Journal of Neuroscience Methods 192 (2010) 277–285

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Assessing the reliability to detect cerebral hypometabolism in probable Alzheimer's disease and amnestic mild cognitive impairment

Xia Wu[a], Kewei Chen[b,c,d,m,o,*], Li Yao[a,m], Napatkamon Ayutyanont[b,o], Jessica B.S. Langbaum[b,o], Adam. Fleisher[b,e,o], Cole Reschke[b,o], Wendy Lee[b,o], Xiaofen Liu[b,o], Gene E. Alexander[g,o], Dan Bandy[b,o], Norman L. Foster[h], Paul M. Thompson[f], Danielle J. Harvey[n], Michael W. Weiner[i,j], Robert A. Koeppe[k], William J. Jagust[l], Eric M. Reiman[b,o], the Alzheimer's Disease Neuroimaging Initiative[1]

[a] School of Information Science and Technology, Beijing Normal University, Beijing, PR China
[b] Banner Alzheimer's Institute, Phoenix, AZ, USA
[c] Departments of Radiology, University of Arizona, Tucson, AZ, USA
[d] School of Mathematics and Statistics, Arizona State University, Tempe, AZ, USA
[e] Department of Neurosciences, University of California, San Diego, USA
[f] School of Medicine, University of California, Los Angeles, USA
[g] Department of Psychology and Evelyn F. McKnight Brain Institute, University of Arizona, Phoenix, AZ, USA
[h] Center for Alzheimer's Care, Imaging and Research and Department of Neurology, University of Utah, USA
[i] Department of Medicine, Radiology, Psychiatry, and Neurology, University of California, San Francisco, USA
[j] Center for Imaging and Neurodegenerative Diseases, San Francisco Veterans Affairs Medical Center, USA
[k] Division of Nuclear Medicine, Department of Radiology, University of Michigan, USA
[l] Helen Wills Neuroscience Institute, School of Public Health, University of California, Berkeley, USA
[m] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, PR China
[n] University of California, Davis, USA
[o] Arizona Alzheimer's Consortium, Phoenix, AZ, USA

ARTICLE INFO

Article history:
Received 8 May 2010
Received in revised form 22 July 2010
Accepted 23 July 2010

Keywords:
Alzheimer's disease
MCI
FDG-PET
Reproducibility of results
Reliability
Bootstrap resampling
Family-wise error
SPM

ABSTRACT

Fluorodeoxyglucose positron emission tomography (FDG-PET) studies report characteristic patterns of cerebral hypometabolism in probable Alzheimer's disease (pAD) and amnestic mild cognitive impairment (aMCI). This study aims to characterize the consistency of regional hypometabolism in pAD and aMCI patients enrolled in the AD neuroimaging initiative (ADNI) using statistical parametric mapping (SPM) and bootstrap resampling, and to compare bootstrap-based reliability index to the commonly used type-I error approach with or without correction for multiple comparisons. Batched SPM5 was run for each of 1000 bootstrap iterations to compare FDG-PET images from 74 pAD and 142 aMCI patients, respectively, to 82 normal controls. Maps of the hypometabolic voxels detected for at least a specific percentage of times over the 1000 runs were examined and compared to an overlap of the hypometabolic maps obtained from 3 randomly partitioned independent sub-datasets. The results from the bootstrap derived reliability of regional hypometabolism in the overall data set were similar to that observed in each of the three non-overlapping sub-sets using family-wise error. Strong but non-linear association was found between the bootstrap-based reliability index and the type-I error. For threshold $p = 0.0005$, pAD was associated with extensive hypometabolic voxels in the posterior cingulate/precuneus and parietotemporal regions with reliability between 90% and 100%. Bootstrap analysis provides an alternative to the parametric family-wise error approach used to examine consistency of hypometabolic brain voxels in pAD and aMCI patients. These results provide a foundation for the use of bootstrap analysis characterize statistical ROIs or search regions in both cross-sectional and longitudinal FDG-PET studies. This approach offers promise in the early detection and tracking of AD, the evaluation of AD-modifying treatments, and other biologically or clinical important measurements using brain images and voxel-based data analysis techniques.

* Corresponding author at: Banner Alzheimer's Institute, Phoenix, AZ, USA. Tel.: +1 602 839 4851.
E-mail addresses: kewei.chen@bannerhealth.com, kchen@math.asu.edu (K. Chen).

0165-0270/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2010.07.030

1. Introduction

[^{18}F]-2-Fluoro-deoxy-D-glucose (FDG) positron emission tomography (PET) measured cerebral metabolic rates for glucose (CMRgl) have been widely used in the Alzheimer's disease (AD) research. In comparison with normal controls, patients with probable Alzheimer's disease (pAD) and amnestic mild cognitive impairment (aMCI) have characteristic and progressive CMRgl reductions in posterior cingulate (PC), temporal (TE), parietal (PA), precuneus (PCu), occipital (OC) (Alexander et al., 2002; Foster et al., 1983; Hoffman et al., 2000; Ibanez et al., 1998; Langbaum et al., 2009; McGeer et al., 1990; Silverman et al., 2001, etc.). In comparisons with non-carriers of the apolipoprotein E ε4 (APOE4) allele, the major genetic risk factor for late-onset AD, carriers of this allele have preferentially reduced CMRgl in the same brain regions, some of which occur decades before the anticipated onset of symptoms (Reiman et al., 1996, 2001, 2004, 2005, etc.).

The FDG-PET data acquired in the multi-center Alzheimer's disease neuroimaging initiative (ADNI) study provided an unprecedented opportunity to confirm and extend the previously reported findings from single-scanner-based studies, many of them were performed with relatively small number of subjects compared to the number of participants in the ADNI study (but see Silverman et al. as an exception) (Silverman et al., 2001).

The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principle Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55–90, to participate in the research – approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years.

