NeuroImage ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

NeuroImage

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

Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and correlation with MMSE

Christopher G. Schwarz [a,*], Matthew L. Senjem [a,b], Jeffrey L. Gunter [a,b], Nirubol Tosakulwong [c], Stephen D. Weigand [c], Bradley J. Kemp [a], Anthony J. Spychalla [a], Prashanthi Vemuri [a], Ronald C. Petersen [d], Val J. Lowe [a], Clifford R. Jack Jr. [a]

[a] Department of Radiology, Mayo Clinic and Foundation, Rochester, MN, United States
[b] Department of Information Technology, Mayo Clinic and Foundation, Rochester, MN, United States
[c] Department of Health Sciences Research, Division of Biostatistics, Mayo Clinic and Foundation, Rochester, MN, United States
[d] Department of Neurology, Mayo Clinic and Foundation, Rochester, MN, United States

ARTICLE INFO

Article history:
Received 16 February 2016
Accepted 26 August 2016

ABSTRACT

Quantitative measurements of change in β-amyloid load from Positron Emission Tomography (PET) images play a critical role in clinical trials and longitudinal observational studies of Alzheimer's disease. These measurements are strongly affected by methodological differences between implementations, including choice of reference region and use of partial volume correction, but there is a lack of consensus for an optimal method. Previous works have examined some relevant variables under varying criteria, but interactions between them prevent choosing a method via combined meta-analysis. In this work, we present a thorough comparison of methods to measure change in β-amyloid over time using Pittsburgh Compound B (PiB) PET imaging.

Methods: We compare 1,024 different automated software pipeline implementations with varying methodological choices according to four quality metrics calculated over three-timepoint longitudinal trajectories of 129 subjects: reliability (straightness/variance); plausibility (lack of negative slopes); ability to predict accumulator/non-accumulator status from baseline value; and correlation between change in β-amyloid and change in Mini Mental State Exam (MMSE) scores.

Results and conclusion: From this analysis, we show that an optimal longitudinal measure of β-amyloid from PiB should use a reference region that includes a combination of voxels in the supratentorial white matter and those in the whole cerebellum, measured using two-class partial volume correction in the voxel space of each subject's corresponding anatomical MR image.

1. Introduction

Since the introduction of amyloid Positron Emission Tomography (PET) Imaging (Klunk et al., 2004), methods for quantitative measurement of these scans have been an important topic of discussion motivated by increasing incorporation of these biomarkers into Alzheimer's Disease (AD) clinical trials and research collaborations. We focus this work on two commonly-used classes of existing methods, originally developed for other PET tracers: Standardized Uptake Value (SUV) and Standardized Uptake Value Ratio (SUVR). SUV, also known as Differential Absorption Ratio or Differential Uptake Ratio, attempts to correct measured values in PET images for variation in the amount of tracer injected and the patient's body weight (Thie, 2004; Zasadny and Wahl, 1993). SUVR uses a ratio of PET uptake values measured in different regions of the image: a "target" region of interest (ROI) containing biology to be measured, divided by a "reference" region that is assumed to be free of the pathology of interest, thus providing a surrogate measure of the amount of tracer present. Traditionally for amyloid PET, this has been a ratio of uptake in the cerebral cortex to uptake in the cerebellum because the cerebellum is known to be relatively unaffected by β-amyloid deposition (Klunk et al., 2004; Zhou et al., 2007). We provide equations for both SUV and SUVR in Table 1.

We focus this work on SUV and SUVR because although PET analysis methods requiring significantly longer scan times are known to be more reliable for quantifying change over time (van Berckel et al., 2013), longitudinal population studies require serial

* Correspondence to: Mayo Clinic, Diagnostic Radiology, 200 First Street SW, Rochester, MN 55905, USA.
E-mail address: schwarz.christopher@mayo.edu (C.G. Schwarz).

http://dx.doi.org/10.1016/j.neuroimage.2016.08.056
1053-8119/

Table 1
Equations for SUV and SUVR.

$$SUVR = \frac{Uptake\ in\ Target\ Region}{Uptake\ in\ Reference\ Region}$$

$$SUV = \frac{(\text{Uptake in Target Region}) \times (\text{Weight})}{(\text{Injected Dose})}$$

scans of many hundreds, or even thousands, of subjects. This requirement excludes the possibility of using only dynamic scans due to cost and scanner access constraints, and subject burden required for full dynamic scanning. For example, the Mayo Clinic Study of Aging, data from which we examine in this work, follows over 1500 subjects longitudinally in order to capture sufficient biologic diversity over the entire range of human aging. Performing dynamic scans on such a scale is not tenable. Thus, we focus this work on optimizing SUV/SUVR methods for analysis of such datasets, for which these methods are the only reasonable options.

SUV and SUVR can be calculated by varying implementations and software packages, affecting their properties in ways that have not been fully explored. For example, both are affected by the precise choice of voxels in the target region, and SUVR is strongly affected by the precise choice of voxels in a chosen reference region. Other factors include use of Partial Volume Correction (PVC), and in what voxel space to perform analysis calculations. In this work we explore the effects of each of these choices on serial measurements of scans using [^{11}C] Pittsburgh Compound B (PiB) (Klunk et al., 2004).

Recently, the amyloid PET community has begun to address the many variations in performing these measurements (Klunk et al., 2015; Schmidt et al., 2015), and many previous works have focused on determining a technique that outperforms others according to a range of criteria. Many have focused on purely cross-sectional criteria for use in cross-sectional studies (Klein et al., 2014; Lopresti et al., 2011; Lowe et al., 2009; Thurfjell et al., 2014). Our present work focuses purely on longitudinal criteria for optimizing change-over-time measures in longitudinal studies; we review the others in this category below. In 2014, Fleisher et. al. evaluated supratentorial white matter (WM) versus cerebellar or pontine references for computing SUVR on [^{18}F] florbetapir scans, using a serial-change-detection criteria, and concluded in favor of supratentorial WM reference regions (Fleisher et al., 2014). Liu et. al. compared measuring SUVR from florbetapir scans with cerebellar gray matter (GM) versus pons references according to their ability to detect any potential treatment effect in subjects treated with bapineuzumab and reported no significant treatment effect using either reference, but also that SUVR values computed from the two methods were qualitatively inconsistent with each other (Liu et al., 2014). Joshi et. al. compared florbetapir scans with supratentorial WM versus cerebellar references according to maximizing serial increase in uptake in subjects considered amyloid positive, and according to a serial variation criteria, and concluded that both criteria were improved with a supratentorial WM reference region (Joshi et al., 2014). In 2015, Brendel et. al. compared cross-sectional group discrimination and serial-change-detection using florbetapir measurements with cerebellar versus brainstem versus WM references, and also examined whether to use full-brain atlas voxels versus GM-segmented atlas voxels versus those PVC-corrected. Both criteria were improved by WM or brainstem rather than cerebellar reference regions, and by using PVC (Brendel et al., 2015). Chen et. al. compared cerebellar versus pontine versus WM reference regions using a PET-only method for measuring pairs of serial florbetapir scans according to reliability, correlation with cognitive declines, and power to detect longitudinal change, also concluding in favor of WM references. (Chen et al., 2015). Landau et. al. compared cerebellar, pontine, and WM reference regions in serial florbetapir measurements. They first identified groups expected to increase over time, versus those expected to remain stable, using concurrent β-amyloid measurements from Cerebrospinal Fluid (CSF). Then, they determined agreement between serial florbetapir measurements and these predictions, also concluding in favor of WM-containing references (Landau et al., 2015). An analysis of PVC was performed by Su et. al., who compared processing methods using two-class PVC, a region-based PVC, and no PVC on a basis of power to detect change using PiB scans, and concluded that region-based PVC was superior to two-class PVC, and both were superior to no PVC (Su et al., 2015).

Many questions remain unsettled. Particularly, most prior comparisons focus on ^{18}F ligands, and their conclusions may not necessarily apply to ^{11}C PiB. Secondly, most focus on a smaller subset of the many methodological variables involved, usually reference region alone, ignoring others such as PVC, segmentation methods, etc. Because it is possible, if not likely, that these variables have interactions, we propose that a comprehensive study must include a combination of all variables involved. Finally, we point out that prior comparisons use a wide variety of criteria: some relating to accuracy, others to precision, some cross-sectional, some longitudinal, that are rarely the same between them. Here we present a comprehensive study examining over 1,024 fully-automated software analysis pipeline implementations (henceforth referred to as "pipelines") with differing combinations of methodological variables for measuring SUV and SUVR values from serial PiB scans, using four criteria addressing different aspects of longitudinal trajectory performance. In the following section we fully describe the aims and scope of this work.

2. Objectives

In calculating a single measure of β-amyloid from an amyloid PET scan, there are potentially infinite methodological variables. In this section, we identify a set of methodological questions that have been the most debated. Answering these will be the objective of this work.

2.1. Question 1: is partial volume correction helpful?

Partial Volume Correction (PVC) attempts to correct for the effect of relative uptake levels in multiple tissue types and/or CSF within each PET voxel. Generally, this uses segmentations from a T1-weighted Magnetic Resonance Image (MRI) in combination with the known resolution of the PET camera used to estimate the fraction of each material expected to lie in each voxel or ROI. In this work we evaluate three options: no PVC; 2-class PVC (PVC2), which attempts to correct for CSF in GM/WM locations; and 3-class PVC (PVC3), which attempts to correct for both CSF and WM in GM locations. Briefly, 2-class PVC uses a binary map of GM+WM locations from MRI segmentation and blurs this by the known resolution of the PET camera (approximately 8 mm^3 for the scanners in our study (Joshi et al., 2009)) to create an approximate fraction of tissue expected in each voxel. The raw PET scan is then divided by this tissue fraction in each voxel. This algorithm depends on the accuracy of the MRI segmentation and registration between PET and MRI (Meltzer et al., 1990). In 3-class PVC, MRI segmentation is instead used to create separate maps of voxels containing GM and those containing WM. First, an atlas is used to locate the centrum semiovale region and calculate the mean PET signal in this ROI, which is then assumed to reflect the true WM signal in all WM voxels. This mean value is assigned to all voxels segmented as WM, then blurred with the approximate PSF and subtracted from the raw PET scan. This step is designed to remove the signal from each voxel that is due to WM. Next, the image

Please cite this article as: Schwarz, C.G., et al., Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and.... NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.08.056

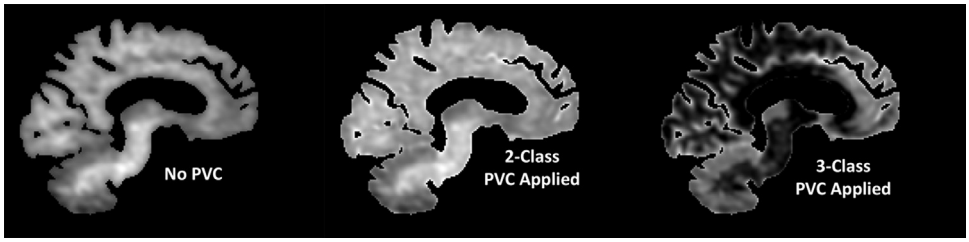

Fig. 1. PVC options tested: An example on a single subject PiB scan with no PVC applied, with 2-class PVC (denoted PVC2) applied, and with 3-class PVC (denoted PVC3) applied. Non-brain voxels were excluded.

follows the same steps as in two-class PVC, which then attempts to correct further partial voluming by CSF (Müller-Gärtner et al., 1992). Because this method is designed to remove WM signal from each voxel, it is not appropriate when WM voxels are included in either reference or target regions for SUVR calculation. For this reason, we omit pipeline combinations that would use 3-class PVC with WM-containing target or reference regions.

