Evaluation of ComBat harmonization for reducing across-tracer biases in regional amyloid PET analyses

Author information

Braden Yang[1] (0000-0002-2558-4132), Tom Earnest[1] (0000-0001-8671-8424), Sayantan Kumar[1] (0000-0001-7213-0734), Deydeep Kothapalli[1], Tammie Benzinger[1] (0000-0002-8114-0552), Brian Gordon[1] (0000-0003-2109-2955), Aristeidis Sotiras[1,2] (0000-0003-0795-8820)

[1] Mallinckrodt Institute of Radiology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA 63110

[2] Institute for Informatics, Data Science and Biostatistics, Washington University School of Medicine in St. Louis, St. Louis, MO, USA 63110

Please address all correspondence to: Braden Yang (b.y.yang@wustl.edu), Aristeidis Sotiras (aristeidis.sotiras@wustl.edu)

Abstract

Background

Differences in amyloid positron emission tomography (PET) radiotracer pharmacokinetics and binding properties lead to discrepancies in amyloid-β uptake estimates. Harmonization of tracer-specific biases is crucial for optimal performance of downstream tasks. Here, we investigated the efficacy of ComBat, a data-driven harmonization model, for reducing tracer-specific biases in regional amyloid PET measurements from [^{18}F]-florbetapir (FBP) and [^{11}C]-Pittsburgh Compound-B (PiB).

Methods

One-hundred-thirteen head-to-head FBP-PiB scan pairs, scanned from the same subject within ninety days, were selected from the Open Access Series of Imaging Studies 3 (OASIS-3) dataset. The Centiloid scale, ComBat with no covariates, ComBat with biological covariates, and GAM-ComBat with biological covariates were used to harmonize both global and regional amyloid standardized uptake value ratios (SUVR). Intraclass correlation coefficient (ICC) and mean standardized absolute error (MsAE) were computed to measure the absolute agreement between tracers. Additionally, longitudinal amyloid SUVRs from an anti-amyloid drug trial were simulated using linear mixed effects modeling. Differences in rates-of-change between simulated treatment and placebo groups were tested, and change in statistical power/Type-I error after harmonization was quantified.

Results

In the head-to-head tracer comparison, the best ICC and MsAE were achieved after harmonizing with ComBat with no covariates for the global summary SUVR. ComBat with no covariates also performed the best in harmonizing regional SUVRs. In the clinical trial simulation, harmonization with both Centiloid and ComBat increased statistical power of detecting true rate-of-change differences between groups and decreased false discovery rate in the absence of a treatment effect. The greatest benefit of harmonization was observed when groups exhibited differing FPB-to-PiB proportions.

Conclusions

ComBat outperformed the Centiloid scale in harmonizing both global and regional amyloid estimates. Additionally, ComBat improved the detection of rate-of-change differences between clinical trial groups. Our findings suggest that ComBat is a viable alternative to Centiloid for harmonizing regional amyloid PET analyses.

Keywords

Positron emission tomography, amyloid-β, harmonization, Centiloid, ComBat

Background

Positron emission tomography (PET) is widely used in clinical and research settings for measuring and monitoring amyloid-β deposition *in vivo* in the brain for patients who are at risk of developing or who already present with Alzheimer's disease (AD). In clinical trials for anti-amyloid drugs, PET is an important tool for screening appropriate candidates who have undergone significant amyloidosis in the brain [1]. Moreover, PET has also been used for monitoring the progression of global amyloid burden longitudinally within these trials, which along with measures of cognitive function serves as a crucial secondary endpoint [2,3]. In research settings, PET is able to resolve the spatial distribution of amyloid within specific regions of the brain, enabling the design of multivariable statistical analyses and predictive models of AD using voxel-wise [4,5] or region-of-interest (ROI) based [6–8] PET biomarkers as multidimensional features.

Several PET radiotracers for imaging brain amyloid pathology have been developed. The first amyloid PET tracer developed for human imaging studies was [^{11}C]-Pittsburg compound B (PiB) [9], but due to its short half-life requires an on-site cyclotron to produce. Consequently, PiB is not accessible by many sites and not appropriate for use in clinical trials. Alternatively, amyloid measurements obtained from ^{18}F-based tracers such as [^{18}F]-florbetapir (FBP) [10–12], [^{18}F]-florbetaben [13] and [^{18}F]-flutemetamol [14,15] have been shown to correlate well with PiB. Coupled with a much longer half-life than PiB, these tracers are a much more suitable option for clinical trials due to their accessibility and ability to be distributed off-site.

Nonetheless, previous studies that performed a head-to-head comparison of amyloid PET tracers have demonstrated significant disparities in dynamic range and non-specific binding properties

between tracers [10,13,16]. Subsequently, this makes it difficult to compare quantitative amyloid measurements between images acquired using different tracers. This may also negatively impact the performance of downstream tasks such as detecting significant treatment effects in anti-amyloid drug trials [17].

To address this, Klunk *et al.* introduced the Centiloid scale [18], which linearly transforms the dynamic range of a global estimate of amyloid burden to a common scale and converts it to Centiloid (CL) units. This involves calibrating the scale to a preselected cohort of amyloid-negative healthy controls and amyloid-positive typical AD patients, where the average global burden of the two groups are set to 0 CL and 100 CL, respectively. However, the calibration process requires at least two PET scans from the same subject within a short time period in order to calibrate conversion equations. Additionally, a single equation is usually derived to operate on the global amyloid estimate, but this cannot address local disparities in amyloid PET signal between tracers. Other methods for tracer harmonization that are based on data-driven and/or machine learning techniques such as principal component analysis [19], non-negative matrix factorization [20], and deep learning [21,22] have been proposed, but like Centiloid they focus on the global amyloid burden.

