Published in final edited form as:

Neuroimage. 2011 April 1; 55(3): 933–941. doi:10.1016/j.neuroimage.2010.12.070.

Inter-rater reliability of manual and automated region-of-interest delineation for PiB PET

Bedda L. Rosario^a, Lisa A. Weissfeld^b, Charles M. Laymon^a, Chester A. Mathis^a, William E. Klunk^c, Michael D. Berginc^a, Jeffrey A. James^a, Jessica A. Hoge^a, and Julie C. Price^a

^a Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian Hospital, 200 Lothrop Street, Pittsburgh, PA 15213, USA

^b Department of Biostatistics, University of Pittsburgh, Graduate School of Public Health, 130 DeSoto Street, Pittsburgh, PA 15261, USA

^c Department of Psychiatry, University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic, 3811 O'Hara Street, Pittsburgh, PA 15213

Abstract

A major challenge in positron emission tomography (PET) amyloid imaging studies of Alzheimer's disease is the reliable detection of early amyloid deposition in human brain. Manual region-of-interest (ROI) delineation on structural magnetic resonance (MR) images is generally the reference standard for the extraction of count-rate data from PET images, as compared to automated MR-template(s) methods that utilize spatial normalization and a single set of ROIs. The goal of this work was to assess the inter-rater reliability of manual ROI delineation for PiB PET amyloid retention measures and the impact of CSF dilution correction (CSF) on this reliability for data acquired in elderly control (n=5) and AD (n=5) subjects. The intraclass correlation coefficient (ICC) was used to measure reliability. As a secondary goal, ICC scores were also computed for

Corresponding author: Bedda L. Rosario, PhD, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, rosariobl@upmc.edu, Phone Number: (412) 605-1568, Fax: (412) 647-0700.

Lisa A. Weissfeld, PhD, Department of Biostatistics, University of Pittsburgh, Graduate School of Public Health, 130 DeSoto Street, 326 Parran Hall, Pittsburgh, PA 15261, lweis@pitt.edu

Charles M. Laymon, PhD, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, laymoncm@upmc.edu

Chester A. Mathis, PhD, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, mathisca@upmc.edu

William E. Klunk, MD, PhD, Department of Psychiatry, University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic Room 1422, 3811 O'Hara Street, Pittsburgh, PA 15213, klunkwe@upmc.edu

Michael D. Berginc, BS, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, bergincm@upmc.edu

Jeffrey A. James, BS, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, jamesja2@upmc.edu

Jessica A. Hoge, MD, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, jessicahoge@gmail.com

Julie C. Price, PhD, Department of Radiology, University of Pittsburgh School of Medicine, Presbyterian University Hospital, B-938, 200 Lothrop Street, Pittsburgh, PA 15213, pricejc@upmc.edu

GE Healthcare holds a license agreement with the University of Pittsburgh based on the technology described in this manuscript. Drs. Klunk and Mathis are co-inventors of PiB and, as such, have a financial interest in this license agreement. GE Healthcare provided no grant support for this study and had no role in the design or interpretation of results or preparation of this manuscript. All other authors have no conflicts of interest with this work and had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

PiB outcome measures obtained by an automated MR-template ROI method and one manual rater; to assess the level of reliability that could be achieved using different processing methods. Fourteen ROIs were evaluated that included anterior cingulate (ACG), precuneus (PRC) and cerebellum (CER). The PiB outcome measures were the volume of distribution (V_T), summed tissue uptake (SUV), and corresponding ratios that were computed using CER as reference (DVR and SUVR). Substantial reliability (ICC = 0.932) was obtained across 3 manual raters for V_T and SUV measures when CSF correction was applied across all outcomes and regions and was similar in the absence of CSF correction. The secondary analysis revealed substantial reliability in primary cortical areas between the automated and manual SUV [ICC = 0.979 (ACG/PRC)] and SUVR [ICC = 0.977/0.952 (ACG/PRC)] outcomes. The current study indicates the following rank order among the various reliability results in primary cortical areas and cerebellum (high to low): 1) V_T or SUV manual delineation, with or without CSF correction; 2) DVR or SUVR manual delineation, with or without CSF correction; 3) SUV automated delineation, with CSF correction; and 4) SUVR automated delineation, with or without CSF correction. The high inter-rater reliability of PiB outcome measures in primary cortical areas (ACG/PRC) is important as reliable methodology is needed for the detection of low levels of amyloid deposition on a cross-sectional basis and small changes in amyloid deposition on a longitudinal basis.

Keywords

Region-of-interest; manual delineation; automated delineation; PiB; PET; atrophy correction; partial volume correction; reliability; intra-class correlation coefficient

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by progressive deterioration of memory and cognitive function that may be accompanied by behavioral symptoms (Nichols et al., 2006). Currently, the definitive diagnosis of AD is based on the presence of amyloid-beta ($A\beta$) plaques and neurofibrillary tangles in cortical regions in the brain at autopsy. Positron emission tomography (PET) imaging is being widely applied for the *in vivo* assessment of $A\beta$ plaque deposition in humans with imaging of cognitively unimpaired normal control (NC), mild cognitive impairment (MCI) and AD subjects. Amyloid-beta deposition can be quantified in the brain using the PET radiotracer Pittsburgh Compound B (PiB), a ^{11}C -labeled thioflavin-T derivative. PiB PET has shown cortical retention in AD and MCI subjects that can exceed NC levels by 2–3 fold (Kempainen et al., 2007; Klunk et al., 2004; Rowe et al., 2007) and significant retention can be detected in 25–40% of NCs (Lopresti et al., 2005; Mintun et al., 2006; Morris et al., 2010; Pike et al., 2007; Rowe et al., 2007). Currently, a major challenge is the reliable detection of the earliest stages of amyloid deposition, ideally before the onset of clinical symptoms, as well as the robust assessment of longitudinal change in the context of disease progression and/or response to therapy.

PET data analysis can provide physiological information for an anatomically relevant region-of-interest (ROI) or volume element (voxel) within the image. As the term implies, manual ROIs are hand-drawn by trained raters on two-dimensional structural MR images that have been co-registered to the corresponding functional PET images. Manual ROI delineation is often considered as the reference standard ROI approach. However, challenges associated with manual ROI delineation include achievement of high reliability across human raters despite intra- and inter-subject variations in anatomical landmarks, and labor-intensive requirements. Additional challenges that arise in the quantification of specific PiB PET retention in brain include non-specific uptake in white matter, CSF dilution of the ROI signal as a result of age- or neurodegeneration-related cerebral atrophy or variations in

manual ROI delineation, as well as determination of the “null” signal in control subjects in the presence of background image noise.

Automated ROI sampling approaches are often needed for efficient and standardized sampling of large data sets and/or multi-center investigations and/or longitudinal follow-up. Several methods have been proposed for automated ROI sampling of PET data (such as, Rusjan et al., 2006; Svarer et al., 2005; Yasuno et al., 2002). Challenges associated with automated ROI delineation include selection of image template, adequate performance of normalization procedure for the particular population of interest (e.g. age, gender or disorder specific), performance of automated method in areas with more individual variation (e.g., sub-cortical regions) and potential bias introduced by pre-processing (e.g. spatial normalization) of the brain images. Furthermore, variations in age- and disease-related atrophy may also influence the reliability of ROI delineation for both manual and automated methods.

The reliability of ROI delineation can be assessed using the intraclass correlation coefficient (ICC, McGraw and Wong, 1996; Shrout and Fleiss, 1979) that is a widely used descriptive statistic to assess reliability across raters for quantitative data. Reliability measures the consistency of a set of measurements; that is, a test is considered reliable if we obtained the same result repeatedly under similar methodology. The reliability of ROI delineation for PET studies has been assessed for several radiotracers and the regional reliability achieved generally ranges from 0.80 to 0.99 in cortical and sub-cortical areas, with automated reliability being similar or less than that measured with manual ROI delineation (Lindgren et al., 1999; Rusjan et al., 2006; Schaefer et al., 2000; Small et al., 1992).