Using FDG-PET data acquired from the ADNI project, this study (a) introduces the use of the bootstrap resampling technique to assess the reliability of the detected regional CMRgl differences between two study groups (and potential use in CMRgl changes over times in future separate study), and (b) illustrates its usage in cross-sectional studies to detect the reliability of regional hypometabolism in AD and aMCI, compared to NC.

To validate the bootstrap resampling approach assessing the reliability and consistency of the brain regions where glucose hypometabolism in patients with pAD or aMCI was observed, analyses were conducted using multiple sub-datasets randomly partitioned from the original dataset to examine consistency across multiple runs and to compare with the reliability obtained via bootstrap resampling. The comparison between the bootstrap reliability and the consistence across multiple runs is with or without the correction for multiple comparisons based on the random-field theory (also called family-wise error, FWE). Finally, we provide our rational for the use of the reliability index in place of the type-I error related indices such as the t or Z scores and the corresponding p-values in the context of multiple comparison correction.

2. Material and methods

2.1. Subjects

As described in Langbaum et al., ADNI subjects aged between 55 and 90 at the time of enrollment. Eligibility criteria for enrollment in each of the three specific groups are as follows. NC participants had a Mini Mental State Exam (MMSE) score of 24 or higher, a Clinical Dementia Rating (CDR) score of 0, and no diagnosis of depression, aMCI, or dementia. AMCI participants had an MMSE score of 24 or higher, a subjective memory complaint, objective memory loss measured by education adjusted scores on the Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, preserved activities of daily living (ADLs), and an absence of dementia (Petersen et al., 2001). Mild AD participants had an MMSE score between 20 and 26 (inclusive), a CDR score of 0.5 or 1.0, and met NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984). As described in the ADNI procedure, each participant provided an informed consent, and was studied under the guidelines approved by Institutional Review Boards at each site. For more information, please refer to the ADNI website (http://www.adni-info.org/index.php). At the initiation of the study, baseline FDG-PET data from 298 ADNI participants (74 pAD, 142 aMCI, and 82 NC) were available for downloading from the ADNI LONI (University of California, Los Angeles) website (http://www.loni.ucla.edu/ADNI/) and were included in this study.

2.2. Data acquisition

FDG-PET scans were acquired according to a standardized protocol. Data presented in this paper were collected using over 22 different PET scanners. The proportion of subjects in each subject group did not differ significantly in the scanners used to acquire their FDG-PET images ($\chi^2(51) = 21.9$; $p = 0.99$). Subjects were instructed to fast for at least 4 h prior to their scans. A six 5-min dynamic emission scan was acquired 30 min after the intravenous injection of 5.0 ± 0.5 mCi of 18F-FDG, as the subjects lay quietly in a dimly lit room with their eyes open and minimal sensory stimulation. Data were corrected for scatter and radiation-attenuation using transmission scans and image reconstruction algorithms specified for each scanner (http://www.loni.ucla.edu/ADNI/Data/ADNI_Data.shtml).

2.3. Centralized FDG-PET image pre-processing

As described in Langbaum et al. in details and briefly re-stated here, the six 5-min emission frames acquired from all subjects were pre-processed at the University of Michigan to help correct for scanner-related differences in FDG uptake: First, the six emission frames for each subject were co-registered to his or her first frame to create an average image. Second, this 30-min emission image was normalized for the variation in absolute voxel intensity and reoriented into an image consisting of 160 × 160 × 96 1.5 mm voxels, such that their horizontal sections were parallel to a section through the anterior and posterior commissures. At this step, none

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analyses or writing of this report. The complete listing of ADNI investigators is available at www.loni.ucla.edu\ADNI\Collaboration\ADNI_Authorship_list.pdf.

of the images was linearly or non-linearly deformed for individual differences in size and shape. Finally, a scanner-specific smooth filter was used to ensure that all of the images had an isotropic spatial resolution of 8 mm full-width-at-half-maximum (FWHM). The specific filter functions were determined from the Hoffman phantom PET scans that were acquired during the certification process.

2.4. SPM spatial pre-processing and statistical analysis

As stated above, the baseline FDG-PET data pre-processed at the University of Michigan were downloaded from the ADNI LONI website SPM5 (http://www.fil.ion.ucl.ac.uk/spm/) was the primary analytic platform for our analyses. The origins of the images were first re-centered so the images are spatially located within the neighborhood of the SPM MNI template. The baseline PET from each subject was spatially deformed, linearly and non-linearly with the SPM5 default settings, to the coordinate space of the SPM brain template with left/right orientation confirmed and normalization results visually inspected during the quality assurance and control process. All images were further smoothed by a Gaussian kernel with the FWHM of 12 mm.

Voxel-by-voxel statistical analyses of various types were then performed. For all analyses, the global count computed by the SPM sub-routine spm_global was accounted for by proportional scaling. Voxel-wise analysis of variance (ANOVA) was used to examine the differences among pAD, aMCI and NC. In the framework of ANOVA (with pooled variance and degree of freedom), Student *t*-test was used to investigate the hypometabolic brain regions of pAD patients or aMCI patients in comparison to NC. In order to understand the results of bootstrap resampling procedure described below, various values of FWE (random-field theory-based multiple comparison corrected type-I error) and uncorrected type-I error were used to examine the SPM *t*-score maps and investigate their relationship to the bootstrap-based reliability.