Alternatively, ComBat [23] is a data-driven harmonization model which has been widely applied in magnetic resonance imaging (MRI) analyses to adjust for differences in scanners and acquisition protocols. It has been used to correct regional volume and cortical thickness measurements from MRI [24–27], and has more recently been applied to [^{18}F]-fluorodeoxyglucose-PET [28] and amyloid PET [29] biomarkers. Much of the current literature on applying ComBat has focused on reducing scanner-level and institutional-level biases. However, it remains unclear whether ComBat is applicable for mitigating across-tracer variance, specifically in regional amyloid PET measurements.

Here, we aimed to evaluate the efficacy of ComBat for harmonizing standardized uptake value ratios (SUVR) from amyloid PET across two tracers - PiB and FBP. Specifically, we addressed two primary inquiries. Firstly, we investigated whether ComBat harmonization may increase the agreement between regional SUVRs obtained from the two tracers. This was accomplished through a head-to-head comparison of PiB and FBP. We selected a set of PiB-FBP scan pairs acquired from the same subject in a short time period and compared measures of the absolute agreement between regional SUVRs before and after ComBat harmonization. Secondly, we explored the utility of ComBat harmonization in the context of clinical tasks. This was examined by simulating a multi-tracer anti-amyloid drug trial where two different amyloid tracers were used to measure brain amyloid deposition, under the assumption that different sites have access to different tracers. We generated longitudinal amyloid PET data of hypothetical treatment and placebo groups with a known underlying treatment effect, and assigned each group a specific proportion of PiB-to-AV45 scans. We then gauged whether ComBat harmonization improves the statistical power of detecting the underlying treatment effect when using two different tracers.

Methods

Participants and data

Data for this study were acquired from the Open Access Series of Imaging Studies 3 (OASIS-3) dataset [30], which consisted of 1098 total participants and their longitudinal imaging data. Of these, we selected 997 who underwent PiB and/or FBP amyloid PET imaging. All available PET scans, including the initial baseline scan and any follow-up scans, were utilized in this study, for a total of 678 FBP scans and 1157 PiB scans. Additionally, each subject's age at scan, sex and apolipoprotein-ε4 (APOE) allele carriership were extracted. Subjects who were missing any of these variables were excluded from further analyses.

Image acquisition and processing

All amyloid PET imaging from OASIS-3 were acquired at Washington University in St. Louis using one of four Siemens scanner models: Biograph mMR PET/MR 3T, Biograph 40 PET/CT, Biograph 128 Vision Edge PET/CT, and ECAT HR+ 962 PET. For PiB PET, participants received a bolus injection of 6-20 mCi of PiB, and a 60-minute dynamic scan was acquired. For FBP PET, participants received a bolus injection of 10 mCi of FBP, and either a 70-minute dynamic scan was acquired, or a 20-minute dynamic scan was acquired at 50-minutes post-injection. Additionally, T1-weighted MRI scans were acquired and utilized for PET processing. All MRI imaging from the OASIS-3 dataset were acquired at Washington University in St. Louis using one of three Siemens scanner models: Vision 1.5T, TIM Trio 3T, and Biograph mMR PET/MR 3T.

PET images were processed using the PET Unified Pipeline (https://github.com/ysu001/PUP), described in [31]. Briefly, raw PET images were smoothed to 8mm spatial resolution, corrected for inter-frame motion, and coregistered to the T1 MRI scan acquired closest in time using a vector-gradient algorithm. T1 images were segmented and parcellated into cortical and subcortical ROIs using FreeSurfer 5.0 or 5.1 for 1.5T scans or FreeSurfer 5.3 for 3T scans. For each ROI, regional SUVRs were computed from the peak time windows of each tracer (30-to-60 minutes post-injection for PiB, 50-to-70 minutes post-injection for FBP). The average of the left and right cerebellar cortex was used as the reference region. Additionally, a summary estimate of global amyloid burden was derived by computing the SUVR of a meta-ROI comprised of lateral and medial orbitofrontal, middle and superior temporal, superior frontal, rostral middle frontal, and precuneus ROIs from both hemispheres. For subsequent analyses, we chose to focus on 68 cortical, 16 subcortical regions, and the global summary region. The full list of regions is given in Supplementary Table 1.

Data harmonization

Three harmonization methods were investigated in the current study: Centiloid [18], ComBat [23,27], and GAM-ComBat [26]. These methods are briefly described below.

Centiloid

The Centiloid scale [18] is a method of linearly transforming global amyloid burden estimates from SUVRs to a scale that is standardized across tracers. Centiloid ranges from 0 to 100, with 0 corresponding to the average amyloid burden of a group of healthy controls, and 100 corresponding to the average amyloid burden of typical AD patients. Note that Centiloids are allowed to fall above 100 CL or below 0 CL.

Although Centiloid is calibrated against and primarily used to harmonize the global summary SUVR, it can also be applied to regional or voxel-wise SUVRs [17,18]. To convert regional SUVRs to Centiloid, we utilized the conversion equations that were previously validated for the OASIS-3 cohort [10,32]:

$$\begin{aligned} CL_{PiB} &= 111.8\, SUVR_{PiB} - 119.3 \\ CL_{FBP} &= 163.6\, SUVR_{FBP} - 181.0 \end{aligned} \qquad (1)$$

ComBat

ComBat is a data-driven method for adjusting data with batch-specific effects [23], where batches refer to any nominal variable(s) which may contribute confounding biases in the target measurement. It utilizes a multivariable linear regression to model measurements in terms of batch-specific shift and scale parameters, as well as other covariates which model variance due to biologically relevant effects. For batch effect i, subject j and feature k, ComBat models the measurement y_{ijk} as:

$$y_{ijk} = \alpha_k + X_j\beta_k + \gamma_{ik} + \delta_{ik}\epsilon_{ijk} \qquad (2)$$

where α_k is the mean measurement across all subjects and all batches, X_j is the vector of biological covariates associated with subject j, and β_k is the vector of coefficients for X_j. The batch-specific shift (additive) and scale (multiplicative) parameters are represented by γ_{ik} and δ_{ik} respectively. These modify the measurement from the group average to account for batch-specific biases. ϵ_{ijk} is the error term, which is assumed to be normally distributed with zero mean and unit variance. γ_{ik} and δ_{ik} are estimated using an empirical Baysean approach, and once estimated, the measurement without batch effects can be recovered by the following:

$$y^*_{ijk} = \frac{y_{ijk} - \alpha_k - X\beta_k\gamma_{ik}}{\delta_{ik}} + \alpha_k + X\beta_k \qquad (3)$$

This adjustment ensures that only variance due to the batch effects is corrected for, while variance due to the covariates is preserved, which is a unique advantage of ComBat over other batch-adjusting techniques. In subsequent experiments, we selected age, sex, and APOE carriership as the covariates of interest to preserve.