Both algorithms were implemented in-house in C++ using the Insight Toolkit (www.itk.org), and were applied using the MRI scans that corresponded to each subject's PET scan at each time-point, in order to allow them to correct for differing amounts of partial volume over time, which is expected due to atrophy and potential pathology. We provide examples in Fig. 1.

2.2. Question 2: which is the optimal reference region?

Most of the debate in this literature has concerned choosing a reference region. We test 55 variations derived from four major regions: cerebellum, pons, midbrain, and supratentorial WM. The 55 variations come from various sub-regions, erosions, and combinations of the above. For supratentorial WM, we include two classes of implementations: those using each subject's individual segmentation from T1-weighted MRI using SPM12 (Ashburner and Friston, 2005), and those using regions from the Johns Hopkins University "Eve" single-subject WM atlas (Oishi et al., 2009). We show examples in Fig. 2.

2.3. Question 3: which is the optimal GM segmentation within the target region?

In this question, we optimize which voxels to include in the cortical target. We are *not* comparing different sets of anatomic brain regions, but only how to apply segmentation to determine a subset of the voxels within these regions for each scan. We test a total of six segmentation methods within a fixed target meta-ROI containing those regions primarily affected by β-amyloid deposition: parietal, cingulate precuneus, prefrontal, orbito-frontal, temporal, and anterior cingulate (Fig. 3). Within these ROIs, we

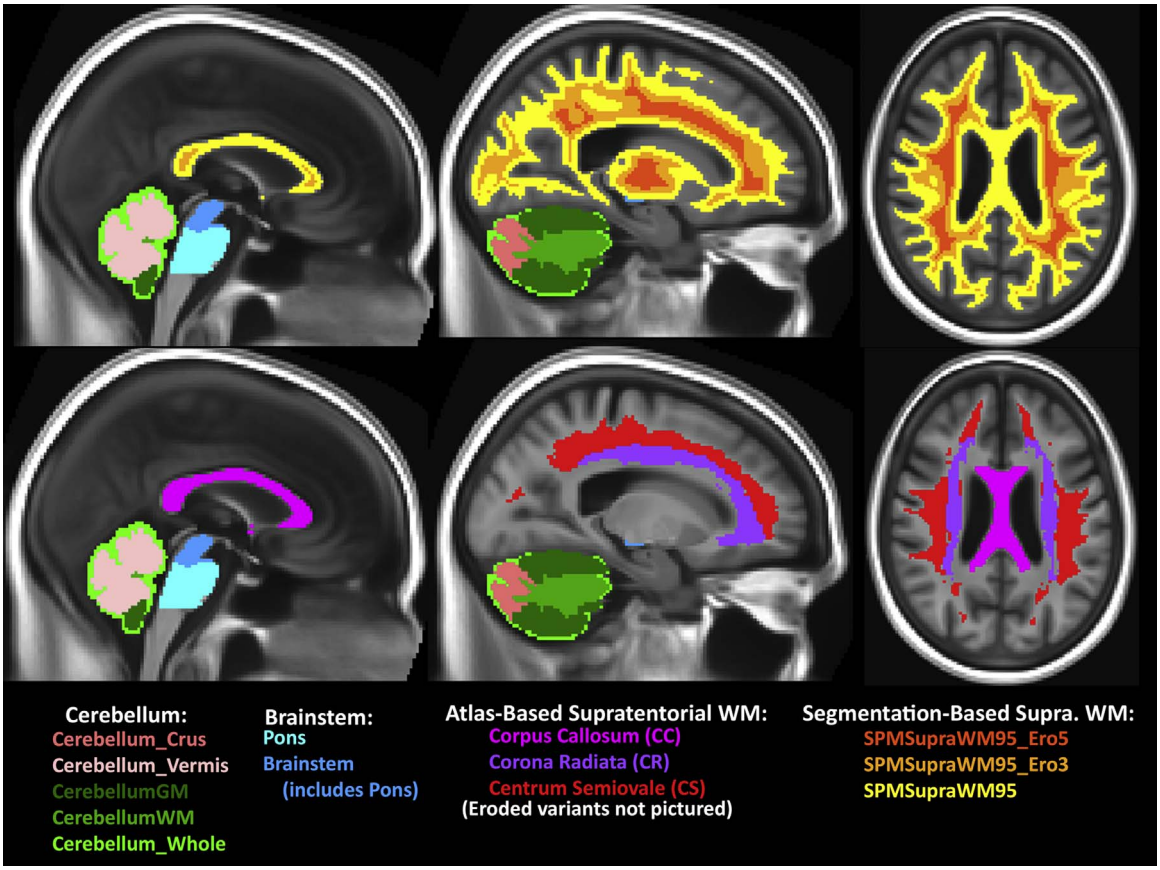

Fig. 2. Reference regions tested: SUVR Reference regions tested in this study include those above, along with various combinations (denoted with +) and erosions of 3 and 5 voxels radius (denoted with _Ero3, _Ero5).

Please cite this article as: Schwarz, C.G., et al., Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and.... NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.08.056

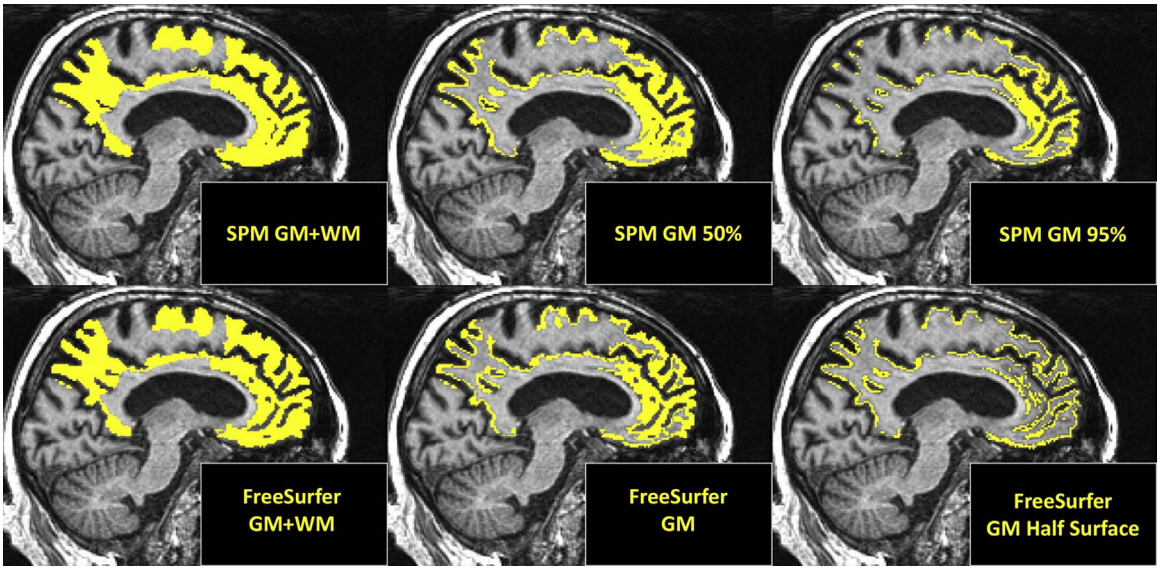

Fig. 3. Target GM Segmentations Tested: In this work, we examine six options for different segmentations within the target ROI. We emphasize that this work does not examine different regions of the brain as targets; instead, we examine options for tissue segmentations that choose sets of voxels within a set of standard anatomical regions: parietal, cingulate precuneus, prefrontal, orbito-frontal, and anterior cingulate.

test three variations with segmentations computed by the Unified Segmentation algorithm in SPM12 (Ashburner and Friston, 2005) with an in-house population-specific set of tissue priors, and three using Freesurfer version 5.3 (Fischl, 2012). For SPM methods, we test three variations: 1) voxels estimated to be either GM or WM together 2) voxels estimated to be GM with at least 50% confidence 3) voxels estimated to be GM with at least 95% confidence. Among Freesurfer methods, we test: 1) voxels segmented as either GM or WM together 2) voxels segmented as GM 3) voxels estimated to lie upon a surface halfway between the GM/WM and GM/CSF (pial) surfaces. Because Freesurfer does not output probabilistic segmentations, we use the output binary segmentations for the GM and GM+WM variations. Freesurfer segmentations are output in the space of its native Talairach template; we resample these back to each subject's native space using the recommended technique (How to Convert from FreeSurfer Space Back to Native Anatomical Space, 2015). The "GM Half Surface" option uses the *mri_surf2vol* tools to estimate those voxels along a surface halfway between the GM/WM and the GM/CSF (pial) surface. These voxels are theoretically those cortical GM voxels furthest from WM. Segmentations produced by all methods were visually examined for all scans, and none had errors sufficient to warrant manual correction or exclusions. All segmentation and registration steps were performed separately for each scan at each timepoint, with the exception of Freesurfer segmentations, which used the longitudinal stream of Freesurfer.

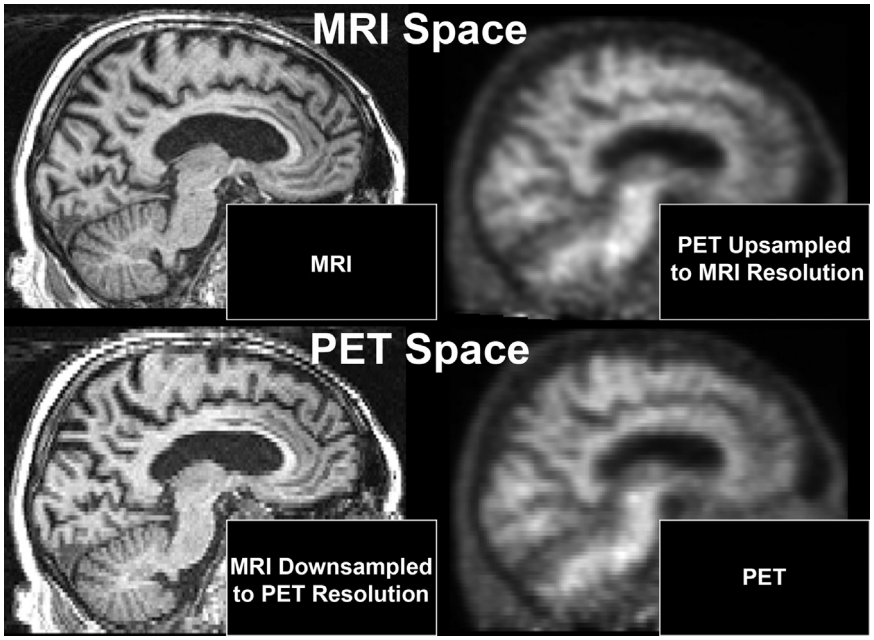

Fig. 4. Analysis space options tested: We examine two options for voxel spaces in which to calculate results: down-sampling MR images to match PET resolution (PET Space) versus up-sampling PET images to match MRI resolution (MRI space).

Please cite this article as: Schwarz, C.G., et al., Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and.... NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.08.056

2.4. Question 4: which is the optimal analysis space?

Typical T1-weighted MRI scans have a resolution of approximately 1 mm^3, while PET scans typically have an effective resolution of approximately 6 mm^3 or larger (Joshi et al., 2009). For this situation, it is necessary to choose sets of voxels defined in some specific voxel space. Thus, resampling one or the other is necessary, but it is unclear which will produce superior results (Fig. 4). We test both options (MRI space and PET space) in this work.

2.5. Question 5: SUV or SUVR?

Among discussion of reference regions, one might consider using SUV to normalize instead of a reference region (Table 1). We include this option in the set of reference methods tested, increasing the count to 56.