2.5. Reliability of regional hypometabolism in pAD and aMCI

As mentioned earlier and primarily to cross-validate our bootstrap approach, we performed two complementary analyses, bootstrap resampling over the full ADNI FDG-PET dataset and SPM runs over several subsets of this entire dataset. First, the entire data set was divided into three non-overlapping sub-datasets of equal size (1/3 of the total cases for each sub-dataset). For each of the 3 datasets, the same analyses described above were used to contrast the group differences between pAD and NC and between aMCI and NC. Using several different significant thresholds (corrected FWE $p = 0.05$ or uncorrected at $p = 0.005$, $p = 0.001$ or $p = 0.0005$), the binary brain region masks of hypometabolism in the pAD or aMCI patients, when compared to NC, were created for each of the 3 datasets. Summing the hypometabolism binary masks, over the 3 runs, for each of the contrasts (pAD vs. NC, aMCI vs. NC) and for each threshold, voxel value of 3, 2 or 1 signifies the fact that the corresponding location (voxel) was detected as hypometabolic in all 3, 2 or 1 of the 3 runs, indicative of the reliability of the detected hypometabolism.

Secondly, the bootstrap resampling approach was used, analyzing the entire dataset, to investigate the reliability of regional hypometabolism in pAD and aMCI. For each of the 1000 resampling iterations with replacement, the same statistical analyses contrasting group differences between pAD and NC, between aMCI and NC were carried out. We developed a SPM batching procedure that automated multiple SPM analyses (primarily setting up the general linear model [GLM], estimating the GLM parameters and conducting the significance assessments). The binary brain region masks of the hypometabolism were generated in the same manner as described above for each of various thresholds at each bootstrap SPM iteration. The sum of these masks over the 1000 iterations for each of the patient groups provides the frequency (or percentage) a given voxel were classified as hypometabolic for a given threshold. For threshold $p = 0.005$, for example, if a voxel was detected as hypometabolic 800 times over the 1000 SPM runs, then the sum of the 1000 masks would have the numerical value of 800 at this voxel location, indicative of the detection reliability as 80% using the corresponding threshold $p = 0.005$.

3. Results

The subject groups' demographic characteristics, MMSE scores, and APOE ε4 carrier status are described in Table 1. The three groups did not differ significantly in their gender distribution. The pAD group was slightly older than the aMCI ($p = 0.01$) and had slightly fewer years of education compared to the NC group ($p = 0.04$). As expected due to enrollment criteria, the pAD group had significantly lower MMSE scores than both the aMCI and NC groups ($p < 0.001$), and the aMCI group had significantly lower MMSE scores than the NC group ($p < 0.001$). Also as expected, the pAD and aMCI groups had a significantly higher proportion of subjects with one or two copies of the APOE ε4 allele ($p < 0.001$) than the NC group, and the three groups differed in the distribution of CDR scores ($p < 0.001$).

Table 1
Probable AD, amnestic MCI and NC group demographics.

	pAD ($n = 74$)	aMCI ($n = 142$)	NC ($n = 82$)	p-value
Age	71.1 ± 10.1	66.3 ± 11.9	68.4 ± 10.3	0.01[a]
Sex (% male/female)	34/66	20/80	22/78	0.08
Education (in years)	14.8 ± 3.3	15.3 ± 2.7	15.9 ± 2.8	0.04[b]
MMSE score	23.2 ± 2.2	27.1 ± 1.7	29.0 ± 1.1	<0.001[a,b,c]
CDR severity score (%)				
0	0	0.7	100	< 0.001[a,b,c]
0.5	37.8	99.3	0	
1.0	62.2	0	0	
APOE ε4 gene dose (%)				
No copies	35.1	45.1	74.4	<0.001[b,c]
One copy	47.3	40.8	23.2	
Two copies	17.6	14.1	2.4	

Age, education and MMSE are in mean ± SD, and all others in percentages. Demographic differences between NC, aMCI, and AD participants were analyzed using one-way analysis of variance (ANOVA), Fisher's exact and Chi-square (χ^2) tests. Scheffé-multiple comparison test was used to compare the differences between each pair of means.

[a] pAD significantly different from aMCI.

[b] pAD significantly different from NC.

[c] aMCI significantly different from NC.