GAM-ComBat

A limitation of the ComBat model is that it is only able to model covariates as linearly related to the target variable. To address this, Pomponio *et al.* [26] developed GAM-ComBat, a variant of ComBat which can model continuous covariates non-linearly using generalized additive models (GAM):

$$y_{ijk} = f(X_j) + \gamma_{ik} + \delta_{ik}\epsilon_{ijk} \qquad (4)$$

where f is the GAM. In subsequent experiments, we explored modeling the age covariate non-linearly using GAMs.

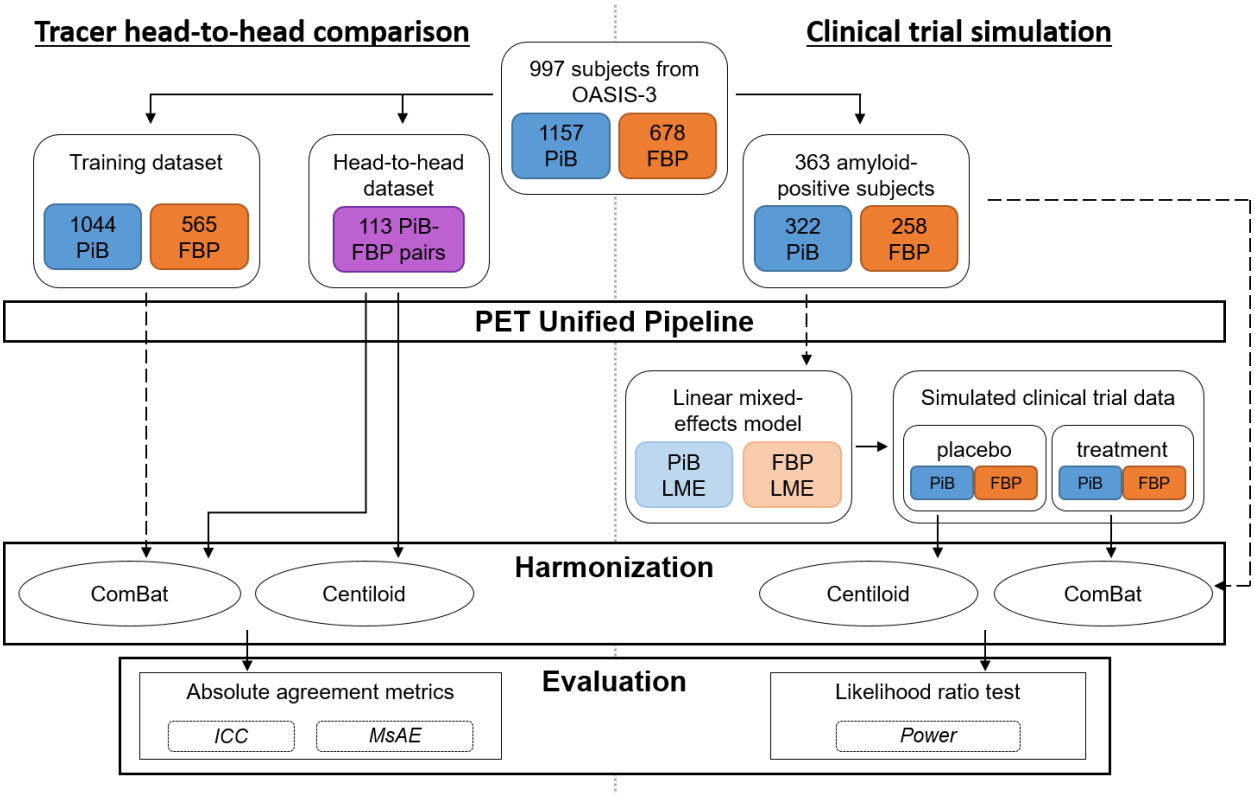

Fig. 1 Flowchart of data for the tracer head-to-head comparison (left) and clinical trial simulation (right). Dotted arrows indicate where data was used to train ComBat or linear mixed effects models

Statistical analysis

Tracer head-to-head comparison

We performed a head-to-head comparison of FBP and PiB measurements and evaluated their absolute agreement after harmonization. We identified 113 FBP-PiB scan pairs across 99 subjects which were acquired within 90 days. All remaining scans were used to train ComBat models, which included 565 FBP and 1044 PiB scans.

Centiloid and three different configurations of ComBat were applied to the global summary SUVR and 84 regional SUVRs from the head-to-head dataset to harmonize tracer differences. We tested ComBat without any covariates, ComBat with age, sex and APOE-ϵ4 carriership as linear covariates, and GAM-ComBat with sex and APOE-ϵ4 carriership as linear covariates and age as a non-linear covariate.