3. Methods

3.1. Subject characteristics

We include scans of 129 subjects from the Mayo Clinic Study of Aging (MCSA) and Mayo Clinic Alzheimer's Disease Research Center (ADRC) studies with three timepoints of serial scans with both PiB and T1-weighted MRI (for a total of 387 scans). Subjects with baseline SUVR >2.5 were excluded, for reasons discussed later. MCSA is an epidemiological study of cognitive aging in Rochester, Olmsted County, Minnesota (Petersen et al., 2010; Roberts et al., 2008). The ADRC study recruits and follows subjects initially seen as patients in the Mayo Clinic Behavioral Neurology practice. All studies were approved by their respective institutional review boards and all subjects or their surrogates provided informed consent compliant with HIPAA regulations. We provide subject characteristics in Table 2. Freesurfer processing failed to produce an output for one subject, who was therefore excluded from Freesurfer-using pipelines.

Table 2
Subject characteristics.

Characteristic	Summary
Number of subjects	**129**
Sex, n (%)	
Female	44 (34%)
Male	85 (66%)
Age at baseline PET, years	76 (71, 80) [41 to 93]
Education, years	14 (12, 16) [7 to 24]
Global cortical PIB, SUVR	1.39 (1.31, 1.81) [1.19 to 2.47]
Diagnosis at baseline, n (%)[a]	
CN	78 (62%)
MCI	25 (20%)
Dementia	23 (18%)
APOE ε4, n (%)	
Carrier	84 (65%)
Non-carrier	45 (35%)
MMSE score	28 (26, 29) [8 to 30]
Time between first and third scan, years	3.1 (2.6, 3.9) [1.7 to 5.1]
Time between corresponding MRI and PET scans, days	11.0 (2.0, 31.0) [0.0 to 148.0]

Ranges are given as: median (1st quartile, 3rd quartile) [min to max]
Abbreviations: **n:** Number of subjects; **CN**: Cognitively Normal; **MCI**: Mild Cognitive Impairment; **APOE**: apolipoprotein E; **MMSE**: Mini-Mental State Exam

[a] Three subjects were given a diagnosis of "Uncertain" and are not included here.

3.2. Scan acquisition parameters

[11C] Pittsburgh Compound B (PiB) PET/CT studies were acquired using GE scanners (models Discovery 690XT and Discovery RX; GE Healthcare, Waukesha, WI). Subjects were injected with PiB (average 625 MBq, range 256–751 MBq) and a low dose CT scan was acquired. Beginning 40 min post-injection, subjects then underwent a 20-min dynamic PET scan with four five-minute frames. Dynamic PET images were generated (256 matrix, 300 mm field of view, 1.17 mm × 1.17 mm × 3.27 mm voxel size) using an iterative reconstruction algorithm. Standard corrections for attenuation, scatter, randoms and decay were applied as well as a 5 mm Gaussian post filter. The images from the four dynamic frames were averaged to create a single static image.

T1-weighted MRI scans (used for atlas normalization/masking, and for PVC where applicable) were acquired on 3 T scanners (models Discovery MR750, Signa HDx, Signa HDxt, and Signa Excite) manufactured by General Electric (GE) using a 3D Sagittal Magnetization Prepared Rapid Acquisition Gradient-Recalled Echo (MP-RAGE) sequence. Repetition time (TR) was ≈7 ms, echo time (TE) ≈3 ms, and inversion time (TI)=900 ms.Voxel dimensions were ≈1.20 mm × 1.015 mm × 1.015 mm.

3.3. Common processing

T1-weighted scans were acquired with sagittal-plane gradient distortion correction performed on the scanner. Through-plane correction was performed as part of image processing (Jovicich et al., 2006). Tissue-class segmentation and inhomogeneity (B0 bias-field) correction were performed using the Unified Segmentation algorithm (Ashburner and Friston, 2005) in SPM12 (revision 6225). Several parameters in SPM12 were modified to produce more accurate segmentations for older-adult populations. Firstly, we used an in-house population-specific template and tissue priors that we call *MCSA202*. This template was created from MRI scans of 202 subjects in the same Mayo Clinic MCSA and ADRC studies from which this study's subjects were selected. Each subject's MRI was segmented using SPM12 and our custom template was created from these segmentations using the DARTEL group-wise registration algorithm in SPM12 (Ashburner, 2007). The tissue probability priors were manually edited to correct common segmentation errors, and 122 ROIs were drawn on the anatomic template. In addition to using the MCSA202 template, we also altered the SPM12 segmentation parameters to use two Gaussians to model WM intensities, instead of one, due to the higher prevalence of WM disease in such populations. We also reduced each of the stiffness penalty parameters of the nonlinear normalization of the tissue class priors of half of their defaults, allowing for increased inter-subject variability due to increased prevalence/severity of atrophy and other pathologies. More detailed information about the MCSA202 template and these parameter alterations is beyond the scope of this text, but are forthcoming in future publication. Images were registered to our MCSA202 template using the ANTs SyN registration algorithm (Avants et al., 2008) version 1.9.x with multiple channels: the post-B0-correction T1-weighted image, segmented tissue probabilities, and a mask of total intracranial volume. When each scan was segmented using Freesurfer 5.3, it was directly entered into Freesurfer version 5.3 (Fischl, 2012) using the *recon-all* pipeline (i.e. not preprocessed with the SPM12 bias-correction described above, because Freesurfer performs its own bias correction step and this would be redundant).

T1 and PiB scans were coregistered using SPM12 with 6 degrees

of freedom (DOF). Resampling between MRI and PET resolutions was performed using ANTs software tools with 3rd-order BSpline interpolation. Atlas ROIs were resampled to subject spaces also using ANTs with nearest-neighbor interpolation.

3.4. 1,024 Pipelines

We compare a total of 1,024 different software pipelines for measuring β-amyloid load in PiB scans. In total, we examine 3 methods of PVC (none, 2-class PVC, and 3-class PVC), 56 intensity normalization methods (55 SUVR reference regions + SUV), 6 methods of cortical target segmentations, and 2 potential analysis spaces (MR-Space versus PET-Space). Multiplying all combinations, we implemented a total of $3 \times 56 \times 6 \times 2 = 2016$ pipelines. We then excluded pipelines in two classes of theoretically-implausible combinations: (1) 3-class PVC used with any of the 52 references containing WM or with either of the 2 targets containing WM $((1 \times 52 \times 6 \times 2 = 624) + (1 \times 56 \times 2 \times 2 = 224) - (1 \times 52 \times 2 \times 2 = 208$ in both$) = 640)$, and (2) target segmentations containing supratentorial WM used with references also containing supratentorial WM $(3 \times 44 \times 2 \times 2 = 528)$. Class 1 was excluded because images corrected with 3-class PVC have signal in WM removed, making it illogical to then attempt to measure signal in these WM regions. Class 2 was excluded in order to prevent use of the same voxels for both target and reference regions, which would at least partially normalize out any signal of interest. After subtracting both classes, $2016 - 640 - 528 + (1 \times 44 \times 2 \times 2 = 176$ in both exclusion groups$) = 1{,}024$ combinations remained, which we analyze in this work.

3.5. Evaluation criteria and statistical methods

In this section we describe the motivations and implementations for each of the four individual criteria, and the weighted, combined metric, used to compare measurement pipelines in this work. Each criterion was designed to address different characteristics desirable in serial PiB measurements. Ideally, one would wish to have an independent, gold-standard measurement of β-amyloid load with which to compare, but the only such measures come from pathological examinations, which are of course impossible for serial measurements. As such, we have chosen criteria based on obtainable data: the plausibility of the serial trajectories produced by each pipeline for each subject. All criteria analyzed are specifically longitudinal, because the goal of this study is to determine the best pipeline for serial measurements, which may or may not also be ideal for cross-sectional studies. Because none of these criterion is flawless nor alone captures all traits desirable in a serial PiB measurement pipeline, and because comparing four separate metrics across different implementations creates a complex array of data, we also present a single metric that is a weighted combination of the four. We describe each below.

3.5.1. Longitudinal reliability

Our Longitudinal Reliability metric is motivated by the intuitive notion that pipelines with less measurement jitter within each subject over time are preferable to those with more. For example, a plotted three-timepoint trajectory shaped like a triangle is less reasonable than one that is approximately linear. From previous Amyloid PET studies, it known that the trajectory of β-amyloid accumulation in AD is a roughly sigmoidal shape where the accumulation phase occurs over a time period of approximately 19 years (Villemagne et al., 2013). Studies of CSF amyloid, which provides an independent source of data for measuring β-amyloid in-vivo, also support the sigmoidal trajectory (Buchhave et al., 2012; Shaw et al., 2009). Based on this evidence, having data from only three timepoints over a time span of ≈ 3 years, it is reasonable to assume that trajectories should be locally linear, i.e. a significant acceleration or deceleration in measured SUVRs is more likely attributed to measurement error than true change in subject β-amyloid.

To quantify this measure, we estimated the "reliability" or "R^2" of each pipeline by fitting a linear mixed-effects regression model with time from baseline as a fixed effect and including random subject-specific intercepts and slopes over time. From the model, we obtained the estimated variances of the subject-specific intercepts, slopes, and errors, denoted by $\sigma^2_{\text{intercept}}$, σ^2_{slope}, and by σ^2_{error}, respectively. We then report the reliability as $1 - \left(\frac{\sigma^2_{\text{error}}}{\sigma^2_{\text{intercept}} + \sigma^2_{\text{slope}} + \sigma^2_{\text{error}}} \right)$. We denote this quantity by R^2 as it can be interpreted as the percentage of the total intra-subject variability in β-amyloid PET SUVR measurements that can be "explained" by the linear model, i.e. the straightness of the trajectory.

One potential limitation of this measure is that it can slightly penalize trajectories that reflect true slightly-nonlinear β-amyloid trajectories, i.e. those that were stable in the first two timepoints but began to accumulate in the third, or those that were accumulating in the first two timepoints but began to plateau in the third. However, our sample is designed to exclude the latter, and the former are expected to be a minority within the sample. Although this measure does slightly penalize small accelerations and decelerations, it much more strongly penalizes triangle-shaped trajectories, which are always implausible for β-amyloid. This limitation could have been partly addressed by using higher-order models to allow nonlinear acceleration/deceleration, but because we have only three measurements available over a time span of ≈ 3 years, we feel it is reasonable to assume amyloid accumulation is locally linear. To be able to effectively model departures from linearity, we suspect that we would need at least five measurements over a 5–10 year period.

3.5.2. Longitudinal plausibility

It is commonly agreed that β-amyloid in AD does not decrease over time; it only increases until approaching a plateau toward the late stage of the disease (Ingelsson et al., 2004). From previous work (Jack et al., 2013), we assume that all β-amyloid measurements over time of subjects whose baseline SUVR value < 2.5, i.e. those much earlier than the disease phase approaching a plateau, should be non-decreasing. Therefore, each must be either stable, or increasing. We excluded from our study all subjects with baseline SUVR ≥ 2.5, thus assuming that any apparently-decreasing trajectories can be attributed to measurement error.

The plausibility criterion was assessed as the percentage of subjects with non-negative slopes. For each pipeline, we fitted a linear regression model for each subject using their three timepoints and obtained an estimate of their individual rate of β-amyloid accumulation (i.e., their slope) and calculated the percentage of subjects with slopes ≥ 0. To minimize potential circularity bias imposed by the choice of method used to determine these SUVR thresholds, we used during subject selection those values computed by an earlier method (Jack et al., 2013), which is not among those considered here and was designed to be a cross-sectional, rather than longitudinal, method. In this earlier method, which we used here only for subject selection criteria, SUVR values were computed using registrations, segmentations, and atlas normalizations using SPM5. Two-class PVC was applied, and cerebellar GM was used as a reference.