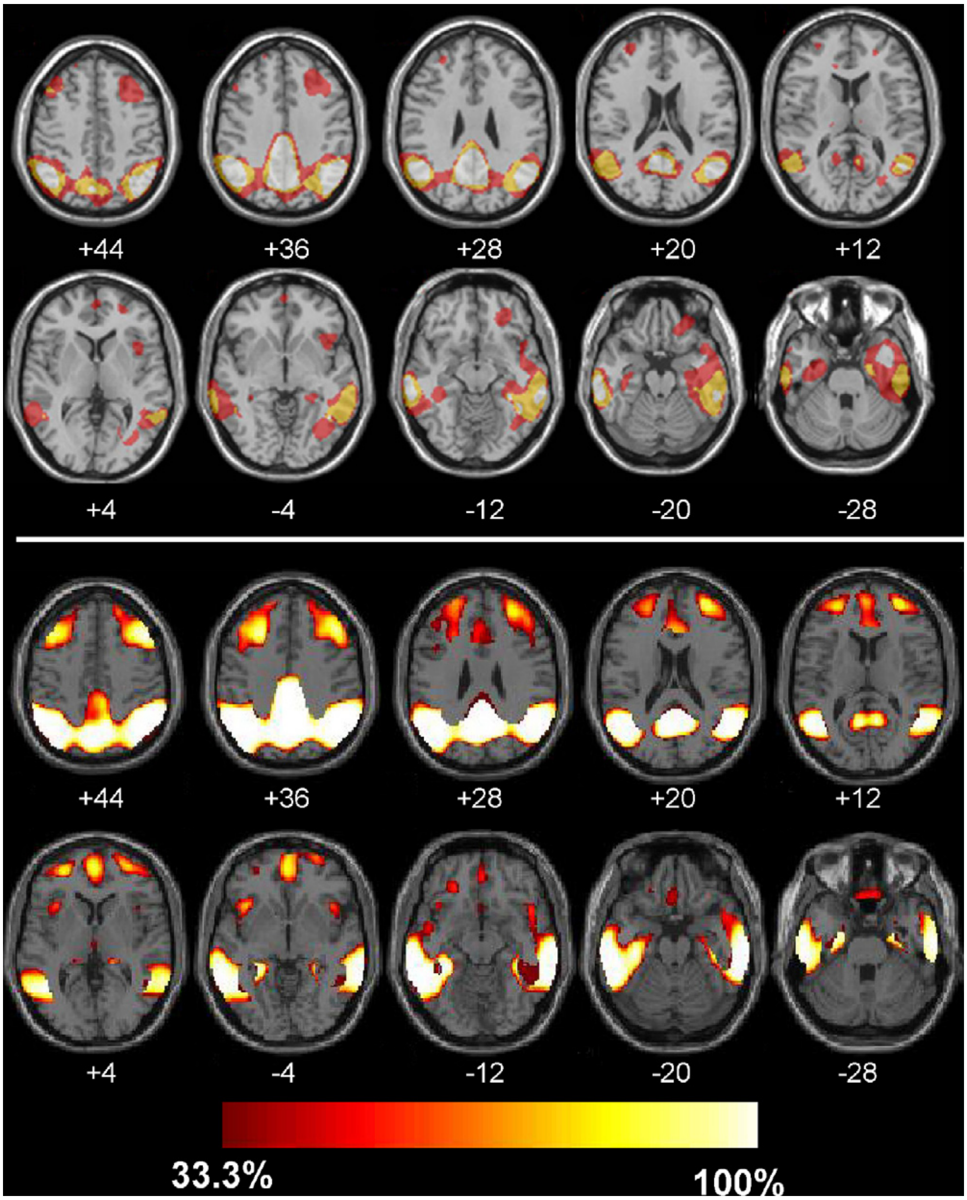

Fig. 1. The hypometabolic pAD brain regions obtained via multi non-overlap sub-datasets and the bootstrap resampling, respectively.

We first examined the consistency of regional hypometabolism for patients with pAD or aMCI over the three separate SPM runs using non-overlap sub-datasets.

3.1. *Comparing reliability obtained using type-I error correction in three non-overlapping datasets to bootstrap resampling in the combined data set*

Using uncorrected $p = 0.005$ to illustrate our general findings, the top panel of Fig. 1 displays the brain regions where the hypometabolism in patients with pAD was observed in all three separate SPM runs (in white), in two of the three runs (in yellow) and in any one of the three runs (in red). For uncorrected $p = 0.005$, $p = 0.001$ and corrected $p = 0.05$, hypometabolic pAD brain regions with the highest reliability (overlap over all three runs) included left and right PCu, PC, middle cingulate (MC), left and right angular gyrus, left and right TE, right inferior PA, left fusiform (only at uncorrected $p = 0.005$), left parahippocampal (only at uncorrected $p = 0.005$), left and right OC, and left supramarginal gyrus. In addition to these regions, hippocampus (only at uncorrected $p = 0.005$) was also detected for the middle-level reliability (2 out of 3 runs). For the lowest reliability (1 out of 3 runs), additional brain regions such as prefrontal and thalamus (only at uncorrected $p = 0.005$) were also detected.

The bootstrap-based reliability maps (shown in the bottom panel of Fig. 1 for pAD and 2 for aMCI, respectively) were compared to the maps estimated from the separate analyses conducted on three independent sub-datasets (shown in the top panels of Figs. 1 and 2). Although there are some differences between the overlap index maps and the bootstrap-based reliability maps, the overall patterns of the two groups of maps are quite similar. Moreover, the reduction in degrees of freedom for each of three separate analyses may contribute to the reduction in overlap maps or reliability maps.

The top panel of Fig. 2 is analogous to the top panel of Fig. 1 but for patients with aMCI. We observed reduced reliability as no voxel survived all 3 runs, even at the uncorrected $p = 0.005$. For $p = 0.001$ and $p = 0.005$, the hypometabolic brain regions for aMCI patients