To evaluate the absolute agreement between FBP and PiB measurements, two metrics were computed. Firstly, intraclass correlation coefficient (ICC) using a fixed rater, single measurement model (i.e. ICC3) was estimated. ICC is roughly the ratio of intraclass variance to total variance, and values closer to 1 indicate better agreement between the two tracers. Secondly, we derived a standardized version of the mean absolute error between PiB and FBP measurements, which we termed mean standardized absolute error (MsAE). Standardization was performed to allow for comparisons of absolute errors between Centiloids and SUVRs, which otherwise have different dynamic ranges. To standardize errors, we used the following formulae:

$$\Delta_i = \text{abs}(PiB_i - FBP_i) \qquad (5)$$

$$MsAE = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta_i}{\sigma_{\Delta_i}} \qquad (6)$$

where PiB_i and FBP_i are the measurements made with PiB and FBP from scan pair i, N is the number of scan pairs, and σ_{Δ_i} is the standard deviation of the absolute differences:

$$\sigma_{\Delta_i} = \sqrt{\frac{\sum_{i=1}^{N}(\Delta_i - \bar{\Delta})}{N-1}} \qquad (7)$$

Paired t-tests were performed to test for significant differences in the distributions of ICC and MsAE between unharmonized SUVRs and each of the four harmonization methods. Additionally, we further subdivided each FreeSurfer region into three groups – regions belonging to the global summary meta-ROI, other cortical regions not part of the summary meta-ROI, and subcortical regions. We then performed paired t-tests for each group separately to compare across harmonization methods. In all statistical tests, Bonferroni correction was applied to correct for multiple comparisons.

Clinical trial simulation

We evaluated Centiloid and ComBat in the context of improving detection of treatment effects in an anti-amyloid drug trial setting, with the assumption that multiple amyloid PET tracers were used due to pooling of data from multiple institutions. To accomplish this, we modeled a simulation experiment after those described in Chen *et al.* [17] to generate data of placebo and treatment groups. We varied the proportion of FBP-to-PiB scans of each group, then tested for group differences of amyloid rate-of-change.

We selected subjects who presented as PET amyloid-positive at least once during their participation in OASIS-3. To mark scans as amyloid-positive, we used a global summary SUVR threshold of 1.31 for PiB and 1.24 for FBP. These thresholds were previously validated for the OASIS-3 cohort [10]. From these criteria, we identified 363 amyloid-positive subjects, from which 258 FBP and 322 PiB scans were selected.

For each tracer and for each region-of-interest (including the global summary region), a linear mixed effects (LME) model was fit on the selected scans to predict longitudinal SUVR. Sex, APOE carriership, baseline age, and time-from-baseline were specified as fixed effects. A random intercept grouped by subject was specified as the only random effect. Fitted LME models were then used to generate new longitudinal data of placebo and treatment groups. For the placebo group, the models were applied as is to generate SUVRs that follow the natural longitudinal trajectory among amyloid-positive subjects in OASIS-3. For the treatment group, we added a negative rate-of-change term to the LME equation to mimic a treatment effect. We tested multiple values of the treatment effect from 0 to -0.03 SUVR, varying in increments of -0.01 SUVR. These values were chosen based off of previously reported clinical trial effect sizes [2,17].

To simulate a single subject's data, the empirical distributions of number of longitudinal scans, age at baseline scan, and interval between scans among the OASIS-3 amyloid-positive cohort

were randomly sampled to generate longitudinal time points. We then randomly assigned each time point a tracer (either PiB or FBP) according to a prespecified tracer mixing proportion. Time points were then fed into the trained LME model to obtain simulated SUVR measurements. For these simulation experiments, we varied the percentage of FBP scans from 0.1 to 0.9 in increments of 0.2 for both clinical trial groups independently. We fixed the number of subjects to 50 per group.

The simulated data was harmonized using either Centiloid, ComBat with no covariates, or ComBat with age, sex and APOE-ε4 carriership as covariates. ComBat was trained on the same data used to train the generative LME models. Note that we omitted GAM-ComBat from this analysis, since the simulated data was generated using a linear age term in the LME. To test for group differences in the rate-of-change in amyloid SUVR between placebo and treatment groups, we first fitted the same LME described previously, but with two additional terms – clinical trial group and interaction of time-from-baseline with trial group. We then tested for statistical significance of the time-from-baseline and clinical trial group interaction term, which would indicate whether the two groups exhibit different rates-of-change. Likelihood ratio tests were used to compare the fit of the full model with a nested model that excludes this term, and significance was determined using $\alpha = 0.05$. Simulations were repeated 1000 times for each permutation of tracer mixing proportions and treatment effect. Statistical power was computed as the proportion of simulation iterations which resulted in a significant finding. Note that for a treatment effect of zero, i.e. the absence of a ground truth treatment effect, this corresponds to the Type-I error rate.

Results

Demographics

Descriptive statistics of each cohort are listed in Table 1. A two-tailed t-test was used to test for differences in age at scan, and Fisher's exact test was used to test for differences in sex, APOE-ε4, and Clinical Dementia Rating (CDR). Significant differences in age were observed between

Table 1 Demographics of each cohort. Note that the reported CDRs are counted by scans, where the closest CDR score in time was assigned to every scan. Statistically significant differences are denoted with asterisks and crosses. Asterisks indicate tracer vs. tracer comparisons within the same cohort, whereas crosses indicate cohort vs. cohort comparisons within the same tracer. The number of symbols indicates the significance level (1 = $p < 0.05$, 2 = $p < 0.01$, 3 = $p < 0.005$, 4 = $p < 1e\text{-}4$)

	Tracer head-to-head comparison				Simulation	
	head-to-head		training			
	FBP	PiB	FBP	PiB	FBP	PiB
Number of subjects	99	99	520	671	213	210
Number of scans	113	113	565	1044	258	322
Mean age at scan (± sd)[1]	69.1 ± 8.48	69 ± 8.48	71.6 ± 8††	68.9 ± 9.37****	73.5 ± 7.18	73.1 ± 7.34
Number of males/females	45/54	45/54	223/297	283/388	77/136	106/104*
APOE noncarriers/carriers[2]	67/32	67/32	337/183	441/230	104/109	78/132*
Number of CDR = 0/CDR = 0.5/CDR > 1[3]	108/4/1	108/4/1	480/67/18†	900/118/26†	189/53/16	232/68/22
Number of scans per subject (± sd)	1.1 ± 0.38	1.1 ± 0.38	1.1 ± 0.29	1.6 ± 0.82	1.2 ± 0.49	1.5 ± 0.88
Mean years between scans (± sd)	1.6 ± 0.54	1.6 ± 0.53	3 ± 0.86	3.8 ± 2.01	2.1 ± 1.06	3.3 ± 1.52