3.5.3. Longitudinal group separability

Based on the same assumptions as in Sections 3.5.1 and 3.5.2, we assume that it is possible to divide non-late-stage subjects into two classes based on their baseline measurements: "*non-*

accumulators", those who do not have significant β-amyloid burden and are not expected to increase over time (i.e. those who do not have AD, or have not yet reached this phase of pathophysiology), and "*accumulators*", those already have measurable β-amyloid burden and are expected to continue increasing (Villain et al., 2012). To assess the separability between these groups, we used a subset of 90 participants categorized as those with baseline SUVR < 1.35 (the expected "non-accumulator group"; $n = 50$) and those with baseline SUVR between 1.5 and 2.2 (the expected "accumulator group"; $n = 40$), using thresholds based on previous work (Jack et al., 2013). We chose 2.2 as the upper cutoff for the accumulator group to obtain subjects expected to increase, rather than re-using the previous work's 2.5 cutoff, which was designed to ensure that subjects would have non-decreasing (but not necessarily still-increasing) amyloid burden. SUVR values used for these thresholds were also computed as in (Jack et al., 2013).

To measure the separability between these groups, we used the area under the receiver operating characteristic curve (AUROC) as a nonparametric effect size estimate and calculated based on group-wise differences in the rate of β-amyloid accumulation, where rates were obtained from the same slopes estimated in the longitudinal plausibility analysis. In this context, the AUROC can be interpreted as the estimated probability that an accumulator would have a greater/more-positive slope than a non-accumulator.

The primary limitation of this criterion is that although it reflects an expected correlation within most subjects that should be reflected by any reasonable pipeline for serial amyloid PET, some subject trajectories naturally deviate from this relationship, and pipelines reflecting this truth are penalized. There is also a potential circularity introduced by having defined the two groups' inclusion thresholds using an SUVR variant that resembles some pipelines being examined; we address this limitation in Section 5.8.

3.5.4. Correlation between ΔSUVR and ΔMMSE

Although cognitive decline in AD is more closely associated with tau pathophysiology than the β-amyloid pathophysiology measured by PiB imaging, a smaller but significant correlation is expected between β-amyloid and cognition (Jack et al., 2008; Whitwell et al., 2008). If changes in a hypothetical measurement of β-amyloid showed no correlation with change in clinical symptoms in any subjects, it would be a poor measure. Therefore, we test this capability of each of our metrics.

To this end, we tested the correlation between the rate of β-amyloid accumulation and the rate of change in Mini-Mental State Exam (MMSE) (Folstein et al., 1975) scores, where both rate measures were obtained from linear regression models using a subject's three timepoints. Because of skewness in the distribution of MMSE scores, and the potential for a non-linear association between change in β-amyloid and change in MMSE, we report Spearman's rank correlation. Generally, the correlation between both rates is negative since greater β-amyloid accumulation would tend to coincide with more rapidly decreasing MMSE scores. Therefore, we report the negative of the Spearman's estimates.

The primarily limitation of this criterion is the same as that in Section 3.5.3; it reflects an assumption that any reasonable pipeline should find true for most subjects, but it can penalize pipelines that correctly measure subjects that do not follow it. Deviation from this assumption is expected for those subjects in the early phase of β-amyloid accumulation, prior to the onset of significant accumulation of tau pathology, neurodegeneration, and clinical decline. However, some degree of correlation is expected

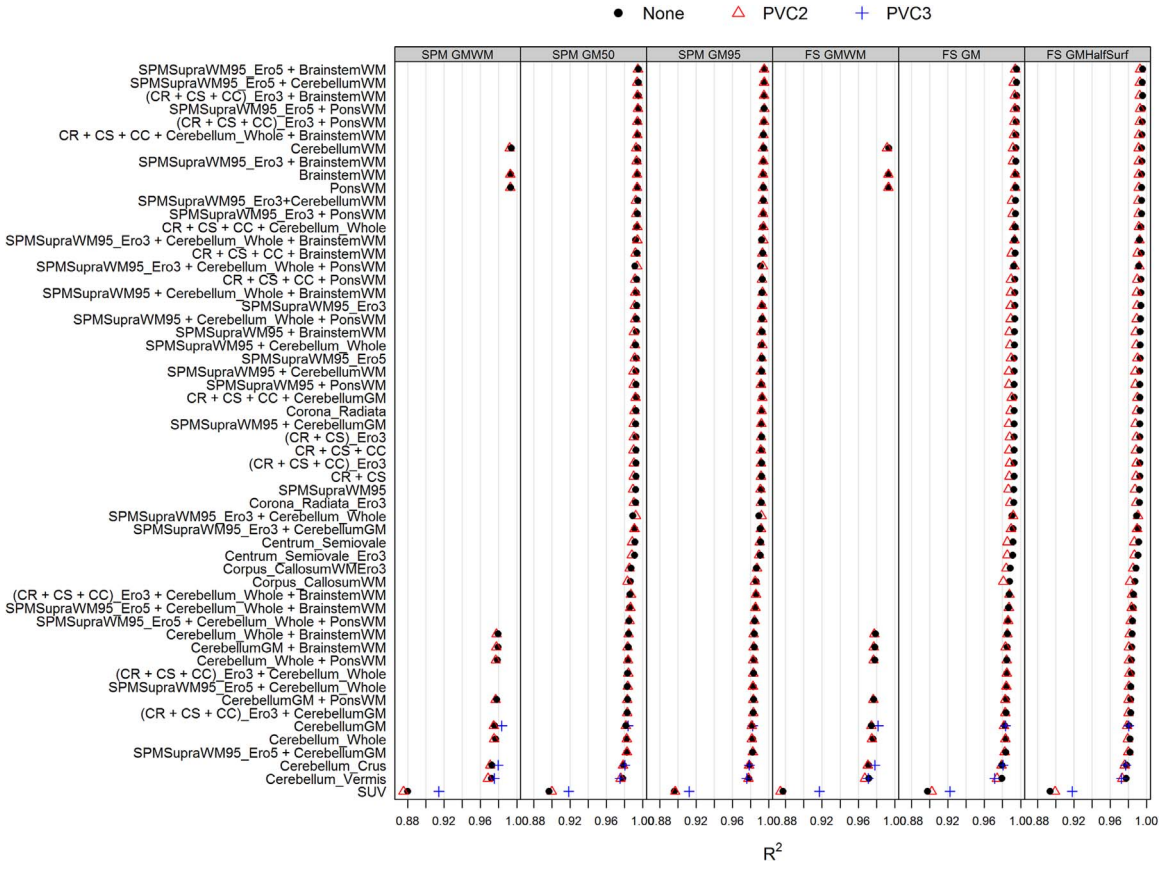

Fig. 5. Plots of the longitudinal reliability (straightness) criterion from a mixed-effects model across pipelines varying in choice of references, target segmentations, and PVC methods, all computed in MRI voxel space. References are ranked according to the maximum R^2 values across all PVC and target segmentation methods.

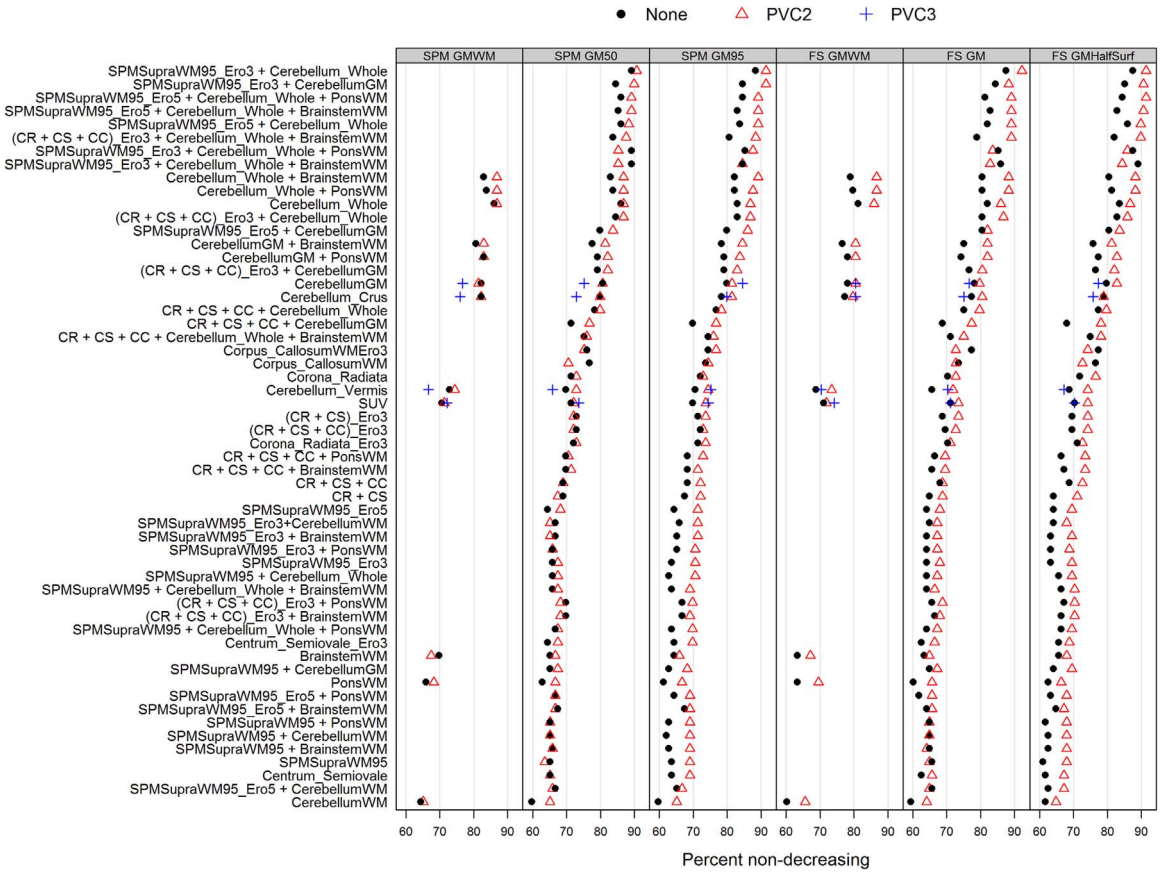

Fig. 6. Plots of the plausibility (percent non-negative) criteria from the linear regression model. References are ranked according to the maximum percent across all PVC and target segmentation methods.

for most subjects in this sample, i.e. those prior to the start of significant β-amyloid accumulation (for which β-amyloid and MMSE will both be stable and thus highly correlated), and those well-within the approximately 19-year period of significant β-amyloid accumulation.

3.5.5. *Combined evaluation criteria*

We created a combined criterion in order to account for the fact that each of these four measures embodies a separate desirable trait for serial PiB measurement that is, in itself, insufficient to determine an ideal pipeline. Therefore, a combined metric can provide a single concise measure while allowing each trait to contribute. As previously discussed, each individual criterion is designed to favor superior pipelines but requires assumptions that are not always valid for all subjects. Therefore, our combined measure is intended to allow each criterion to weigh in while also compensating for each other's shortcomings.

To arrive at a single value combining all four quality metrics, we first normalized each metric using rank-based scaling with 1 being the highest rank and 0 being the lowest rank. Next, we applied pre-specified weights reflecting the level of importance we deemed for each quality metric. We used weights of 0.4 for reliability, 0.4 for plausibility, 0.1 for group separability, and 0.1 for correlation between the rates of β-amyloid and MMSE. We chose these weights *a priori* based on the assumption that the reliability and plausibility metrics are criteria that should be true for all trajectories (i.e. all included subjects' trajectories should be reliable and should not have decreasing slopes), whereas the separability and MMSE criteria are expected to be looser correlations that apply to most subjects but from which some subjects will naturally deviate (see sections 3.5.3 and 3.5.4). Finally, we added the normalized, weighted values of the four quality metrics together. Using this method, a pipeline that performed best across all metrics would get an overall score of 1.0 while a pipeline that performed worst across all metrics would get a 0.0.