The primary aim of the current study was to assess the inter-rater reliability of manual ROI delineation for the generation of regional PiB PET retention measures; and the extent to which this reliability is impacted by atrophy-related correction for CSF dilution (CSF) in elderly subjects. In addition, the reproducibility of the manual ROI delineation was evaluated relative to an automated MR-template based approach for which a single set of ROIs was used to sample PiB retention images across all subjects. The automated MR-template (MCI template) approach was used for automated ROI sampling for the Alzheimer's Disease Neuroimaging Initiative (ADNI) PiB PET data (www.loni.ucla.edu/ADNI; Jagust et al., 2009, 2010). The reliability of both methods (manual and automated) was tested in AD and NC subjects. The motivation for this work lies in ongoing and future research, in aging and neurodegeneration, that relies on ROI delineation to detect early amyloid deposition and/or changes in deposition that can occur over time and in response to therapy.

2. Methods

2.1 Human Subjects

A subset of ten subjects, five cognitively unimpaired NC [78 ± 3 yrs (range 75–81 yrs), 28 ± 1 MMSE, 4M:1F] and five AD [68 ± 11 yrs (range 54–77 yrs), 22 ± 4 MMSE, 4M:1F], was randomly selected from a larger data base of 31 PiB PET studies (11 AD, 20 NC) for which arterial blood sampling, PiB PET and MR imaging were performed. Data for some or all of these ten subjects have also been included in previous PiB PET evaluations (Aizenstein et al., 2008; Cohen et al., 2009; Lopresti et al., 2005; Price et al., 2005; Wolk et al., 2009). All subjects or caregivers provided informed consent to participate in the study in accordance with the local Institutional Review Board.

2.2 Neuroimaging

Structural MR images (1.5 Tesla G.E. Signa) were acquired in the coronal plane using a volumetric spoiled gradient recall (SPGR) sequence, as previously described (Price et al., 2005). The SPGR MR images were used for MR/PET co-registration, ROI definition, correction for CSF dilution and spatial normalization.

PiB radiosynthesis was performed using a simplified method (Wilson et al., 2004). An intravenous catheter was inserted into the radial artery for arterial blood sampling. Head motion was minimized using an immobilization device. PiB was injected intravenously (14.3 ± 2.2 mCi, 1.4 ± 0.8 Ci/ μ mol) as a slow bolus over 20 sec. PET scanning was performed using a Siemens/CTI ECAT HR+ scanner (3D mode, 15.2 cm field of view, 63 planes) with 34 dynamic frames acquired over 90 min. Thirty five hand-drawn samples (0.5 mL) were collected over 90 min (20 samples over first 2 min). Additional blood samples were collected at 5 to 6 times during the scan for the measurement of radiolabeled metabolites. PET data were reconstructed by filtered back-projection with correction for attenuation, scatter and radioactive decay (final image resolution ~ 6 mm FWHM). Additional imaging details can be found in earlier works (Aizenstein et al., 2008; Lopresti et al., 2005; Price et al., 2005).

2.3 Image Co-registration and ROI Delineation

2.3.1 Manual ROI Delineation—Prior to image co-registration, voxels corresponding to scalp and calvarium were manually stripped from the SPGR MR images. The alignment process began with a reorientation of the MR along the anterior-posterior commissure (AC-PC line). The image alignment was performed using an automated method for centering (Minoshima et al. 1992) and alignment and re-slicing (Woods et al. 1993; Automated Image Registration, AIR, Version 3.0). The PET image data were aligned to the SPGR MR data to yield registration parameters. The MR data were then re-sliced using the inverse transformation of the PET-to-MR alignment to match PET image space ($128 \times 128 \times 63$, pixel size = $2.06 \times 2.06 \times 2.43$ mm). The data were visually inspected for subject motion and inter-frame motion was corrected by applying a more extensive registration procedure on a frame-by-frame basis.

ROI placement was based on guidance criteria developed locally that is not wholly anatomical and relies to some extent on our in-house imaging capabilities and software. The criteria do refer to anatomical landmarks, number of contiguous planes for ROI definition and physical ROI shape but exact region shapes, sizes and number of contiguous planes are generally ultimately determined by the individual's anatomy. Manual ROIs were generated for fourteen regions by three raters (JAH, MDB and JAJ) on each co-registered MR image. The ROIs included anterior cingulate (ACG, pregenual and subgenual, 10 planes), anterior ventral striatum (AVS, caudate and anterior putamen, 5 planes), cerebellum (CER, anterior plane (see Figure 1) near the top of fourth ventricle, 3 planes), frontal cortex (FRC, dorsal and ventral, 10 planes), lateral temporal cortex (LTC, 5 planes), mesial temporal cortex (MTC, amygdala and hippocampus, 2 planes), occipital cortex (OCC, primary visual cortex, 5 planes), occipital pole (OCP, 3–5 planes), parietal cortex (PAR, 5 planes), precuneus cortex (PRC, 15 planes), pons (PON, 5 planes), sensory motor cortex (SMC, 4 planes), sub-cortical white matter (WM, 2 planes), and thalamus (THL, 3 planes). The CER ROIs were drawn to reduce sampling of high atrophy areas, minimize white matter uptake, and limit sampling of the more inferior slices within the PET field-of-view where the 3D sensitivity of the tomograph is lower. A “Global cortical” PiB retention measure was also computed across five cortical ROIs (ACG, FRC, LTC, PAR, PRC). Pictorial examples of the ROI placement are provided in Figure 1 and previous publications (Lopresti et al., 2005; McNamee et al., 2009; Price et al., 2005). The ROIs were used to sample the co-

registered PET images to obtain regional time-activity data using a modified version of ROIttool (CTI PET Systems, Knoxville, TN, USA).

2.3.2 Automated ROI Delineation (MCI Template)—A single-subject MR image template (MCI template) was used for automated ROI sampling. For this method, the fourteen ROIs listed above were hand-drawn (Rater 2, MDB) on a high-resolution MR image of a single elderly MCI subject (age: 79 years with mild atrophy and ventricular enlargement) who was scanned at the University of Pittsburgh PET facility and thoroughly studied with PiB PET on a manual ROI basis. The same criteria used for manual ROI definition were used to guide ROI delineation for the MCI template (the automated CER ROIs were drawn over 7 planes), although ROI sizes were somewhat smaller and more elliptical for general application across varying individual anatomy (Svarer et al., 2005).

Each subject's native PET and MR image data were co-registered using Statistical Parametric Mapping (SPM8) software. The individual's co-registered MR scan was then spatially normalized to the MCI template using the SPM8 Normalize option, with default parameter settings (Ashburner and Friston, 1999). All normalized images were written out using a bounding box of $-90:90$, $-126:90$, $-72:108$. These transformation parameters were used to normalize the PET image data to the MCI template. The PET data were then re-sliced to the same dimensions ($146 \times 177 \times 152$, pixel size = $1 \times 1 \times 1$ mm) as the MCI template.

2.4 Data Analysis and Outcome Measures

The PiB retention outcomes that were evaluated were based on both the Logan graphical analysis volume of distribution (V_T) and the summed tissue uptake (or standardized uptake value, SUV) measures. The Logan analysis utilized metabolite-corrected arterial data as input (Logan et al., 1990) and the regression was performed using 8 points over the 40–90 min post-injection integration intervals. The SUVs were determined over 50–70 min (SUV70) post injection intervals and normalized to injected dose and body mass. For the manual ROI data, volume of distribution ratios (DVR) and summed tissue uptake value ratios (SUVr) were generated using cerebellum as reference. The automated template-based ROI sampling was performed only on summed PiB PET SUV uptake and SUV ratio (CER as reference) images.

