal condition during the follow-up (MCI to normal - M: F = 1:1; age: 77±0). All healthy controls remained stable (HC - M: F = 18:23; age: 67±1). This dataset optimally reflects the characteristic variability of clinical reality, inclusive of pathological subjects as well as of worried-well subjects.

2.2 Imaging

FDG-PET scans were performed within two months from baseline clinical-neuropsychological examination and were performed following the European Association of Nuclear Medicine (EANM) procedural guidelines [14]. 185-250 MBq of FDG were injected into the subject intravenously. The acquisition of 3D emission scans started at least 30 minutes post-injection and lasted at least 10 minutes. The image reconstruction was performed via an ordered subset expectation maximization (OSEM) algorithm in all centers but Amsterdam (FBP). Attenuation correction was based on CT in Brescia, Genoa, and Marseille, and on transmission scan, in Munich and Amsterdam. Scatter correction was applied using standard manufacturer software. DICOM files were exported and converted to Analyze format. Anonymous scans and clinical information were uploaded in a dedicated

secured file transfer protocol (FTP) online platform. Structural neuroimaging was available for all patients (mainly, magnetic resonance imaging (MRI) or, as an alternative, computed tomography (CT) when no MRI was possible because of contraindications or patient intolerance). The study was approved by the local Medical Ethics Committee in each center and all the recruited subjects provided written informed consent [12].

2.3 Image preprocessing with SPMGrid

Baseline FDG-PET images were processed with the SPMGrid analysis tool. SPMGrid is the implementation of SPM8 [15] on the online e-infrastructure for neuroimaging, neuGRID [16]. SPMGrid provides an automated PET-based pipeline to process and test FDG-PET scans voxel-wise for hypometabolism against a control group of 107 healthy subjects from the EADC dataset with a mean age of 67.2 (SD: 6.7). Four regressors were fitted for the analysis (1. Indicator function for the test case; 2. Indicator function for the control group; 3. Subject age; 4. Total FDG uptake), yielding a total number of 103 degrees of freedom. In order to test for hypometabolism, the t-contrast [-1 1 0 0] was applied. The result of the analysis was not thresholded tridimensional Students t-map of the brain, in the ICBM152 atlas space [17]. Higher t-values were less consistent with the null hypothesis, and the voxel-wise difference between the individual and the control group was equal to zero, indicating that the alternative hypothesis of hypometabolism got stronger as the t-value was increased.

2.4 Image preprocessing with Hermes BRASS

Baseline FDG-PET images were processed with the Brain Analysis Software v3.5 (BRASS, Hermes Medical Solutions, Stockholm, Sweden) on a Hermes workstation (Nuclear Diagnostics, Stockholm, Sweden) [18]. BRASS provides an automated PET-based pipeline to process and test FDG-PET scans, voxel-wise for hypometabolism against a control group of 30 normal individuals with a mean age of 62.2 (SD: 7.8). Individual data are scaled to the total count in the brain volume, identified with an implicit masking of 0.8 of the maximum value. Age is not taken into account for the analysis. The result of the analysis is an unthresholded tri-dimensional Gaussian z-map of the brain in the reference template database. The more extreme the z-value, the more the tested voxel differs from the reference one.

2.5 Comparison of statistical maps

A linear transformation from the BRASS reference template space to the ICBM152 atlas space was calculated with a registration tool in FSL, FLIRT (FMRIB's Linear Imaging Registration Tool), by using the default values (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT). FLIRT is an automated tool allowing a robust and accurate registration of brain images, as previously shown with quantitative experiments on different modalities [19]. The transformation was applied to the BRASS z-maps, resulting in images in a common reference space and numerically equivalent to the SPMGrid t-maps. Mean SPMGrid t-values and BRASS z-values were extracted for every anatomical region of interest (ROI) defined in the AAL atlas [20]. The average BRASS z-value was plotted against the average SPMGrid t-value for each ROI and for each subject in this study. Pearson's R correlation coefficient was computed as a measure of agreement.

For the evaluation of the level of concordance between the maps, Dice scores were computed.

A Dice score for binary variables A and B is defined as the intersection of A and B divided by the union of A and B. The Dice value will, therefore, be equivalent to 1 if A and B have the same logical value in every pixel, and a value of 0 if they always disagree. In our case, the binary variables of interest were the volumetric z-maps and t-maps thresholded at a p-value threshold of 0.05, which were uncorrected for multiple comparisons. The two binary maps had a value of 1 for all significantly hypometabolic voxels and a value of 0 for the rest of the volume. Thus, the Dice scores provide a measure of the amount of spatial overlap of the hypometabolic patterns identified by the two software, relative to the sum of the hypometabolic regions detected by each [21]. This indicator can, therefore, quantitatively complement the visual interpretation of the abnormal maps provided by the three readers.