We investigated the sampling variability of the quality metrics using bootstrap procedures by randomly sampling participants with replacement and performing all calculations described above on the bootstrap sample. The process was repeated 1000 times and resulted in 1000 estimates of the quality metrics and 1000 estimates of the combined scores. We used these bootstrap estimates to obtain confidence intervals based on the quantiles of the sampling distribution. We then selected 60 top-performing pipelines according to the combined values and plotted the estimates that were obtained from our original data. To facilitate comparisons across pipelines, we used a heuristic that non-overlapping 83% confidence intervals (obtained from the 8.5th and 91.5th percentiles of the bootstrap distribution) provide an informal indication of significant differences at $p < 0.05$ (Knol et al., 2011).

To more formally compare pipelines, we report p-values for tests of differences between pipelines based on the confidence intervals for the bootstrap difference between pipelines. To do this we note that a 95% confidence interval that does not include zero can be interpreted as providing a statistical test that is significant at $p < 0.05$. Generalizing this, for each pairwise comparison we estimated confidence intervals of the difference in the combined values and identified the widest confidence interval (i.e., one with the highest level) that still excluded the null value of zero. For example, if a 98% confidence interval did not include zero, but a 99% confidence interval included zero, we define the p value as $p = 0.02$.

To address the concern that our analysis may be sensitive to the

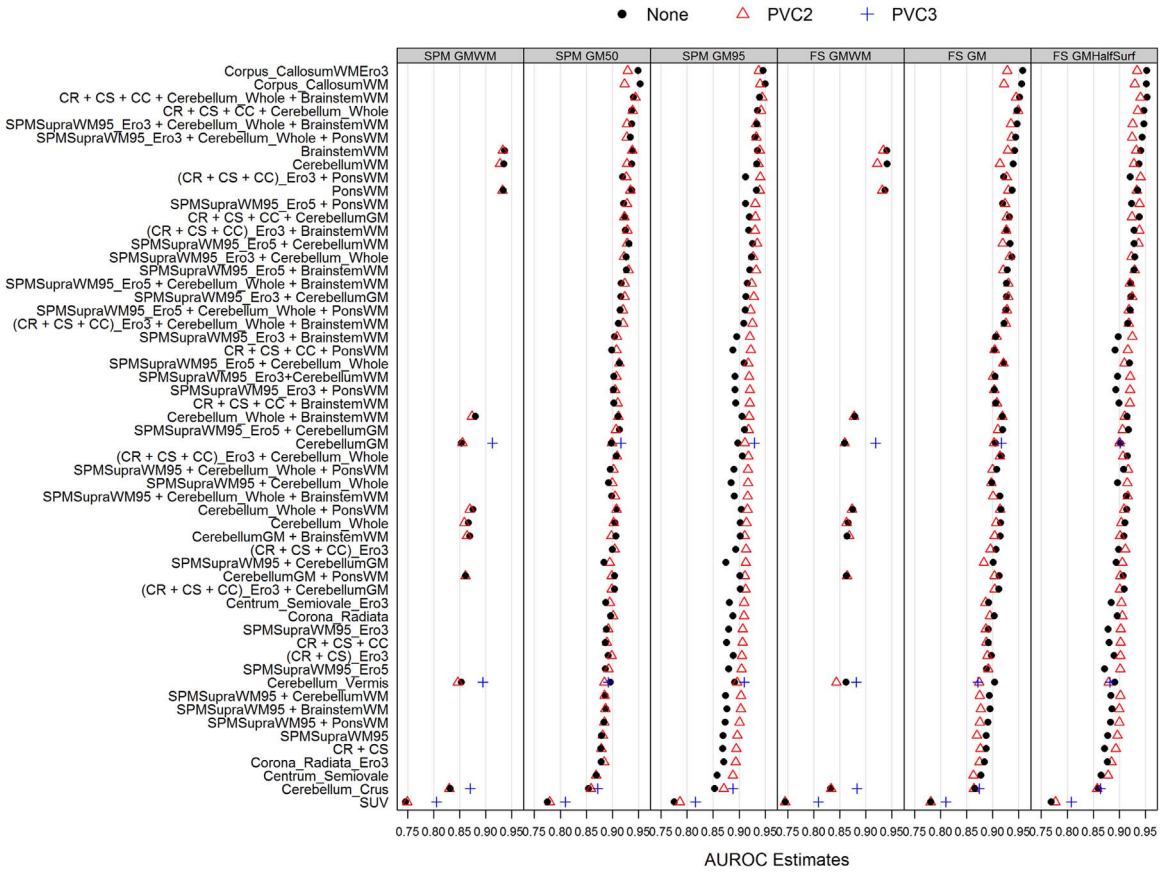

Fig. 7. Plots of AUROC estimates. Subjects are separated into 2 groups using baseline global PiB; 50 subjects with PiB < 1.35 and 40 subjects with PiB between 1.5 and 2.2. Rates of change in PiB are from linear regression. References are ranked according to the maximum AUROC estimates across all PVC and target segmentation methods.

specific choice of weightings for each criterion, we also examined the effects of using equal weightings for each criterion, and of reversing the weightings for each criteria (assigning 0.1 to those regularly assigned 0.4, and vice versa). In that analysis, we found that most major conclusions were not altered by either permutation, and thus the analysis is not sensitive to alterations in these weights. For space reasons, we present this data only in the Supplementary materials.

4. Results

Results of each individual evaluation criterion are presented in Fig. 5–8, and of the combined, weighted criterion in Fig. 9. In each of these figures, we list reference ROIs from best to worst, ranking each according to its best performing combination of PVC and target-segmentation (i.e. top-scoring plotted point within that row). For space reasons, we present only variations where calculations were performed with PET images upsampled to match MRI's voxel space, because in almost every case, for all criteria, performance of these were slightly better than their corresponding PET-space counterparts where all else was the same (data not shown). In this section we discuss the best- and worst-performing pipelines according to each plot. In the following section, we discuss these results in aggregate and give overall recommendations.

We present results of the longitudinal reliability (straightness) criterion in Fig. 5. By this criterion, differences across segmentations (columns) and across PVC variations were generally minimal, so we focus on reference region differences. SUV-using pipelines performed worse than all SUVR-using pipelines. The top five candidates all included some form of eroded supratentorial white matter combined with some area of infratentorial WM. Each of these combined supra/infra variants performed better than its individual components. Cerebellar WM alone also performed well, achieving the seventh position. Some SUVR variants where supratentorial WM was combined with the whole cerebellum (with or without brainstem) also performed near the top of the list.

In Fig. 6 we present the results of the non-negative (plausibility) criteria. Again, target ROI segmentation method did not have a large influence among top-performing reference ROIs. In most instances, 2-class PVC outperformed variants without PVC in otherwise-equivalent pairings. Of the top-10 reference choices, all but one contained the whole cerebellum, and seven of these were in combination with some form of supratentorial WM. Again, references using supratentorial WM combined with infratentorial regions performed well, and better than their individual components. Cerebellar WM alone, which performed well in Fig. 5, was the worst performer in this criterion.

We present the results of our separability (accumulator group versus non-accumulator group) analysis in Fig. 7. Corpus callosum references performed best, followed by combinations of supratentorial WM and whole cerebellum (optionally with pons/brainstem), and brainstem/pons/Cerebellar WM alone. Cerebellar GM only performed reasonably when using 3-class PVC, but was among the worst performers otherwise. SUV pipelines performed worse than all others.

Fig. 8 shows the results of testing correlation between change in measured uptake and change in MMSE. In this criterion, variations with 2-class PVC were either equivalent to or outperformed those without PVC and those with 3-class PVC in almost all instances. Differences across target spaces were comparatively small. Corpus callosum references performed near the top, in agreement

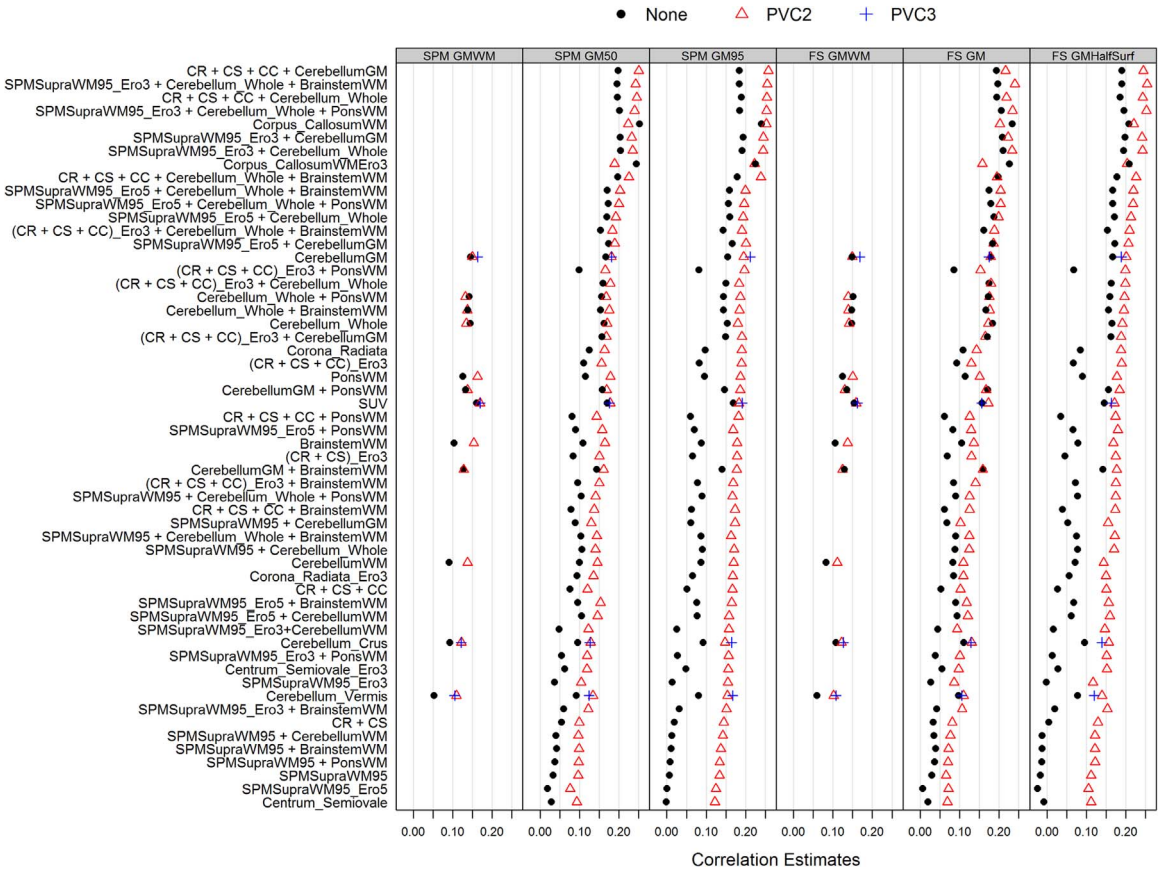

Fig. 8. Plots of Spearman's correlation estimates between rates of change in PiB and rates of change in MMSE (both are from linear regression). References are ranked according to the maximum correlation estimates across all PVC and target segmentation methods.

with the separability criteria (Fig. 7). Otherwise, top performing reference methods primarily included a combination of supratentorial WM and cerebellum (GM or whole), with or without brainstem.

