HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 September 01.

Published in final edited form as:

Neuroimage. 2021 September ; 238: 118259. doi:10.1016/j.neuroimage.2021.118259.

Selecting software pipelines for change in flortaucipir SUVR: Balancing repeatability and group separation

*Corresponding author. schwarz.christopher@mayo.edu (C.G. Schwarz).

Credit authorship contribution statement

Christopher G. Schwarz: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Terry M. Therneau:** Methodology, Formal analysis, Visualization, Writing - review & editing. **Stephen D. Weigand:** Methodology, Formal analysis, Writing - review & editing. **Jeffrey L. Gunter:** Software, Writing - review & editing. **Val J. Lowe:** Resources, Data curation, Writing - review & editing. **Scott A. Przybelski:** Data curation, Writing - review & editing. **Matthew L. Senjem:** Software, Writing - review & editing. **Hugo Botha:** Data curation, Writing - review & editing. **Prashanthi Vemuri:** Resources, Writing - review & editing, Funding acquisition. **Kejal Kantarci:** Resources, Writing - review & editing, Funding acquisition. **Bradley F. Boeve:** Resources, Writing - review & editing, Funding acquisition. **Jennifer L. Whitwell:** Resources, Writing - review & editing, Funding acquisition. **Keith A. Josephs:** Resources, Writing - review & editing, Funding acquisition. **Ronald C. Petersen:** Resources, Writing - review & editing, Funding acquisition. **David S. Knopman:** Resources, Writing - review & editing, Funding acquisition. **Clifford R. Jack Jr:** Conceptualization, Resources, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

Christopher Schwarz receives research support from the NIH.

Terry Therneau reports no disclosures.

Stephen Weigand reports no disclosures.

Jeffrey Gunter reports no disclosures.

Val Lowe consults for Bayer Schering Pharma, Piramal Life Sciences, Eisai, Inc., and Merck Research and receives research support from GE Healthcare, Siemens Molecular Imaging, AVID Radiopharmaceuticals and the NIH (NIA, NCI).

Scott Przybelski reports no disclosures.

Prashanthi Vemuri is funded by the NIH.

Matthew Senjem has owned stock in medical related companies, unrelated to the current work within the past 12 months: Align Technology, Inc., Inovio Pharmaceuticals, Inc., LHC Group, Inc., Mesa Laboratories, Inc., Natus Medical Inc., and Varex Imaging Corporation. He has also owned stock in these medical related companies within the past three years, unrelated to the current work: CRISPR Therapeutics, Gilead Sciences, Inc., Ionis Pharmaceuticals, Johnson & Johnson, Medtronic, Inc.

Hugo Botha is funded by the NIH.

Kejal Kantarci serves on the data safety monitoring board for Takeda Global Research and Development Center, Inc., receives research support from Avid Radioparmaceuticals and Eli Lilly, and receives funding from NIH and Alzheimer's Drug Discovery Foundation.

Bradley Boeve has served as an investigator for clinical trials sponsored by Axovant and Biogen. He receives royalties from the publication of a book entitled Behavioral Neurology Of Dementia (Cambridge Medicine, 2009, 2017). He serves on the Scientific Advisory Board of the Tau Consortium. He receives research support from the NIH, the Mayo Clinic Dorothy and Harry T. Mangurian Jr. Lewy Body Dementia Program and the Little Family Foundation.

Jennifer Whitwell is funded by the NIH.

Keith Josephs is funded by the NIH.

Ronald Petersen is a consultant for Roche, Inc., Biogen, Inc., and Eisai, Inc., served on a DSMB for Genentech, Inc.; receives royalties from publishing Mild Cognitive Impairment (Oxford University Press, 2003) and UpToDate; and receives research support from the NIH (P30 AG062677 (PI) and U01-AG006786 (PI), R01-AG011378 (Co-I), and U01–024904 (Co-I)).

David Knopman serves on a Data Safety Monitoring Board for the DIAN study; is an investigator in clinical trials sponsored by Biogen and Lilly Pharmaceuticals; and receives research support from the NIH.

Clifford Jack serves on an independent data monitoring board for Roche, has served as a speaker for Eisai, and consulted for Biogen, but he receives no personal compensation from any commercial entity. He receives research support from NIH and the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic.

Data for reference

Raw calculations of repeatability and separability for all tested methods, and p-values for all comparisons, are provided in supplementary material. In-house software will be made available by request to the corresponding author. External software (FreeSurfer, GTM-Tau) are made available by their respective authors. MRI and PET images from the Mayo Clinic Study of Aging and the Alzheimer's Disease Research Center are available to qualified academic and industry researchers by request to the Mayo Clinic Study of Aging/Alzheimer's Disease Research Center Executive Committee. Images from LEFFTDS and ARTFL can be requested via allftd.org. Images from the NRG group can be requested from its principal investigators (Josephs/Whitwell).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118259.

Christopher G. Schwarz[a,*], **Terry M. Therneau**[b], **Stephen D. Weigand**[b], **Jeffrey L. Gunter**[a,c], **Val J. Lowe**[a], **Scott A. Przybelski**[b], **Matthew L. Senjem**[a,c], **Hugo Botha**[d], **Prashanthi Vemuri**[a], **Kejal Kantarci**[a], **Bradley F. Boeve**[d], **Jennifer L. Whitwell**[a], **Keith A. Josephs**[d], **Ronald C. Petersen**[d], **David S. Knopman**[d], **Clifford R. Jack Jr**[a]

[a]Department of Radiology, Mayo Clinic and Foundation, 200 First Street SW, Rochester 55905, MN, USA

[b]Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic and Foundation, Rochester, MN, USA

[c]Department of Information Technology, Mayo Clinic and Foundation, Rochester, MN, USA

[d]Department of Neurology, Mayo Clinic and Foundation, Rochester, MN, USA

Abstract

Since tau PET tracers were introduced, investigators have quantified them using a wide variety of automated methods. As longitudinal cohort studies acquire second and third time points of serial within-person tau PET data, determining the best pipeline to measure change has become crucial. We compared a total of 415 different quantification methods (each a combination of multiple options) according to their effects on a) differences in annual SUVR change between clinical groups, and b) longitudinal measurement repeatability as measured by the error term from a linear mixed-effects model. Our comparisons used MRI and Flortaucipir scans of 97 Mayo Clinic study participants who clinically either: a) were cognitively unimpaired, or b) had cognitive impairments that were consistent with Alzheimer's disease pathology. Tested methods included cross-sectional and longitudinal variants of two overarching pipelines (FreeSurfer 6.0, and an in-house pipeline based on SPM12), three choices of target region (entorhinal, inferior temporal, and a temporal lobe meta-ROI), five types of partial volume correction (PVC) (none, two-compartment, three-compartment, geometric transfer matrix (GTM), and a tau-specific GTM variant), seven choices of reference region (cerebellar crus, cerebellar gray matter, whole cerebellum, pons, supratentorial white matter, eroded supratentorial WM, and a composite of eroded supratentorial WM, pons, and whole cerebellum), two choices of region masking (GM or GM and WM), and two choices of statistic (voxel-wise mean vs. median). Our strongest findings were: 1) larger temporal-lobe target regions greatly outperformed entorhinal cortex (median sample size estimates based on a hypothetical clinical trial were 520–526 vs. 1740); 2) longitudinal processing pipelines outperformed cross-sectional pipelines (median sample size estimates were 483 vs. 572); and 3) reference regions including supratentorial WM outperformed traditional cerebellar and pontine options (median sample size estimates were 370 vs. 559). Altogether, our results favored longitudinally SUVR methods and a temporal-lobe meta-ROI that includes adjacent (juxtacortical) WM, a composite reference region (eroded supratentorial WM + pons + whole cerebellum), 2-class voxel-based PVC, and median statistics.

Keywords

AV-1451; Flortaucipir; Tau PET; Partial volume correction; PVC; GTM; Geometric transfer matrix; RSF; Region spread function; SUVR; Change over time; Precision; Reference region; Bias correction; Inhomogeneity correction

1. Introduction

Measurements of change-over-time in tau PET SUVR, over relatively short time periods, are especially challenging because the annual rate of change in amyloid-positive cognitively impaired individuals (~3%/year) (Jack et al., 2018; Lowe et al., 2018) is small when compared with estimates of test-retest reproducibility (4.6%) (Devous et al.,2018). Existing tau PET studies have used a variety of automated techniques, but an exhaustive comparison of this family of methods for measuring change over time has not been performed. We designed this study to fulfill that need.

Here, we compare measurements produced by software pipelines based on SPM12 and on FreeSurfer 6.0, each with both cross-sectional and longitudinally stabilized variants. For tau PET tracers, the optimal target region and the optimal reference region are both active areas of research (Gordon et al., 2019; Harrison et al., 2019; Jack et al., 2018; Johnson et al., 2016; Lowe et al., 2018; Maass et al., 2017; Ossenkoppele et al., 2018; Schultz et al., 2018; A. J. Schwarz et al., 2018; Southekal et al., 2018; Sperling et al., 2019; Tetzloff et al., 2018). Therefore, we compare three choices of target region of interest (ROI): 1) entorhinal cortex; 2) inferior temporal cortex; 3) a temporal-lobe meta-ROI including bilateral amygdala, fusiform, middle/inferior temporal, entorhinal, and parahippocampal regions; and seven choices of reference region: 1) cerebellar crus; 2) cerebellar gray matter; 3) whole cerebellum; 4) pons; 5) supratentorial white matter; 6) eroded supratentorial WM; and 7) a composite of eroded supratentorial WM, pons, and whole cerebellum.

Analyses of tau PET images have also varied in their use of partial volume correction (PVC). Tau PET studies have increasingly applied Gaussian Transfer Matrix (GTM) PVC (Rousset et al., 1998) (Gordon et al., 2019; Hanseeuw et al., 2019; Harrison et al., 2019; Maass et al., 2017; Mattsson et al., 2017; Mishra et al., 2017; Ossenkoppele et al., 2018; Schöll et al., 2016). These studies often seek to measure tau in relatively small, individual regions rather than across large “global” cortical measures like those popular for amyloid PET. Current tau PET tracers also have problematic off-target binding in brain (e.g. basal ganglia, choroid plexus) and non-brain regions (skull, meninges, etc.) (Baker et al., 2017; Choi et al., 2018; Lee et al., 2018; Marquié et al., 2017b; Wolters et al., 2018) that are spatially near to disease-relevant regions of interest (ROIs) such as the medial temporal lobe. GTM attempts to model signal spill-out and spill-in between nearby regions (i.e. can attempt to correct for nearby off-target binding), which is impossible with traditional voxel-based PVC or without PVC. In this study, we compare multiple implementations of voxel- and region-based PVC techniques (5 variants overall) for measuring change in AV-1451 tau PET SUVR.

In total, we compare the suitability of 415 measurement pipelines varying in their use of two major software packages (SPM and FreeSurfer, each with cross-sectional and longitudinal variants), three temporal-lobe target regions, seven reference regions, five methods of partial volume correction, two variants of ROI masking (GM or GM and WM), and two statistical summary measures (voxel-wise mean or median) for measuring change over time in serial Flortaucipir SUVR. We present a summary of the options for each design choice in Table 1.

2. Materials and methods

2.1. Study design

We designed this study to evaluate each SUVR measurement method using two criteria: 1) repeatability, and 2) longitudinal group separability. Repeatability (similar to stability or precision) is the inverse of measurement noise at each time point. For longitudinal studies measuring change over time (i.e. slope) from one measurement at each time point, noisy individual measurements reduce the precision of these slope measurements, particularly with shorter time intervals. We estimated each measurement's repeatability as the error term from a per-subject linear fit, in unit of percentage SUVR. However, an important corollary to repeatability is that to be a useful biomarker each measurement must also correlate with some biological feature(s) of interest; a hypothetical method that always gives the same value for any scan is perfectly repeatable, but it has no dynamic range and does not *measure* anything. Therefore, for each SUVR method we also measured its group separation ability (as a t-score, analogous to an effect size) between SUVR change-over-time of participants with and without clinically significant cognitive impairment. Our assumption was that both of these metrics are equally important; we ranked all methods first by each of these two criteria individually, then additionally using the average of its ranks from each of the two criteria as a combined criterion to create a final overall ranking of methods. We acknowledge that these two criteria are not entirely independent because the repeatability of individual measures also affects the accuracy of slope measurements, and because the t statistic we use for evaluating group separation also depends on the estimate's standard error (which varies with measurement error). A reader might consider using group separation exclusively and ignoring repeatability, but repeatability is the more objective and generalizable property because it does not depend on the underlying participants, their groupings, or the comparisons/questions being studied.

We used group separation of cognitively impaired vs. unimpaired individuals because this is a straightforward task with high face validity that was feasible with our available data. It is plausible, however, that SUVR methods would perform differently for other subsets, e.g., unimpaired participants versus those with mild cognitive impairment, impaired participants versus those with dementia, across multiple pathologies or clinical phenotypes. For example, differing distributions of tau in the brain across different groups could affect choice of target or reference region, and varying degrees of atrophy could affect the importance of partial volume correction. For these reasons, we included both criteria: repeatability, to objectively estimate measurement noise; and group separation, to estimate their biological utility. We further describe exactly how we measured each of these criteria in Statistical Methods below.