A 2-component MR-based approach was used to adjust the regional PET measures for the dilutional effects of expanded CSF spaces (Meltzer et al., 1990). This approach has been applied for over 10 years in our laboratory in PET studies of cerebral blood flow, neuroreceptor binding and PiB binding (Bailer et al., 2007; Cohen et al., 2009; Fisher et al., 2009; Lopresti et al., 2005; Meltzer et al., 1999, 2000; Price et al., 2005) to correct for dilution effects of CSF on quantification of binding measures in brain tissue. Note that this correction can apply to anatomical variations that may be associated with normal aging- and disease-related cerebral atrophy, as well as to variations in the amount of CSF included in the manual ROI delineation. For each slice, an intensity-based histogram is generated that provides good separation of brain (i.e., gray and white matter) and non-brain (i.e., CSF) voxels. The MR voxel intensities are fit to two Gaussian distributions and these results are used to segment the MR image into brain and non-brain components. The segmentation thresholds correspond to the mid-point between peak intensities for brain and non-brain. A binary image is created by assigning brain voxels a value of unity and non-brain voxels a value of zero. The resulting binary image is convolved with a Gaussian smoothing kernel that is equivalent to the point-spread function of the PET scanner. For regional measures, the binary image is sampled using each subject's ROIs. A voxel average CSF correction is obtained for each ROI that varies from 0 to 1, where 0 corresponds to only CSF. The regional PiB retention measure is divided by this correction factor to generate a final

measure that reflects retention in brain tissue (i.e., gray and white matter only) in that region. For a given regional retention measure (RM), the correction was applied as follows: 1) $RM_{ROI-CORR} = RM_{ROI-RAW}/CSF_{ROI}$ and 2) $RM_{RATIO-CORR} = [RM_{ROI-RAW}/CSF_{ROI}] / [RM_{CER-RAW}/CSF_{CER}]$, where RM could be V_T or SUV. Outcomes corrected for CSF dilution will be denoted as V_{T-CSF} , SUV_{CSF} , DVR_{CSF} and $SUVR_{CSF}$.

2.5 Statistical Analysis

Descriptive statistics, including mean and coefficient of variation ($CV\% = SD/Mean \cdot 100$) were computed for CSF factor values and regional PiB outcome measures. The inter-rater reliability for all outcome measures, with and without CSF correction, was computed as an intraclass correlation coefficient (ICC, McGraw and Wong, 1996) using a two-way mixed effects model for absolute agreement [$ICC_A = (MS_R - MS_E)/(MS_R + (k-1)MS_E + k/n(MS_C - MS_E))$] and consistency [$ICC_C = (MS_R - MS_E)/(MS_R + (k-1)MS_E)$] where n is the number of subjects, k is the number of raters, MS_R is the mean square for rows/subjects, MS_E is the mean square due to error, and MS_C is the mean square for columns/raters. An absolute agreement definition refers to the fact that the ratings do not differ in absolute value. A consistency definition refers to the extent that an additive transformation serves to equate the ratings/values across raters, retaining subject ranking ratings. The difference between absolute agreement and consistency definition is that the rater variance [$k/n(MS_C - MS_E)$] is excluded from the denominator variance for the consistency definition but not for the absolute agreement definition. Perfect absolute agreement between two measurements would yield a slope of 1.0 and a y-intercept of 0.0; perfect consistency would yield a slope of 1.0 but the y-intercept could be any number. An ICC score ranges from 0 to 1, with the score reflecting the level of agreement as follows: virtually none (0.00, 0.10); slight (0.11, 0.40); fair (0.41, 0.60); moderate (0.61, 0.80); and substantial (0.81, 1.00) (Shrout, 1998). Confidence intervals were also calculated for each ICC score. To test if the reliability was sensitive to CSF dilution correction, the Wilcoxon signed rank test ($\alpha = 0.05$, 2-sided, exact inference) was used to compare reliability with and without correction for CSF dilution. All analyses were completed using SPSS statistical software (Version 17, SPSS Inc., Chicago, IL).

3. Results

3.1 CSF Dilution Correction Factors

The descriptive statistics for CSF dilution correction factors are summarized in Table 1 by group and rater. For manual delineation, the regions with greatest CSF dilution for AD subjects were ACG, FRC and PRC, while ACG and FRC were greatest for NC subjects. There was slightly more CSF dilution in the CER region for NC subjects (~2%). The largest differences for correction for CSF dilution between the manual raters were seen for FRC and PRC for the AD subjects (~4–5%) and for ACG and OCP for NC subjects (~3–4%). For automated delineation, the regions with greatest CSF dilution for AD subjects were ACG, MTC, PRC and SMC, while ACG, MTC and SMC were greatest for NC subjects.

3.2 Manual ROI Delineation

The descriptive statistics for the PiB PET outcome measures for five primary cortical ROIs, Global cortical and CER are listed by group, rater, correction for CSF dilution, and ROI delineation method, in Table 2. Statistics for the remaining 8 regions will be limited to the range of the mean values determined by the 3 raters (no CSF correction). For AD V_T , AVS: 7.35–7.50, MTC: 4.54–4.71, OCC: 6.40, OCP: 6.48–6.67, SMC: 6.35–6.59, PON: 5.43–5.46, SWM: 5.64–5.71, THL: 5.87–5.93. For AD SUV70, AVS: 1.48–1.50, MTC: 0.88–0.91, OCC: 1.18–1.19, OCP: 1.24–1.27, SMC: 1.20–1.23, PON: 1.12–1.13, SWM: 1.11–1.15, THL: 1.19–1.21. For NC V_T , AVS: 3.76–3.83, MTC: 3.33–3.36, OCC: 4.04–4.11,

OCP: 4.01–4.13, SMC: 3.79–3.88, PON: 5.26–5.28, SWM: 4.96–5.16, THL: 4.20–4.40. For NC SUV70, AVS: 0.90–0.91, MTC: 0.86, OCC: 0.93–0.95, OCP: 0.94–0.98, SMC: 0.93–0.97, PON: 1.43–1.44, SWM: 1.36–1.37, THL: 1.02–1.09.

3.2.1 Distribution Volume (V_T)—PiB retention was as much as 118% greater for AD subjects when compared to NC subjects (e.g. ACG). The variability for CER in terms of V_T values was higher for NC subjects than for ADs. When comparing the V_T values, with and without correction for CSF dilution, the regions with the highest mean percent difference for AD subjects were the ACG, FRC, PRC and the SMC, and for NCs these regions were the ACG, FRC and the SMC. The variability remained the same or decreased (~1–44%) without CSF correction for AD subjects in all regions except CER and PON. In contrast, the variability for NC subjects remained the same or decreased (~4–30%) without correction for CSF dilution in all regions except CER, ACG, OCC and OCP.

The reliability (i.e. ICC_A) and 95% confidence intervals for regional V_{T-CSF} values are summarized in Figure 2. The ICC_A scores for V_{T-CSF} values ranged from 0.934 to 0.999. Without correction for CSF dilution, the lowest and highest ICC_A for V_T values were for the WM ($ICC=0.922$) and the ACG, LTC, and the PRC ($ICC=0.999$), respectively. The reliability for several cortical regions, namely ACG, LTC, OCC, OCP, and PRC, and CER increased in the absence of CSF dilution correction. No statistically significant difference in reliability was observed for V_T values with and without correction for CSF dilution (Wilcoxon Signed Rank Test, $p=0.615$). As a result of the small sample sizes, only a preliminary examination of V_T ICC agreement scores could be performed for the separate NC and AD subject groups. A high level of agreement ($ICC_A = 0.989$) was evident in ACG and PRC for both subject groups but somewhat lower values were found for AD subjects in CER ($ICC_A = 0.825$) and MTC ($ICC_A = 0.845$), relative to controls ($ICC_A = 0.976$).

3.2.2 Standardized Uptake Value (SUV)—PiB retention was as much as 78% greater for AD subjects when compared to NC subjects (e.g. ACG). The inter-subject variation was greater for SUV70 measures than for V_T measures for AD subjects, when correction for CSF dilution was applied. Similar to V_T values, when comparing with and without correction for CSF dilution, the regions with the highest mean percent difference for SUV70 measures for AD subjects were the ACG, FRC, PRC and the SMC, and for NCs these regions were the ACG, FRC and the SMC. The variability for SUV70 measures remained the same or decreased (~1–39%) in the absence of CSF dilution correction for AD subjects, with the exception of CER (increase ~4–7%). For NC subjects, the variability remained the same or increased (~2–22%) in the absence of CSF dilution correction, with the exception of LTC, MTC, PRC and SMC (decrease ~3–25%).

The ICC_A reliability scores for the regional SUV70_{CSF} values are shown in Figure 2. The ICC_A scores for SUV70_{CSF} ranged from 0.969 to 0.999. The ICC_A for several cortical regions, namely the ACG, OCC, OCP and PRC, increased in the absence of CSF dilution correction. No statistically significant difference in reliability was found for SUV70 values with and without correction for CSF dilution (Wilcoxon Signed Rank Test, $p=0.457$).