2.6 Visual rating of statistical maps

Statistical maps were overlapped to a template MRI scan in the ICBM152 atlas space and were displayed as a fixed multislice sequence at $p < 0.001$, uncorrected for multiple comparisons. Three blinded independent expert readers evaluated the multislice set for visual rating. Based on the identified hypometabolic pattern, a neuro-imaging diagnosis of normal or abnormal distribution was assigned.

Raters' instructions were operationalized as follows: if negative (or only small scattered areas of abnormality), the image was considered normal, and otherwise abnormal. An agreement was reached on the neuro-imaging diagnosis based on the majority of the rates on the individual case.

Figure 1 shows examples of the SPM and BRASS output in 3 individual cases.

2.7 Statistical analysis

The assigned neuro-imaging diagnosis on FDG-PET images acquired at baseline was compared to the clinical diagnosis at follow-up, assumed as gold-standard, in order to evaluate the performance of the two tests in predicting progression to AD dementia in MCI patients. MCI patients that converted to AD at follow-up (clinical diagnosis of AD) were considered as true positive (TP) if a hypometabolic pattern was diagnosed from neuro-imaging at baseline, as a false negative otherwise (normal distribution). Each MCI patient that remained stable or reverted to normal at follow-up (clinical diagnosis) was considered as a true negative (TN) if a normal metabolic pattern was radiologically diagnosed at baseline and as a false positive (FP) otherwise. Sensitivity, specificity, accuracy, and other relevant statistical values were computed for the software-assisted visual rating (both for SPMGrid and BRASS), for each rater, and the majority agreement among rater neuro-imaging diagnosis, as a measure of prediction of accuracy when considering conversion to dementia. The McNemar $\chi 2$ test was applied: 1. To neuro-imaging diagnosis among MCI converting to AD only to compare the sensitivities of the two software tools; 2. To neuro-imaging diagnosis among healthy only individuals to assess the significance of the difference between specificities [22].

The Intraclass Correlation Coefficient (ICC) was calculated to assess inter-rater reliability [23].

3 Results

3.1 Comparison of the output of the two tools and ROI correlation

The t-map and z-map calculated with SPMGrid and BRASS, respectively, showed a good correlation for the majority of individual cases ($R > .50$ for 128 subjects, out of 131), and for the majority of the ROIs, as identified in the AAL atlas ($R > .50$ for 98 ROIs, out of 116). A plot selection of the average BRASS z-value versus the average SPMGrid t-value, per subject and per ROI, have been reported in **Figure 2**, **3,** and **4** (selection of relevant ROIs for

the prognostic evaluation of MCI subjects).
However, a low average Dice score (36%) was measured when comparing the regions of the hypometabolic patterns from the two analyses, indicating that, on average, the hypometabolic regions estimated by the two software have poor percentage overlap. Consistently, poor agreement for the individual classification was in general observed when comparing the majority agreement neuro-imaging diagnosis for the two software tools (ICC = .54).

3.2 Computer-assisted Diagnostic performance by the two software tools

The results of the software-assisted visual rating for the two software tools have been reported in **Table 1**. Sensitivity was tested significantly higher for BRASS, for the three single raters and for the majority agreement (McNemar test on the diseased subject at follow-up), whereas specificity was tested significantly higher for SPMGrid, for the single raters and for the majority agreement (McNemar test over the healthy subject at follow-up).
A good to excellent inter-rater reliability was observed for SPMGrid when considering two categories of neuro-imaging diagnosis (0.86, 95% CI: 0.82-0.9), which wwere higher than the inter-rater reliability calculated for BRASS (0.76, 95% CI: 0.68-0.82).

4 Discussion

The most recent definition of AD, as well as other types of dementia, associates the evidence for specific underlying pathophysiology of the disease with its traditional clinical definition. FDG-PET stands among the biological markers included in contemporary diagnostic criteria for AD pathology, also in the MCI stage [6, 24]. Reimbursement of FDG-PET for the differential diagnosis of AD and FTD is nowadays authorized by the majority of national health services in Europe for the image acquisition and conventional visual reading [9]. Despite international recommendations have agreed on the fact that automated or semi-automated techniques should complement the traditional visual rating [14], no standardization has been carried out yet regarding procedures for assessing the typical hypometabolic AD pattern. The lack of standardized criteria can lead clinicians to use biomarkers in a subjective way, based on their own practice and experience, rather than clinical and evidence-based objective assumptions. The inhomogeneity of the analytic approaches explains a large part of the variability of the diagnostic performance of FDG in MCI. It has also been one of the main criticisms that a recent Cochrane review has identified [4, 8]. The lack of a standardized threshold for abnormality is a major milestone in the process of the validation of the clinical utility of FDG PET as a biomarker [5, 9]. For this reason, being associated with the increasing availability of biomarkers, the need for guidelines and recommendations regulating the systematic and automated use of biomarkers for the diagnosis of AD in the prodromal stage becomes urgent.