In Fig. 9 we present the results of analysis with our combined criterion, which is a weighted combination of the other four (results with alternative weightings, the major conclusions from which are highly consistent with the presented weighting, are provided in Supplementary material). By the combined criteria, the best performing references were those with a combination of supratentorial WM and whole cerebellum, sometimes also containing brainstem and/or pons. Each of these combinations outperformed their individual components. This result is consistent with the fact that such pipelines were among the top performers by many of the individual criteria. Also highly consistent with the individual criteria, SUV pipelines were the worst performers, and differences across target segmentations were relatively small. 2-class PVC outperformed no-PVC in most otherwise-equivalent pairings. 3-class PVC, where applicable, had mixed performance versus other PVC methods.

We plot the statistical significance of differences between pipelines in Fig. 10. By this analysis, many top-performing variations of combined supra-/infratentorial references (e.g. including versus excluding brainstem, eroded segmentation versus atlas implementations) did not differ significantly from each other; however, the top-performing pipeline does differ significantly from all pipelines using non-WM regions only, or using supra-/infratentorial regions alone. Therefore, we consider this combined reference region significantly superior to these other options.

5. Discussion and conclusions

In this section, we first discuss the results as they relate to our objective questions. Then, we give our final pipeline recommendations and discuss the strengths and limitations of our study.

5.1. Question 1: is partial volume correction helpful?

Our results suggest that under most combinations of other factors, two-class (Meltzer) PVC improved results according to all of our criteria when compared to no-PVC pipelines. Three-class PVC was not applicable to a large majority of pipelines tested because these include WM, which 3-class PVC attempts to remove, in the reference regions. Three-class PVC was often superior among the pipelines with GM-only references, although pipelines using these GM-only references were otherwise not among the top candidates.

5.2. Question 2: which is the optimal reference region?

This question has received the most attention in prior literature, and also most strongly impacted our analyses. Overall, our results strongly suggest the superiority of reference regions containing both supratentorial WM and whole cerebellum, optionally also including the pons or entire brainstem. Supratentorial WM may be implemented equivalently either by segmenting each individual T1 and eroding the WM voxels with a radius of 3 or 5 voxels, or by using an atlas where the relevant WM ROIs exclude WM near the cortex. Using individual segmentations without such erosion was generally worse. These combined-reference-ROI

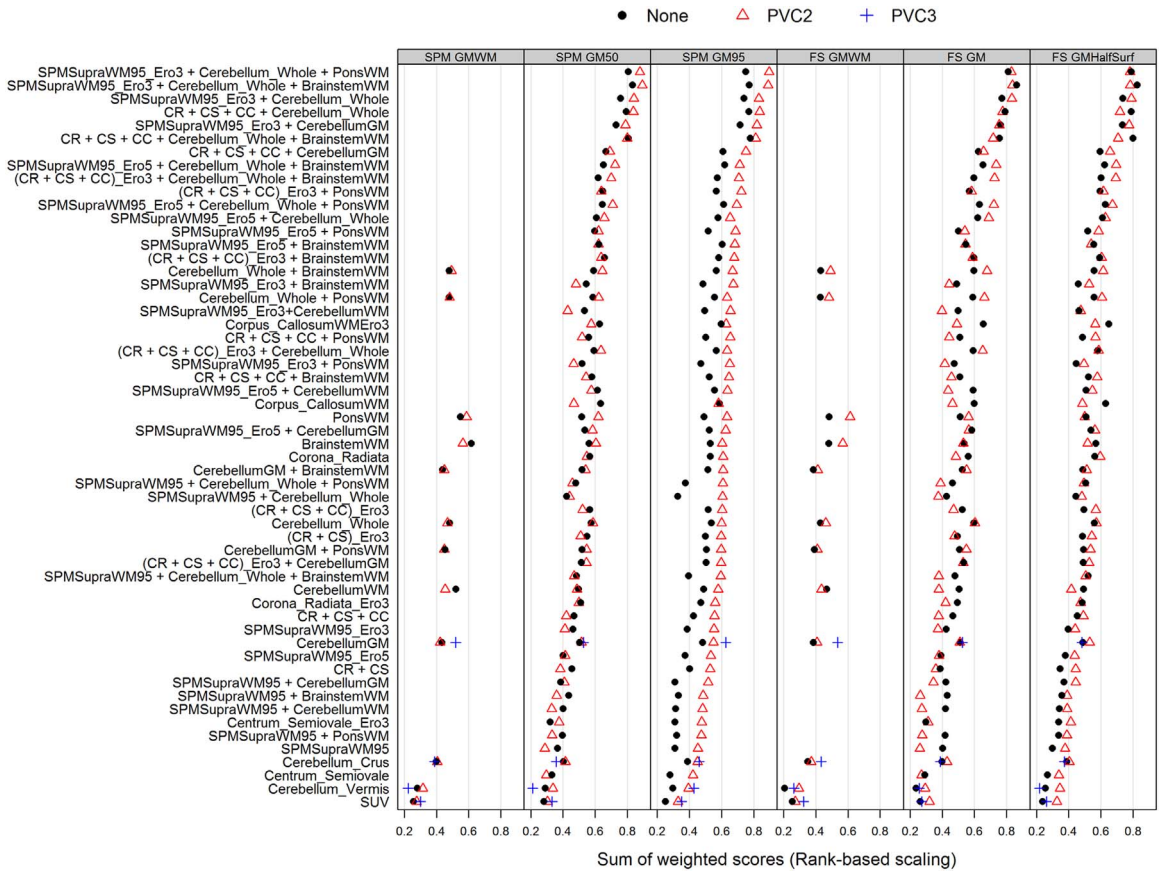

Fig. 9. Plots of sum of weighted scores. Data were normalized using rank-based scaling. (0.4*reliability + 0.4*trending non-negative + 0.1*AUROC estimates + 0.1*correlation between Δ PiB and Δ MMSE). References are ranked according to the maximum correlation estimates across all PVC and target segmentation methods. Results using alternate weightings are plotted in Supplementary material.

approaches significantly outperform those with any of their components individually (in Fig. 10, the top-performing pipeline, which uses this combined-reference approach, differs significantly from all pipelines using non-WM or supra-/infratentorial regions alone). We visualize these winning variations in Fig. 11.

5.3. Question 3: which is the optimal GM segmentation within the target region?

Among the top-performing pipelines, differences between target segmentation methods were not significant (Fig. 10). Across all pipelines, this had a relatively small impact on our quality criteria compared to other methodological factors.

5.4. Question 4: which is the optimal analysis space?

In almost all cases, pairings of pipelines differing on only this variable trended toward slightly better performance in MRI-space. We omitted showing this data for space reasons.

5.5. Question 5: SUV or SUVR?

Pipelines using SUV, rather than a reference region (SUVR), were by far the worst performing on the straightness and separability criteria, and they were only mid-level performers on the other two. Using the combined criterion, even the worst-performing SUVR pipelines generally outperformed the best-performing SUV pipelines. Therefore, we conclude that use of SUVR with any reasonable reference region is superior to SUV.

5.6. Overall recommended pipelines

Our study suggests that an optimal pipeline to measure change in β-amyloid from PiB scans should use two-class PVC, perform calculations in the voxel space of the MRI, and use SUVR with a reference region containing both supratentorial WM and the whole cerebellum, optionally also including pons/brainstem (Fig. 11).

6. Discussion

Our analysis is unique for its finding that references containing supratentorial WM and whole cerebellum together are significantly superior to either alone for analysis of serial PiB scans. Some previous comparisons using florbetapir have also examined these “composite” reference regions (Landau et al., 2015), but such results could theoretically not be the same for PiB. Some other works favoring supratentorial WM references have hypothesized that the cerebellum is noisier due to its relatively peripheral location in the field of view, where scanner sensitivity is lower, or due to its relatively smaller size (Chen et al., 2015; Landau et al., 2015). Our data supports the hypothesis that cerebellar ROIs are noisier than supratentorial WM references (In Fig. 5, the longitudinal reliability criteria, references containing only cerebellar ROIs are among the worst performers). However, the statistical equivalence of including brainstem, and the superiority of some smaller WM variants and some larger WM variants, suggests that larger references are not always superior, perhaps due to diminishing returns, or because larger WM regions tend to include more

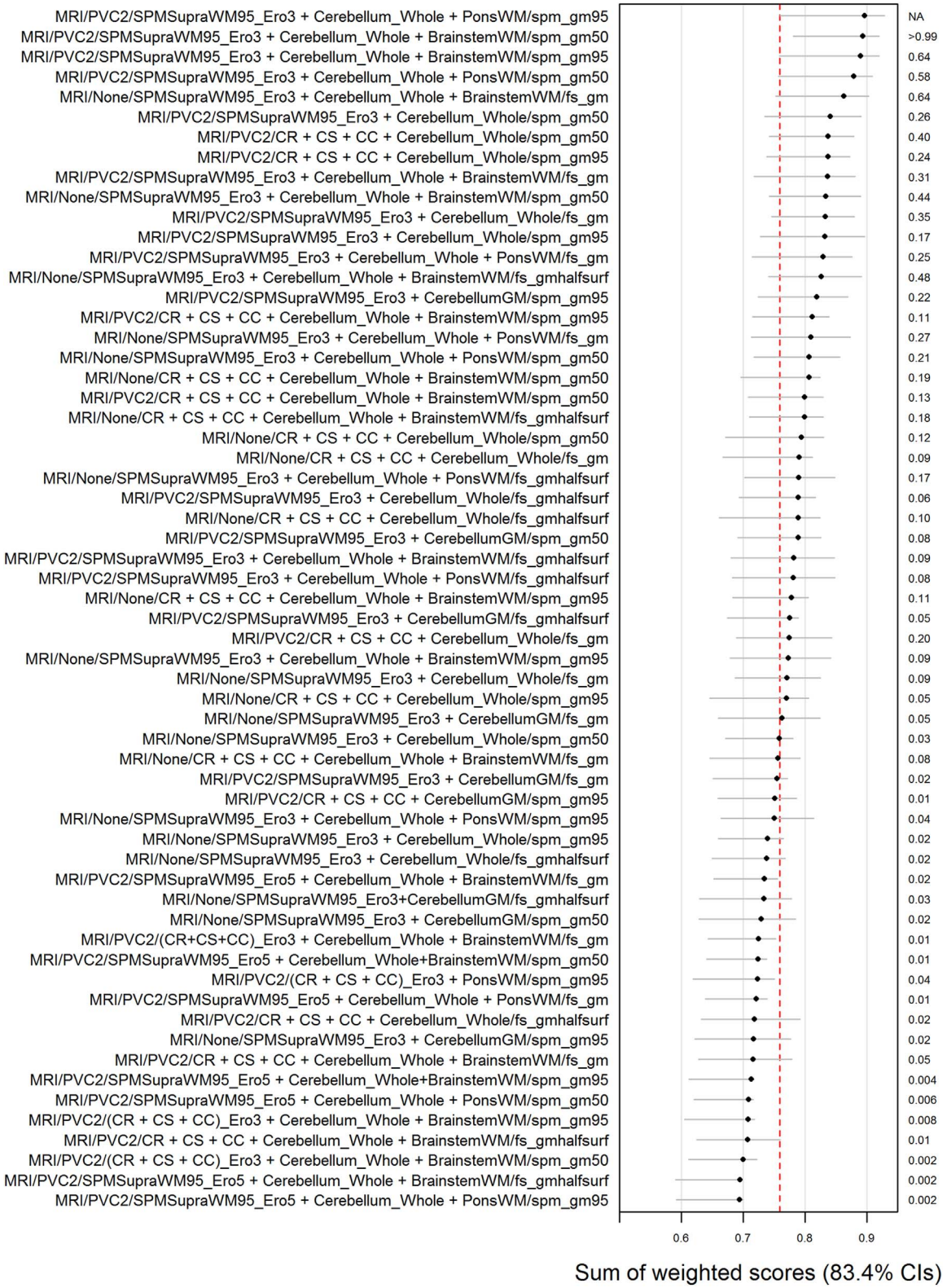

Fig. 10. Statistical Significance of Differences Between Top Pipelines: Top-performing pipelines according to the weighted-sum combined criterion are shown with their 83.4% confidence intervals according to 1000 bootstrap replicates. The rightmost column lists *p*-values of differences between the best pipeline (first row) and each subsequent row. The vertical, red dotted line follows the lower end of the best pipeline's confidence interval, and represents the approximate point at which differences become significant at $p < 0.05$.

contamination from adjacent GM.