3.2.3 Distribution Volume Ratio (DVR) and Standardized Uptake Value Ratio (SUVR)—Regional DVR and SUVR70 values for manual ROI delineation were then generated using the cerebellum as reference region. The V_T and SUV70 outcome measures were divided by measure-specific manual CER V_T and CER SUV70 value to obtain DVR or SUVR70 values that are commonly used PiB retention measures. An ICC analysis showed substantial agreement and consistency ($ICC = 0.831$) for the manual ratio outcome measures with correction for CSF dilution (DVR_{CSF} and SUVR70_{CSF}) using CER as reference. The ICC_A for DVR_{CSF} values ranged from 0.906 to 0.991. In the absence of CSF correction, the

ICC_A scores for DVR values reached 0.900 or higher for all regions except MTC (ICC=0.882) and WM (ICC=0.889). The ICC_A scores for ACG, OCP, PRC, SMC and PON increased in the absence of CSF dilution correction. No statistically significant difference in reliability was found for DVR values with and without correction for CSF (Wilcoxon Signed Rank Test, $p=0.506$). The ICC_A SUVR70_{CSF} values ranged from 0.831 to 0.986. In the absence of CSF dilution correction, the minimum and maximum ICC_A score for SUVR70 were obtained for the MTC (ICC=0.794) and the AVS (ICC=0.983). The ICC_A scores for SUVR70 for several regions, namely ACG, OCP, PAR, PRC, SMC and PON, increased in the absence of CSF dilution correction. No statistically significant difference in reliability was found for SUVR70 values with and without correction for CSF (Wilcoxon Signed Rank Test, $p=0.576$).

3.3 Automated ROI Delineation versus Single Manual Rater

The average SUV70 values for the Global ROI for the automated method, with and without CSF correction, are shown in Table 2. The reliability (absolute agreement and consistency) for all ROIs in terms of SUV70 values comparing one independent rater (Rater 2, MDB) and the automated method as a rater is summarized in Table 3. Results for SUVR70_{CSF} show substantial agreement (ICC_A > 0.871) for all regions between SUVs derived from the manual and automated methods for all regions except for OCP (ICC_A = 0.803). Without correction for CSF dilution, the ICC_A scores show substantial agreement (ICC_A > 0.868) for all regions except for OCP and THL (moderate agreement, ICC_A > 0.783). In addition, there is substantial consistency (ICC_C > 0.863) for SUVR70 measures between the manual and automated methods for all regions except OCP (ICC_C > 0.783). The reliability for all regions decreased or remained the same in the absence of CSF dilution correction. A statistically significant difference in reliability was observed for SUVR70 values with and without correction for CSF dilution (Wilcoxon Signed Signed Rank Test, $p=0.001$).

Regional SUVR70 values were then generated using the cerebellum as reference region. An ICC analysis of the automated and manual SUVR70_{CSF} outcomes showed substantial levels of agreement and consistency in primary cortical areas, but the ICC scores were generally lower overall and more variable across all regions when compared to ICC scores for SUVR70_{CSF}. The agreement ranged from 0.964 in ACG/PRC to 0.403 in OCP, as shown in Table 4. The ICC_A scores for FRC, OCC and PRC decreased in the absence of CSF dilution correction. No statistically significant difference in reliability was found for SUVR70 values with and without correction for CSF dilution (Wilcoxon Signed Signed Rank Test, $p=0.046$).

4. Discussion

The in vivo measurement of amyloid beta deposition through PiB PET imaging is ongoing in numerous single- and multi-site studies that utilize cross-sectional and longitudinal study designs to better understand Alzheimer's disease progression, preclinical amyloid beta deposition, and normal aging. However, no prior studies have reported on the reproducibility of the manual ROI delineation method for PiB PET studies. Our experience with PiB PET analyses has been based on hand-drawn ROIs with subsequent CSF correction of the PiB outcomes. This has been the preferred method at our center, although it is common not to apply CSF correction in other PiB PET studies. The current work reports on a step-wise assessment of the reliability of the manual ROI outcomes by including a reliability comparison between more quantitative outcomes and simpler semi-quantitative summed tissue uptake measures (with and without CSF correction). In addition, we assessed the inter-rater reliability that can be achieved when methods are based on different processing streams by comparing our standard manual method and a more recent automated method (MCI MR template, with and without CSF correction).

The main findings in our study indicate that there is a high level of agreement between manual ROI analyses by the three independent raters. For each ROI, the ICC_A scores for the PiB V_{T-CSF} and SUV_{CSF} measures were in the range of 0.934–0.999, and suggest that the manual method is a highly reliable method for defining ROIs. Additionally, the reliability was greater than 0.980 for regions that can exhibit high PiB PET retention levels and usually show amyloid deposition earliest in NC subjects (e.g. ACG, FRC and PRC) (Aizenstein et al., 2008; Mintun et al., 2006). The reliability for manual ROI delineation for V_T and SUV was not influenced by the absence of CSF correction.

Even though the cerebellum and sub-cortical white matter regions showed substantial reliability (ICC_A > 0.922) for V_{T-CSF}, V_T, SUV_{CSF} and SUV, the decrease in reliability compared to other regions may be partly explained by low signal, highly variable atrophy (cerebellum) and highly variable spillover (white matter). This is particularly concerning for cerebellum since it is often the region used as the reference for all other regions in ratio methods. Despite this, substantial-to-moderate agreement was observed for the manual ratio outcome measures with and without CSF correction using CER as reference. For manual ratio outcome measures (DVR_{CSF} and SUVR_{CSF}), the ICC_A scores for each ROI were in the range of 0.794–0.991. The reliability for manual ROI delineation for ratio outcome measures was not influenced by the absence of CSF correction.

The goal of the CSF correction is to yield an outcome measure that reflects retention in brain tissue within a region-of-interest, in the absence of CSF dilution. In this manner, the CSF correction could provide a more accurate assessment of radioligand binding in brain tissue. While results suggest that CSF dilution correction (sometimes called atrophy correction or partial volume correction) is not necessary to obtain a reliable inter-rater reliability for manual delineation, correction for variable CSF dilution may be necessary to achieve accurate determination of the in vivo amyloid load per unit of brain tissue. That is, the importance of the CSF correction is very apparent when comparing subjects with low-to-moderate gray matter loss to those with substantial gray matter loss and CSF signal dilution. For the latter cases, the regional PiB PET retention measure can be very low because of the CSF dilution.

These findings imply that studies that require manual ROI delineation can be highly reliable. The comparison of this work with other reliability studies (Lindgren et al., 1999; Rusjan et al., 2006; Schaefer et al., 2000; Small et al., 1992) is complicated by the use of different tracers, ROIs and subject groups, but further indicates the high level of reproducibility that can be achieved using manual ROI delineation.

Comparison of manual versus automated (MCI template) methods revealed ICC scores for absolute agreement for SUV_{70CSF} and SUV₇₀ measures in the range of 0.803–0.997 and 0.783–0.996, respectively. The ICC scores for agreement for SUVR_{70CSF} and SUVR₇₀ measures were similar but more variable across regions with ranges of 0.403–0.964 and 0.482–0.977, respectively. That is, the agreement between manual and automated methods for ROI delineation in terms of SUV₇₀ was to some extent less than the agreement across the three different manual raters. Further assessment of the results show that the decrease in reliability in terms of absolute agreement, when the automated method is considered as a rater, was influenced by the higher between-rater variability. That is, the decrease in absolute agreement was a result of taking into account the raters' variance. As an example, when comparing the manual and automated methods the rater variance was approximately 2-fold higher than for the manual method (3 raters) for the CER. The differences in inter-rater reliability that could potentially arise from basic differences in the automated and manual methods are: 1) automated method: use of MR-template and associated ROIs, spatial normalization, and subsequent sampling of summed PET image; 2) manual sampling: use of

individual's native MR to generate ROIs, sampling of regional PET time-activity data, and generation of summed outcome measures. As mentioned above, correcting for CSF dilution did not influence the reliability of manual ROI delineation. In contrast, the reproducibility of the automated method for SUV70 measures was influenced by the absence of CSF correction. The reliability for SUV70_{CSF} measures was higher than for SUV70 measures.