* = significance of tracer vs. tracer comparison (within same cohort)
† = significance of cohort vs. cohort comparison (within same tracer)
[1] sd = standard deviation
[2] APOE = apolipoprotein-ε4
[3] CDR = Clinical Dementia Rating

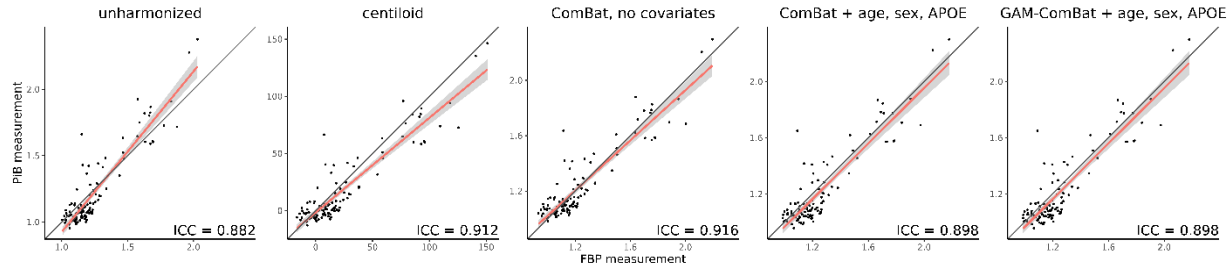

Fig. 2 Global summary measures computed from PiB and FBP scans in the tracer head-to-head dataset. The red line indicates the best fit line from OLS linear regression, the gray area indicates confidence interval of the slope, and the black line represents the identity line

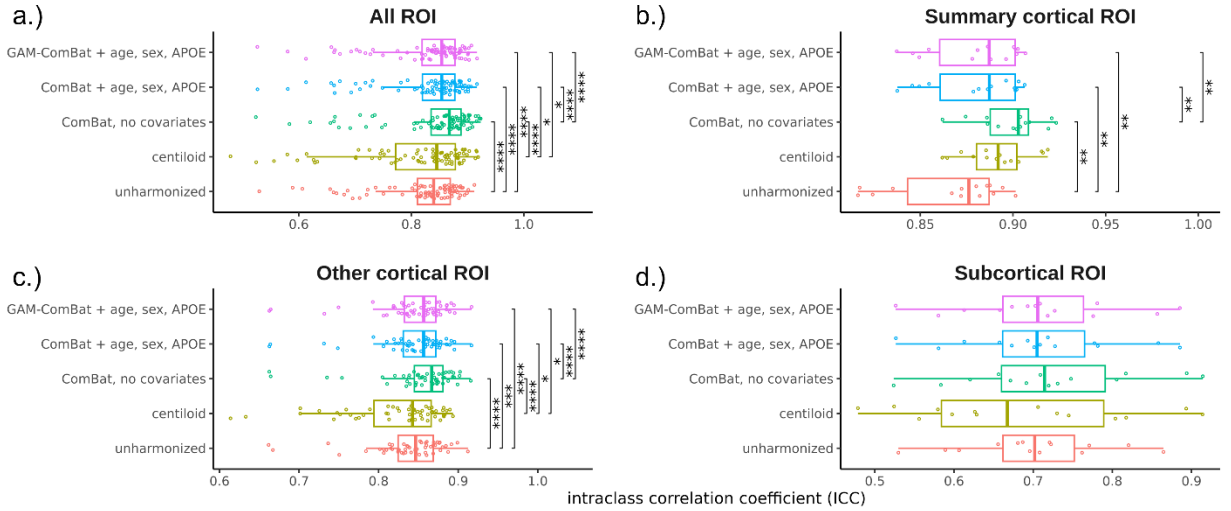

Fig. 3 Distribution of regional ICC across all ROI, grouped by harmonization method and ROI subgroup. Each point represents a single ROI. Significance levels from paired t-tests with Bonferroni correction are indicated for each pair of harmonization methods (* = p < 0.05, ** = p < 0.01, *** = p < 0.005, **** = p < 1e-4)

FBP and PiB groups within the training set (p < 1e-4), and between the training and head-to-head cohorts in FBP only (p < 0.01). CDR status also differed between training and head-to-head cohorts in both FBP and PiB (p < 0.05). In the simulation cohort, sex and APOE-ε4 were different between FBP and PiB groups (p < 0.05).

Tracer head-to-head comparison

We evaluated the ability of Centiloid and ComBat to improve the absolute agreement between FBP and PiB using the head-to-head dataset. For the global summary region, absolute agreement, as measured by ICC, increased after harmonization with either Centiloid ($ICC =$