2.2. Participant characteristics

Participants ($n = 97$) were selected retrospectively from multiple longitudinal Mayo Clinic studies on the basis of having three imaging visits, each with AV-1451 tau PET and 3T MRI. These Mayo Clinic studies included: the Mayo Clinic Study of Aging (MCSA) (Roberts et al., 2008); Mayo Clinic Alzheimer's Disease Research Center (ADRC); Advancement of Research and Treatment in Frontotemporal Lobar Degeneration (ARTFL);

Longitudinal Evaluation of Familial Frontotemporal Dementia Subjects (LEFFTDS); Longitudinal Imaging Biomarkers of Disease Progression in DLB; and studies from the Neurodegenerative Research Group (NRG).We required three time points because we are jointly estimating within-participant change (over a relatively short-term interval) and measurement error, which can only be assessed properly with at least three measurements. All studies were approved by their respective institutional review boards and all participants or their surrogates provided informed consent compliant with HIPAA regulations.

All selected participants had a clinical status at baseline of either: a) cognitively unimpaired, or b) had only clinical features that were suggestive of Alzheimer's disease pathology. We specifically excluded all participants with impairments suggesting non-AD pathology because Flortaucipir was designed to measure tau proteins associated with AD and has relatively low binding affinity for non-AD tauopathies and for non-tauopathies (Mathis et al., 2017). Therefore, we excluded these participants because we did not want to confound our study with the effects of off-target binding of uncertain origin. We could have made these exclusions based on amyloid or tau PET, but we used clinical diagnosis instead to avoid circularity in our study design.

2.3. Scan parameters

T1-weighted MRIs, which we used for atlas normalization, masking, and for PVC where applicable, were acquired using 3T General Electric (GE) scanners (models Discovery MR750, Signa HDx, Signa HDxt, and Signa Excite; GE Healthcare, Waukesha, WI) and 3T Siemens Prisma (Siemens, Erlangen, Germany) scanners each using 3D Sagittal Magnetization Prepared Rapid Acquisition Gradient-Recalled Echo (MP-RAGE) sequences.

[18F]AV-1451 tau PET scans were acquired using GE PET/CT scanners (models Discovery 690XT and Discovery MI; GE Healthcare, Waukesha, WI). Participants were injected with Flortaucipir (370 MBq (range 333–407 MBq)) and a low-dose CT scan was acquired for attenuation correction. At 80 minutes post-injection, participants underwent a 20-min dynamic PET scan with four five-minute frames. Dynamic PET images were reconstructed on-scanner (256 matrix, 300 mm field of view) using fully 3D (Iatrou et al., 2004) or Fourier-rebinned (Stearns and Fessler, 2002) OSEM iterative algorithms with 3 iterations and 35 subsets. A 5 mm Gaussian post-reconstruction filter was applied, along with standard corrections for attenuation, scatter, random coincidences, and decay. Four-frame dynamic PET images were co-registered with a group-wise rigid registration to correct for cross-frame motion, and averaged to produce a single static (summed) PET image.

2.4. SUVR measurement pipelines

We compare four SUVR measurement pipelines in this work. We compared cross-sectional and longitudinally stabilized variant pipelines from two major sources: 1) in-house Mayo pipelines based on SPM12 and 2) PetSurfer pipelines in FreeSurfer 6.0 (Greve et al., 2016, 2014). We have previously shown SUVR measurements from these pipelines to have a very high level of agreement (Schwarz et al., 2018) . Therefore, our primary goal for including both implementations in our comparisons was not to directly compare the two, but to

support and validate each other's findings about other questions e.g. target region, reference region, PVC, and cross-sectional vs. longitudinal pipeline design.

2.4.1. Mayo SUVR pipelines—Our in-house Mayo pipelines are based on SPM12 Unified Segmentation (Ashburner and Friston, 2005) with population-optimized templates, priors, and settings from the Mayo Clinic Adult Lifespan Template (Schwarz et al., 2017a) (MCALT; https://www.nitrc.org/projects/mcalt) (Schwarz et al., 2017a) and atlas spatial normalization using Advanced Registration Tools (ANTs) (Avants et al., 2008). Both pipelines were updated and enhanced, as described below, since our previous publications.

In the cross-sectional variant, PET scans were rigidly registered and resampled to the corresponding MRI with SPM12. MRI were segmented and corrected for intensity inhomogeneity using Unified Segmentation (Ashburner and Friston, 2005) in SPM12 with MCALT tissue priors and settings. Normalization parameters between MRI and the MCALT template were computed using ANTs, and MCALT atlases were resampled into the MRI space using GenericLabel interpolation. Transformed atlas regions were masked to include only voxels estimated to primarily contain tissue (GM or WM), or to include GM voxels only.

In the longitudinal variant, all MRI and PET scans for all of a participant's time points were processed simultaneously. It was designed to minimize variance resulting from segmentation and parcellation of each MRI, and from registration between PET and MRI (Schwarz et al., 2017b). All MRI were group-wise co-registered using affine spm_coreg, and an in-house differential bias correction (DBC) (Vemuri et al., 2015) algorithm was applied to harmonize MRI field inhomogeneity artifacts across the entire FOV of all scans. The resulting corrected outputs were then used to create a nonlinear mean-space template image using ANTs *buildTemplateParallel* (Avants et al., 2010). This image was then input to the cross-sectional pipeline (as if it were a standard single-time-point MRI scan) to perform bias correction and produce the final T1-weighted single-subject template (denoted *T1-SST*). ANTs was used to compute nonlinear registration parameters between the MCALT_T1 template and T1-SST, then resample the MCALT_ADIR122 atlas into T1-SST space using GenericLabel interpolation. Each of the co-registered, DBC-corrected MRI were transformed (affine) and resampled (BSpline) to this T1-SST space, then individually segmented using the standard cross-sectional methods above to produce tissue-class segmentations that were used for each time point.

All PET scans were used to create a mean-space template (denoted *PET-SST*) using group-wise rigid registration with *spm_coreg*. A single rigid registration was computed between this PET-SST and T1-SST, and these parameters were used to resample (BSpline) each PET scan to the space of the T1-SST. Each resampled PET scan was then analyzed in T1-SST space to produce mean and median SUVR measurements for each atlas region using the atlas parcellations (identical for all time points, calculated from T1-SST) and tissue-class masks (from individually segmented DBC-corrected MRI in the common T1-SST space) described above.

In summary, we designed this pipeline to minimize all possible sources of software quantification-related variance by performing all of each participant's processing and measurements in a common single-subject template space with a single rigid registration between all MRI and PET images, and a single atlas registration/parcellation across all time points. The MRI for each time point were segmented individually (after group-wise registration, differential bias correction, and resampling into this common space) and these segmentations were used to account for changing levels of atrophy across time by removing non-tissue or non-GM voxels from each atlas region.

2.4.2. FreeSurfer SUVR pipelines—FreeSurfer is a popular software package for surface-based analyses of brain MRI. It can be run using standard cross-sectional mode (Fischl, 2012), or using a longitudinal variant that segments all time points simultaneously for increased repeatability (Reuter et al., 2012). The PETSurfer module (Greve et al., 2014) is designed to segment PET scans and calculate SUVR either with or without GTM PVC. We used standard recommended settings when running all FreeSurfer/PETSurfer pipelines. There is no longitudinal variant of the PETSurfer portion itself (i.e. longitudinal PET images are not coregistered intramodally before quantification), but we used the officially recommended approach of running PETSurfer on original PET images with outputs from the longitudinal-variant MRI segmentations (Greve, 2016), and we refer to this as the longitudinal FreeSurfer PET pipeline.

2.5. Partial volume correction methods

We compared multiple implementations of a total of five PVC variants (including no PVC). We used in-house implementations of two-compartment (Meltzer-style; PVC2)(Meltzer et al., 1990) and three-compartment (Müller-Gärtner-style; PVC3) (Müller-Gärtner et al., 1992) PVC. Both of these voxel-based methods correct for varying amounts of non-tissue in brain tissue voxels. Three-compartment PVC additionally corrects for signal spilling into GM voxels from nearby WM voxels. For GTM (Rousset-style) PVC (Rousset et al., 1998), we compared the implementation in FreeSurfer 6.0 directly with our in-house Mayo implementation. We have previously shown that values from these two pipelines are highly correlated (Schwarz et al., 2018). Detailed reviews of these methods have been previously published (Erlandsson et al., 2012; Schwarz et al., 2018; Thomas et al., 2016). For all methods, we assumed a point spread function (PSF) of 8 mm, previously determined from an internal study using an F-18 point-source at the center of a water phantom with the same set of scanners and identical reconstruction methods.

In addition to the above PVC methods compared in our previous amyloid PET study, here we also add an additional, publicly available variant of GTM that was designed specifically for AV-1451 tau PET (Baker et al., 2017), which we will refer to as *GTM-Tau*. Unlike traditional GTM implementations that determine region parcellations entirely from MRI, this method attempts to also detect and correct for focal regions of off-target binding throughout the PET image. Region parcellations are primarily based on FreeSurfer analysis of the corresponding MRI, then modified using tissue-class segmentations from SPM in conjunction with the input tau PET image itself. Contiguous sets of voxels with relatively high SUVR in off-target locations in the skull and meninges, in the choroid plexus, and in

the (cerebellar) reference region are each grouped into individual regions that are separated from their existing regions and added as new ones before performing a traditional GTM analysis (Baker et al., 2017). The implementation was designed to use an inferior-cerebellum reference region, but it can be given masks of any subset of the cerebellar GM as input. To match as many of our other pipelines as possible without modifying its source code, we tested this method using the cerebellar crus (a subset of the inferior cerebellar region) and (whole) cerebellar GM reference regions. Although *GTM-Tau* used segmentations from both the FreeSurfer pipelines and our SPM-based pipelines, we chose to plot it among the FreeSurfer-based pipelines in our comparisons because it is a region-based PVC method with regional parcellations based much more on FreeSurfer segmentations than they are on the SPM-based pipelines.

2.6. Target and reference regions

The FreeSurfer and Mayo pipelines use different templates and atlases, but we used the closest analogous regions possible for each tested target and reference region.

We compared SUVR measurements using three choices of target region: 1) entorhinal cortex; 2) inferior temporal cortex; 3) a temporal-lobe composite region including bilateral amygdala, fusiform, middle and inferior temporal, entorhinal, and parahippocampal regions; and seven choices of reference region: 1) cerebellar crus; 2) cerebellar gray matter; 3) whole cerebellum; 4) pons; 5) supratentorial white matter (SWM); 6) eroded SWM; and 7) a composite of eroded SWM, pons, and whole cerebellum. For FreeSurfer, the composite reference used un-eroded SWM because PETSurfer does not have an analogous eroded SWM region.

2.7. Statistical methods

2.7.1. Quality control and preprocessing—Our initial analyses began with data from 117 participants, selected consecutively without exclusions. We examined the data for all methods graphically and detected that a small number of strongly outlying SUVR measurements were preventing sensible fits by the mixed-effects models. We decided to use data only from participants for which all method-combinations produced valid data for all time points. In total, we removed 16 participants for which at least one method was missing results for at least one time point, and we removed 4 participants for which at least one method had at least one invalid measurement (SUVR, after any PVC, of <0.5 or NaN i.e. no voxels were in the localized region) for at least one time point. Of these 20 (total, 16+4) participants removed, 9 were cognitively unimpaired and 11 were impaired. From the above criteria, 18/20 were excluded due to FreeSurfer-based pipelines, 1/20 was excluded due to our in-house Mayo pipelines, and 1/20 was excluded due to both pipelines. We attribute this large number of participant removals to the very large number of SUVR methods compared (415) and our stringent requirement that every measurement by all of them was valid for all three time points (a total of approximately 1500 individual measurements per method per participant), recognizing that mixed-effects models can be particularly sensitive to outliers. Further analyses continued with the remaining $n = 97$ participants ($n = 46$ unimpaired, 51 impaired) who had valid data across all methods compared, and we present their characteristics in the results section.

2.7.2. Statistical modelling—In this section we detail the statistical methods by which we comparatively evaluate SUVR pipelines. All statistical analyses were performed using R version 3.6.2 (R Development Core Team, 2008).

Our comparisons excluded methodological combinations that would be conceptually flawed (e.g. three-compartment PVC with reference regions containing WM), or that could not be implemented without modifying source code of externally created pipelines (e.g. FreeSurfer/PETSurfer does not support voxel-based PVC, eroded SWM, or cerebellar crus regions, and GTM-Tau does not support non-cerebellar reference regions.

To estimate both quantities of interest (reproducibility and group separation) for each SUVR method, we fit a linear mixed-effects model with the lme4 package version 1.1–24 (Bates et al., 2015), and we used the bootMer function to estimate 95% confidence intervals from 1000 posterior simulations. We used the model: where *impaired* indicates presence of cognitive impairment (i.e. diagnostic group), and *years* denotes each scan's time elapsed since baseline, in years, after centering (to improve model stability). This model allowed for individual SUVR variation at both baseline (intercept) and slope, but without inducing correlations between the two. We fit *log(SUVR)* rather than direct SUVR for several reasons. Key to our analysis, the standard deviation of log-transformed values is equal to the coefficient of variation of the original distribution. This allows us to interpret the residual standard deviation estimate from the model as a measure of repeatability having a percentage error interpretation. The log transformation also addressed right skew in the SUVR values and non-constant variance in which residual variability tends to increase with higher SUVR values. Demographic variables (e.g. age, sex, and education) were not included in the model because their differences across groups (see: Results) were not significant and not considered clinically meaningful.