It is important to note potential limitations of the current study with respect to the generalization of the results. A major limitation of the present work is the small sample size that may bias the inter-rater reliability results. Although we only utilized 10 subjects, the inter-rater reliability was quite high for both manual and automated methods, particularly in primary cortical areas. We think these findings are encouraging and that a larger sample would likely confirm the present results, rather than provide substantially different results. Nonetheless, we realize that the inclusion of more subjects is needed for our sample to be more representative of the neuroanatomical variability that is inherent in an elderly subject sample of this type. We believe that the current subject group provides a reasonable initial sampling of such variation given that the subjects were randomly chosen from a larger subject pool of elderly control and AD subjects. For the automated method, based on a single representative subject, it is important to note that there are more current and widely used methods for spatial normalization and automated ROI template sampling that could have been applied in this work, including the SPM5/8 unified normalization method (Ashburner and Friston, 2005), a probability template approach (Svarer et al., 2005) and automated anatomical labeling (Tzourio-Mazoyer et al., 2002). It is possible that the spatial normalization algorithm did not perform at the level required to achieve high agreement between the automated and manual results across all regions of the elderly diseased brain. Previous examination of the default parameters within the SPM Normalize option performed by our group on aging brains indicated that the parameter of importance was the degree of regularization (Rosario et al. 2008). Based on this, we examined different degree of regularization and all other parameters as default. Results suggest that this parameter can affect the inter-rater reliability, specifically for high variable regions such as CER, OCP, and THL (Results not shown). In addition, it is possible that differences in co-registration methods for the manual and automated delineation methods might affect the reliability results. It has been shown that the co-registration algorithm in SPM does not have the highest precision when compared to the co-registration algorithm in AIR software (West et al., 1997). However, the performance of the AIR and SPM co-registration algorithms has been shown to be comparable on scalp-edited MR images (Kiebel et al., 1997). Currently, there is no recent paper that assesses and compares the precision of current available versions of AIR and SPM. It is possible that the single subject MCI MR template was not representative or "average" enough to address age- and neurodegeneration variations across subjects examined herein. We performed a preliminary comparison and evaluated the inter-rater reliability between the average of three templates (single-subject AD MR, single-subject MCI MR and single-subject NC MR) and manual delineation. The inter-rater reliability between manual and the average of the three templates resulted in substantial reliability (ICC > 0.810, Results not shown). Motivated by these concerns, we are currently evaluating the reliability of several automated ROI sampling methods for PiB PET studies, in larger subject samples. Ideally, such a reliability study would include a larger group of subjects to yield analysis results that are less vulnerable to potential outliers, examine and compare different approaches for obtaining CSF dilution correction factors beyond the 2-component method applied herein, examine the effect of different image processing steps (e.g. co-registration and spatial normalization) on the reliability, examine different templates or approaches (e.g. multiple templates and probability based approaches) where results from different normalizations can be combined and include subjects with a wider range of cerebral atrophy.

In summary, the current study indicates the following rank order among the various reliability results in primary cortical areas and cerebellum (high to low): 1) V_T or SUV manual delineation, with or without CSF correction; 2) DVR or SUVR manual delineation, with or without CSF correction; 3) SUV automated delineation, with CSF correction; and 4) SUVR automated delineation, with or without CSF correction. This rank order of reliability was generally similar for other areas of cortex and white-matter rich areas with respect to V_T or SUV but not so for ratio outcomes for which lower performance was noted.

The high inter-rater reliability of PiB outcome measures in primary cortical areas (ACG/PRC) where amyloid may be detected earliest is important as reliable methodology is needed for the detection of low levels of amyloid deposition on a cross-sectional basis and small changes in amyloid deposition on a longitudinal basis. Reliable methodology is also important to enable robust definition of amyloid positivity thresholds and determination of relationships between in vivo ante-mortem and postmortem assessments of amyloid-beta load. Despite methodological differences in the manual and automated approaches, good agreement (i.e., ICC = 0.8) was achieved in primary cortical areas and the cerebellar reference region, although this comparison resulted in a wider range of regional ICC scores than the manual rater comparison.

Acknowledgments

We thank our University of Pittsburgh colleagues at the Alzheimer Disease Research Center and the PET facility for their efforts in conducting and analyzing these studies. We also thank the volunteers and their families for their commitment to further discovery into the causes and treatment of AD through their efforts in this and other related studies. We would like to thank Andrew Redfield and Davneet Minhas for help with automated template processing. This work was supported by grants from the National Institutes of Health (R01 AG018402, P50 AG005133, K02 AG001039, R01 MH070729, R37 AG025516, P01 AG025204, K02 AG027998, and R01 AG033042), Dana Foundation, Alzheimer's Association (TLL-01-3381) and U.S. Department of Energy (DE-FD02-03, ER63590).

References

- Aizenstein HJ, Nebes RD, Saxton JA, Price JC, Mathis CA, Tsopelas ND, Ziolkowski SK, James JA, Snitz BE, Houck PR, Bi W, Cohen AD, Lopresti BJ, DeKosky ST, Halligan EM, Klunk WE. Frequent Amyloid Deposition Without Significant Cognitive Impairment Among the Elderly. *Archives of Neurology* 2008;65(11):1509–1517. [PubMed: 19001171]
- Ashburner J, Friston KJ. Nonlinear spatial normalization using basis functions. *Human Brain Mapping* 1999;7(4):254–266. [PubMed: 10408769]
- Ashburner J, Friston KJ. Unified Segmentation. *Human NeuroImage* 2005;26:839–851.
- Bailer UF, Frank GK, Henry SE, Price JC, Meltzer CC, Mathis CA, Wagner A, Thornton L, Hoge J, Ziolkowski SK, Becker CR, McConaha CW, Kaye WH. Exaggerated 5-HT1A but normal 5-HT2A receptor activity in individuals ill with anorexia nervosa. *Biological Psychiatry* 2007;61:1090–1099. [PubMed: 17241616]
- Cohen AD, Price JC, Weissfeld LA, James J, Rosario BL, Bi W, Nebes RD, Saxton JA, Snitz BE, Aizenstein HJ, Wolk DA, DeKosky ST, Mathis CA, Klunk WE. Basal Cerebral Metabolism May Modulate the Cognitive Effects of A beta in Mild Cognitive Impairment: An Example of Brain Reserve. *Journal of Neuroscience* 2009;29(47):14770–14778. [PubMed: 19940172]
- Fisher PM, Meltzer CC, Price JC, Coleman RL, Ziolkowski SK, Becker C, Moses-Kolko E, Berga SL, Hariri AR. Medial prefrontal cortex 5-HT2A density is correlated with amygdala reactivity, Response Habituation and Functional Coupling. *Cerebral Cortex* 2009;19(11):2499–2507. [PubMed: 19321655]
- Folstein M, Folstein S, McHugh PR. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975;12 (3):189–198. [PubMed: 1202204]