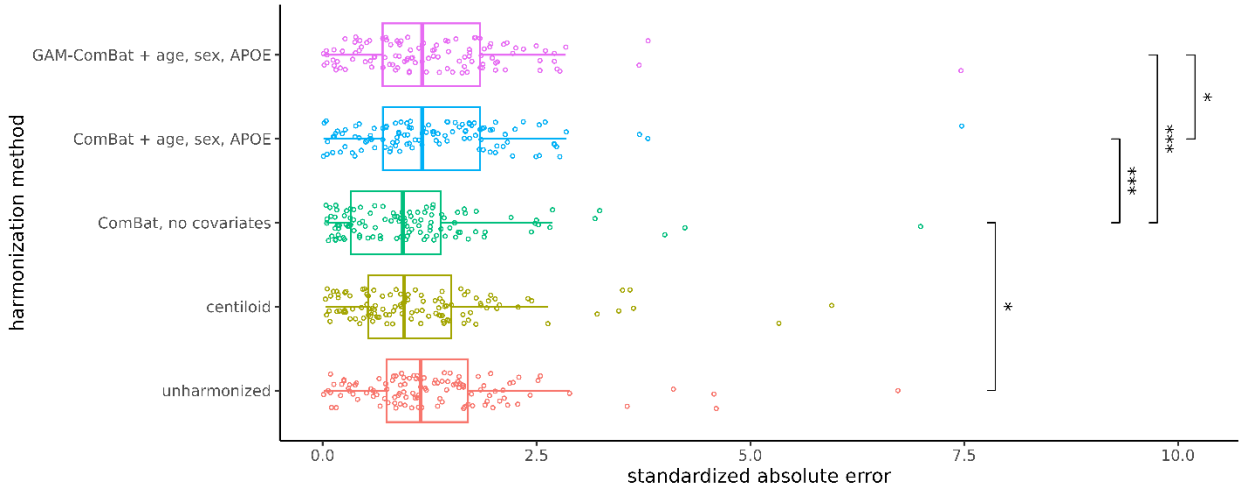

Fig. 4 Standardized absolute error of the global summary measure from each PiB and FBP scan pair in the tracer head-to-head dataset. Each point represents a single scan pair. Significance levels from paired t-tests with Bonferroni correction are indicated for each pair of harmonization methods (* = p < 0.05, ** = p < 0.01, *** = p < 0.005, **** = p < 1e-4)

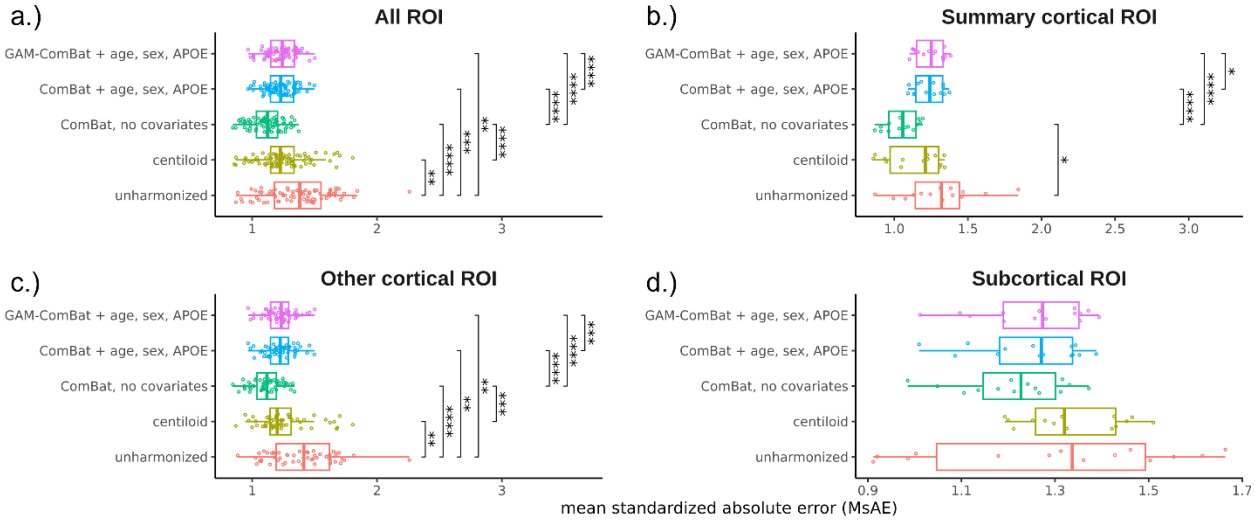

Fig. 5 Distribution of regional MsAE across all ROI, grouped by harmonization method and ROI subgroup. Each point represents a single ROI. Significance levels from paired t-tests with Bonferroni correction are indicated for each pair of harmonization methods (* = p < 0.05, ** = p < 0.01, *** = p < 0.005, **** = p < 1e-4)

0.912) or ComBat with no covariates ($ICC = 0.916$), compared to the unharmonized SUVR ($ICC = 0.882$) (Table 2, Fig. 2). ICC also increased slightly after harmonization with ComBat with covariates and GAM-ComBat ($ICC = 0.898$), albeit not to the same degree as Centiloid or ComBat with no covariates. For ROI measurements, all three ComBat harmonization methods led to a statistically significant increase in average ICC among all ROIs compared to unharmonized SUVR

(p < 1e-4), with ComBat with no covariates again performing the best ($\overline{ICC} = 0.838$) (Table 2, Fig. 3a). Additionally, ComBat with no covariates performed the best within the summary cortical ROIs ($\overline{ICC} = 0.899$) and other cortical ROIs ($\overline{ICC} = 0.853$) (Fig. 3b-c). No method was effective at improving across-tracer agreement for the subcortical ROIs (Fig. 3d). When focusing on each ROI individually, Centiloid resulted in a decrease in ICC in the bilateral occipital and sensorimotor regions, and in the left temporal and parietal cortices (Supp. Fig. 1b). In contrast, none of the ComBat variants led to such decrease in these regions (Supp. Fig. 1c-e).