2.7.3. Estimating repeatability for each SUVR method—To estimate repeatability, interpreted here the magnitude of within-participant variation for each measurement method, we used the standard deviation of the residual or error term from fitting the above model, which estimates the average variation of a participant's measures around a (per participant) linear regression line. By a property of the log-normal distribution, the standard deviation (SD) of log-transformed values equals the coefficient of variation (CV = SD/mean) of the original distribution. We therefore interpret the residual standard deviation from our mixed model as an estimate of the CV on the SUVR scale. A key advantage of the CV is that it is a measure of relative error rather than absolute error.

Mixed models for longitudinal data analysis account for correlation among repeated measurements within an individual and, relatedly, allow for both a population mean to change over time and for person-specific changes over time (via random effects). The mixed model does not separately estimate person-specific regression parameters but instead models them, essentially estimating the distribution of subject-specific parameters. These are penalized, or shrunk, toward the population average which reduces overfitting and provides better predictions at the individual level (Fitzmaurice et al., 2011). While mixed models do not depend on a balanced design, we required all individuals to have three time points so that each individual is contributing information about both inter- and intra-individual variation.

With any statistical modeling exercise both the systematic component and the random component of the model must be chosen purposefully based on several factors including sample size, the patterns seen in the observed data, and background knowledge. A common choice for the structural component of a longitudinal mixed effects model is that the population mean and subject-specific means are a linear function of time, i.e. there is an underlying "trend" in the population along with subject-specific trends. Based on observing up-and-down variation in SUVR measurements within individuals over time and the assumption that over an interval of several years we would not expect marked nonlinear changes in SUVR values, we found it appropriate to make this linear assumption. Normal or Gaussian errors on the log(SUVR) scale also seemed appropriate. If we had measurements made more often over a longer timeframe, we could explore more complicated systematic components using quadratic terms or spline function, but given the current data set this linear assumption is reasonable. We note that the linear assumption includes the case that the population as a whole is flat over time and doesn't prescribe an upward trend on average, nor does it prescribe a common trend across all individuals.

Given these considerations, our model can perhaps be thought of as identifying the systematic component of SUVR measurements over time within a person, "detrending" them without overfitting, and then quantifying the magnitude of the remaining random variation.

We refer to this quantity as *residual error %*, an estimate of measurement *repeatability*. Although this residual term may contain other components in addition to measurement error (e.g. biological tau nonlinearity, variation in off-target signal or tracer perfusion, systematic error due to scanner drift or use of different scanners across time), we believe that this quantity is primarily driven by error in the measurement process itself (including software factors) because it is so strongly impacted by choices in software pipelines, and because its measurements agree very well with external test-retest error measurements using scan-rescan data (see: Discussion). Further, these other sources of error would affect all pipelines equally, so they would not affect our usage of using these estimates to compare and rank pipelines relative to each other.

2.7.4. Estimating group separation for each SUVR method—We also quantified our ability to detect separation in annualized SUVR slopes between the clinical groupings (cognitively impaired vs. unimpaired) using the t statistic from the *impaired:years* term in the same model fit above. The coefficient from this term corresponds to the estimated difference in group-wise rates of change and the corresponding t statistic reflects our power to detect these differences. We chose this as a measure of effect size because Cohen's *d* effect sizes are not well-defined for longitudinal models. However, our t statistic is conceptually very similar in that it is ratio of group-wise differences over a measure of uncertainty, and thus it can be interpreted analogously to a "longitudinal Cohen's d" effect size.

We recognize that some unimpaired individuals may have tau accumulation, and this would contribute to reduced group-separation equally across all methods. However, because the relationship between tau PET and cognitive impairment is stronger than that of amyloid

PET (Aschenbrenner et al., 2018; Maass et al., 2018; Pontecorvo et al., 2017), we assume that average tau accumulation rate should be greater in impaired participants, as in our previous findings (Jack et al., 2018). We also recognize that the cognitive impairments in some individuals may not be driven by the AD-associated forms of tau that are imaged by Flortaucipir, but we tried to minimize this by excluding all participants who had clinical impairments that were inconsistent with AD pathology (see Participant Characteristics, above).

2.7.5. Sample size estimates for each SUVR method—To assist readers with interpreting these metrics in practical terms, we also computed sample size estimates for each SUVR method for a hypothetical interventional clinical trial designed for 80% power to reduce SUVR accumulation by 20% (vs. placebo). The trial assumed three PET scans over three years. We used only the cognitively impaired participants ($n = 46$) for this estimate because these are more likely to have SUVR accumulation. For computation, we used the *lmmpower* function from the *longpower* package (Donohue, 2021) on the existing linear mixed-effects model described above. Because this study was not designed to imitate a clinical trial, and because there are any number of important methodological and statistical considerations when designing a trial, we treat these estimates cautiously and do not use the resulting estimates as one of our primary criteria to compare methods, but rather only as an additional metric for discussion to provide practical estimates of the value of using methods that are ranked more highly by our other criteria.

3. Results

3.1. Participant characteristics

We present a table of participant baseline characteristics in Table 2. 46 participants were cognitively unimpaired, including 3 who had clinical evidence of REM sleep behavior disorder (RBD) but no other features of an underlying neurodegenerative disease. Among the cognitively impaired group ($n = 51$), the clinical diagnoses were as follows: 23 AD clinical syndrome, 11 amnestic mild cognitive impairment (MCI), 11 multi-domain amnestic MCI, 5 probable posterior cortical atrophy, and 1 logopenic variant primary progressive aphasia. When comparing the cognitively impaired vs. unimpaired groups, we found the expected significant differences in baseline amyloid and tau PET SUVR, prevalence of APOE E4 carriers, and MMSE score. We use error as a percentage in repeatability measures, to account for differences in baseline tau. Study time durations were significantly longer for the unimpaired group, but we used annualized SUVR change over time to account for this. Unimpaired participants also had slightly longer time intervals between their MRI and PET scans, but the differences between these medians was only two days, which we feel is negligible.

3.2. Effects of target region

Our approach for this section is to present several scatter plots where each point shows the estimated measurement error (x-axis) and group separability (y-axis). To combine the two criteria, we ranked all methods by each criterion individually, and used the mean of their ranks in these two criteria to create a joint ranking. When ties occurred (two methods had

the same mean ranking), our joint ranking gave the smaller ranking to whichever method was better in the group separability criterion. In each plot, most points are unlabeled out of necessity to avoid over-plotting; instead we label and color certain subsets for discussion. In supplementary material, we also provide a table of all 415 SUVR methods with each one's ranks and estimates by all the criteria. We then present a series of box plots showing the effects of each option for each major methodological decision and explore their effects across all the pipelines that use them.

In Fig. 1, we plot all 415 SUVR methods on both axes, colored by their choice of temporal-lobe target region. Methods using the inferior temporal and temporal meta-ROI target regions were largely comparable in both criteria, but those using the entorhinal cortex target region performed poorly: they had much worse group separation and some-what larger measurement error. Consequently, to reduce over-plotting we removed the 152 methods using the entorhinal cortex from subsequent plots, leaving only 304 methods using either the inferior temporal or temporal-lobe composite target region.

3.3. Overall best and worst SUVR methods

We present our primary findings in Fig. 2, which labels a subset of methods/points that were arguably among the "best" (top/left) or "worst" (bottom/right) by each criterion. We also labelled three popular/standard methods for comparison. In subsequent sections we will attempt to break down the reasons (i.e. underlying properties) that affected the performance of each method (itself a combination of component decisions).

Best group separation: The method with the largest group separation was FSLong_GM_TemplMeta_CereGM_GTM–Tau_Mean with a t score of 6.19 using the FreeSurfer longitudinal pipeline with GTM-Tau PVC and the cerebellar GM reference regions; however, its measurement error was among the worst methods at 4.51% (rank 371), giving it an overall rank of of 189 by both criteria combined.

Consistencies among the top 10 methods: The top 10 methods by the combined criteria (top-left corner of Fig. 2) all used the Mayo SPM-based pipeline with a composite (7/10) or eroded SWM (3/10) reference region. Among them, 9/10 used the longitudinal variant of this pipeline, and 7/10 used 2-compartment PVC (the other 3/10 used no PVC). Also, 7/10 used the temporal-lobe meta-ROI (the other 3/10 used inferior temporal), 8/10 used GM+WM tissue masking, and 8/10 used median statistics.

Comparing best overall vs. best group separation: Compared to the method with the best group separation (rank 189 overall), the method ranked #1 overall (Mayo-Long_GMWM_TemplMeta_Composite_PVC2_Median, rank 6 repeatability, rank 10 group-separation) reduced measurement error by 68% (1.45% vs. 4.51%) in return for only 9% smaller group separation (t score 5.64 vs. 6.19). The method with the second highest group separation was rank 5 overall (Mayo-Long_GMWM_TemplMeta_SupraWMero3_PVC2_Median, rank 41 repeatability, rank 2 group-separation). Compared to the method with the highest group separation (rank 189), it reduced measurement error by 58% (1.90% vs. 4.51%) in return for only 4% smaller group

separation (t score 5.95 vs. 6.19). Compared to the method ranked #1 overall, it had 32% higher measurement error (1.91% vs. 1.45%) and 6% worse group separation (t score 5.64 vs. 5.95).

The worst methods: After omitting those using the ERC target region, the methods with the worst measurement error each used cross-sectional pipelines with eroded or uneroded supratentorial WM reference regions and GTM PVC, while those with the worst group separation used cross-sectional pipelines with eroded supratentorial WM or pontine reference regions and GTM or no PVC.

Comparing with our standard internal approach: The SUVR method used by our group in prior tau PET publications is "MayoCX_GMWM_TemplMeta_CereCrus_None_Median", and it performed at rank 202 with middling performance on both criteria (rank 239 repeatability; rank 154 group separation). Compared to the method ranked #1, its measurement error was 93% larger (2.80% vs. 1.45%), and its group separation was 17% worse (t score 4.69 vs. 5.64). In terms of sample size estimates for a hypothetical intervention trial on the cognitively impaired participants, this translated to a reduction from roughly n=633 participants (per arm) to roughly 312 participants.

Comparing with standard FreeSurfer methods: We also labelled what we believe to be the most-standard FreeSurfer based methods in the literature: "FSLong_GM_TemplMeta_CereGM_GTM_Mean" (rank 171 overall). It performed well for group separation (rank 17; t score 5.50) but poorly for measurement error (rank 333; 3.62%). Compared to the method ranked #1, its measurement error was 150% larger (3.62% vs. 1.45%), and its group separation was 2% worse (t score 5.50 vs. 5.64). Stated another way, one could switch quantification methods from this FreeSurfer standard (rank 171) to our method ranked #1 to reduce measurement noise by 60% while also improving group separation by 2%. In terms of sample size estimates, they were similar (n = 312 per arm for rank 1, vs n = 243 for rank 171).

Compared to the method with the best group separation (rank 189 overall), which is otherwise identical except uses GTM-Tau PVC instead of GTM, its measurement error was 25% smaller (3.62% vs. 4.51%), but its group separation was 13% worse (t score 5.50 vs. 6.19). Mean-while, the corresponding cross-sectional "standard" FreeSurfer pipeline "FSCX_GM_TemplMeta_CereGM_GTM_Mean" (rank 247) performed similarly to its longitudinal counterpart in terms of measurement error (3% smaller; 3.52 vs. 3.62) but had 15% worse group separation (t score 4.68 vs. 5.50). Compared to these two FreeSurfer methods, our "standard" Mayo method (rank 202 overall) ranked between the two for group separation but had better repeatability (lower measurement error) than both.

3.4. Impacts of individual methodological choices

In Fig. 3 (2 pages) we present a series of six sets of three box plots (18 plots total). Each point on each of these plots represents an individual SUVR measurement method, which itself is a combination of six component methodological decisions: target region, reference

region, PVC, software package, tissue masking, and voxel statistic. Each of the six sets of plots (rows) explores the effects of each possible choice for one of these decisions. For each row, the first column plots repeatability, the second plots group separation, and the last plots a combination of the two (average of its ranking in each of the two criteria individually). The y axes on the first and third plots are flipped so that greater y axis indicates better results, consistently across all three columns. Within each plot, line segments between points indicate sets where all other variables (choices) were the same except for that row's variable of interest. Points that would have had no such line segments are omitted from the plot; for example, the mean vs. median comparisons include no points from FreeSurfer-based methods because FreeSurfer does not support median statistics and thus there are no valid direct paired comparisons for these methods. Because many combinations are not possible or not valid, the number of points on each plot varies. We also omit all methods using the entorhinal target region from all plots after the first row, to reduce plotting density by eliminating the worst performing methods first. We recognize that it is impossible to separate each of these effects individually, but any specific pipeline inherently requires a choice for each of them, and we feel that this data provides the best possible way to make these choices. We describe these results below. To provide p-values, we performed paired Wilcoxon signed rank tests between only those pairs of points that were otherwise the same (i.e. those with line segments connecting them).

Target region: In all criteria the temporal lobe meta-ROI slightly outperformed the inferior temporal and greatly outperformed entorhinal cortex target regions (all comparisons $p \leq 0.019$). Across all other variables, median sample size estimates (per arm) were 1720 for entorhinal, 526 for inferior temporal, and 520 for the temporal lobe meta-ROI.