- Jagust WJ, Landau SM, Shaw LM, Trojanowski JQ, Koeppe RA, Reiman EM, Foster NL, Petersen RC, Weiner MW, Price JC, Mathis CA. For the Alzheimer's Disease Neuroimaging Initiative. Relationships between biomarkers in aging and dementia. *Neurology* 2009;73:1193–1199. [PubMed: 19822868]
- Jagust WJ, Bandy D, Chen K, Foster NL, Landau SM, Mathis CA, Price JC, Reiman EM, Skovronsky D, Koeppe RA. Alzheimer's Disease Neuroimaging Initiative. The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. *Alzheimer's and Dementia* 2010;6:221–229.
- Kemppainen NM, Aalto S, Wilson IA, Nägren K, Helin S, Brück A, Oikonen V, Kailajärvi M, Scheinin M, Viitanen M, Parkkola R, Rinne JO. PET amyloid ligand [11C]PIB uptake is increased in mild cognitive impairment. *Neurology* 2007;68:1603–1606. [PubMed: 17485647]
- Kiebel SJ, Ashburner J, Poline J, Friston KJ. MRI and PET Coregistration – A Cross Validation of Statistical Parametric Mapping and Automated Image Registration. *NeuroImage* 1997;5:271–279. [PubMed: 9345556]
- Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, Bergstrom M, Savitcheva I, Huang GF, Estrada S, Ausen B, Debnath ML, Barletta J, Price JC, Sandell J, Lopresti BJ, Wall A, Koivisto P, Antoni G, Mathis CA, Langstrom B. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Annals of Neurology* 2004;55(3):306–19. [PubMed: 14991808]
- Lindgren KA, Larson CL, Schaefer SM, Abercrombie HC, Ward RT, Oakes TR, Holden JE, Perlman SB, Benca RM, Davidson RJ. Thalamic Metabolic Rate Predicts EEG Alpha Power in Healthy Control Subjects But Not in Depressed Patients. *Biological Psychiatry* 1999;45(8):934–952. [PubMed: 10202584]
- Logan J, Fowler J, Volkow N, Wolf A, Dewey S, Schlyer D, Macgregor R, Hitzmann R, Bendriem B, Gatley S, Christman D. Graphical analysis of reversible radioligand binding from time-activity measurements applied to [N-11C-methyl-(–)-cocaine PET studies in human subjects. *Journal of Cerebral Blood Flow & Metabolism* 1990;10:740–747. [PubMed: 2384545]
- Lopresti BJ, Klunk WE, Mathis CA, Hoge JA, Ziolkowski SK, Lu X, Meltzer CC, Schimmel K, Tsopelas ND, DeKosky ST, Price JC. Simplified Quantification of Pittsburgh Compound B Amyloid Imaging PET Studies: A Comparative Analysis. *Journal of Nuclear Medicine* 2005;46(12):1959–1972. [PubMed: 16330558]
- McGraw KO, Wong SP. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods* 1996;1(1):30–46.
- McNamee RL, Yee S, Price JC, Klunk WE, Rosario B, Weissfeld L, Ziolkowski S, Berginc M, Lopresti BJ, DeKosky S, Mathis CA. Consideration of Optimal Time Window for Pittsburgh Compound B PET Summed Uptake Measurements. *Journal of Nuclear Medicine* 2009;50:348–355. [PubMed: 19223409]
- Meltzer CC, Leal JP, Mayberg HS, Wagner HN Jr, Frost JJ. Correction of PET data for partial volume effects in human cerebral cortex by MR imaging. *Journal of Computer Assisted Tomography* 1990;14(4):561–570. [PubMed: 2370355]
- Meltzer CC, Kinahan PE, Greer PJ, Nichols TE, Comtat C, Cantwell MN, Lin MP, Price JC. Comparative Evaluation of MR-Based Partial-Volume Correction Schemes for PET. *Journal of Nuclear Medicine* 1999;40:2053–2065. [PubMed: 10616886]
- Meltzer CC, Cantwell MN, Greer P, Ben-Eliezer D, Smith GS, Frank G, Kaye W, Houck PR, Price JC. Does cerebral blood flow decline in healthy aging? A PET study with partial volume correction. *Journal of Nuclear Medicine* 2000;41(11):1842–1848. [PubMed: 11079492]
- Mintun MA, Larossa GN, Sheline YI, Dence CS, Lee SY, Mach RH, Klunk WE, Mathis CA, DeKosky ST, Morris JC. [11C]PIB in a nondemented population: potential antecedent marker of Alzheimer disease. *Neurology* 2006;67(3):446–452. [PubMed: 16894106]
- Minoshima S, Frey KA, Foster NL, Kuhl DE. An automated method for rotational correction and centering of three-dimensional functional images. *Journal of Nuclear Medicine* 1992;33:1579–1585. [PubMed: 1634959]
- Morris JC, Roe CM, Xiong C, Fagan AM, Goate AM, Holtzman DM, Mintun MA. APOE Predicts A β but not Tau Alzheimer's Pathology in Cognitively Normal Aging. *Annals of Neurology* 2010;67:122–131. [PubMed: 20186853]

- Nichols L, Pike VW, Cai L, Innis RB. Imaging and In Vivo Quantification of β -Amyloid: An Exemplary Biomarker for Alzheimer's Disease? *Biological Psychiatry* 2006;59:940–947. [PubMed: 16487944]
- Pike KE, Savage G, Villemagne VL, Ng S, Moss SA, Maruff P, Mathis CA, Klunk WE, Masters CL, Rowe CC. β -amyloid imaging and memory in non-demented individuals: evidence for preclinical Alzheimer's disease. *Brain* 2007;130(11):2837–2844. [PubMed: 17928318]
- Price JC, Klunk WE, Lopresti BJ, Lu X, Hoge JA, Ziolk SK, Holt DP, Meltzer CC, DeKosky ST, Mathis CA. Kinetic modeling of amyloid binding in humans using PET imaging and Pittsburgh Compound-B. *Journal of Cerebral Blood Flow & Metabolism* 2005;25:1528–1547. [PubMed: 15944649]
- Rosario BL, Ziolk SK, Weissfeld LA, Price JC. Assessment of parameter settings for SPM5 spatial normalization of structural MRI data: Application to type 2 diabetes. *NeuroImage* 2008;41(2):363–370. [PubMed: 18394922]
- Rowe CC, Ackermann U, Gong SJ, Pike K, Savage G, Cowie TF, Dickinson KL, Maruff P, Darby D, Smith C, Woodward M, Merory J, Tothom-Danguy H, O'Keefe G, Klunk WE, Mathis CA, Price JC, Masters CL, Villemagne VL. Imaging β -amyloid burden in aging and dementia. *Neurology* 2007;68(20):1718–1725. [PubMed: 17502554]
- Rusjan P, Mamo D, Ginovart N, Hussey D, Vitcu I, Yasuno F, Tetsuya S, Houle S, Kapur S. An automated method for the extraction of regional data from PET images. *Psychiatry Research: Neuroimaging* 2006;147:79–89.
- Schaefer SM, Abercrombie HC, Lindgren KA, Larson CL, Ward RT, Oakes TR, Holden JE, Perlman SB, Turski PA, Davidson RJ. Six-Month Test-Retest Reliability of MRI-Defined PET Measures of Regional Cerebral Glucose Metabolic Rate in Selected Subcortical Structures. *Human Brain Mapping* 2000;10:1–9. [PubMed: 10843513]
- Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin* 1979;86:420–428. [PubMed: 18839484]
- Small GW, Stern CE, Mandelkern MA, Fairbanks LA, Min CA, Guze BH. Reliability of drawing regions of interest for positron emission tomographic data. *Psychiatry Research: Neuroimaging* 1992;45(3):177–185.
- Svarer C, Madsen K, Hasselbalch SG, Pinborg LH, Haugbol S, Frokjaer VG, Holm S, Paulson OB, Knudsen GM. MR-based automatic delineation of volumes of interest in human brain PET images using probability maps. *NeuroImage* 2005;24:969–979. [PubMed: 15670674]
- 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;15:273–289. [PubMed: 11771995]
- West J, Fitzpatrick JM, Wang MY, Dawant BM, Maurer CR Jr, Kessler RM, Maciunas RJ, Barillot C, Lemoine D, Collignon A, Maes F, Suetens P, Vandermeulen D, van den Elsen PA, Napel S, Sumanaweera TS, Harkness B, Hemler PF, Hill DLG, Hawkes DJ, Studholme C, Maintz JBA, Viergever MA, Malandain G, Pennec X, Noz ME, Maguire GQ Jr, Pollack M, Pelizzari CA, Robb RA, Hanson D, Woods RP. Comparison and evaluation of retrospective intermodality image registration techniques. *Journal of Computer Assisted Tomography* 1997;21:554–566. [PubMed: 9216759]
- Wilson AA, Garcia A, Chestakova A, Kung H, Sylvain Houle S. A rapid one-step radiosynthesis of the β -amyloid imaging radiotracer N-methyl- $[^{11}\text{C}]2$ -(4'-methylaminophenyl)-6-hydroxybenzothiazole ($[^{11}\text{C}]$ -6-OH-BTA-1). *Journal of Labelled Compounds and Radiopharmaceuticals* 2004;47:679–682.
- Wolk DA, Price JC, Saxton JA, Snitz BE, James JA, Lopez OL, Aizenstein HJ, Weissfeld LA, Mathis CA, Klunk WE, DeKosky ST. Amyloid Imaging in Mild Cognitive Impairment Subtypes. *Annals of Neurology* 2009;65(5):557–568. [PubMed: 19475670]
- Woods RP, Mazziotta JC, Cherry SR. MRI-PET registration with automated algorithm. *Journal of Computed Assisted Tomography* 1993;17:536–546.
- Yasuno F, Hasnine AH, Suhara T, Ichimiya T, Sudo Y, Inoue M, Takano A, Ou T, Ando T, Toyama H. Template-Based Method for Multiple Volumes of Interest of Human Brain PET Images. *NeuroImage* 2002;16:577–586. [PubMed: 12169244]

Figure 1.