Within this context, the aim of this study has been to investigate and compare the capability of two different automated voxel-wise methods (SPMGrid and BRASS) in identifying those MCI subjects with prodromal dementia progressing to AD in a follow-up time of three years and excluding metabolic deficits in normal subjects and stable MCI. The concordance of the two analyses was tested by correlating the quantitative parameters measured by the two software tools (z- and t-scores, respectively), and by measuring the overlap of the abnormalities identified by the two methods via the Dice score. Three independent expert readers were assigned a diagnosis based on the pattern of hypometabolism detected by the two tools at the baseline, and Inter-rater reliability was also assessed for each software.

Multiple software solutions exist, but systematic cross-validation and comparison is still lacking, even if mandatory for their validated use in clinical practice. Our comparison showed that two widely used voxel-wise methods provide outputs that are significantly correlated in a majority of subjects and regions, however with systematic differences, resulting in a low Dice score for the distribution of abnormalities and consequent differences in the diagnostic decision making.

We used the McNemar test to compare sensitivity and specificity of two diagnostic tests applied to a given set of individuals, some with the disease, and some without the disease, as recommended in the literature. In our dataset, SPMGrid showed a significantly higher specificity, and BRASS showed a significantly higher sensitivity, resulting in a similar AUC, with a trend for a higher performance of SPMGrid. The two softwares showed a good correlation between average t- and z-values both per ROI and per subject.

It should be noted that the performance of the two reported softwares uses reading guidelines as defined at the beginning of the study and sets the cut-off for normal/abnormal images at the $p < 0.001$ threshold. This could be adapted in specific settings if a higher sensitivity is required, for example, in a ”screening” perspective. As an example, the subject in **Figure 5** is an MCI progressing to AD, categorized as normal in this study, and who would have been categorized as AD pattern if the threshold $p < 0.01$ was chosen.

The ICC was computed to assess the reliability among the three raters in this study. A good to excellent ICC was retrieved for SPMGrid, which was higher than BRASS.

There are a number of issues that can explain the systematic differences observed between the two software tools. For the sake of comparability, the output from the two softwares was matched into the same unit interval and mapping space. However, we voluntarily allowed for a number of other differences, namely the control group, which is specific for each software, and the algorithm used for spatial normalization, as we decided to test each software in its standard configuration, clinically used. A larger control group in SPMGrid might explain the higher specificity and ICC, for example (107 HC for SPMGrid versus 30 HC for BRASS).

Our findings are consistent with the literature, observing that different softwares have different outputs for the identification of hypometabolic patterns; however, in a smaller or less homogeneous group [25].

The study has some limitations worth mentioning. The readers had a fixed set of images to analyze, rather than the whole volume to browse, as it would be in clinical practice. They also did not have access to the original unprocessed images or to MR data, which could have helped in, e.g., correctly identifying some abnormalities due to atrophy. This choice was made to focus on the different impacts of different software, rather than evaluating the added value of automated approaches to visual reading, as this question has already been addressed by multiple previous studies [12, 26]. We included only MCI subjects converting to AD, asking the readers to identify a normal/abnormal distribution. The main reason for this choice was to adopt a gold standard that could be ascertained clinically, namely the conversion to AD dementia, avoiding the evaluation of different etiological diagnoses in the

absence of a neuropathological verification. However, this approach did not allow us to evaluate the ability of the automated methods to identify specific patterns evoking other forms of degenerative processes. In addition, this selection process reduced the size of our dataset, and thus our population is not the largest MCI population as reported in the literature. However, our MCI sample was in the upper range of the sample sizes reported in the literature for similar studies, as recently reviewed [27], and was sufficiently large enough to obtain significant findings. Finally, the parameters for FDG-PET scan acquisition and reconstruction were different between centers, and this might have affected the differentiall analyses between the two automated methods.

The present evidence provides a validation of the SPMGrid and BRASS procedures for their use in assisting the interpretation of FDG-PET images in predicting the conversion to AD. Despite an overall similar diagnostic performance, the two methods showed significant differences, resulting in a limited overlap of the abnormality maps provided by each method. These discrepancies in the hypometabolic patterns had an impact on the interpretation of the maps by expert readers, resulting in different sensitivity and specificity and inter-rater agreement, with a tendency for a higher AUC and a higher agreement among raters when the visual rating was computer-assisted by the SPMGrid software tool.

Our results emphasize the need to establish a specific procedure for the use of statistical mapping tools to analyze FDG-PET images in clinical practice. Indeed, it would be crucial to ensure that results obtained across different centers using different statistical mapping tools are comparable and reliable. For this purpose, cross-validation across software would be a necessary step. Given the increasing availability of large, freely accessible databases, a performance test against a common reference dataset should become a requirement for introducing statistical mapping tools in clinical practice to ensure harmonization and comparability of results.