Although we had no strong reason to believe *a priori* that our findings would agree with those of previous studies using Fluorine ligands, our findings were consistent with previous studies using Florbetapir in suggesting that references containing supratentorial WM are superior to purely-infratentorial references (Chen et al., 2015; Landau et al., 2015). This agreement with prior studies using differing ligands, populations, and measurement criteria adds to a

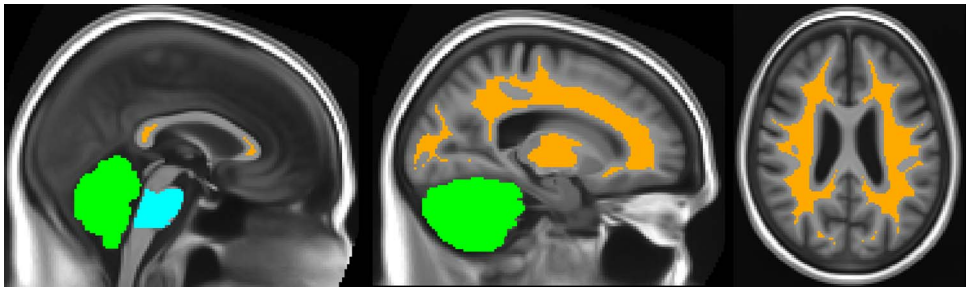

Fig. 11. Winning Reference ROI: The top-performing reference ROI according to our combined criterion includes all voxels colored above. Orange: Supratentorial WM segmented by SPM12, thresholded at 0.95, and eroded by 3 voxels Green: Whole Cerebellum Blue: Pons (masked to include only voxels segmented as WM) Not pictured: other variations that were statistically equivalent, including atlas-based supratentorial WM variations, and possible omission of pons and/or brainstem.

growing body of evidence that supratentorial WM references are superior to purely-infratentorial references for serial Amyloid PET measurements regardless of tracer used. This agreement across studies using differing criteria also suggests that this finding is strong enough to be seen consistently despite the lack of gold-standard validation measures and despite the limitations of each criterion that has been used in their place.

6.1. Strengths and limitations of current study

The primary strength of this work is its large scope, attempting to test most reasonable combinations of methods across several variables according to four different longitudinal criteria. We use three timepoints for all subjects, giving increased power and enabling more analysis criteria, such as trajectory reliability, versus previous studies using two. This work also addresses a relative lack of such comparisons using PiB, as opposed to ^{18}F-based amyloid PET ligands. It is important to stress that our work applies only to longitudinal measurements of PiB. We do not assume that our findings necessarily apply to cross-sectional measurements; follow-up studies will be needed to examine this question. We also leave for future work a more exhaustive comparison of PVC methods with more variations, e.g. ROI-based methods (Rousset et al., 1998) and comparisons with methods using data-driven selections of individual voxels (Borghammer et al., 2009; Carbonell et al., 2015; Razifar et al., 2009). We also did not examine PET-only methods (Bilgel et al., 2015; Fripp et al., 2008; Raniga et al., 2007; Zhou et al., 2014) because MRI is generally available in most studies that include amyloid PET.

One limitation of our study is the potential for circularity in some criteria: although we examine different varieties of SUVR measurements, it was necessary to use cross-sectional SUVR measurements for study subject selection, and for their division into subgroups for the group separability criterion. Theoretically, such circularity could bias results toward favoring methods that most resemble the selection criteria. Although we attempted to minimize this potential by using a method that was not among those tested, the selection method was still based on SUVR with two-class PVC and a cerebellar gray reference, and this could have biased results toward similar pipelines. However, methods using cerebellar GM references performed relatively poorly in our analysis, suggesting that this potential bias was not strong enough to incorrectly favor these methods. For the group separability analysis, we also selected thresholded ranges that excluded many subjects between the extremes to minimize the ability for subtle changes in SUVR measurements to impact group selection.

Because external, gold-standard measurements (i.e. autopsy) of change over time in ß-amyloid do not exist, our study is primarily limited by the assumptions of each criterion used instead (see Section 3.5). In addition to carefully designing each criterion to minimize the impact of these limitations, we also created the combined criterion that allows each to compensate for the other's shortcomings, and we weighted each metric in proportion with our confidence in the universality of its assumptions in our dataset. While these weights were chosen *a priori*, our major findings were largely consistent across all four measurement criteria, and thus across different weightings of the combined criteria (see Supplementary Material). This consistency of our findings across criteria that measured fundamentally different properties of each pipeline suggests that these criteria were each reasonable, despite their varying assumptions and limitations, and adds confidence to our findings overall.

Our study adds to the growing evidence in favor of analysis methods that include WM voxels within reference regions. One potential caveat of WM-containing references for longitudinal analyses is the potential for a confound by changes in WM myelination, which have been shown to affect uptake of ß-amyloid PET ligands (Stankoff et al., 2011; Veronese et al., 2015). It is possible that such WM changes over long periods of time could affect the value of a WM-containing reference ROI. However, this effect would likely be reduced by the inclusion of infratentorial voxels in composite references, which are those that performed best in this analysis.

Acknowledgements

The authors would like to thank these funding sources: NIH R01-AG011378; NIH R00-AG37573; NIH R01-AG041851; NIH U01-AG24904; NIH U01-AG06786; NIH P50-AG16574; NIH R01-AG034676; The Alexander Family Professorship of Alzheimer's Disease Research, Mayo Clinic; the GHR Foundation, Elsie and Marvin Dekelboum Family Foundation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuroimage.2016.08.056.

References

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95–113. http://dx.doi.org/10.1016/j.neuroimage.2007.07.007.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851. http://dx.doi.org/10.1016/j.neuroimage.2005.02.018.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41. http://dx.doi.org/10.1016/j.media.2007.06.004.

Bilgel, M., Carass, A., Resnick, S.M., Wong, D.F., Prince, J.L., 2015. Deformation field correction for spatial normalization of PET images. Neuroimage 119, 152–163. http://dx.doi.org/10.1016/j.neuroimage.2015.06.063.

Borghammer, P., Aanerud, J., Gjedde, A., 2009. Data-driven intensity normalization