Example of regions-of-interest (ROIs) defined by Rater 2 on an SPGR magnetic resonance (MR) image acquired from a NC subject (left, manual) and a MCI subject (right, automated, MCI template). The images were acquired for a 80 year-old NC subject and for a 79-year-old single elderly MCI subject, respectively. Shown in white are examples of ROIs examined in this study: CER = cerebellum; ACG = anterior cingulate; FRC = frontal; LTC = lateral temporal; MTC = mesial temporal; OCC = occipital; PAR = parietal; PRC = precuneus.

Figure 2.

Intraclass correlation coefficient scores for absolute agreement (ICC_A) for regional PiB PET retention measures. The ICC_A was determined for manual ROI delineation (three independent raters) with (gray bars) and without (white bars) correction for CSF dilution (CSF) for all subjects ($n=10$). The error bars represent the corresponding 95% confidence intervals for each ICC_A score.

Table 1

Regional CSF dilution correction factors for each group and rater

AD Subjects (n=5)											
Rater	Region										
	CER	ACG	FRC	LTC	MTC	OCC	OCP	PAR	PRC	SMC	
Rater 1											
Mean	0.97	0.80	0.86	0.88	0.88	0.91	0.89	0.87	0.80	0.80	
SD	0.03	0.08	0.05	0.04	0.07	0.04	0.07	0.05	0.08	0.07	
CV (%)	3	10	6	5	8	4	8	6	10	9	
Rater 2											
Mean	0.97	0.83	0.81	0.87	0.87	0.89	0.91	0.86	0.80	0.81	
SD	0.04	0.08	0.05	0.04	0.08	0.05	0.07	0.05	0.08	0.06	
CV (%)	4	10	6	5	9	6	8	6	10	7	
Rater 3											
Mean	0.98	0.83	0.86	0.88	0.88	0.90	0.91	0.88	0.84	0.83	
SD	0.03	0.07	0.05	0.03	0.05	0.04	0.05	0.04	0.07	0.03	
CV (%)	3	8	6	3	6	4	5	5	8	4	
Rater 4											
Mean	0.98	0.86	0.90	0.92	0.84	0.91	0.95	0.88	0.81	0.81	
SD	0.02	0.09	0.04	0.03	0.08	0.05	0.03	0.05	0.07	0.06	
CV (%)	2	10	4	3	10	5	3	6	9	7	
NC Subjects (n=5)											
Rater	Region										
	CER	ACG	FRC	LTC	MTC	OCC	OCP	PAR	PRC	SMC	
Rater 1											
Mean	0.95	0.82	0.84	0.90	0.86	0.89	0.90	0.88	0.88	0.82	
SD	0.03	0.05	0.03	0.01	0.05	0.05	0.03	0.02	0.03	0.03	
CV (%)	3	6	4	1	6	6	3	2	3	4	
Rater 2											
Mean	0.96	0.85	0.82	0.90	0.85	0.90	0.93	0.88	0.89	0.84	

AD Subjects (n=5)												
Rater	Region											
	CER	ACG	FRC	LTC	MTC	OCC	OCP	PAR	PRC	SMC		
SD	0.02	0.04	0.04	0.02	0.05	0.04	0.03	0.03	0.03	0.03		
CV (%)	2	5	5	2	6	4	3	3	3	4		
Rater 3												
Mean	0.95	0.84	0.83	0.89	0.86	0.89	0.93	0.87	0.90	0.82		
SD	0.02	0.02	0.03	0.01	0.05	0.04	0.03	0.02	0.02	0.02		
CV (%)	2	2	4	1	6	4	3	2	2	2		
Rater 4												
Mean	0.97	0.86	0.90	0.93	0.83	0.93	0.93	0.90	0.91	0.82		
SD	0.01	0.05	0.03	0.01	0.03	0.02	0.05	0.02	0.02	0.03		
CV (%)	1	6	3	1	4	2	5	2	2	4		

* AVS, PON, SWM and THL are not included because CSF factors were equal or close to 1. SD = standard deviation, CV = coefficient of variation (SD/Mean * 100)

Table 2

PiB PET measures obtained by the manual and automated methods

AD Subjects (n=5)										
Region	Measure	Rater 1		Manual Rater 2		Rater 3		Automated Rater 4		No CSF
		CSF	No CSF	CSF	No CSF	CSF	No CSF	CSF	No CSF	
CER	V _T	4.03 (5%)	3.91 (7%)	4.08 (4%)	3.97 (7%)	4.14 (8%)	4.05 (10%)	NA	NA	NA
	SUV70	0.69 (21%)	0.67 (23%)	0.71 (22%)	0.69 (24%)	0.72 (23%)	0.70 (24%)	0.73 (18%)	0.72 (19%)	
ACG	V _T	10.55 (14%)	8.41 (10%)	10.19 (13%)	8.37 (10%)	10.11 (12%)	8.31 (10%)	NA	NA	NA
	SUV70	2.09 (32%)	1.65 (25%)	2.02 (31%)	1.64 (24%)	2.00 (29%)	1.63 (24%)	1.98 (32%)	1.67 (25%)	
FRC	V _T	9.15 (15%)	7.83 (14%)	9.29 (17%)	7.51 (11%)	9.14 (16%)	7.82 (14%)	NA	NA	NA
	SUV70	1.81 (32%)	1.55 (29%)	1.84 (33%)	1.48 (27%)	1.81 (32%)	1.55 (29%)	1.84 (31%)	1.65 (27%)	
LTC	V _T	8.72 (19%)	7.69 (18%)	8.80 (20%)	7.59 (17%)	8.68 (20%)	7.61 (18%)	NA	NA	NA
	SUV70	1.73 (33%)	1.52 (31%)	1.75 (34%)	1.50 (30%)	1.72 (33%)	1.50 (31%)	1.74 (31%)	1.59 (29%)	
PAR	V _T	8.79 (15%)	7.63 (14%)	8.93 (15%)	7.64 (12%)	8.91 (14%)	7.81 (13%)	NA	NA	NA
	SUV70	1.72 (32%)	1.49 (28%)	1.75 (33%)	1.49 (28%)	1.74 (29%)	1.52 (27%)	1.83 (30%)	1.61 (28%)	
PRC	V _T	10.38 (17%)	8.24 (14%)	10.40 (16%)	8.26 (14%)	9.71 (15%)	8.16 (13%)	NA	NA	NA
	SUV70	2.03 (32%)	1.60 (27%)	2.02 (32%)	1.60 (27%)	1.90 (30%)	1.60 (27%)	2.01 (30%)	1.62 (27%)	
Global	V _T	9.52 (17%)	7.96 (13%)	9.52 (17%)	7.87 (13%)	9.31 (15%)	7.94 (13%)	NA	NA	NA
	SUV70	1.88 (30%)	1.56 (26%)	1.88 (30%)	1.54 (25%)	1.83 (29%)	1.56 (25%)	1.88 (29%)	1.63 (25%)	
NC Subjects (n=5)										
Region	Measure	Rater 1		Manual Rater 2		Rater 3		Automated Rater 4		No CSF
		CSF	No CSF	CSF	No CSF	CSF	No CSF	CSF	No CSF	
CER	V _T	3.47 (12%)	3.31 (13%)	3.50 (11%)	3.36 (12%)	3.54 (10%)	3.37 (11%)	NA	NA	NA
	SUV70	0.76 (17%)	0.73 (18%)	0.77 (17%)	0.74 (18%)	0.78 (13%)	0.74 (15%)	0.71 (15%)	0.69 (15%)	
ACG	V _T	4.85 (15%)	3.99 (10%)	4.77 (14%)	4.07 (10%)	4.76 (14%)	4.03 (10%)	NA	NA	NA
	SUV70	1.15 (12%)	0.95 (18%)	1.15 (13%)	0.98 (17%)	1.15 (12%)	0.97 (15%)	1.11 (12%)	0.96 (17%)	
FRC	V _T	4.75 (12%)	4.00 (14%)	4.78 (13%)	3.92 (11%)	4.76 (11%)	3.94 (14%)	NA	NA	NA
	SUV70	1.16 (10%)	0.98 (13%)	1.16 (10%)	0.95 (12%)	1.16 (9%)	0.96 (12%)	1.18 (11%)	1.06 (14%)	
LTC	V _T	4.45 (13%)	3.98 (18%)	4.42 (13%)	4.00 (17%)	4.43 (13%)	3.92 (18%)	NA	NA	NA
	SUV70	1.09 (13%)	0.97 (13%)	1.08 (13%)	0.98 (14%)	1.08 (13%)	0.95 (13%)	1.14 (11%)	1.06 (12%)	