Conflict of interest

All authors declare that he/she has no conflict of interest.

Acknowledgments

This work was partly funded by the Swiss National Science Foundation (SNSF) (grant number 320030 169876).

References

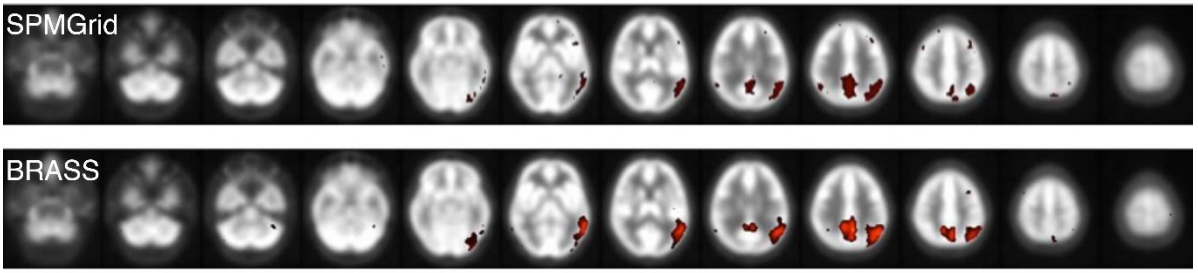
1. Chetelat G, Arbizu J, Barthel H, Garibotto V, Law I, Morbelli S, van de Giessen E, Agosta F, Barkhof F, Brooks DJ *et al*: **Amyloid-PET and (18)F-FDG-PET in the diagnostic investigation of Alzheimer's disease and other dementias**. *Lancet Neurol* 2020, **19**(11):951-962.
2. Jack CR, Jr., Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD *et al*: **Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers**. *Lancet Neurol* 2013, **12**(2):207-216.
3. Frisoni GB, Bocchetta M, Chetelat G, Rabinovici GD, de Leon MJ, Kaye J, Reiman EM, Scheltens P, Barkhof F, Black SE *et al*: **Imaging markers for Alzheimer disease: which vs how**. *Neurology* 2013, **81**(5):487-500.
4. Morbelli S, Garibotto V, Van De Giessen E, Arbizu J, Chetelat G, Drezgza A, Hesse S, Lammertsma AA, Law I, Pappata S *et al*: **A Cochrane review on brain [(1)(8)F]FDG PET in dementia: limitations and future perspectives**. *Eur J Nucl Med Mol Imaging* 2015, **42**(10):1487-1491.
5. Garibotto V, Herholz K, Boccardi M, Picco A, Varrone A, Nordberg A, Nobili F, Ratib O: **Clinical validity of brain fluorodeoxyglucose positron emission tomography as a biomarker for Alzheimer's disease in the context of a structured 5-phase development framework**. *Neurobiol Aging* 2017, **52**:183-195.
6. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC *et al*: **The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease**. *Alzheimers Dement* 2011, **7**(3):270-279.
7. Nobili F, Arbizu J, Bouwman F, Drzezga A, Agosta F, Nestor P, Walker Z, Boccardi M, Disorders E-ETFftPoF-PfDN: **European Association of Nuclear Medicine and European Academy of Neurology recommendations for the use of brain (18) F-fluorodeoxyglucose positron emission tomography in neurodegenerative cognitive impairment and dementia: Delphi consensus**. *Eur J Neurol* 2018, **25**(10):1201-1217.
8. Smailagic N, Vacante M, Hyde C, Martin S, Ukoumunne O, Sachpekidis C: **(1)(8)F-FDG PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI)**. *Cochrane Database Syst Rev* 2015, **1**:CD010632.
9. Frisoni GB, Boccardi M, Barkhof F, Blennow K, Cappa S, Chiotis K, Demonet JF, Garibotto V, Giannakopoulos P, Gietl A *et al*: **Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers**. *Lancet Neurol* 2017, **16**(8):661-676.
10. Cerami C, Della Rosa PA, Magnani G, Santangelo R, Marcone A, Cappa SF, Perani D: **Brain metabolic maps in Mild Cognitive Impairment predict heterogeneity of progression to dementia**. *Neuroimage Clin* 2015, **7**:187-194.
11. Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J *et al*: **NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease**. *Alzheimers Dement* 2018, **14**(4):535-562.
12. Morbelli S, Brugnolo A, Bossert I, Buschiazzo A, Frisoni GB, Galluzzi S, van Berckel BN, Ossenkoppele R, Perneczky R, Drzezga A *et al*: **Visual versus semi-quantitative analysis of 18F-FDG-PET in amnestic MCI: an European Alzheimer's Disease Consortium (EADC) project**. *J Alzheimers Dis* 2015, **44**(3):815-826.
13. Caroli A, Prestia A, Chen K, Ayutyanont N, Landau SM, Madison CM, Haense C, Herholz K, Nobili F, Reiman EM *et al*: **Summary metrics to assess Alzheimer disease-related hypometabolic pattern with 18F-FDG PET: head-to-head comparison**. *J Nucl Med* 2012, **53**(4):592-600.
14. Varrone A, Asenbaum S, Vander Borght T, Booij J, Nobili F, Nagren K, Darcourt J, Kapucu OL, Tatsch K, Bartenstein P *et al*: **EANM procedure guidelines for PET brain imaging using [18F]FDG, version 2**. *Eur J Nucl Med Mol Imaging* 2009, **36**(12):2103-2110.
15. Friston KJ, Holmes AP, Worsley KJ, Poline J, Frith CD, Frackowiak RS: **Statistical parametric maps in functional imaging: A general linear approach**. *Hum Brain Mapp* 1994, **2**(4):189-210.
16. Redolfi A, Bosco P, Manset D, Frisoni GB: **Brain investigation and brain conceptualization**.