Reference region: The composite reference region outperformed all others in terms of repeatability (all $p \leq 0.008$) and combined (all $p \leq 0.016$). In terms of group separation, the composite reference region was statistically equivalent to the non-eroded SWM ($p = 0.461$) and significantly outperformed all other reference regions ($p \leq 0.011$). The non-eroded SWM reference region performed the worst for repeatability ($p = 0.125$ for eroded SWM and cerebellar crus, $p < 0.008$ for all others), but this reference is only valid (and thus only included/plotted) when using GTM PVC, so this finding is likely due to GTM PVC rather than any property of the region itself. Among cerebellar and pontine reference regions, the whole cerebellum outperformed the others in the repeatability (all $p < 0.001$) and combined (all $p \leq 0.012$) criteria, and for group separation it outperformed cerebellar crus ($p < 0.001$) and cerebellar GM ($p = 0.018$) but tied with the pons ($p = 0.179$). Across all other variables, median sample size estimates (per arm) were 373 for the composite reference region, 385 for eroded SWM, 503 for the pons, 523 for the whole cerebellum, and 633 for the cerebellar crus. Altogether, cerebellar and pontine reference regions had a median sample size estimate of 559, while those including SWM (including the composite) were 370 (34% less).

PVC: Partial Volume Correction was the variable where the repeatability and group separation criteria disagreed the most. GTM-Tau had by far the worst repeatability but also by far the best group separation. This disagreement made it only 4th best by the combined criterion, but because only 4 methods use GTM-Tau, none of its comparisons

were significant. Among the rest, for repeatability, none > PVC2 > PVC3 > GTM, and all adjacent pairs in this ordering were significant ($p \leq 0.008$). For group separation, PVC2 was better than all others (GTM-Tau excluded), but this comparison was only significant for no PVC ($p < 0.001$ vs. no PVC; $p = 0.065$ vs. GTM; $p = 0.94$ vs. PVC3). PVC2 was also better than all others (GTM-Tau excluded) for the combined criteria ($p \leq 0.016$). Across all other variables, median sample size estimates (per arm) were 385 for GTM-Tau, 490 for GTM, 499 for PVC2, 523 for none, and 631 for PVC3.

Software Package/Design: Within the Mayo SPM-based pipelines, the longitudinal design outperformed the cross-sectional design (for this longitudinal study) in all criteria ($p < 0.001$ for error and combined; $p = 0.099$ for group separation). Within the FreeSurfer-based pipelines, the longitudinal design outperformed the cross-sectional design in the group separation criterion ($p < 0.001$), and this drove its superiority in the combined criteria ($p < 0.001$), but both designs were statistically similar for repeatability ($p = 0.312$). Our primary goal for including both the Mayo and FreeSurfer pipelines in our comparisons was not to directly compare them, but to support and validate each other's findings regarding choices of regions, PVC, etc. Such direct comparisons were also under-powered because it was only possible to implement 16 methods consistently across all four, but we found some significant differences between them anyway. For repeatability, both Mayo variants outperformed both FreeSurfer variants; this was significant for the Mayo longitudinal variant vs. both FreeSurfer variants ($p \leq 0.001$) but not for the Mayo cross-sectional variant ($p > 0.32$). For group separation, both Mayo variants outperformed the FreeSurfer cross-sectional variant ($p \leq 0.013$). The FreeSurfer longitudinal variant had the largest group separation on average, but comparisons were only significant for the FreeSurfer cross-sectional variant ($p < 0.001$). For the combined criteria, the Mayo longitudinal variant outperformed all others on average; this was significant for both cross-sectional pipelines ($p < 0.001$) but not for the FreeSurfer longitudinal pipeline ($p = 0.205$). Across all other variables, median sample size estimates (per arm) were 583 and 497 (15% smaller) for the Mayo cross-sectional and longitudinal variants respectively, and 485 and 325 (33% smaller) for the FreeSurfer cross-sectional and longitudinal variants respectively. Combining both cross-sectional variants and both longitudinal variants, median sample size estimates were 572 for cross-sectional and 483 for longitudinal (16% smaller).

Tissue Masking: Computing SUVRs from ROIs masked to include both GM and WM voxels (i.e. including adjacent juxtacortical WM) outperformed using only GM voxels, in all criteria (all $p \leq 0.001$). All these comparisons were within the Mayo SPM-based pipelines only, because PETSurfer always uses GM ROIs for PET. Across all other variants, median sample size estimates (per arm) were the same (521) but paired comparisons favored GM and WM together.

Voxel Statistic: Computing SUVRs using the median statistic across voxels in each ROI greatly outperformed using the mean in the group separability ($p < 0.001$) criterion. This drove median's superiority in the combined criterion ($p < 0.001$), despite mean outperforming median in repeatability by a very small but significant ($p < 0.001$) margin. All these comparisons were within the Mayo SPM-based pipelines only, because PETSurfer

always uses mean values for PET regions. Across all other variants, median sample size estimates (per arm) were 497 for median and 539 for mean.

3.5. Best choices with only cerebellar SUVR

Although reference regions including supratentorial WM (SWM) performed very well in the above analyses according both to repeatability and effect-sizes criteria, we recognize that there is some uncertainty with the origin of the signal in these locations, and thus hesitance to use them as a reference region. We will review those reasons in the discussion, but here we also present the results of our analyses when methods are limited only to variants of the traditional cerebellar reference regions.

In Fig. 4, we present a modification of Fig. 2 that includes only the cerebellar reference regions. In this plot, the methods that performed best (made up the top-left corner of Fig. 2) are all missing, because these used either SWM or composite reference regions. The highest ranked cerebellar-reference method (Mayo-Long_GM_TemplMeta_CereWhole_PVC2_Median) was ranked 52 across all methods. Compared to the method ranked #1 overall, its measurement error was 62% larger (2.34% vs. 1.45%), and its group separation was 11% worse (t score 5.04 vs. 5.64). When limited to the methods in Fig. 4, there are no methods that performed well in both the repeatability and effect-sizes criteria, and thus users must choose which is more important to them. If favoring repeatability (or using our combined criterion), one would choose among a large cluster of methods that were largely equivalent. Among the top 5, 4/5 used the Mayo pipelines and 4/5 used longitudinal variants. If favoring group separation, one would choose method 68 (FSLong_GM_TemplMeta_CereGM_GTM-Tau_Mean, rank 189 across all methods), which was also the method with the highest group separation across all methods and uses longitudinal FreeSurfer with GTM-Tau PVC and a cerebellar GM reference region.

4. Discussion

4.1. Discussion of results

Target regions: Our strongest finding was that AV-1451 tau PET SUVR in the entorhinal cortex may be less stable and have worse separation of tau accumulation rates in cognitively impaired vs. unimpaired participants, when compared to inferior temporal or temporal-lobe composite regions. The lower repeatability may be attributable to its smaller size (fewer voxels) compared to larger target regions. As defined in our *MCALT_ADIR122* atlas, the entorhinal region has only about 17% of the volume of the inferior temporal region and only about 5% of the volume of the larger temporal composite region that includes them both. The lower performance for group separation may be explained by the presence of early-stage tau accumulation in the cognitively unimpaired group, i.e. signal that occurs in both groups because entorhinal tau pathology occurs in most adults irrespective of amyloid pathology (Braak and Braak, 1991; Crary et al., 2014; Price and Morris, 1999). Some previous studies have also reported lower group separability cross-sectionally (Johnson et al., 2016; Ossenkoppele et al., 2018; Schultz et al., 2018). Although entorhinal cortex tau PET measurements may have worse group separation and repeatability, it has also been found that they are associated with memory performance (Knopman et al., 2019; Lowe et

al., 2019; Maass et al., 2018), and also for separating unimpaired participants according to amyloid status (Vemuri et al., 2017).

Reference Regions: Optimizing the reference region for tau PET SUVR is a highly active area of research (Devous et al., 2018; Harrison et al., 2019; Ossenkoppele et al., 2018; Southekal et al., 2018; Timmers et al., 2019). We found improved repeatability with eroded supratentorial WM and composite reference regions, which is consistent with previous findings for amyloid PET (Chen et al., 2015; Fleisher et al., 2017; Landau et al., 2015; Schwarz et al., 2017c) and for tau PET (Devous et al., 2018; Harrison et al., 2019; Southekal et al., 2018). It differs from one other FTP study that found statistically similar repeatability between cerebellar and eroded WM reference regions (Timmers et al., 2019). In our study, these regions also had improved longitudinal group separation vs. cerebellar or pontine regions, which made the composite reference strongly favored by our combined criterion. These gains in group separation are consistent with some previous cross-sectional and longitudinal studies (Southekal et al., 2018). However, despite our strong findings here we are hesitant to unconditionally recommend these reference regions containing supratentorial WM because they contain off-target signal of unknown and/or mixed origin. This off-target WM signal is partly correlated with cortical GM signal (Baker et al., 2019), which would suggest that SWM reference regions should *reduce* group separation. Since the opposite occurs, it may be that the correlated component of cortical signal is actually additive off-target signal within the target region and thus cancelling it out via the SWM reference may improve the accuracy of tau measurements, as previously hypothesized (Baker et al., 2019). Until this hypothesis can be confirmed via larger studies comparing in-vivo scans with quantitative histopathology, reference regions containing only SWM should only be used with careful consideration of these limitations.

PVC: Partial volume correction methods inherently trade repeatability for accuracy, i.e. improve group separation while harming repeatability. The question of whether these gains make up for the losses can be subjective. Our study design attempts to give both criteria an equal weighting. We found that the GTM-Tau variant had by-far the best group separation, but by-far the worst repeatability, which made it the fourth best PVC in our combined criterion. This variant was designed specifically for AV-1451 tau PET and attempts to locate and correct for regions of off-target binding by parcellating them into separate regions prior to performing an otherwise-standard implementation of GTM (Baker et al., 2017). Thus, we were not surprised to find that compared to standard GTM it further improved group separation at cost of further reduction in repeatability, making it the “strongest” form of PVC tested. Standard GTM had the worst repeatability after GTM-Tau, but its group separation was on average substantially worse than those of two-compartment voxel-based PVC (PVC2) ($p = 0.065$), which had repeatability that was only slightly but significantly ($p < 0.001$) worse than no PVC. By the combined criterion, PVC2 was the clear winner, giving second-best group separation and second-best repeatability. Since the best choice for repeatability (no PVC) was among the worst choices for group separation, and the best choice for group separation (GTM-Tau) was the worst choice for repeatability, these cancelled out for PVC2 (second best in both) to win the combined ranking.

Our findings for standard GTM largely mirror our previous findings in amyloid PET (Schwarz et al., 2018), where it greatly reduced repeatability without substantial gains in group separation. The consistency of this result across PET tracers and implementations suggests that the major source of relative instability in GTM PVC is not specific to any particular PET tracer, but is intrinsic to the method itself. This result was consistent across two distinct implementations of GTM PVC, suggesting that it is not specific to any particular software implementation.

Our finding of improved group separation with PVC2 vs. no PVC mirrors our previous study that found larger longitudinal group separation for two-compartment voxel-based PVC when examining a larger set of participants each with shorter, two-time-point trajectories (Jack et al., 2018), and mostly agrees with our previous study of PVC in amyloid PiB PET (Schwarz et al., 2018).

Although we found a clear benefit for using PVC2 (or GTM-Tau if repeatability is not a concern), we are hesitant to recommend it unconditionally because a previous study with PiB (amyloid) PET has shown that all these PVC methods did not improve, and may minimally reduce, correlations between SUVR and the gold standard of quantitative protein measurements at autopsy (Minhas et al., 2018). Therefore, while it is true that PVC generally improves group separation and PVC2 can do this with only a very modest penalty to repeatability, these improvements may not reflect more accurate quantitation of the underlying pathology but may be driven instead by boosting signal from atrophy via the underlying T1-weighted MRI. If true, this would suggest that PVC is beneficial in studies that analyze PET measurements only but increases redundancy and correlation between MRI- and PET-based measurements in studies that use both tests as independent measures of different underlying pathologic processes.

Cross-sectional vs. Longitudinal Pipeline Variants: Longitudinal pipeline variants are designed specifically to stabilize serial measurements and consequently can improve group separation of change over time. The Mayo longitudinal pipeline showed small but significant improvements in repeatability and in group separation, vs. its cross-sectional variants. The FreeSurfer longitudinal variant outperformed the cross sectional variant in the group separation and the combined criteria, but they were statistically equivalent in repeatability. The FreeSurfer longitudinal variant (PETSurfer) does not include any option to perform coregistration of the PET images across time (it only uses longitudinally processed segmentations from MRI) and this may explain why repeatability was equivalent. In total, these results showed clear benefits to using longitudinal pipeline designs when measuring tau PET SUVR change over time.

Longitudinal processing designs have the drawbacks of additional complexity/runtime and that addition of subsequent time points will alter all previous measurements. For most clinical trial designs with finite end points, these increased logistical costs would be much smaller than savings from reduced enrollment sample sizes. For example, our median sample size estimates (per arm) for the FreeSurfer pipelines were 325 for longitudinal and 485 for cross-sectional (32% improvement). It has been estimated that the average cost per patient enrolled in US clinical trials of new therapeutic agents is over $40,000 (Moore

et al.,2020). With this estimate, assuming two arms per trial, the cost savings of using longitudinal processing software would be approximately $13 million (2 × (485–325) × 40,000). However, for long-running observational studies, and especially those with rolling data releases, these added complexities may be more challenging.