AD Subjects (n=5)									
Region	Measure	Rater 1		Manual Rater 2		Rater 3		Automated Rater 4	
		CSF	No CSF	CSF	No CSF	CSF	No CSF	CSF	No CSF
PAR	V _T	4.97 (12%)	4.35 (14%)	5.00 (14%)	4.42 (12%)	4.94 (13%)	4.32 (13%)	NA	NA
	SUV70	1.20 (9%)	1.05 (11%)	1.22 (11%)	1.08 (14%)	1.20 (11%)	1.05 (13%)	1.19 (16%)	1.07 (17%)
PRC	V _T	5.05 (17%)	4.46 (14%)	4.98 (17%)	4.40 (14%)	4.97 (15%)	4.48 (13%)	NA	NA
	SUV70	1.21 (17%)	1.07 (15%)	1.20 (17%)	1.06 (15%)	1.21 (15%)	1.09 (14%)	1.22 (16%)	1.11 (14%)
Global	V _T	4.81 (14%)	4.16 (14%)	4.79 (14%)	4.16 (14%)	4.77 (13%)	4.14 (14%)	NA	NA
	SUV70	1.16 (12%)	1.00 (14%)	1.16 (13%)	1.01 (14%)	1.16 (12%)	1.00 (14%)	1.17 (13%)	1.05 (14%)

Mean (CV% = coefficient of variation (SD/Mean * 100)), CSF = correction for CSF dilution, VT = distribution volume, SUV = standardized uptake values, NA = not available

Table 3

Intraclass correlation coefficients and corresponding 95% confidence interval for regional PiB PET standardized uptake value (SUV70) measure including one manual rater (Rater 2) and the automated method

ROI	Manual versus Automated Rater (SUV70)					
	CSF			No CSF		
	ICC _A	95% CI	ICC _C	95% CI	ICC _A	95% CI
CER	0.871	(0.582,0.966)	0.867	(0.555,0.965)	0.868	(0.559,0.965)
ACG	0.996	(0.973,0.999)	0.997	(0.988,0.999)	0.996	(0.982,0.999)
FRC	0.995	(0.981,0.999)	0.995	(0.979,0.999)	0.936	(0.050,0.989)
LTC	0.991	(0.964,0.998)	0.991	(0.963,0.998)	0.971	(0.391,0.995)
MTC	0.943	(0.786,0.985)	0.937	(0.768,0.984)	0.919	(0.708,0.979)
OCC	0.977	(0.913,0.994)	0.975	(0.904,0.994)	0.953	(0.796,0.988)
OCP	0.803	(0.390,0.947)	0.791	(0.360,0.944)	0.789	(0.375,0.943)
PAR	0.973	(0.899,0.993)	0.971	(0.889,0.993)	0.959	(0.841,0.990)
PRC	0.997	(0.990,0.999)	0.997	(0.988,0.999)	0.990	(0.940,0.998)
SMC	0.991	(0.966,0.998)	0.991	(0.964,0.998)	0.980	(0.923,0.995)
AVS	0.991	(0.963,0.998)	0.993	(0.971,0.998)	0.980	(0.927,0.995)
PON	0.980	(0.926,0.995)	0.980	(0.922,0.995)	0.979	(0.920,0.995)
SWM	0.979	(0.919,0.995)	0.977	(0.911,0.994)	0.977	(0.915,0.994)
THL	0.945	(0.804,0.986)	0.943	(0.788,0.985)	0.783	(0.158,0.947)
Global	0.999	(0.995,1.000)	0.999	(0.995,1.000)	0.992	(0.708,0.999)

ICC_A = intraclass correlation coefficient for absolute agreement, ICC_C = intraclass correlation coefficient for consistency, CI = confidence interval; Shaded areas correspond to moderate mean ICC_A values less than 0.80.

Table 4

Intraclass correlation coefficients and corresponding 95% confidence interval for regional PiB PET standardized uptake value ratio (SUVR70) measure including one manual rater (Rater 2) and the automated method

ROI	Manual versus Automated Rater (SUVR70)									
	CSF					No CSF				
	ICC _A	95% CI	ICC _C	95% CI	ICC _A	95% CI	ICC _C	95% CI	ICC _C	95% CI
ACG	0.964	(0.866,0.991)	0.961	(0.853,0.990)	0.977	(0.908,0.994)	0.975	(0.899,0.994)	0.975	(0.899,0.994)
FRC	0.962	(0.858,0.990)	0.959	(0.844,0.990)	0.935	(0.280,0.987)	0.971	(0.883,0.993)	0.971	(0.883,0.993)
LTC	0.948	(0.817,0.987)	0.948	(0.807,0.987)	0.955	(0.717,0.990)	0.971	(0.882,0.993)	0.971	(0.882,0.993)
MITC	0.414	(-0.286,0.816)	0.396	(-0.271,0.806)	0.644	(-0.443,0.912)	0.634	(-0.474,0.909)	0.634	(-0.474,0.909)
OCC	0.868	(0.574,0.965)	0.865	(0.548,0.965)	0.850	(0.443,0.962)	0.864	(0.453,0.966)	0.864	(0.453,0.966)
OCP	0.403	(-0.264,0.809)	0.395	(-0.272,0.805)	0.482	(-0.960,0.869)	0.484	(-1.077,0.872)	0.484	(-1.077,0.872)
PAR	0.945	(0.806,0.986)	0.946	(0.798,0.986)	0.953	(0.817,0.988)	0.958	(0.832,0.990)	0.958	(0.832,0.990)
PRC	0.964	(0.866,0.991)	0.961	(0.853,0.990)	0.952	(0.819,0.988)	0.951	(0.804,0.988)	0.951	(0.804,0.988)
SMC	0.919	(0.725,0.979)	0.922	(0.718,0.980)	0.934	(0.736,0.984)	0.929	(0.713,0.982)	0.929	(0.713,0.982)
AVS	0.938	(0.783,0.984)	0.937	(0.768,0.984)	0.957	(0.830,0.989)	0.954	(0.813,0.988)	0.954	(0.813,0.988)
PON	0.775	(0.352,0.938)	0.776	(0.326,0.939)	0.870	(0.464,0.968)	0.859	(0.434,0.965)	0.859	(0.434,0.965)
SWM	0.726	(0.224,0.924)	0.713	(0.195,0.920)	0.841	(0.327,0.961)	0.827	(0.303,0.957)	0.827	(0.303,0.957)
THL	0.696	(0.130,0.916)	0.673	(0.119,0.907)	0.785	(0.211,0.945)	0.818	(0.266,0.955)	0.818	(0.266,0.955)
Global	0.983	(0.933,0.996)	0.981	(0.925,0.995)	0.964	(0.846,0.991)	0.971	(0.882,0.993)	0.971	(0.882,0.993)

ICC_A = intraclass correlation coefficient for absolute agreement, ICC_C = intraclass correlation coefficient for consistency, CI = confidence interval; Shaded areas correspond to moderate mean ICC_A values less than 0.80.