Funct Neurol 2013, **28**(3):175-190.

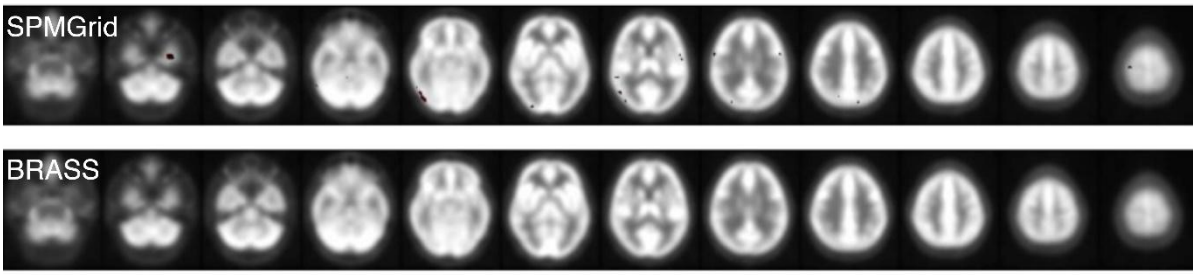
17. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B *et al*: **A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)**. *Philos Trans R Soc Lond B Biol Sci* 2001, **356**(1412):1293-1322.
18. Garibotto V, Montandon ML, Viaud CT, Allaoua M, Assal F, Burkhard PR, Ratib O, Zaidi H: **Regions of interest-based discriminant analysis of DaTSCAN SPECT and FDG-PET for the classification of dementia**. *Clin Nucl Med* 2013, **38**(3):e112-117.
19. Jenkinson M, Bannister P, Brady M, Smith S: **Improved optimization for the robust and accurate linear registration and motion correction of brain images**. *Neuroimage* 2002, **17**(2):825-841.
20. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M: **Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain**. *Neuroimage* 2002, **15**(1):273-289.
21. Presotto L, Ballarini T, Caminiti SP, Bettinardi V, Gianolli L, Perani D: **Validation of (18)F-FDG-PET Single-Subject Optimized SPM Procedure with Different PET Scanners**. *Neuroinformatics* 2017, **15**(2):151-163.
22. Trajman A, Luiz RR: **McNemar chi2 test revisited: comparing sensitivity and specificity of diagnostic examinations**. *Scand J Clin Lab Invest* 2008, **68**(1):77-80.
23. Shrout PE, Fleiss JL: **Intraclass correlations: uses in assessing rater reliability**. *Psychol Bull* 1979, **86**(2):420-428.
24. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R *et al*: **Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria**. *Lancet Neurol* 2014, **13**(6):614-629.
25. Lim H: **Automated Brain FDG-PET Analysis in Patients Suspected with Alzheimer's Disease: Comparison between Computer-aided Reading using SPM and BRASS**. 2016.
26. Perani D, Della Rosa PA, Cerami C, Gallivanone F, Fallanca F, Vanoli EG, Panzacchi A, Nobili F, Pappata S, Marcone A *et al*: **Validation of an optimized SPM procedure for FDG-PET in dementia diagnosis in a clinical setting**. *Neuroimage Clin* 2014, **6**:445-454.
27. Smailagic N, Lafortune L, Kelly S, Hyde C, Brayne C: **18F-FDG PET for Prediction of Conversion to Alzheimer's Disease Dementia in People with Mild Cognitive Impairment: An Updated Systematic Review of Test Accuracy**. *J Alzheimers Dis* 2018, **64**(4):1175-1194.