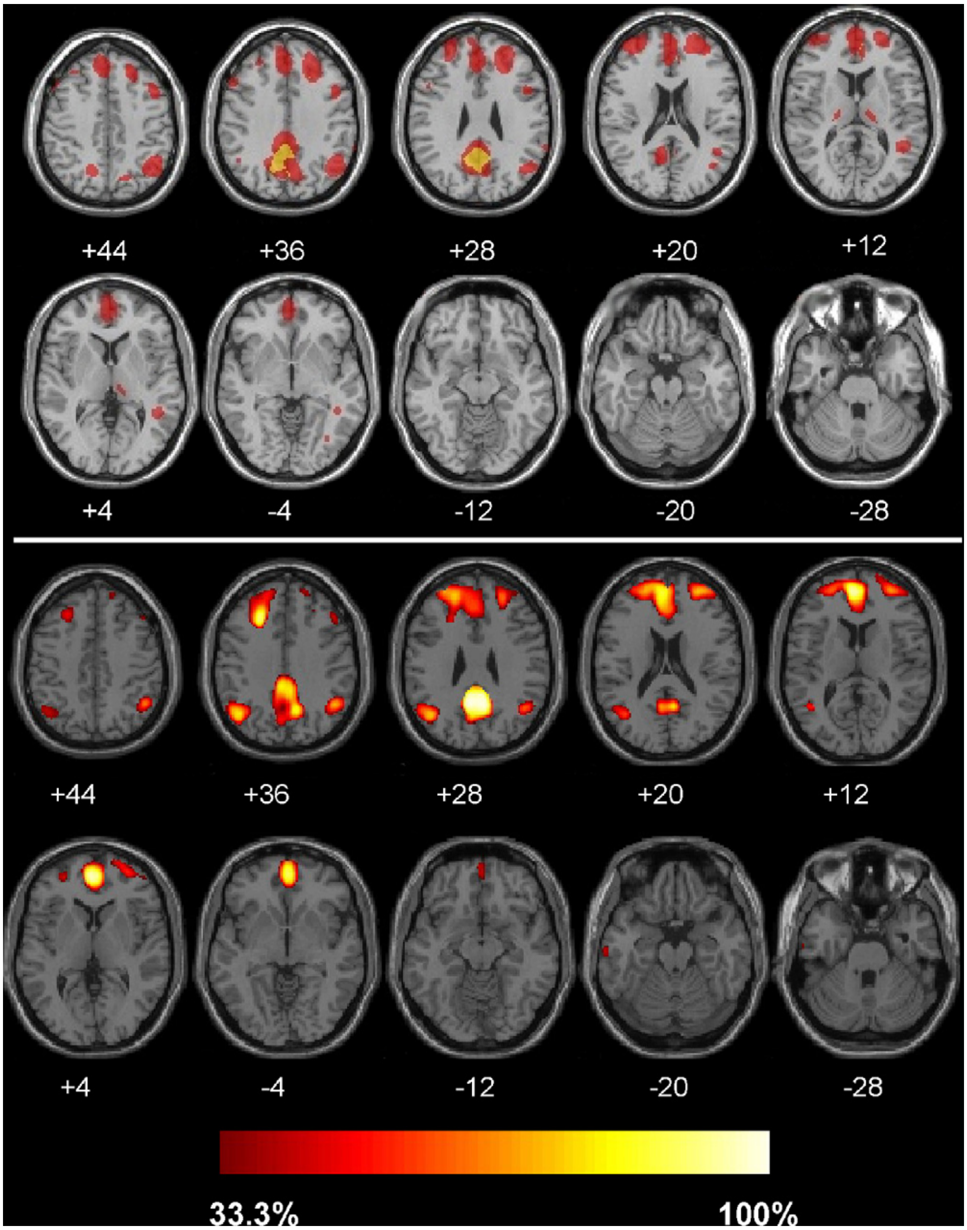

Fig. 2. The hypometabolic aMCI brain regions obtained via multi non-overlap sub-datasets and bootstrap resampling, respectively.

at middle-level reliability (2 out of 3 runs) included PCu, PC, prefrontal (only at $p = 0.005$), anterior cingulate (AC) and MC (only at $p = 0.005$). The low-level reliability (1 out of 3 runs) hypometabolic brain regions included additional prefrontal, parietal temporal, occipital (only at $p = 0.005$), fusiform (only at $p = 0.005$), parahippocampal (only at $p = 0.005$) and thalamus.

The degree of overlap out of the 3 separate sub-analyses was strongly correlated with the t-scores in examining, based on a single SPM analysis on the entire dataset, the hypometabolism in pAD patients compared to NC as shown in Fig. 3 (linear trend $R^2 = 0.527$, $p < 1.1e{-}16$). This raises the possibility of using the overall t-score or the corresponding p-value as an alternative reliability index (see more discussion on this below).

3.2. Bootstrap results

The bootstrap procedure was run 1000 times to contrast the group differences between patients with pAD and NC subjects,

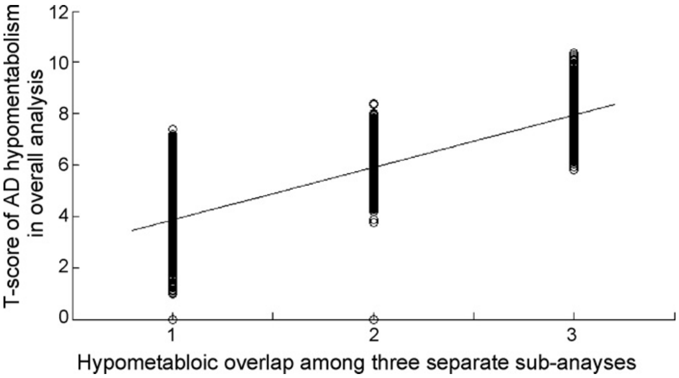

Fig. 3. The correlation between the degree of hypometabolic overlap out of the 3 separate sub-analyses and t-score in overall analysis.

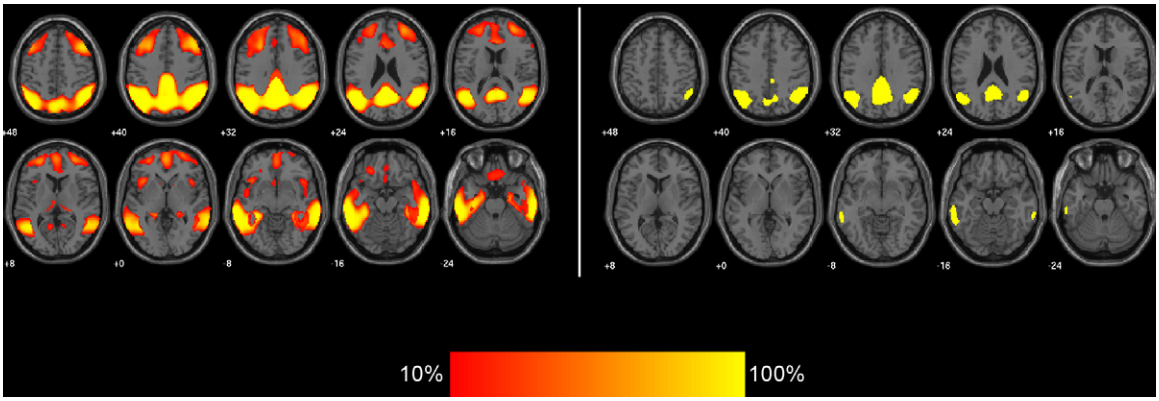

Fig. 4. The maps of hypometabolic voxels with detection frequency of ≥10% and 100% of the 1000 runs, respectively reliable at $p = 0.0005$ in pAD patients compared to NC using bootstrapping.