Tissue-Class Masking: Our findings showed that including adjacent (juxtacortical) WM voxels in target ROIs greatly improved measurement repeatability and improved group separation. Improved repeatability is likely due to a larger set of voxels being averaged, consistent with repeatability improvements when using larger target and reference regions. Improved group separation may be due to inclusion of low-level signal in juxtacortical WM that has been associated with several non-AD pathologies that affect cognition (Josephs et al., 2016; Lowe et al., 2016; Marquié et al., 2017a) and may be concomitant in our sample even though we tried to exclude them. The option to include juxtacortical WM voxels in SUVR measurements is not available in FreeSurfer, and we hope that our findings will encourage its addition to this and other software.

Statistic: Our findings showed significant and substantial benefits to group separation from computing SUVRs using median values instead of mean values across the included voxels. These improvements may stem from reducing the influence of outliers in these measurements. The option to use median values instead of mean is not available in PETSurfer, and our literature searches suggest it is generally uncommon in PET measurement software. We hope that our findings will encourage its more widespread implementation.

Overall Tau PET SUVR Repeatability: Our data suggests that the measurement error in Flortaucipir SUVR in a temporal-lobe meta-ROI with a cerebellar reference region and no PVC is approximately 2–3%. This value is considerably lower than one previously published estimate of 4.6% using test-retest data with comparable target and reference regions (Devous et al., 2018). Another study estimated 3.5% for a medial temporal meta-ROI and 0.7% for a lateral temporal meta-ROI (Timmers et al., 2019); our meta-ROI includes ROIs from both, so this is difficult to compare directly. Our estimates may differ from others' due to differences in study design (unexplained longitudinal variance in our study vs. short-time test-retest), differences in the SUVR quantification software used, or simply the different populations tested. Using our same methodology, we previously measured the unexplained variance in PiB (amyloid) PET SUVR in a cortical composite region as approximately 2–4% depending on the reference region used (Schwarz et al., 2018). Our findings here suggest that the repeatability of Flortaucipir SUVR is roughly comparable to that of PiB.

4.2. Strengths and limitations of current study

To our knowledge, this is the first study to analyze deviation from linearity as a measure of measurement noise in tau PET SUVR. Our study is most similar to (Harrison et al., 2019), which compared a smaller set of approaches using a smaller cohort ($n = 42$) with two-time-point trajectories. Another previous study used test-retest repeatability in 21 participants to compare several target regions and two reference regions, but it did not study the effects of partial volume correction and other methodological variables, nor effects on longitudinal

group separation (Devous et al., 2018). Another measured repeatability in 14 participants and additionally compared SUVR with fully quantitative methods from dynamic scans, but it also compared only two reference regions and did not assess other methodological variables or longitudinal group separation.

Our sample size of 97 participants is relatively modest when compared to some previous longitudinal studies of Flortaucipir SUVR, but we restricted our cohort to include only participants with at least 3 longitudinal time points in order to use our repeatability metric, and we only included impaired participants with amnestic phenotypes to exclude participants who were most likely to have pathologies that are not targeted by Flortaucipir. Most of our findings were statistically significant, and we believe that these findings provide the best available guidance to those presently analyzing serial tau PET data to make data- driven methodological decisions. To lend practical interpretability to our findings, we also calculated sample size estimates for a hypothetical clinical trial using each SUVR method, but we these values should be interpreted with caution because we computed them retrospectively for this study that was not carefully designed to imitate a clinical trial. Future work will formally explore the effects of our software pipeline findings on clinical trial design.

We acknowledge that our conclusions regarding analytical approach would best generalize to other studies with similar populations and length of follow-up. For example, studies including participants with more-rapid atrophy, or those with a longer period of follow-up, could potentially benefit more from partial volume correction than what we measured in our current study. We also acknowledge that our data used SUVR, which is inherently a semi-quantitative measurement compared to quantitative metrics from dynamic PET scans (Timmers et al., 2019). It is possible that our findings may not generalize to quantitative tau PET data, but studies using late-uptake SUVR measures are also much more prevalent due to their substantially reduced scan times and patient burden. It is also possible that our findings may not generalize to studies using other tau PET tracers.

It is possible that cross-scanner differences contributed partly to our estimates of measurement noise, but this source of variance would similarly affect all pipelines. We also repeated our analyses using only the participants who did not switch between GE and Siemens MRI across their three time points (n = 18: 7 CU + 11 CI), and we found that the exact same method was select as rank 1 despite this much smaller sample.

5. Conclusions

Combining all our findings, an optimal software method for measuring change in tau PET SUVR should use a longitudinal design with a temporal-lobe composite target region including juxtacortical WM, a composite (eroded supratentorial WM + pons + whole cerebellum) reference region, 2-class voxel-based PVC, and median statistics. The strongest among these findings were: 1) larger temporal-lobe ROIs greatly outperformed entorhinal cortex (median sample size estimates were 520–526 vs. 1740); 2) longitudinal pipelines outperformed cross-sectional pipelines (median sample size estimates were 483 vs. 572);

and 3) reference regions including supratentorial WM outperformed traditional cerebellar and pontine options (median sample size estimates were 370 vs. 559).

We came to each of the above conclusions individually by comparing performance across each variable, and when we compared all 415 valid combinations, the top-ranked combination was the one that used each of these individually optimal choices. Compared with the most standard FreeSurfer approach, this top-ranked method reduced measurement error from 3.62% to 1.45% (60% smaller) while also slightly improving group separation from a t-score 5.50 to 5.64 (2% larger).

The GTM-Tau PVC method (Baker et al., 2017) achieved extremely high group-discrimination but had poor repeatability. It is also based primarily on the FreeSurfer architecture, which had worse repeatability than our SPM-based pipelines and lacks other options that improved repeatability, like median statistics, eroded WM and composite reference regions, and including juxtacortical WM in target regions. Future work should aim to implement more of these repeatability-improving options in FreeSurfer/PETSurfer, and implement GTM-Tau within the SPM-based architecture, and compare these new variants toward producing a more optimal combined method. These approaches should also be compared against methods that are not based on pre-defined regions of interest but rather data-driven sets of voxels (Bourgeat et al., 2021; Lilja et al., 2019; Whittington and Gunn, 2018), and PET-only approaches (Bourgeat et al., 2017; Doré et al., 2019; Lilja et al., 2019). Future work could also compare methods using other tau PET tracers, or amyloid PET tracers, or compare methods using a new criteria to determine which maximize similarity across tracers using tracer cross-over datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors give their thanks to all the volunteers, participants, and coordinators who contributed to this research. We gratefully thank our funding sources that made this work possible: NIH grants R37 AG011378, R01 AG041851, R56 AG068206, U01 AG006786, P50 AG016574, P30 AG062677, R01 AG034676, R01 NS097495, U01 AG045390, U54 NS092089, U01 NS100620, R01 NS89757, R01 DC12519, R21 NS94684, R01 AG50603, U01 NS100620, Gerald and Henrietta Rauenhorst Foundation, Elsie and Marvin Dekelboum Family Foundation, Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic, Liston Award, Schuler Foundation, and Mayo Foundation for Medical Education and Research. We also thank Brad Kemp for his assistance with details of the nuclear medicine acquisitions. We also thank AVID Radiopharmaceuticals, Inc., for their support in supplying AV-1451 precursor, chemistry production advice and oversight, and FDA regulatory cross-filing permission and documentation needed for this work.

References

Aschenbrenner AJ, Gordon BA, Benzinger TLS, Morris JC, Hassenstab JJ, 2018. Influence of tau PET, amyloid PET, and hippocampal volume on cognition in Alzheimer disease. Neurology91, e859–e866. doi:10.1212/WNL.0000000000006075. [PubMed: 30068637]

Ashburner J, Friston KJ, 2005. Unified segmentation. Neuroimage26, 839–851. doi:10.1016/j.neuroimage.2005.02.018. [PubMed: 15955494]

Avants BB, Epstein CL, Grossman M, Gee JC, 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal12, 26–41. doi:10.1016/j.media.2007.06.004. [PubMed: 17659998]

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC, 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage49, 2457–2466. doi:10.1016/j.neuroimage.2009.09.062. [PubMed: 19818860]

Baker SL, Harrison TM, Maass A, La Joie R, Jagust WJ, 2019. Effect of off-target binding on 18 F-flortaucipir variability in healthy controls across the life span. J. Nucl. Med60, 1444–1451. doi:10.2967/jnumed.118.224113. [PubMed: 30877180]

Baker SL, Maass A, Jagust WJ, 2017. Considerations and code for partial volume correcting [18F]-AV-1451 tau PET data. Data Brief15, 648–657. doi:10.1016/j.dib.2017.10.024. [PubMed: 29124088]

Bates D, Mächler M, Bolker BM, Walker SC, 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw67. doi:10.18637/jss.v067.i01.

Bourgeat P, Doré V, Doecke J, Ames D, Masters CL, Rowe CC, Fripp J, Villemagne VL, 2021. Non-negative matrix factorisation improves Centiloid robustness in longitudinal studies. Neuroimage226, 117593. doi:10.1016/j.neuroimage.2020.117593. [PubMed: 33248259]

Bourgeat P. Villemagne VL, Dore V. Masters CL, Ames D. Rowe CC, Salvado O. Fripp J, 2017. PET-Only 18F-AV1451 tau quantification. Biomed. Imaging, Int. Symp1173–1176.

Braak H, Braak E, 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82, 239–259. doi:10.1007/BF00308809. [PubMed: 1759558]

Chen K, Roontiva A, Thiyyagura P, Lee W, Liu X, Ayutyanont N, Protas H, Luo JL, Bauer R, Reschke C, Bandy D, Koeppe RA, Fleisher AS, Caselli RJ, Landau S, Jagust WJ, Weiner MW, Reiman EM, 2015. Improved power for characterizing longitudinal amyloid-β PET changes and evaluating amyloid-modifying treatments with a cerebral white matter reference region. J. Nucl. Med56, 560–566. doi:10.2967/jnumed.114.149732. [PubMed: 25745091]

Choi JY, Cho H, Ahn SJ, Lee JH, Ryu YH, Lee MS, Lyoo CH, 2018. Off-target 18F-AV-1451 binding in the basal ganglia correlates with age-related iron accumulation. J. Nucl. Med59, 117–120. doi:10.2967/jnumed.117.195248. [PubMed: 28775201]

Crary JF, Trojanowski JQ, Schneider JA, Abisambra JF, Abner EL, Alafuzoff I, Arnold SE, Attems J, Beach TG, Bigio EH, Cairns NJ, Dickson DW, Gearing M, Grinberg LT, Hof PR, Hyman BT, Jellinger K, Jicha GA, Kovacs GG, Knopman DS, Kofler J, Kukull WA, Mackenzie IR, Masliah E, McKee A, Montine TJ, Murray ME, Neltner JH, Santa-Maria I, Seeley WW, Serrano-Pozo A, Shelanski ML, Stein T, Takao M, Thal DR, Toledo JB, Troncoso JC, Vonsattel JP, White CL, Wisniewski T, Woltjer RL, Yamada M, Nelson PT, 2014. Primary age-related tauopathy (PART): a common pathology associated with human aging. Acta Neuropathol. 128, 755–766. doi:10.1007/s00401-014-1349-0. [PubMed: 25348064]

Devous MD, Joshi AD, Navitsky M, Southekal S, Pontecorvo MJ, Shen H, Lu M, Shankle WR, Seibyl JP, Marek K, Mintun MA, 2018. Test-retest reproducibility for the tau PET imaging agent flortaucipir F 18. J. Nucl. Med59, 937–943. doi:10.2967/jnumed.117.200691. [PubMed: 29284675]

Donohue MC, 2021. Longpower: Power and Sample Size Calculations for Linear Mixed Models.

Doré V, Bullich S, Rowe CC, Bourgeat P, Konate S, Sabri O, Stephens AW, Barthel H, Fripp J, Masters CL, Dinkelborg L, Salvado O, Villemagne VL, De Santi S, 2019. Comparison of 18 F-florbetaben quantification results using the standard Centiloid, MR-based, and MR-less CapAIBL ® approaches: Validation against histopathology. Alzheimer's Dement. 1–10. doi:10.1016/j.jalz.2019.02.005. [PubMed: 30195482]

Erlandsson K, Buvat I, Pretorius PH, Thomas BA, Hutton BF, 2012. A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. Phys. Med. Biol57, R119–R159. doi:10.1088/0031-9155/57/21/R119. [PubMed: 23073343]

Fischl B, 2012. FreeSurfer. Neuroimage62, 774–781. doi:10.1016/j.neuroimage.2012.01.021. [PubMed: 22248573]

Fitzmaurice GM, Laird NM, Ware JH, 2011. Applied Longitudinal Analysis, second ed. Wiley.