Legends to Figure

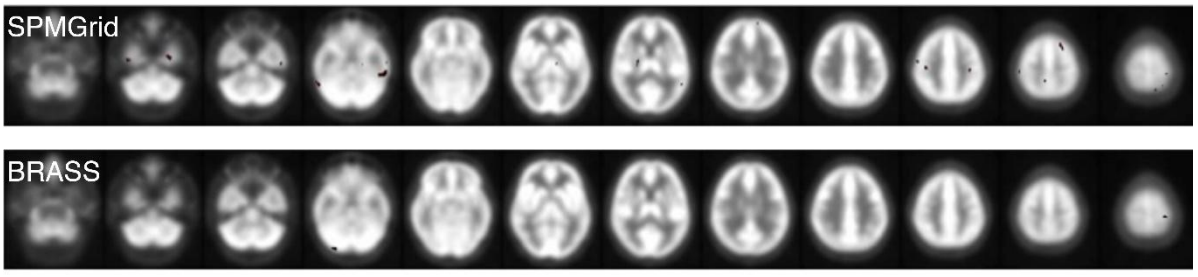
Fig. 1: Individual examples of the maps obtained with the two automated voxel-based analysis methods were tested here, namely SPMGrid and BRASS, overlapped to a standard template and displayed as multislice sequences. The threshold adopted is the same for both softwares and corresponds to a significance level of p-value < 0.001. Subfigure 1(a): Mild Cognitive Impairment (MCI) patient at baseline progressing to Alzheimer's Disease (AD) at follow-up; subfigure 1(b): Mild Cognitive Impairment (MCI) patient at baseline, stable at follow-up; subfigure 1(c): Healthy control (HC).