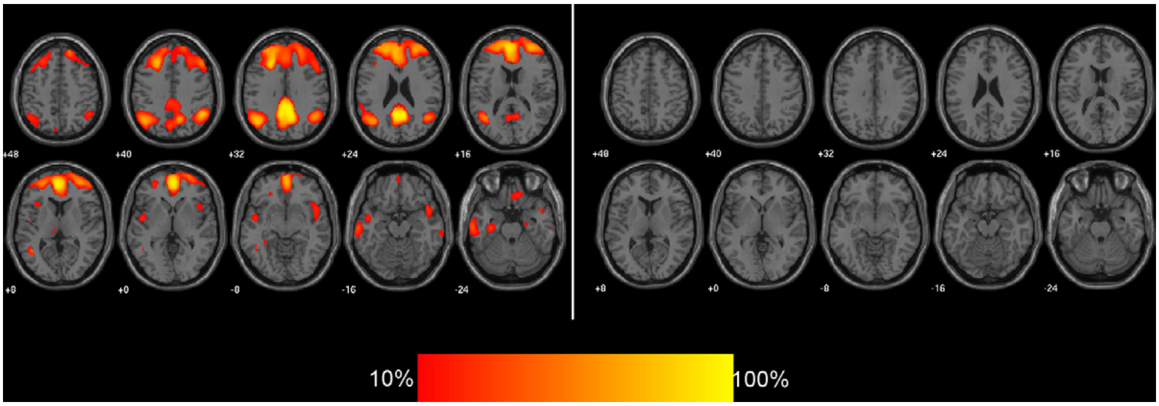

Fig. 5. The maps of hypometabolic voxels with detection frequency of ≥10% and 100% of the 1000 runs, respectively reliable at $p = 0.0005$ in aMCI patients compared to NC using bootstrapping.

and between aMCI and NC subjects. Various thresholds for type-I error, corrected and uncorrected, were used at each bootstrap iteration to examine if the CMRgl in patients with pAD or MCI was lower than that in NC at each voxel location. Fig. 4 shows the maps of voxels found to be hypometabolic 10% reliable or higher (left panel) and 100% reliable at $p = 0.0005$ in pAD patients compared to NC, illustrating the generally consistent findings for these different thresholds. At $p = 0.0005$, the number of times out of 1000 iterations a given voxel is detected as hypometabolic (i.e., CMRgl is lower in pAD than in NC for $p = 0.0005$ uncorrected for multiple comparisons) was calculated and recorded as intensity value of the voxel. The left and right panels of Fig. 4 display the maps of hypometabolic voxels with detection frequency of ≥10% and 100% of the 1000 runs, respectively. The collection of voxels detected as hypometabolic over the map with a detection frequency, f_1, was a subset of hypometabolic voxels over the map with a frequency f_2 which is lower than f_1.

Fig. 5 shows the hypometabolic detection frequency results for aMCI patients compared to NC. Note that for $p = 0.0005$, no voxel was detected as hypometabolic 100% over the 1000 runs.

We note that both p-values (FWE corrected and uncorrected) and their corresponding t-scores can be treated as an index of reliability. By definition, the more significant a given type-I error or the higher the t-score is, the less likely the detected group difference is a false positive; therefore such group difference is more reliable. Thus the numerical values of the reliability index assessed by the bootstrap procedure should be positively correlated with the t-scores and negatively correlated with p-values as shown in Fig. 6. However, this relationship is not linear, nor is it one-to-one (numerically). From Fig. 6, the mid portion of the curve depicting the relationship between the number of times a voxel is detected as hypometabolic and its t-score is gradual and monotonic (between the range of 50 and 950). At the lower and upper extremes of the curve, the increase is accelerating with more vertical appearance in plotting. In fact, for all t-scores ≥8, we have 100% reliability by the bootstrap procedure. The corresponding t-scores for FWE cor-

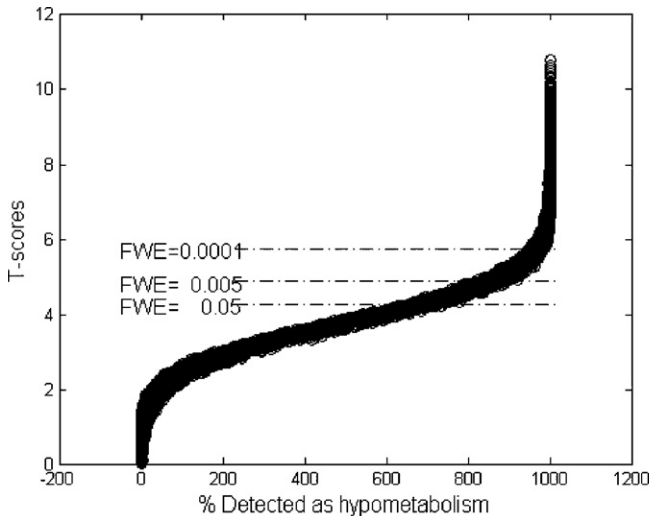

Fig. 6. The numerical values of the reliability index assessed by the bootstrap procedure are positively correlated with the t-scores and negatively correlated with p-values.

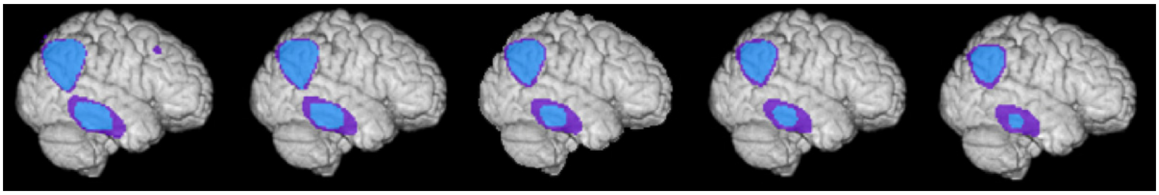

Fig. 7. The hypometabolic brain regions detected using bootstrap for 100% of the runs (in purple) and those detected using FWE (in blue) for the uncorrected $p = 0.01$, 0.005, 0.001,0.0005 and 0.0001, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