These trends were consistent when assessing the standardized absolute error between FBP and PiB measurements of the global summary region (Table 2, Fig. 4). ComBat with no covariates was the only method that reduced the mean error with statistical significance ($MsAE = 1.07$, p < 0.05) compared to unharmonized SUVRs ($MsAE = 1.335$). For ROI measurements, all methods significantly reduced the average MsAE among all ROIs (p < 0.01), but ComBat with no covariates resulted in the greatest reduction ($\overline{MsAE} = 1.12$) (Table 2, Fig. 5a). ComBat with no covariates also performed the best within the summary cortical ROI ($\overline{MsAE} = 1.048$) and other cortical ROIs

Table 2 Mean ± standard deviation of ICC and MsAE from the tracer head-to-head comparison. The mean across subjects was computed for the global summary region, whereas the mean across ROIs was computed for the FreeSurfer ROI. Bold values indicate the best performing harmonization method. Asterisks indicate statistical significance of paired t-tests comparing each harmonization method with unharmonized (* = p < 0.05, ** = p < 0.01, *** = p < 0.005, **** = p < 1e-4). Note that the standard deviation of standardized absolute error (second column) is always one by design (see Equations $\Delta_i = \text{abs}(PiB_i - FBP_i)$ (5 through $\sigma_{\Delta_i} = \sqrt{\frac{\sum_{i=1}^{N}(\Delta_i - \bar{\Delta})}{N-1}}$ (7)

			All ROI		Summary cortical ROI		Other cortical ROI		Subcortical ROI	
Harmonization method	ICC	MsAE	ICC	MsAE	ICC	MsAE	ICC	MsAE	ICC	MsAE
unharmonized	0.882	1.335 ± 1	0.819 ± 0.077	1.369 ± 0.26	0.866 ± 0.03	1.308 ± 0.255	0.838 ± 0.048	1.403 ± 0.259	0.701 ± 0.089	1.299 ± 0.26
centiloid	0.912	1.181 ± 1	0.811 ± 0.098	1.249 ± 0.196**	0.891 ± 0.018	1.139 ± 0.181	0.822 ± 0.063	1.254 ± 0.206**	0.686 ± 0.137	1.341 ± 0.108
ComBat, no covariates	**0.916**	**1.07** ± 1*	**0.838** ± 0.083****	**1.12** ± 0.123****	**0.899** ± 0.018**	**1.048** ± 0.11*	**0.853** ± 0.05****	**1.115** ± 0.116****	**0.72** ± 0.111	**1.212** ± 0.112
ComBat + age, sex, APOE	0.898	1.325 ± 1	0.827 ± 0.08****	1.229 ± 0.112***	0.881 ± 0.024**	1.233 ± 0.098	0.844 ± 0.049***	1.223 ± 0.116**	0.707 ± 0.099	1.248 ± 0.115
GAM-ComBat + age, sex, APOE	0.898	1.326 ± 1	0.827 ± 0.08****	1.234 ± 0.113**	0.881 ± 0.024**	1.241 ± 0.096	0.844 ± 0.049****	1.227 ± 0.117**	0.707 ± 0.099	1.255 ± 0.118