(a) MCI to AD

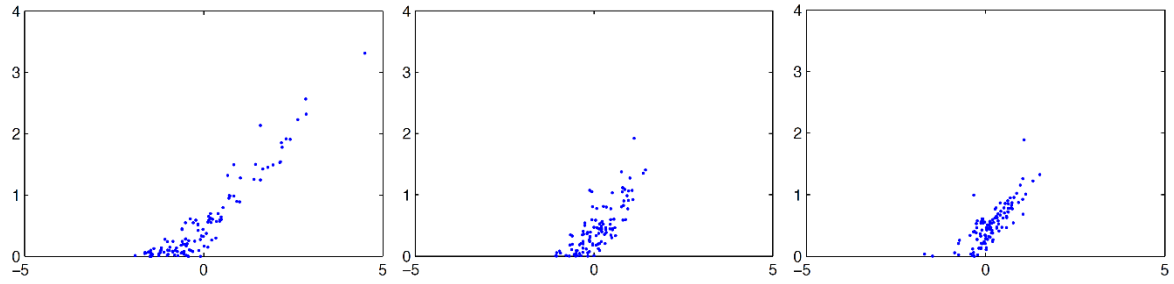


(b) MCI stable at follow-up



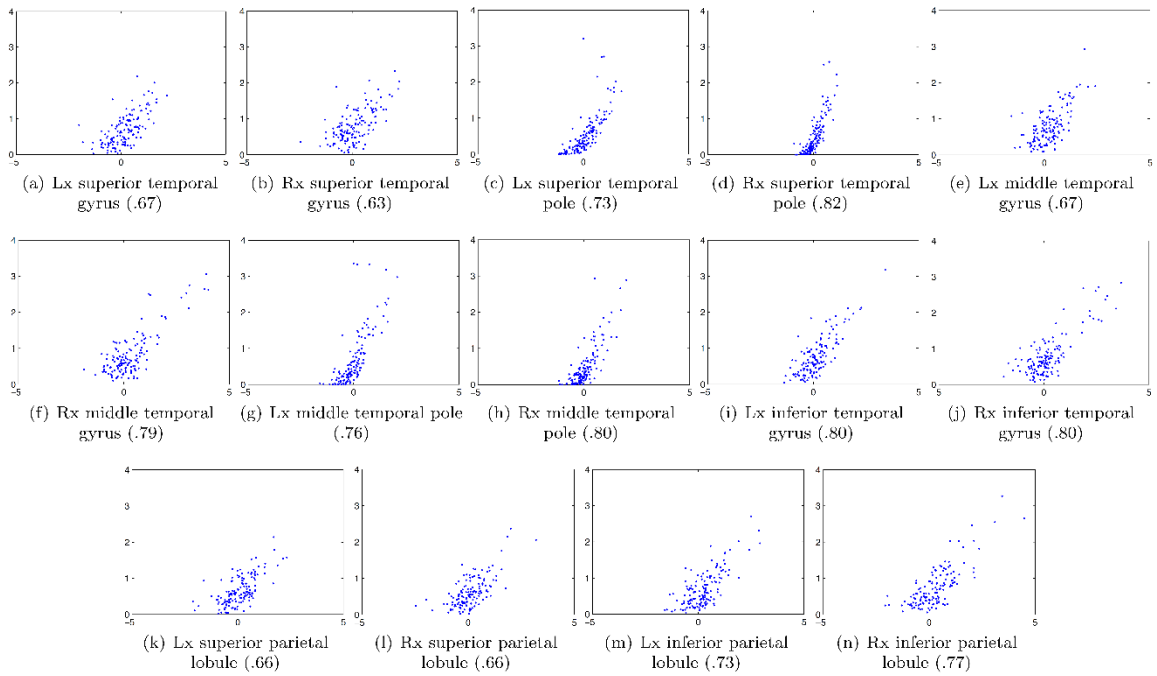
(c) HC

Fig. 2: Average abnormality values obtained with the two automated voxel-based analysis methods, SPMGrid and BRASS, in three cases. The plot shows BRASS z-value (y-axis) against the average SPMGrid t-value (x-axis) and Pearson's R correlation coefficient (in parenthesis). Subfigure 2(a): Mild Cognitive Impairment (MCI) patient at baseline progressing to Alzheimer's Disease (AD) at follow-up; subfigure 2(b): Mild Cognitive Impairment (MCI) patient at baseline, stable at follow-up; subfigure 2(c): Healthy control (HC).



(a) MCI to AD (.94) (b) MCI to MCI (.79) (c) HC (.79)

Fig. 3: Average abnormality values obtained with the two automated voxel-based analysis methods, SPMGrid and BRASS, in temporal and parietal Regions of Interests (ROIs): these ROIs are relevant for the prognostic assessment of MCI subjects. Average BRASS z-value (y-axis) against average SPMGrid t-value (x-axis) and Pearson's R correlation coefficient (in parenthesis). Lx: left, Rx: right.



(a) Lx superior temporal gyrus (.67) (b) Rx superior temporal gyrus (.63) (c) Lx superior temporal pole (.73) (d) Rx superior temporal pole (.82) (e) Lx middle temporal gyrus (.67)

(f) Rx middle temporal gyrus (.79) (g) Lx middle temporal pole (.76) (h) Rx middle temporal pole (.80) (i) Lx inferior temporal gyrus (.80) (j) Rx inferior temporal gyrus (.80)

(k) Lx superior parietal lobule (.66) (l) Rx superior parietal lobule (.66) (m) Lx inferior parietal lobule (.73) (n) Rx inferior parietal lobule (.77)

Fig. 4: Average abnormality values obtained with the two automated voxel-based analysis methods, SPMGrid and BRASS, in frontal Regions of Interest (ROIs). These ROIs are relevant for the prognostic assessment of MCI subjects. Average BRASS z-value (y-axis) against average SPMGrid t-value (x-axis) and Pearson's R correlation coefficient (in parenthesis). Lx: left, Rx: right.

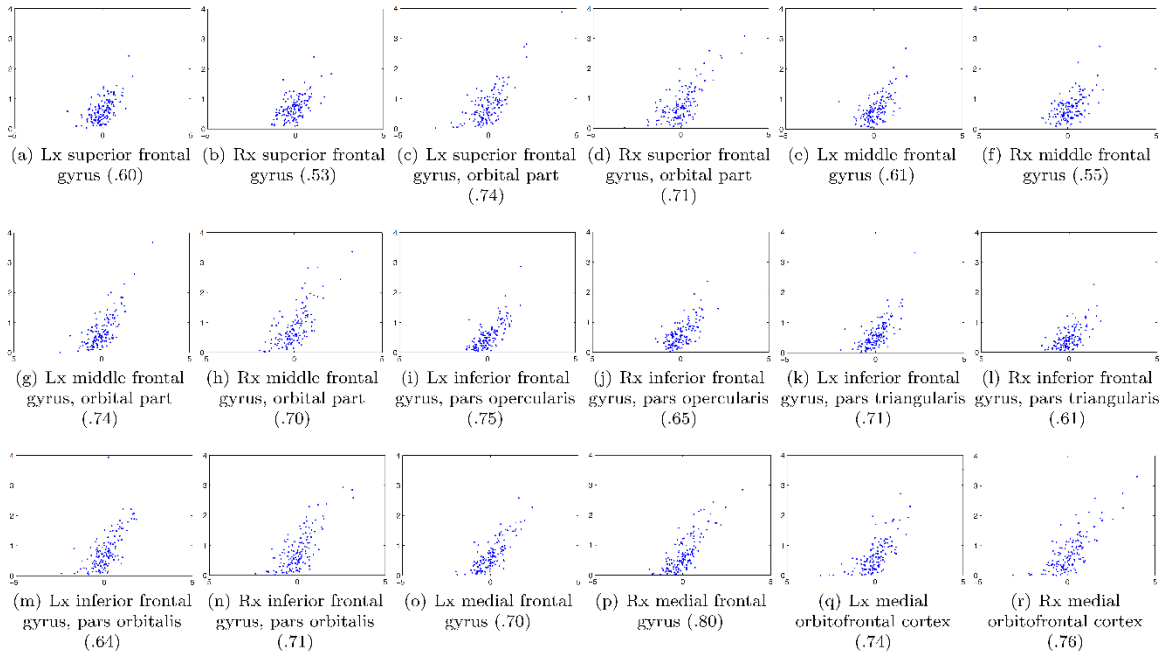
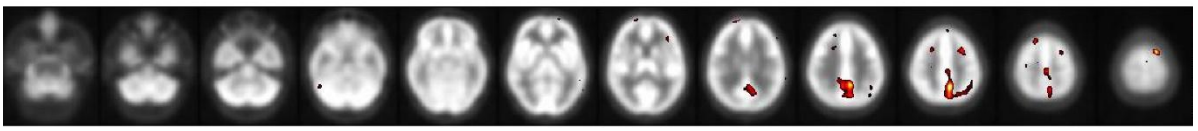
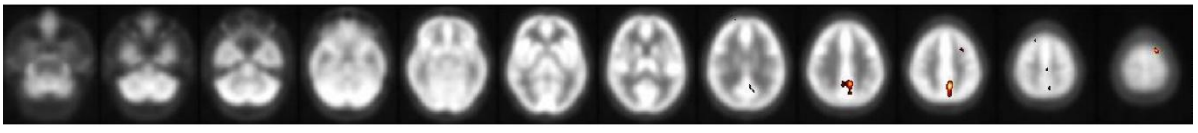