Fleisher AS, Joshi AD, Sundell KL, Chen Y-F, Kollack-Walker S, Lu M, Chen S, Devous MD, Seibyl J, Marek K, Siemers ER, Mintun MA, 2017. Use of white matter reference regions for detection of change in florbetapir positron emission tomography from completed phase 3 solanezumab trials. Alzheimer's Dement. 13, 1117–1124. doi:10.1016/j.jalz.2017.02.009. [PubMed: 28365320]

Gordon BA, Blazey TM, Christensen J, Dincer A, Flores S, Keefe S, Chen C, Su Y, Mcdade EM, Wang G, Li Y, Hassenstab J, Aschenbrenner A, Hornbeck R, Jack CRJ, Ances BM, Berman SB, Brosch JR, Galasko D, Gauthier S, Lah JJ, Masellis M, van Dyck CH, Mintun MA, Klein G, Ristic S, Cairns NJ, Marcus DS, Xiong C, Holtzman DM, Raichle ME, Morris JC, Bateman RJ, Benzinger TLS, 2019. Tau PET in Autosomal Dominant Alzheimer's Disease: Relationship With Cognition. Dementia and other biomarkers, Brain doi:10.1093/brain/awz019.

Greve DN, 2016. [Freesurfer] Longitudinal surface analysis of PET data [WWW Document]. Free. Mail. List URL https://mail.nmr.mgh.harvard.edu/pipermail///freesurfer/2016-September/047921.html (accessed 1.1.16).

Greve DN, Salat DH, Bowen SL, Izquierdo-Garcia D, Schultz AP, Catana C, Becker JA, Svarer C, Knudsen G, Sperling RA, Johnson KA, 2016. Different partial volume correction methods lead to different conclusions: an 18F-FDG PET Study of aging. Neuroimage132, 334–343. doi:10.1016/j.neuroimage.2016.02.042. [PubMed: 26915497]

Greve DN, Svarer C, Fisher PM, Feng L, Hansen AE, Baare W, Rosen B, Fischl B, Knudsen GM, 2014. Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data. Neuroimage92, 225–236. doi:10.1016/j.neuroimage.2013.12.021. [PubMed: 24361666]

Hanseeuw BJ, Betensky RA, Jacobs HIL, Schultz AP, Sepulcre J, Becker JA, Cosio DMO, Farrell M, Quiroz YT, Mormino EC, Buckley RF, Papp KV, Amariglio RA, Dewachter I, Ivanoiu A, Huijbers W, Hedden T, Marshall GA, Chhatwal JP, Rentz DM, Sperling RA, Johnson K, 2019. Association of amyloid and tau with cognition in preclinical alzheimer disease: a longitudinal study. JAMA Neurol. doi:10.1001/jamaneurol.2019.1424,02114.

Harrison TM, La Joie R, Maass A, Baker SL, Swinnerton K, Fenton L, Mellinger TJ, Edwards L, Pham J, Miller BL, Rabinovici GD, Jagust WJ, 2019. Longitudinal tau accumulation and atrophy in aging and alzheimer disease. Ann. Neurol. 85, 229–240. doi:10.1002/ana.25406. [PubMed: 30597624]

Iatrou M, Ross SG, Manjeshwar RM, Stearns CW, 2004. A fully 3D iterative image reconstruction algorithm incorporating data corrections. IEEE Nucl. Sci. Symp. Conf. Rec. 4, 2493–2497. doi:10.1109/NSSMIC.2004.1462761.

Jack CR, Wiste HJ, Schwarz CG, Lowe VJ, Senjem ML, Vemuri P, Weigand SD, Therneau TM, Knopman DS, Gunter JL, Jones DT, Graff-Radford J, Kantarci K, Roberts RO, Mielke MM, Machulda MM, Petersen RC, 2018. Longitudinal tau PET in ageing and Alzheimer's disease. Brain141, 1517–1528. doi:10.1093/brain/awy059. [PubMed: 29538647]

Johnson KA, Schultz A, Betensky RA, Becker JA, Sepulcre J, Rentz D, Mormino E, Chhatwal J, Amariglio R, Papp K, Marshall G, Albers M, Mauro S, Pepin L, Alverio J, Judge K, Philiossaint M, Shoup T, Yokell D, Dickerson B, Gomez-Isla T, Hyman B, Vasdev N, Sperling R, 2016. Tau positron emission tomographic imaging in aging and early Alzheimer disease. Ann. Neurol79, 110–119. doi:10.1002/ana.24546. [PubMed: 26505746]

Josephs KA, Whitwell JL, Tacik P, Duffy JR, Senjem ML, Tosakulwong N, Jack CR, Lowe V, Dickson DW, Murray ME, 2016. [18F]AV-1451 tau-PET uptake does correlate with quantitatively measured 4R-tau burden in autopsy-confirmed corticobasal degeneration. Acta Neuropathol. 132, 931–933. doi:10.1007/s00401-016-1618-1. [PubMed: 27645292]

Knopman DS, Lundt ES, Therneau TM, Vemuri P, Lowe VJ, Kantarci K, Gunter JL, Senjem ML, Mielke MM, Machulda MM, Boeve BF, Jones DT, Graff-Radford J, Albertson SM, Schwarz CG, Petersen RC, Jack CRJ, 2019. Entorhinal cortex tau, amyloid-β, cortical thickness and memory performance in nondemented subjects. Brain1–13. doi:10.1093/brain/awz025. [PubMed: 30596908]

Landau SM, Fero A, Baker SL, Koeppe R, Mintun M, Chen K, Reiman EM, Jagust WJ, 2015. Measurement of longitudinal B-amyloid change with 18F-florbetapir PET and standardized uptake value ratios. J. Nucl. Med56, 567–574. doi:10.2967/jnumed.114.148981. [PubMed: 25745095]

Lee CM, Jacobs HIL, Marquié M, Becker JA, Andrea NV, Jin DS, Schultz AP, Frosch MP, Gómez-Isla T, Sperling RA, Johnson KA, 2018. 18F-flortaucipir binding in choroid plexus: related to race

and hippocampus signal. J. Alzheimer's Dis62, 1691–1702. doi:10.3233/JAD-170840. [PubMed: 29614677]

Lilja J, Leuzy A, Chiotis K, Savitcheva I, Sörensen J, Nordberg A, 2019. Spatial normalization of 18 F-flutemetamol PET images using an adaptive principal-component template. J. Nucl. Med60, 285–291. doi:10.2967/jnumed.118.207811. [PubMed: 29903930]

Lowe VJ, Bruinsma TJ, Wiste H, Min H-K, Fang P, Senjem ML, Weigand SD, Therneau TM, Boeve BF, Josephs KA, Pandey M, Murray ME, Kantarci K, Jones D, Vemuri P, Graff-Radford J, Schwarz CG, Machulda M, Mielke MM, Roberts RO, Knopman DS, Petersen RC, Jack CR, 2019. Cross-sectional associations of Tau-PET signal with cognition in cognitively unimpaired adults. Neurology doi:10.1212/WNL.0000000000007728.

Lowe VJ, Curran G, Fang P, Liesinger AM, Josephs KA, Parisi JE, Kantarci K, Boeve BF, Pandey MK, Bruinsma T, Knopman DS, Jones DT, Petrucelli L, Cook CN, Graff-Radford NR, Dickson DW, Petersen RC, Jack CR, Murray ME, 2016. An autoradiographic evaluation of AV-1451 Tau PET in dementia. Acta Neuropathol. Commun4, 1–19. doi:10.1186/s40478-016-0315-6. [PubMed: 26727948]

Lowe VJ, Wiste HJ, Senjem ML, Weigand SD, Therneau TM, Boeve BF, Josephs KA, Fang P, Pandey MK, Murray ME, Kantarci K, Jones DT, Vemuri P, Graff-Radford J, Schwarz CG, Machulda MM, Mielke MM, Roberts RO, Knopman DS, Petersen RC, Jack CR, 2018. Widespread brain tau and its association with ageing, Braak stage and Alzheimer's dementia. Brain141, 271–287. doi:10.1093/brain/awx320. [PubMed: 29228201]

Maass A, Landau S, Baker SL, Horng A, Lockhart SN, La Joie R, Rabinovici GD, Jagust WJ, 2017. Comparison of multiple tau-PET measures as biomarkers in aging and Alzheimer's Disease. Neuroimage157, 448–463. doi:10.1016/j.neuroimage.2017.05.058. [PubMed: 28587897]

Maass A, Lockhart SN, Harrison TM, Bell RK, Mellinger T, Swinnerton K, Baker SL, Rabinovici GD, Jagust WJ, 2018. Entorhinal tau pathology, episodic memory decline, and neurodegeneration in aging. J. Neurosci38, 530–543. doi:10.1523/JNEUROSCI.2028-17.2017. [PubMed: 29192126]

Marquié M, Normandin MD, Meltzer AC, Siao Tick Chong M, Andrea NV, Antón-Fernández A, Klunk WE, Mathis CA, Ikonomovic MD, Debnath M, Bien EA, Vanderburg CR, Costantino I, Makaretz S, DeVos SL, Oakley DH, Gomperts SN, Growdon JH, Domoto-Reilly K, Lucente D, Dickerson BC, Frosch MP, Hyman BT, Johnson KA, Gómez-Isla T, 2017a. Pathological correlations of [F-18]-AV-1451 imaging in non-alzheimer tauopathies. Ann. Neurol81, 117–128. doi:10.1002/ana.24844. [PubMed: 27997036]

Marquié M, Verwer EE, Meltzer AC, Kim SJW, Agüero C, Gonzalez J, Makaretz SJ, Siao Tick Chong M, Ramanan P, Amaral AC, Normandin MD, Vanderburg CR, Gomperts SN, Johnson KA, Frosch MP, Gómez-Isla T, 2017b. Lessons learned about [F-18]-AV-1451 off-target binding from an autopsy-confirmed Parkinson's case. Acta Neuropathol. Commun5, 75. doi:10.1186/s40478-017-0482-0. [PubMed: 29047416]

Mathis CA, Lopresti BJ, Ikonomovic MD, Klunk WE, 2017. Small-molecule PET tracers for imaging proteinopathies. Semin. Nucl. Med47, 553–575. doi:10.1053/j.semnuclmed.2017.06.003. [PubMed: 28826526]

Mattsson N, Schöll M, Strandberg O, Smith R, Palmqvist S, Insel PS, Häger-ström D, Ohlsson T, Zetterberg H, Jögi J, Blennow K, Hansson O, 2017. 18 F-AV-1451 and CSF T-tau and P-tau as biomarkers in Alzheimer's disease. EMBO Mol. Med46, 1–12. doi:10.15252/emmm.201707809.

Meltzer CC, Leal JP, Mayberg HS, Wagner HNJ, Frost JJ, 1990. Correction of PET data for partial volume effects in human cerebral cortex by MR imaging. J. Comput. Assist. Tomogr14, 561–570. [PubMed: 2370355]

Minhas DS, Price JC, Laymon CM, Becker CR, Klunk WE, Tudorascu DL, Abrahamson EE, Hamilton RL, Kofler JK, Mathis CA, Lopez OL, Ikonomovic MD, 2018. Impact of partial volume correction on the regional correspondence between in vivo [C-11]PiB PET and postmortem measures of Aβ load. NeuroImage Clin. 19, 182–189. doi:10.1016/j.nicl.2018.04.007. [PubMed: 30023168]

Mishra S, Gordon BA, Su Y, Christensen J, Friedrichsen K, Jackson K, Hornbeck R, Balota DA, Cairns NJ, Morris JC, Ances BM, Benzinger TLS, 2017. AV-1451 PET imaging of tau pathology in preclinical Alzheimer disease: defining a summary measure. Neuroimage161, 171–178. doi:10.1016/j.neuroimage.2017.07.050. [PubMed: 28756238]

Moore TJ, Heyward J, Anderson G, Alexander GC, 2020. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015-2017: a cross-sectional study. BMJ Open10, 1–5. doi:10.1136/bmjopen-2020-038863.

Müller-Gärtner HW, Links JM, Prince JL, Bryan RN, McVeigh E, Leal JP, Davatzikos C, Frost JJ, 1992. Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. J. Cereb. Blood Flow Metab12, 571–583. doi:10.1038/jcbfm.1992.81. [PubMed: 1618936]

Ossenkoppele R, Rabinovici GD, Smith R, Cho H, Schöll M, Strandberg O, Palmqvist S, Mattsson N, Janelidze S, Santillo A, Ohlsson T, Jögi J, Tsai R, La Joie R, Kramer J, Boxer AL, Gorno-Tempini ML, Miller BL, Choi JY, Ryu YH, Lyoo CH, Hansson O, 2018. Discriminative accuracy of [18F]flortaucipir positron emission tomography for Alzheimer disease vs other neurodegenerative disorders. JAMA320, 1151–1162. doi:10.1001/jama.2018.12917. [PubMed: 30326496]

Pontecorvo MJ, Devous MD, Navitsky M, Lu M, Salloway S, Schaerf FW, Jennings D, Arora AK, McGeehan A, Lim NC, Xiong H, Joshi AD, Siderowf A, Mintun MA, 2017. Relationships between flortaucipir PET tau binding and amyloid burden, clinical diagnosis, age and cognition. Brain140, 748–763. doi:10.1093/brain/aww334. [PubMed: 28077397]

Price JL, Morris JC, 1999. Tangles and plaques in nondemented aging and “preclinical” Alzheimer’s disease. Ann. Neurol45, 358–368. [PubMed: 10072051]

R Development Core Team, 2008. R: a Language and Environment for Statistical Computing [WWW Document]. URL http://www.r-project.org.