rected $p = 0.05$, $p = 0.005$ and $p = 0.0001$ are indicated by the dotted horizontal lines. Note that, for example, FWE of 0.005 does not correspond to 100% detection by the bootstrap resampling procedure with uncorrected $p = 0.005$ at each iteration. In fact, the bootstrap assessed reliability is 81.5% (815 out of 1000 runs) for FWE = 0.005. For further examination of the threshold of 0.005 used in bootstrap, the most liberal p-value for those voxels over which 100% overlap was observed is 2.7934e−07, more restricted than 1.24e−5 corresponding to FWE of 0.005. The hypometabolic regions defined using the more restricted p-value threshold should overlap precisely with the bootstrap defined 100% reliable hypometabolic regions. Also note that the 100% detection with the use of $p = 0.005$ threshold in each bootstrap iteration is with FWE $p = 0.001853$.

Fig. 7 shows the pattern of regional hypometabolism detected using bootstrap for 100% of the runs (in purple) and those detected using FWE (in blue) for the uncorrected $p = 0.01$, 0.005, 0.001, 0.0005 and 0.0001, respectively. The figure clearly demonstrates the more restricted nature of the p-value used in the bootstrap iteration compared to the FWE. As an example, the rightmost image shows the hypometabolic regions that survive FWE $p = 0.0001$ which is larger than the hypometabolic regions detected 100% of the times at $p = 0.0001$ threshold at each bootstrap iteration. Since the re-sampled datasets are not independent of each other by the virtue of the bootstrap resampling with replacement technique, the non-linear relationship with the t-score (or uncorrected or corrected p-value) with the bootstrap derived reliability index might partially due to the data dependency.

4. Discussion

In this study, we proposed the bootstrap resampling technique to assess the reliability of the detected hypometabolic brain regions in patients with pAD or aMCI in comparison to NC. We believe that this approach offers the potential to reliably detect either a statistical ROI or search region in cross-sectional studies, helping in the early detection of AD (e.g., the differential diagnosis of AD, the prediction of subsequent rates of cognitive decline, and the enrichment of clinical trials for those individuals most likely to have AD or show subsequent cognitive decline. It also has potential to detect either a statistical ROI or search region in longitudinal studies, helping to track AD and evaluate AD-modifying treatments with optimal statistical power and freedom from the type-I error associated with multiple comparisons (Chen et al., 2010).

Researchers have proposed using brain imaging techniques, such as FDG-PET and volumetric MRI, to evaluate AD-modifying treatments with better statistical power than clinical endpoints. Unlike the region of interest (ROI)-based method, the widely used voxel-based approach fully utilizes the richness of image-wise information. However, the latter approach has the issue of the inflated type-I error when multiple tests are performed simultaneously (i.e., multiple comparisons). We have recently shown how empirically derived statistical ROI's of CMRgl decline can be used to help optimize statistical power to evaluate AD-modifying treatments free of multiple comparisons (Chen et al., 2010). The method (using family-wise error in the previous publication or bootstrap with resampling [unpublished data] permits us to capitalize on the entire data set, without relying on preconceived ROIs. The bootstrap resampling technique allows us to assess the reliability of the observed hypometabolism and thus define a collection of *reliable* voxels as the functional ROI. Although we used cross-sectional data in the current study to introduce this new approach and compared it with the FWE approach, all the proposed procedures can be applied to longitudinal data to define functional ROI. Future work will focus on using the defined and validated functional ROI in power estimation. Moreover, we plan to define the functional ROI with 100% reliability via the bootstrap approach with one dataset and then perform the power analysis using an independent dataset. We believe our planned analyses fully utilize the richness of the neuroimaging data, while at the same time being free of multiple comparison concerns.

In applying this method to examine the longitudinal CMRgl decline, one should be also aware of the effects of the use of global counts as a reference by proportional scaling. Our own previous study showed the global CMRgl decline in AD patients (Alexander et al., 2002). The regional CMRgl decline relative to global over time could be artificially reduced due to the decrease of the global measurement for the follow up scans. In this study, we followed the conventional use of global counts the same as in a number of previous studies, aiming primarily to introduce and cross-validate the Bootstrap resampling based reliability index of the AD related hypometabolism for cross-sectional studies. In one recent separate but related study by our group (Chen et al., 2010), we examined the effects of using different reference regions for their sensitivity and statistical powers to characterize the 12-month CMRgl decline. The reference regions we examined in that study included the global counts, sensory-motor, thalamus, pontine, cerebellum and the relatively spared brain regions over the 12-month period. Results from that study showed that the use of the spared region gave the highest sensitivity in detecting the longitudinal declines. We note that additional studies are needed to examine the combined effects of using one of these reference regions, the degree of additional smoothing to the images, the threshold to define the decline and spared regions and other settings in conjunction with the generalization of the proposed reliability index defined by the Bootstrap resampling procedure to longitudinal data.

In the statistical inference stage of the neuroimaging data analysis, multiple comparison correction is conducted over a specified brain volume (the search volume is the whole brain by default). A volume chosen for multiple comparison correction is based on previous findings or expert knowledge about its involvement with the biological process of interest (such as hypometabolism in patients with AD or activation in response to a stimulus). Each location or voxel in the volume is treated equally in this multiple comparison correction protocol. We are interested in incorporating the differential involvement each location into the correction. For the AD study, for example, the differential involvement can be in the form of the assessed degree of reliability of the hypometabolic voxel. Voxels with lower reliability should have been given less weight than those with higher reliability in the correction. The present study was conducted with the goal of establishing a technique to estimate the variation in reliability. In future studies we will address the issue of how to integrate the estimated reliability into the multiple comparison correction.