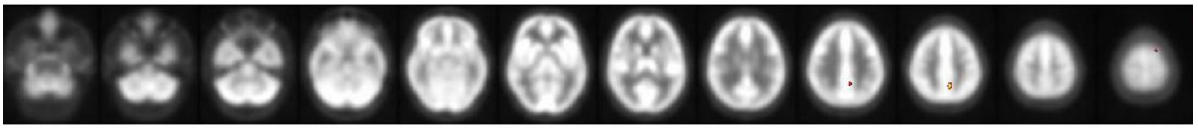
Fig. 5: Example case of a Mild Cognitive Impairment (MCI) patient progressing to Alzheimer's Disease (AD) at follow up. The case was categorized as normal by the readers in this study, but it would have been presumably categorized as pathological if the threshold $p < 0.01$ or $p < 0.05$, showing a typical AD pattern, was instead adopted.



(a) $p < 0.05$



(b) $p < 0.01$



(c) $p < 0.001$

Table 1: Diagnostic accuracy of SPMGrid and BRASS across raters versus clinical diagnosis at follow-up, in the whole study population (N = 131): sensitivity, specificity , positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-), area under the ROC curve (AUC). 95% confidence intervals in brackets.

McNemar test for sensitivity and specificity to compare SPMGrid and BRASS: *: $p \leq 0.05$; **: $p \leq 0.001$.

	Rater1		**Rater 2**		**Rater 3**		**Majority**	
	SPMGrid	**BRASS**	**SPMGrid**	**BRASS**	**SPMGrid**	**BRASS**	**SPMGrid**	**BRASS**
Sensitivity	.57	.79*	.55	.79**	.68	.87**	.59	.82**
	[.43,.70]	[.65,.88]	[.42,.68]	[.65,.88]	[.54,.79]	[.75,.94]	[.45,.72]	[.69,.91]
Specificity	.84**	.47	.72**	.37	.80**	.52	.87**	.52
	[.73,.91]	[.35,.58]	[.60,.81]	[.27,.49]	[.69,.88]	[.40,.64]	[.76,.93]	[.40,.64]
PPV	.73	.52	.60	0.48	.72	.58	.77	.56
	[.57,.85]	[.41,.63]	[.45,.73]	[.38,.59]	[.57,.83]	[.46,.68]	[.61,.88]	[.45,.67]
NPV	.72	.75	.68	.70	.77	.85	.74	.80
	[.62,.81]	[.59,.86]	[.57,.78]	[.53,.83]	[.66,.85]	[.71,.93]	[.63,.82]	[.65,.89]
LR+	3.57	1.47	1.98	1.25	3.39	1.82	4.42	1.71
	[2.02,6.29]	[1.15,1.90]	[1.28,3.05]	[1.00,1.57]	[2.08,5.52]	[1.41,2.35]	[2.38,8.19]	[1.31,2.23]
LR-	.51	.46	.62	.57	.40	.24	.47	.34
	[.37,.69]	[.27,.78]	[.46,.84]	[.33,.99]	[.27,.59]	[.12,.49]	[.34,.65]	[.19,.62]
AUC	.71	.62	.64	.58	.74	.70	.73	.67