Reuter M, Schmansky NJ, Rosas HD, Fischl B, 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage61, 1402–1418. doi:10.1016/j.neuroimage.2012.02.084. [PubMed: 22430496]

Roberts RO, Geda YE, Knopman DS, Cha RH, Pankratz VS, Boeve BF, Ivnik RJ, Tangalos EG, Petersen RC, Rocca WA, 2008. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology30, 58–69. doi:10.1159/000115751. [PubMed: 18259084]

Rousset OG, Ma Y, Evans AC, 1998. Correction for partial volume effects in PET: principle and validation. J. Nucl. Med39, 904–911. [PubMed: 9591599]

Schöll M, Lockhart SN, Schonhaut DR, O’Neil JP, Janabi M, Ossenkoppele R, Baker SL, Vogel JW, Faria J, Schwimmer HD, Rabinovici GD, Jagust WJ, 2016. PET imaging of tau deposition in the aging human brain. Neuron89, 971–982. doi:10.1016/j.neuron.2016.01.028. [PubMed: 26938442]

Schultz SA, Gordon BA, Mishra S, Su Y, Perrin RJ, Cairns NJ, Morris JC, Ances BM, Benzinger TLS, 2018. Widespread distribution of tauopathy in preclinical Alzheimer’s disease. Neurobiol. Aging72, 177–185. doi:10.1016/j.neurobiolaging.2018.08.022. [PubMed: 30292840]

Schwarz AJ, Shcherbinin S, Slieker LJ, Risacher SL, Charil A, Irizarry MC, Fleisher AS, Southekal S, Joshi AD, Devous MD, Miller BB, Saykin AJ, 2018. Topographic staging of tau positron emission tomography images. Alzheimer’s Dement. 10, 221–231. doi:10.1016/j.dadm.2018.01.006.

Schwarz CG, Gunter JL, Lowe VJ, Weigand S, Vemuri P, Senjem ML, Petersen RC, Knopman DS, Jack CRJ, 2018. A Comparison of Partial Volume Correction Techniques for Measuring Change in Serial Amyloid PET SUVR. J. Alzheimer’s Dis67, 181–195. doi:10.3233/JAD-180749.

Schwarz CG, Gunter JL, Ward CP, Vemuri P, Senjem ML, Wiste HJ, Petersen RC, Knopman DS, Jack CR, 2017a. The Mayo Clinic Adult lifespan template: better quantification across the lifespan. Alzheimer’s Dement. 13, P792. doi:10.1016/j.jalz.2017.06.1071.

Schwarz CG, Jones DT, Gunter JL, Lowe VJ, Vemuri P, Senjem ML, Petersen RC, Knopman DS, Jack CR, 2017b. Contributions of imprecision in PET-MRI rigid registration to imprecision in amyloid PET SUVR measurements. Hum. Brain Mapp38, 3323–3336. doi:10.1002/hbm.23622. [PubMed: 28432784]

Schwarz CG, Senjem ML, Gunter JL, Tosakulwong N, Weigand SD, Kemp BJ, Spychalla AJ, Vemuri P, Petersen RC, Lowe VJ, Jack CRJ, 2017c. Optimizing PiB-PET SUVR change-over-time measurement by a large-scale analysis of longitudinal reliability, plausibility, separability, and correlation with MMSE. Neuroimage144, 113–127. doi:10.1016/j.neuroimage.2016.08.056. [PubMed: 27577718]

Southekal S, Devous MD, Kennedy I, Navitsky M, Lu M, Joshi AD, Pontecorvo MJ, Mintun MA, 2018. Flortaucipir F 18 quantitation using parametric estimation of reference signal intensity. J. Nucl. Med59, 944–951. doi:10.2967/jnumed.117.200006. [PubMed: 29191858]

Sperling RA, Mormino EC, Schultz AP, Betensky RA, Papp KV, Amariglio RE, Hanseeuw BJ, Buckley R, Chhatwal J, Hedden T, Marshall GA, Quiroz YT, Donovan NJ, Jackson J, Gatchel JR, Rabin JS, Jacobs H, Yang H, Properzi M, Kirn DR, Rentz DM, Johnson KA, 2019. The impact of amyloid-beta and tau on prospective cognitive decline in older individuals. Ann. Neurol85, 181–193. doi:10.1002/ana.25395. [PubMed: 30549303]

Stearns CW, Fessler JA, 2002. 3D PET reconstruction with FORE and WLS-OS-EM. In: IEEE Nuclear Science Symposium Conference Record. IEEE, pp. 912–915. doi:10.1109/NSSMIC.2002.1239472.

Tetzloff KA, Graff-Radford J, Martin PR, Tosakulwong N, Machulda MM, Duffy JR, Clark HM, Senjem ML, Schwarz CG, Spychalla AJ, Drubach DA, Jack CR, Lowe VJ, Josephs KA, Whitwell JL, 2018. Regional distribution, asymmetry, and clinical correlates of tau uptake on [18F]AV-1451 PET in atypical Alzheimer's disease. J. Alzheimers Dis62, 1713–1724. doi:10.3233/JAD-170740. [PubMed: 29614676]

Thomas BA, Cuplov V, Bousse A, Mendes A, Thielemans K, Hutton BF, Erlandsson K, 2016. PETPVC: a toolbox for performing partial volume correction techniques in positron emission tomography. Phys. Med. Biol61, 7975–7993. doi:10.1088/0031-9155/61/22/7975. [PubMed: 27779136]

Timmers T, Ossenkoppele R, Visser D, Tuncel H, Wolters EE, Verfaillie SCJ, van der Flier WM, Boellaard R, Golla SSV, van Berckel BNM, 2019. Test-retest repeatability of [18F]Flortaucipir PET in Alzheimer's disease and cognitively normal individuals. J. Cereb. Blood Flow Metab40, 2464–2474. doi:10.1177/0271678X19879226. [PubMed: 31575335]

Vemuri P, Lowe VJ, Knopman DS, Senjem ML, Kemp BJ, Schwarz CG, Przybelski SA, Machulda MM, Petersen RC, Jack CR, 2017. Tau-PET uptake: Regional variation in average SUVR and impact of amyloid deposition. Alzheimer's Dement. Diagnosis Assess. Dis. Monit6, 21–30. doi:10.1016/j.dadm.2016.12.010.

Vemuri P, Senjem ML, Gunter JL, Lundt ES, Tosakulwong N, Weigand SD, Borowski BJ, Bernstein MA, Zuk SM, Lowe VJ, Knopman DS, Petersen RC, Fox NC, Thompson PM, Weiner MW, Jack CR, 2015. Accelerated vs. unaccelerated serial MRI based TBM-SyN measurements for clinical trials in Alzheimer's disease. Neuroimage113, 61–69. doi:10.1016/j.neuroimage.2015.03.026. [PubMed: 25797830]

Whittington A, Gunn RN, 2018. Amyloid Load – a more sensitive biomarker for amyloid imaging. J. Nucl. Med118, 210518. doi:10.2967/jnumed.118.210518.

Wolters EE, Golla SSV, Timmers T, Ossenkoppele R, van der Weijden CWJ, Scheltens P, Schwarte L, Schuit RC, Windhorst AD, Barkhof F, Yaqub M, Lammertsma AA, Boellaard R, van Berckel BNM, 2018. A novel partial volume correction method for accurate quantification of [18F] flortaucipir in the hippocampus. EJNMMI Res. 8, 79. doi:10.1186/s13550-018-0432-2. [PubMed: 30112620]

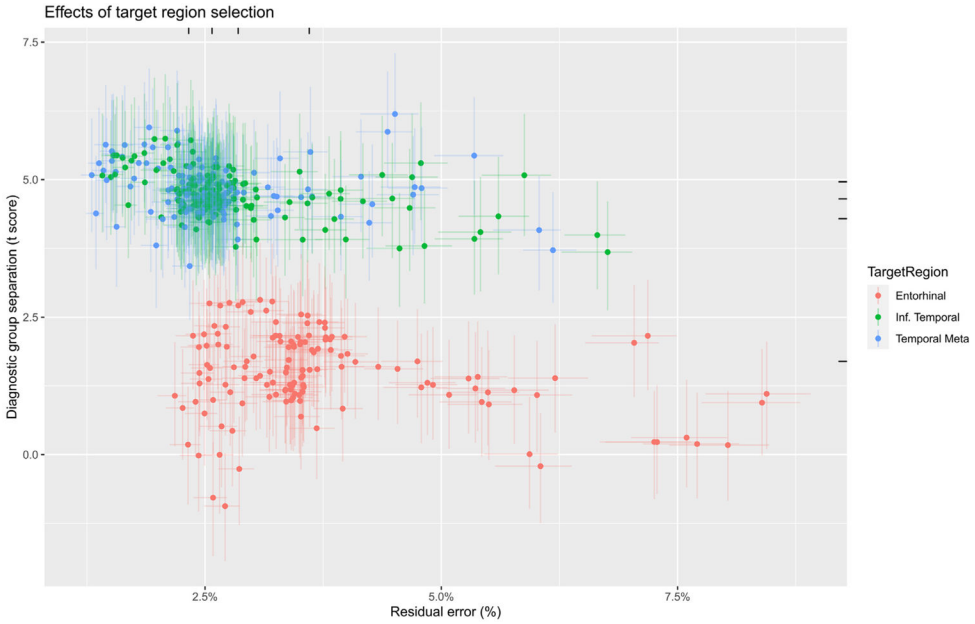


Fig. 1.

Each point shows the estimated relative residual error (x-axis) and longitudinal group separation effect size (*y*-axis; t statistic) for a given SUVR method, across all participants. Methods in the top-left are best by both criteria. The crossbars show 95% confidence intervals for each point's position on each axis. Points are colored by that method's choice of target region. Ticks in the margins indicate the 20, 40, 60, and 80th percentiles for each axis.

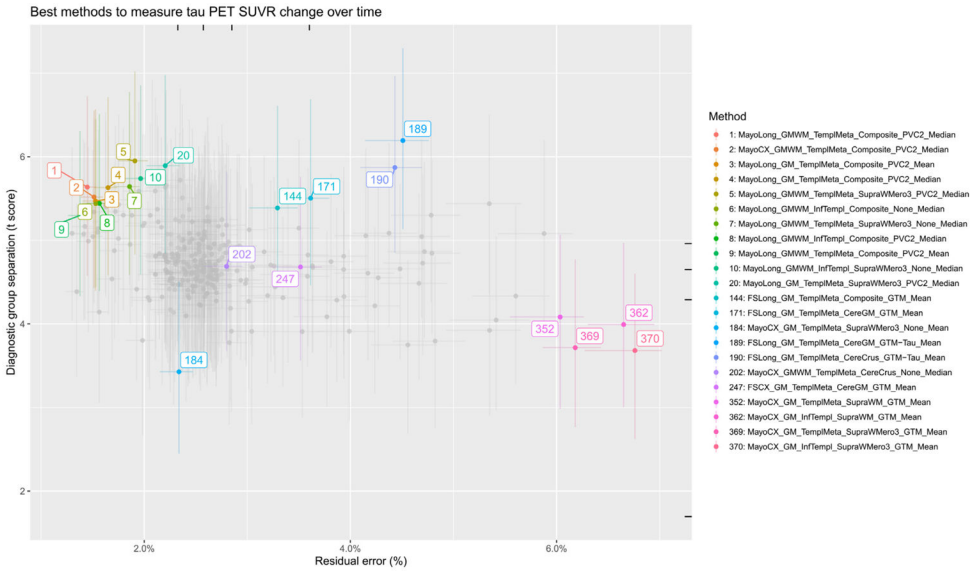
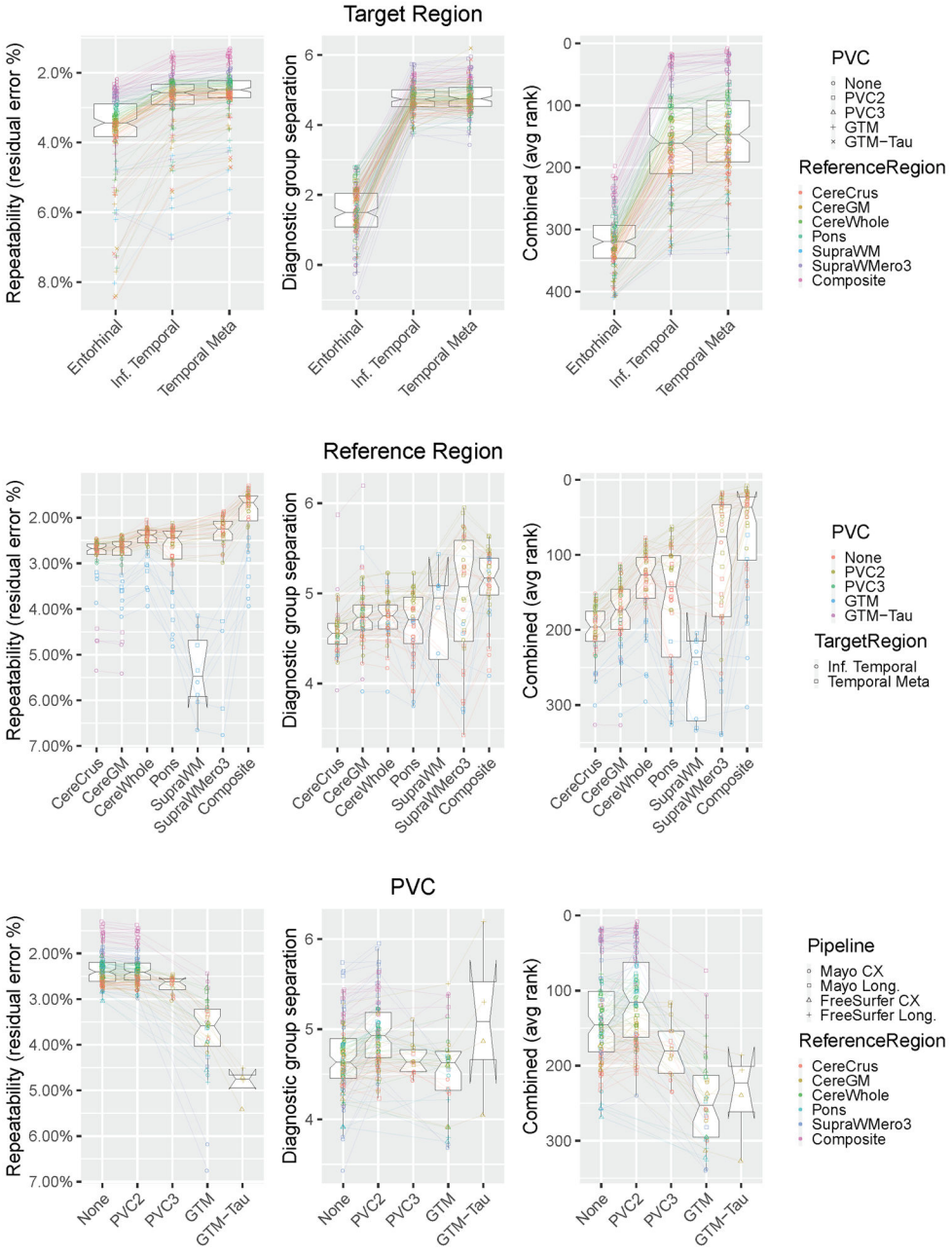