It is important not to confuse the chance that a voxel is repeatedly observed as hypometabolic with the FWE. For instance, when using bootstrap resampling with a given uncorrected *p*-value such as 0.005, the corresponding FWE is high. In our analysis (given the brain mask and the smoothness applied), $P_{\text{FWE}} = 0.896$ for uncorrected $p = 0.005$. One may argue that given such high false positive rate (0.896), the chance to see the same group difference in three independent runs is as high as 0.7193 (=0.896^3). This reasoning, however, is inadequate since the random-field based FWE is the probability of existence of *at least one* brain location/voxel at which the magnitude is higher than the threshold corresponding to the given probability, this probability does not address the repeated observations at the *same* voxel locations over multiple analyses. When a location is observed as significant in the first analysis, the probability to observe significance at the same location again in an independent dataset, under the null hypothesis, is the uncorrected *p*-value and there is no need to correct for the multiple comparisons for this approach.

Please note that the reliability in the present study was estimated using repeated resampling with replacement from a single dataset. With the exact theoretical statistical nature yet to be fully understood, we caution its use and equating it with conventional perception of reliability. However, our bootstrap results were quite similar to that for each of the three non-overlapping data sub-sets. Moreover, we related the reliability index to the well-understood type-I error (FWE or uncorrected) with a positive and significant association between the two.

Apparently, the computational expense is high for the bootstrap resampling procedure compared to the generation of the statistical *parametric* maps. However, it is not impractical with the efficient batching procedure in place even for a modern personal computer. Both approaches can be used to define statistical region of interests for examining the cross-sectional group differences and for investigating the longitudinal declines in each subject group and the decline differences among different groups. Though the results of both approaches could be similar, results from Bootstrap approach could be easier to interpret (see more details below) and potentially with higher power as our on-going internal investigations suggested and as the possibility of the violation of the normal distribution assumption for the *parametric* approach. In that, we note that *p*-value threshold used in the bootstrap procedure was not for statistical inference purpose.

In place of the type-I error concept, we proposed the notion of reliability for potential power estimation and for a more precise multiple comparison correction taking the degree of confidence of prior knowledge into consideration. We believe this tactic is straightforward, intuitive, and can potentially be used for power estimation and for correction of multiple comparisons. One might think the use of $1 - \rho$ as reliability index (ρ is FWE) as reasoned in the current study relating FWE to the bootstrap defined reliability. FWE is defined under the null hypothesis of no group difference. Thus, the direct interpretation for $1 - \rho$ is the probability that there is no voxel/location in the examined volume that exceeds the threshold determined by ρ. With this strict and technical definition, relating to observation reliability needs abstract and difficult reasoning, a motivation for our current undertaking efforts.

Acknowledgments

The authors thank Patti Aguilar, David Branch, Sandra Goodwin, Debbie Intorcia, Jennifer Keppler, Les Mullen, Anita Prouty, Stephanie Reeder, Sydney Schilcher, Desiree Van Egmond, and Justin Venditti for their technical assistance.

This work was supported by the Key Program of National Natural Science Foundation of China (60931003) and National Natural Science Foundation of China (60905063, 60805040, and 90820019). It is also supported by the National Institute of Mental Health, US (R01 MH57899), the National Institute on Aging, US (9R01AG031581-10 and P30 AG19610, k23 AG24062) and the state of Arizona.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., and Wyeth, as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization 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, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

Alexander GE, Chen K, Pietrini P, Rapoport SI, Reiman EM. Longitudinal PET evaluation of cerebral metabolic decline in dementia: a potential outcome measure in Alzheimer's disease treatment studies. Am J Psychiatry 2002;159(5):738–45.

Chen K, Langbaum J, Fleisher A, Ayutyanont N, Reschke C, Lee L, et al. Twelve-month metabolic declines in probable Alzheimer's disease and amnestic mild cognitive impairment assessed using an empirically pre-defined statistical region-of-interest: findings from the Alzheimer's disease neuroimaging initiative. Neuroimage 2010;51:654–64.

Foster NL, Chase TN, Fedio P, Patronas NJ, Brooks RA, Di Chiro G. Alzheimer's disease: focal cortical changes shown by positron emission tomography. Neurology 1983;33:961–5.

Hoffman JM, Welsh-Bohmer KA, Hanson M, Crain B, Hulette C, Earl N, et al. FDG PET imaging in patients with pathologically verified dementia. J Nucl Med 2000;41(11):1920–8.

Ibanez V, Pietrini P, Alexander GE, Furey ML, Teichberg D, Rajapakse JC, et al. Regional glucose metabolic abnormalities are not the result of atrophy in Alzheimer's disease. Neurology 1998;50(6):1585–93.

Langbaum JBS, Chen KW, Lee W, Reschke C, Bandy D, Fleisher AS, et al. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). NeuroImage 2009;45:1107–16.

McGeer EG, Peppard RP, McGeer PL, Tuokko H, Crockett D, Parks R, et al. 18Fluorodeoxyglucose positron emission tomography studies in presumed Alzheimer cases, including 13 serial scans. Can J Neurol Sci 1990;17(1):1–11.

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34(7):939–44.

Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, et al. Current concepts in mild cognitive impairment. Arch Neurol 2001;58(12): 1985–92.

Reiman EM, Caselli RJ, Yun LS, Chen K, Bandy D, Minoshima S, et al. Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E. New Engl J Med 1996;334(12):752–8.

Reiman EM, Caselli RJ, Chen K, Alexander GE, Bandy D, Frost J. Declining brain activity in cognitively normal apolipoprotein E epsilon 4 heterozygotes: a foundation for using positron emission tomography to efficiently test treatments to prevent Alzheimer's disease. Proc Natl Acad Sci U S A 2001;98(6): 3334–9.

Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, et al. Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. Proc Natl Acad Sci U S A 2004;101(1):284–9.

Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, et al. Correlations between apolipoprotein E epsilon4 gene dose and brain-imaging measurements of regional hypometabolism. Proc Natl Acad Sci U S A 2005;102(23):8299–302.

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