of PET group comparison studies is superior to global mean normalization. Neuroimage 46, 981–988. http://dx.doi.org/10.1016/j.neuroimage.2009.03.021.
Brendel, M., Högenauer, M., Delker, A., Sauerbeck, J., Bartenstein, P., Seibyl, J., Rominger, A., 2015. Improved longitudinal [18F]-AV45 amyloid PET by white matter reference and VOI-based partial volume effect correction. Neuroimage 108, 450–459. http://dx.doi.org/10.1016/j.neuroimage.2014.11.055.
Buchhave, P., Minthon, L., Zetterberg, H., Wallin, A.K., Blennow, K., Hansson, O., 2012. Cerebrospinal fluid levels of β-amyloid 1-42, but not of tau, are fully changed already 5 to 10 years before the onset of Alzheimer dementia. Arch. Gen. Psychiatry 69, 98–106. http://dx.doi.org/10.1001/archgenpsychiatry.2011.155.
Carbonell, F., Zijdenbos, A.P., Charil, A., Grand'Maison, M., Bedell, B., 2015. Optimal Target Region for Subject Classification based on Amyloid PET Images. J. Nucl. Med., 1351–1358. http://dx.doi.org/10.2967/jnumed.115.158774.
Chen, K., Roontiva, A., Thiyyagura, P., Lee, W., Liu, X., Ayutyanont, N., Protas, H., Luo, J.L., Bauer, R., Reschke, C., Bandy, D., Koeppe, R.A., Fleisher, A.S., Caselli, R.J., Landau, S., Jagust, W.J., Weiner, M.W., Reiman, E.M., 2015. Improved power for characterizing longitudinal amyloid-β PET changes and evaluating amyloid-modifying treatments with a cerebral white matter reference region. J. Nucl. Med. 56, 560–566. http://dx.doi.org/10.2967/jnumed.114.149732.
Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781. http://dx.doi.org/10.1016/j.neuroimage.2012.01.021.
Fleisher, A.S., Roontiva, A., Reschke, C., Bandy, D., Reiman, E.M., Protas, H., Luo, J., Chen, K., Weiner, M.W., Ayutyanont, N., Thiyyagura, P., Caselli, R.J., Baur, R.I., Koeppe, R., Landau, S., Lee, W., Jagust, W., Liu, X., 2014. Improving the Power to Track Fabrillar Amyloid PET Measurements and Evaluate Amyloid-Modifying Treatments using a Cerebral White Matter Reference Region-of-Interest, in: Alzheimer's Association International Conference (AAIC). Elsevier, Copenhagen, Denmark.
Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state" A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198, doi:0022-3956(75)90026-6 [pii].
Fripp, J., Bourgeat, P., Raniga, P., Acosta, O., Villemagne, V., Jones, G., O'Keefe, G., Rowe, C., 2008. MR-less high dimensional spatial normalization of 11C PiB PET images on a population of elderly, mild cognitive impaired and alzheimer disease patients. Med. Image Comput. Comput. Interv. 5241, 442–449. http://dx.doi.org/10.1007/978-3-540-85988-8_53.
How to Convert from FreeSurfer Space Back to Native Anatomical Space [WWW Document], 2015. URL ⟨http://surfer.nmr.mgh.harvard.edu/fswiki/FsAnat-to-NativeAnat⟩ (accessed 01.01.15).
Ingelsson, M., Fukumoto, H., Newell, K.L., Growdon, J.H., Hedley-Whyte, E.T., Frosch, M.P., Albert, M.S., Hyman, B.T., Irizarry, M.C., 2004. Early Abeta accumulation and progressive synaptic loss, gliosis, and tangle formation in AD brain. Neurology 62, 925–931. http://dx.doi.org/10.1212/01.WNL.0000115115.98960.37.
Jack, C.R.J., Lowe, V.J., Senjem, M.L., Weigand, S.D., Kemp, B.J., Shiung, M.M., Knopman, D.S., Boeve, B.F., Klunk, W.E., Mathis, C.A., Petersen, R.C., 2008. 11C PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment. Brain 131, 665–680. http://dx.doi.org/10.1093/brain/awm336.
Jack, C.R.J., Wiste, H.J., Lesnick, T.G., Weigand, S.D., Knopman, D.S., Vemuri, P., Pankratz, V.S., Senjem, M.L., Gunter, J.L., Mielke, M.M., Lowe, V.J., Boeve, B.F., Petersen, R.C., 2013. Brain β-amyloid load approaches a plateau. Neurology 80, 890–896. http://dx.doi.org/10.1212/WNL.0b013e3182840bbe.
Joshi, A., Koeppe, R.A., Fessler, J.A., 2009. Reducing between scanner differences in multi-center PET studies. Neuroimage 46, 154–159. http://dx.doi.org/10.1016/j.neuroimage.2009.01.057.
Joshi, A., Kennedy, I.A., Mintun, M., Pontecorvo, M., Navitsky, M.A., Devous, M.D., 2014. Measuring change in beta amyloid burden over time using florbetapir-PET and a subcortical white matter reference region, in: Alzheimer's Association International Conference (AAIC). Elsevier, Copenhagen, Denmark, p. 2014.
Jovicich, J., Czanner, S., Greve, D., Haley, E., Van Der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., Dale, A., 2006. Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. Neuroimage 30, 436–443. http://dx.doi.org/10.1016/j.neuroimage.2005.09.046.
Klein, G., Chiao, P., Barakos, J., Purcell, D., Sampat, M., Oh, J., Sevigny, J., Suhy, J., 2014. Concordance of Quantitative SUVR Methods with Visual Assessment of Florbetapir PET Screening Results, in: Alzheimer's Association International Conference (AAIC). Elsevier, Copenhagen, Denmark.
Klunk, W.E., Koeppe, R.A., Price, J.C., Benzinger, T.L., Devous, M.D., Jagust, W.J., Johnson, K.A., Mathis, C.A., Minhas, D., Pontecorvo, M.J., Rowe, C.C., Skovronsky, D.M., Mintun, M.A., 2015. The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. Alzheimer's Dement 11, 1–15. http://dx.doi.org/10.1016/j.jalz.2014.07.003.
Klunk, W.E., Engler, H., Nordberg, A., Wang, Y., Blomqvist, G., Holt, D.P., Bergström, M., Savitcheva, I., Huang, G.F., Estrada, S., Ausén, B., Debnath, M.L., Barletta, J., Price, J.C., Sandell, J., Lopresti, B.J., Wall, A., Koivisto, P., Antoni, G., Mathis, C.A., Långström, B., 2004. Imaging brain amyloid in Alzheimer's disease with pittsburgh compound-B. Ann. Neurol. 55, 306–319. http://dx.doi.org/10.1002/ana.20009.
Knol, M.J., Pestman, W.R., Grobbee, D.E., 2011. The (mis)use of overlap of confidence intervals to assess effect modification. Eur. J. Epidemiol. 26, 253–254. http://dx.doi.org/10.1007/s10654-011-9563-8.
Landau, S.M., Fero, A., Baker, S.L., Koeppe, R., Mintun, M., Chen, K., Reiman, E.M., Jagust, W.J., 2015. Measurement of Longitudinal B-Amyloid Change with 18F-Florbetapir PET and Standardized Uptake Value Ratios. J. Nucl. Med. 56, 567–574. http://dx.doi.org/10.2967/jnumed.114.148981.
Liu, E., Di, J., Booth, K., Brashear, R.H., Novak, G., Margolin, R., 2014. Evaluation of Cerebral Gray Matter and Pons as Reference Regions for Amyloid PET: Results from a Bapineuzumab Subcutaneous Phase 2 Trial, in: Alzheimer's Association International Conference (AAIC). Elsevier, Copenhagen, Denmark, p. 2014.
Lopresti, B., Klunk, W., Bi, W., Cohen, A., Mathis, C., Price, J., 2011. Use of PONS as a normalizing region for [c-11]pib-pet scans: Effect on subject classification. Alzheimer's Dement 7, S61. http://dx.doi.org/10.1016/j.jalz.2011.05.093.
Lowe, V.J., Kemp, B.J., Jack, C.R., Senjem, M., Weigand, S., Shiung, M., Smith, G., Knopman, D., Boeve, B., Mullan, B., Petersen, R.C., 2009. Comparison of 18F-FDG and PiB PET in cognitive impairment. J. Nucl. Med. 50, 878–886. http://dx.doi.org/10.2967/jnumed.108.058529.
Meltzer, C.C., Leal, J.P., Mayberg, H.S., Wagner, H.N.J., Frost, J.J., 1990. Correction of PET data for Partial Volume Effects in Human Cerebral Cortex by MR Imaging. J. Comput. Assist. Tomogr. 14, 561–570.
Müller-Gärtner, H.W., Links, J.M., Prince, J.L., Bryan, R.N., McVeigh, E., Leal, J.P., Davatzikos, C., Frost, J.J., 1992. Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. J. Cereb. Blood Flow Metab. 12, 571–583. http://dx.doi.org/10.1038/jcbfm.1992.81.
Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J.T., Miller, M.I., van Zijl, P.C.M., Albert, M., Lyketsos, C.G., Woods, R., Toga, A.W., Pike, G.B., Rosa-Neto, P., Evans, A., Mazziotta, J., Mori, S., 2009. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. Neuroimage 46, 486–499. http://dx.doi.org/10.1016/j.neuroimage.2009.01.002.
Petersen, R.C., Roberts, R.O., Knopman, D.S., Geda, Y.E., Cha, R.H., Pankratz, V.S., Boeve, B.F., Tangalos, E.G., Ivnik, R.J., Rocca, W.A., 2010. Prevalence of mild cognitive impairment is higher in men. The Mayo Clinic Study of Aging. Neurology 75, 889–897. http://dx.doi.org/10.1212/WNL.0b013e3181f11d85.
Raniga, P., Bourgeat, P., Ourselin, S., Villemagne, V., O'Keefe, G., Rowe, C., 2007. PIB-PET Segmentation for Automatic SUVR Normalisation without MR Information. Biomed. Imaging, Int. Symp., 348–351. http://dx.doi.org/10.1109/ISBI.2007.356860.
Razifar, P., Engler, H., Ringheim, A., Estrada, S., Wall, A., Långström, B., 2009. An automated method for delineating a reference region using masked volume-wise principal-component analysis in 11C-PIB PET. J. Nucl. Med. Technol. 37, 38–44. http://dx.doi.org/10.2967/jnmt.108.054296.
Roberts, R.O., Geda, Y.E., Knopman, D.S., Cha, R.H., Pankratz, V.S., Boeve, B.F., Ivnik, R.J., Tangalos, E.G., Petersen, R.C., Rocca, W.A., 2008. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology 30, 58–69. http://dx.doi.org/10.1159/000115751.
Rousset, O.G., Ma, Y., Evans, A.C., 1998. Correction for partial volume effects in PET: principle and validation. J. Nucl. Med. 39, 904–911.
Schmidt, M.E., Chiao, P., Klein, G., Matthews, D., Thurfjell, L., Cole, P.E., Margolin, R., Landau, S., Foster, N.L., Mason, N.S., De Santi, S., Suhy, J., Koeppe, R.A., Jagust, W., 2015. The influence of biological and technical factors on quantitative analysis of amyloid PET: Points to consider and recommendations for controlling variability in longitudinal data. Alzheimers Dement 11, 1050–1068. http://dx.doi.org/10.1016/j.jalz.2014.09.004.
Shaw, L.M., Vanderstichele, H., Knapik-Czajka, M., Clark, C.M., Aisen, P.S., Petersen, R.C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V.M.-Y., Trojanowski, J.Q., 2009. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann. Neurol. 65, 403–413. http://dx.doi.org/10.1002/ana.21610.
Stankoff, B., Freeman, L., Aigrot, M.S., Chardain, A., Dollé, F., Williams, A., Galanaud, D., Armand, L., Lehericy, S., Lubetzki, C., Zalc, B., Bottlaender, M., 2011. Imaging central nervous system myelin by positron emission tomography in multiple sclerosis using [methyl-11C]−2-(4-methylaminophenyl)− 6-hydroxybenzothiazole. Ann. Neurol. 69, 673–680. http://dx.doi.org/10.1002/ana.22320.
Su, Y., Blazey, T.M., Snyder, A.Z., Raichle, M.E., Marcus, D.S., Ances, B.M., Bateman, R.J., Cairns, N.J., Aldea, P., Cash, L., Christensen, J.J., Friedrichsen, K., Hornbeck, R.C., Farrar, A.M., Owen, C.J., Mayeux, R., Brickman, A.M., Klunk, W., Price, J.C., Thompson, P.M., Ghetti, B., Saykin, A.J., Sperling, R.A., Johnson, K.A., Scho, P.R., Buckles, V., Morris, J.C., Benzinger, T.L.S., Alzheimer, I., 2015. Partial volume correction in quantitative amyloid imaging. Neuroimage 107, 55–64. http://dx.doi.org/10.1016/j.neuroimage.2014.11.058.
Thie, J.A., 2004. Understanding the standardized uptake value, its methods, and implications for usage. J. Nucl. Med. 45, 1431–1434, doi:45/9/1431 [pii].
Thurfjell, L., Lilja, J., Lundqvist, R., Buckley, C., Smith, A., Vandenberghe, R., Sherwin, P., 2014. Automated quantification of 18F-flutemetamol PET activity for categorizing scans as negative or positive for brain amyloid: concordance with visual image reads. J. Nucl. Med. 55, 1623–1628. http://dx.doi.org/10.2967/jnumed.114.142109.
van Berckel, B.N.M., Ossenkoppele, R., Tolboom, N., Yaqub, M., Foster-Dingley, J.C., Windhorst, A.D., Scheltens, P., Lammertsma, A.A., Boellaard, R., 2013. Longitudinal amyloid imaging using 11C-PiB: methodologic considerations. J. Nucl. Med. 54, 1570–1576. http://dx.doi.org/10.2967/jnumed.112.113654.
Veronese, M., Bodini, B., García-Lorenzo, D., Battaglini, M., Bongarzone, S., Comtat, C., Bottlaender, M., Stankoff, B., Turkheimer, F.E., 2015. Quantification of [(11)C] PIB PET for imaging myelin in the human brain: a test-retest reproducibility study in high-resolution research tomography. J. Cereb. Blood Flow Metab. 35, 1771–1782. http://dx.doi.org/10.1038/jcbfm.2015.120.
Villain, N., Chételat, G., Grassiot, B., Bourgeat, P., Jones, G., Ellis, K.A., Ames, D., Martins, R.N., Eustache, F., Salvado, O., Masters, C.L., Rowe, C.C., Villemagne, V.L., 2012. Regional dynamics of amyloid-β deposition in healthy elderly, mild cognitive impairment and Alzheimer's disease: a voxelwise PiB-PET

Please cite this article as: Schwarz, C.G., et al., Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and.... NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.08.056

longitudinal study. Brain 135, 2126–2139. http://dx.doi.org/10.1093/brain/aws125.

Villemagne, V.L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K.A., Salvado, O., Szoeke, C., Macaulay, S.L., Martins, R., Maruff, P., Ames, D., Rowe, C.C., Masters, C.L., 2013. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. Lancet Neurol. 12, 357–367. http://dx.doi.org/10.1016/S1474-4422(13)70044-9.

Whitwell, J.L., Josephs, K.A., Murray, M.E., Kantarci, K., Przybelski, S.A., Weigand, S. D., Vemuri, P., Senjem, M.L., Parisi, J.E., Knopman, D.S., Boeve, B.F., Petersen, R.C., Dickson, D.W., Jack, C.R., 2008. MRI correlates of neurofibrillary tangle pathology at autopsy: A voxel-based morphometry study. Neurology 71, 743–749. http://dx.doi.org/10.1212/01.wnl.0000324924.91351.7d.

Zasadny, K.R., Wahl, R.L., 1993. Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: variations with body weight and a method for correction. Radiology 189, 847–850. http://dx.doi.org/10.1148/radiology.189.3.8234714.

Zhou, L., Salvado, O., Dore, V., Bourgeat, P., Raniga, P., Macaulay, S.L., Ames, D., Masters, C.L., Ellis, K.A., Villemagne, V.L., Rowe, C.C., Fripp, J., 2014. MR-less surface-based amyloid assessment based on 11C PiB PET. PLoS One 9, e84777. http://dx.doi.org/10.1371/journal.pone.0084777.

Zhou, Y., Resnick, S.M., Ye, W., Fan, H., Holt, D.P., Klunk, W.E., Mathis, C.A., Dannals, R., Wong, D.F., 2007. Using a reference tissue model with spatial constraint to quantify [11C]Pittsburgh compound B PET for early diagnosis of Alzheimer's disease. Neuroimage 36, 298–312. http://dx.doi.org/10.1016/j.neuroimage.2007.03.004.

Please cite this article as: Schwarz, C.G., et al., Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and.... NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.08.056