Fig. 2.

Each point shows the estimated relative residuals error (*x*-axis) and longitudinal group separation effect size (*y*-axis; t statistic) for a given SUVR method, across all participants. Methods in the top-left are best by both criteria. Selected points are labelled, for discussion purposes, using numbered labels (their ranking by the average of their ranks on both criteria). The cross-bars show 95% confidence intervals for each point's position on each axis. To reduce over-plotting, methods using the entorhinal cortex target region ($n = 152$, which performed strictly worse than the other target regions) are omitted from the plot, leaving 304 total points/methods here. Ticks in the margins indicate the 20, 40, 60, and 80th percentiles for each axis.



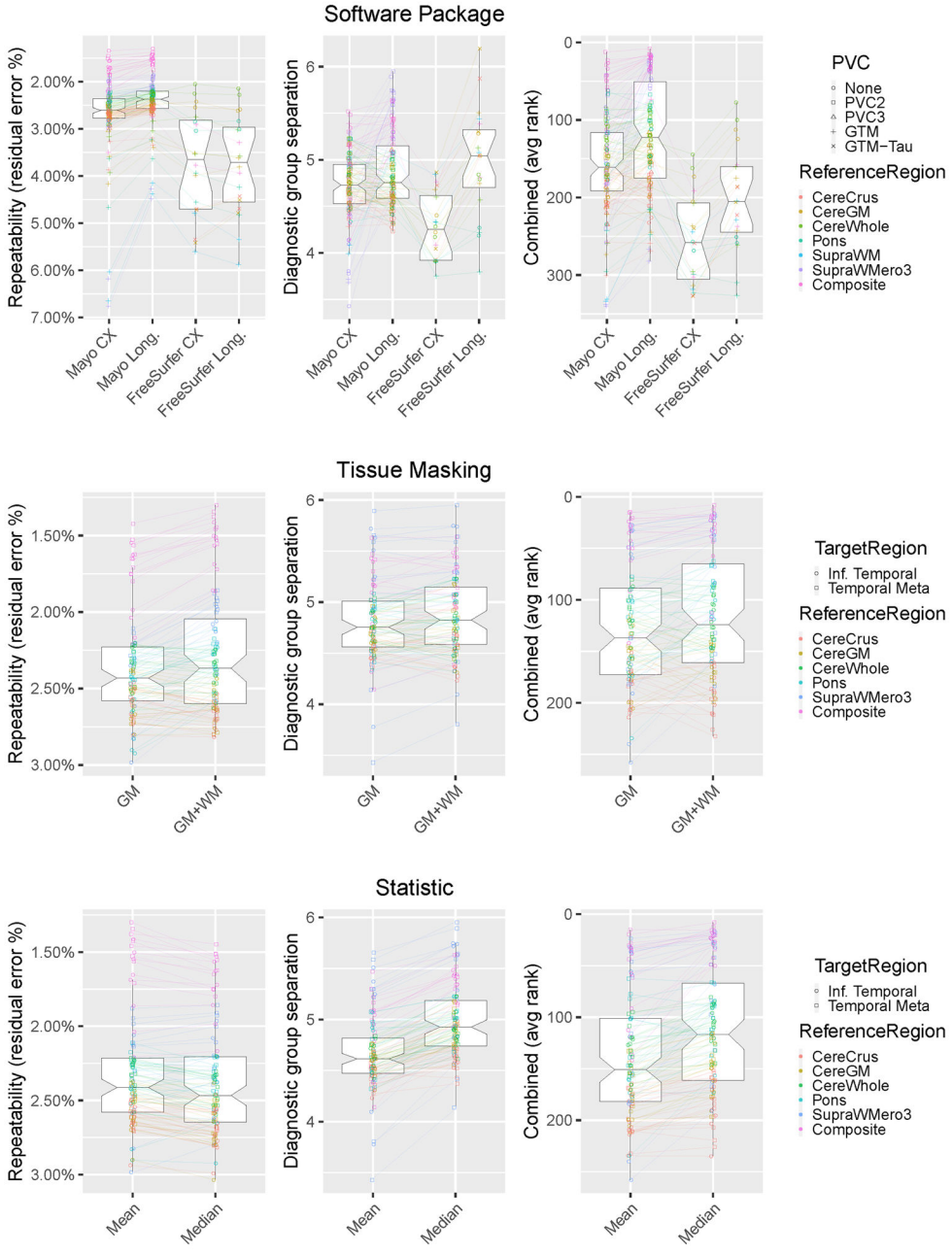


Fig. 3.

(part 1): Contributions of individual methodological choices. Each point is an SUVR method, and lines indicate sets where all other variables were consistent. Only valid combinations are plotted, and points without at least one valid direct comparison are omitted. Axes are flipped as needed so that higher values are better. The combined criterion is the average of each method's rank according to each of the two criteria individually. Methods using the entorhinal target are omitted from all plots except target region.

Fig. 3 (part 2): Contributions of individual methodological choices. Each point is an SUVR method, and lines indicate sets where all other variables were consistent. Only valid combinations are plotted, and points without at least one valid direct comparison are omitted. Axes are flipped as needed so that higher values are better. The combined criterion is the average of each method's rank according to each of the two criteria individually. Methods using the entorhinal target are omitted from all plots except target region.

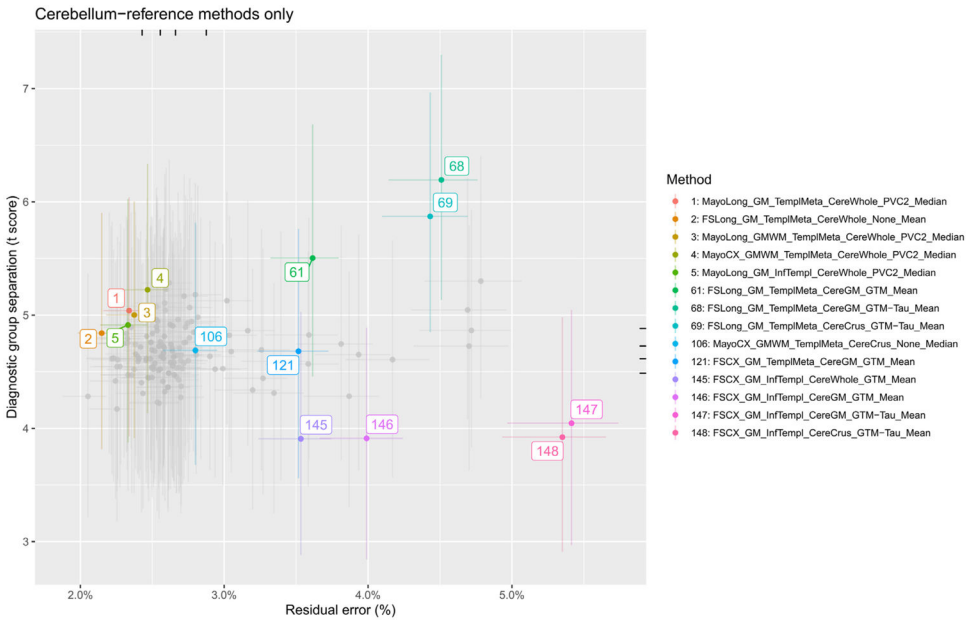


Fig. 4.

Cerebellum-reference subset of Fig. 2: Each point shows the estimated relative residual error (*x* axis) and longitudinal group separation effect size (*y* axis; t statistic) for a given SUVR method, across all participants. Methods plotted in the top-left are best by both criteria. Selected points are labelled, for discussion purposes, using numbered labels (their ranking by the average of their ranks on both criteria). The cross-bars show 95% confidence intervals for each point's position on each axis. To reduce over-plotting, methods using the entorhinal cortex target region ($n = 74$, which performed strictly worse than the other target regions) are omitted from the plot, leaving 148 total points/methods here. Ticks in the margins indicate the 20, 40, 60, and 80th percentiles for each axis.

Table 1

Options tested (rows) for each methodological choice (columns).

Target Region	Reference Region	Partial Volume Correction (PVC)	Software Package	Target Tissue Masking	Statistic
• Entorhinal Cortex • Inferior Temporal • Temporal Meta-ROI	• Cerebellar Crus (CereCrus) • Cerebellar GM (CereGM) • Whole Cerebellum (CereWhole) • Pons • Supratentorial WM (SupraWM) • Supratentorial WM eroded 3mm (SupraWMEro3) • Composite (SupraWMEro3 + Pons + CereWhole)	• None • 2-Compartment (PVC2) • 3-Compartment (PVC3) • Gaussian Transfer Matrix (GTM) • GTM with FTP-specific enhancements (GTM-Tau)	• Mayo Cross-Sectional • Mayo Longitudinal • FreeSurfer 6.0 Cross-Sectional • FreeSurfer 6.0 Longitudinal	• GM • GM + WM	• Mean • Median

Table 2

Participant Demographics.

Characteristic	All participants	Unimpaired	Impaired	P-value
Number of subjects	97	46	51	—
Sex, n (%)				
Female	35 (36%)	16 (35%)	19 (37%)	0.80
Male	62 (64%)	30 (65%)	32 (63%)	
Age at baseline PET, years	69 (62, 76) [45, 94]	70.5 (61, 76) [45, 87]	67 (63, 76) [52, 94]	0.84
Education, years {2}	16 (14, 18) [7, 24]	16 (14,18, [12, 20]	16 (12, 17) [7, 24]	0.11
Global cortical PIB, SUVR	1.82 (1.36,2.41) [1.19, 3.38]	1.38 (1.32, 1.56) [1.19, 3.22]	2.28 (1.87, 2.59) [1.32, 3.38]	<0.001
Global cortical Tau, SUVR	1.24 (1.17, 1.51) [0.99, 2.87]	1.19 (1.14, 1.24) [1.01, 1.50]	1.50 (1.24, 1.91) [0.99, 2.87]	<0.001
Diagnosis at baseline, n (%)				
Cognitively unimpaired	46 (47%)	46 (100%)	0	—
Impaired	51 (53%)	0	51 (100%)	
APOE $\varepsilon 4$, n (%) {2}				
Carrier	51 (54%)	19 (41%)	32 (65%)	0.02
Non-carrier	44 (46%)	27 (59%)	17 (35%)	
MMSE score {4}	28 (25, 29) [13, 30]	29 (28, 29) [25, 30]	25 (24, 28) [13, 29]	< 0.001
Time between first and third scan, years	2.5 (2.1, 3.7) [1.6, 4.3]	3.7 (2.5, 3.9) [1.6, 4.3]	2.2 (2.0, 2.6) [1.8, 4.3]	< 0.001
Time between corresponding MRI and Tau PET scans, days[a]	1 (1, 7) [0, 55]	3 (1, 8) [0, 36]	1 (1, 2) [0, 55]	0.002

Values are given as: median (1st quartile, 3rd quartile) [min to max] or number (percent)

Abbreviations: n: Number of subjects; CU: Clinically Unimpaired; MCI: Mild Cognitive Impairment; APOE: apolipoprotein E; MMSE: Mini-Mental State Exam

[a]Based on all 291 scans for all 97 individuals{} Brackets in the characteristics column indicate the number of subjects missing this particular variable.P-values are from either a Wilcoxon two-sample two-sided rank sum test or chi-squared test.