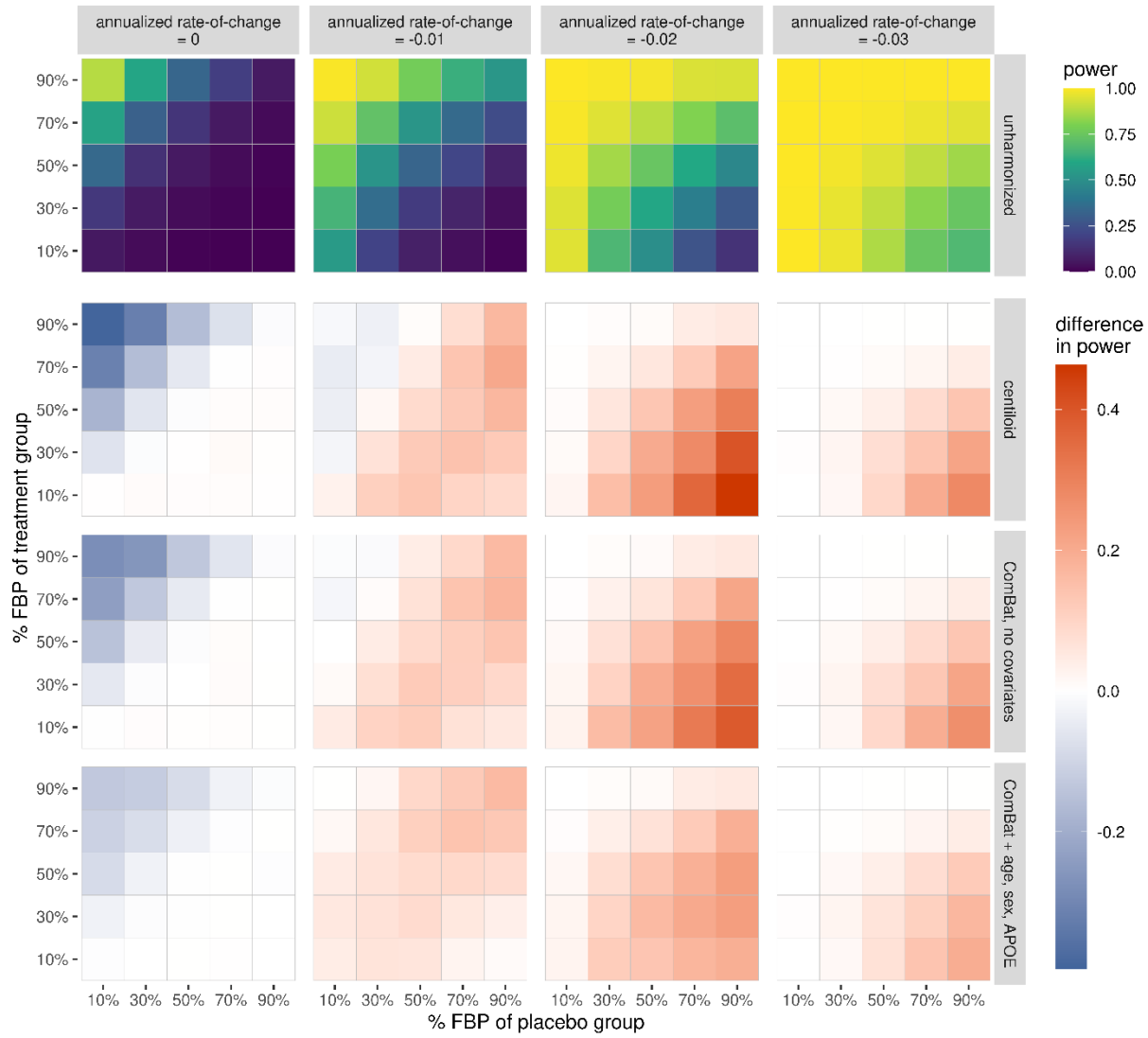

Fig. 6 Statistical power of detecting group differences in rate-of-change of the global summary SUVR between treatment and placebo groups, computed as the proportion of significant findings over 1000 iterations. Power is plotted for unharmonized SUVR, while difference in power relative to unharmonized is plotted for all harmonization methods. The true underlying rate-of-change is varied across columns. The proportion of FBP scans in the placebo and treatment groups are varied across the horizontal and vertical axes, respectively. Note that for annualized rate-of-change equal to zero, the proportion of significant findings corresponds to Type-I error rate

($\overline{MsAE} = 1.115$) (Fig. 5b-c). Again, no method was able to significantly reduce error in the subcortical regions (Fig. 5d).

Clinical trial simulation

We performed simulations to test for group differences in amyloid rate-of-change between treatment and placebo groups in a hypothetical clinical trial, and evaluated whether harmonization improved the ability of detecting these differences. For the global summary SUVR, both Centiloid and ComBat resulted in overall increases in statistical power after harmonization in the presence of a treatment effect (i.e. for rate-of-change $\in \{-0.01, -0.02, -0.03\}$), primarily when the placebo group had high FBP composition and the treatment group had low FBP composition (Fig. 6). The greatest benefit of harmonization was observed for rate-of-change = -0.02. In the absence of a treatment effect (i.e. rate-of-change $= 0$), Centiloid and ComBat achieved decreases in Type-I error, primarily when the placebo group had low FBP composition and the treatment group had high FBP composition. Similar patterns of change in power were observed when using ComBat with covariates, albeit to a lesser magnitude. For regional SUVRs, surface plots of the statistical power reveal similar patterns of change in power for most regions (Supp. Fig. 3-4).

Discussion

We demonstrated that ComBat may effectively harmonize amyloid PET measurements across FBP and PiB. Notably, ComBat with no covariates outperformed Centiloid in increasing absolute agreement between tracers in both the global summary and regional measurements, and resulted in a comparable improvement in detecting group differences in the simulated clinical trial. As more studies shift focus from using a global summary metric of amyloid burden to using the spatial distribution of regional amyloid as features [6,7], harmonization techniques like ComBat that can be applied to multiple regions become appealing for pooling PET data across multiple tracers.

ComBat poses several methodological advantages over Centiloid. Firstly, whereas calibration of Centiloid requires *a priori* selection of representative individuals from healthy control and typical

AD cohorts, training ComBat requires no such step. This *a priori* cohort selection may introduce bias into the calibration process. Especially if the selected sample is small and/or captures only a subset of the overall population (e.g. biased towards a single ethnicity group), Centiloid may not generalize well to heterogeneous or out-of-sample datasets. ComBat circumvents this requirement, which allows it to learn a robust harmonization on a potentially more varied dataset which consists of controls, AD patients and “in-between” subjects. Furthermore, much like Centiloid, a trained ComBat model may then be used to harmonize out-of-sample data. Secondly, Centiloid requires at least two PET scans of different tracers for each subject in the calibration cohort, one of which should be acquired using PiB. In contrast, ComBat can train using just one scan per subject and does not require PiB to be used. Thirdly, a region-specific harmonization is important for addressing sources of tracer bias which variably affect different regions, such as non-specific binding [33]. However, as suggested by Klunk et al. [18], a region-specific Centiloid calibration is not ideal, since it would fix different SUVRs of different regions to the same Centiloid value. In contrast, ComBat independently removes the tracer-specific variance from the target measurements without scaling the dynamic range of each region to fixed points, making it a more suitable technique for regional harmonization.

Lastly, ComBat has the advantage of being able to preserve covariate relationships in the target measurement, which may be useful in downstream analyses such as in predictive models which take into account biologically-related variance to make accurate predictions. However, in the context of purely evaluating the absolute agreement between tracers, we observed that including covariates into the ComBat model led to worse ICC and absolute errors. This may be partially due to differences in covariate distributions between the training and head-to-head cohorts, of which age and CDR differed with statistical significance. Although CDR was not explicitly included as a ComBat covariate, it may have indirectly contributed to a biased ComBat model which does not generalize well to testing data with different covariate characteristics.

It was noted that no harmonization method investigated in this study performed well for the subcortical regions. Notably, these regions lie close to white matter regions, and thus may be affected by non-specific binding more so than cortical regions. This may contribute to more noise in the subcortical regions, which batch harmonization methods such as ComBat are not able to mitigate. One potential area of investigation is to evaluate whether partial volume correction [33] would have an effect on regional harmonization of PET SUVRs, especially for regions which experience high amounts of signal spill-over from neighboring white matter regions.

Our simulation experiments revealed the importance of harmonization in settings where multiple tracers are utilized to track brain amyloid deposition in clinical trial participants. Particularly, harmonization was the most beneficial when trial groups exhibited differing proportions of tracer data. In these cases, tracer biases contributed to a substantial confounding effect across clinical trial groups, resulting in either a reduction of power in detecting the true underlying treatment effect, or an increase in Type-I error in the case when no treatment effect exists. Harmonization effectively served to mitigate these confounding effects due to tracer differences. This was consistent with previous reports that found significant differences in amyloid rates-of-change across different tracers within real clinical trial groups, and that these differences were subsequently removed after harmonization [17]. Little change was observed in power when clinical trial groups exhibited the same proportion of tracer data, likely due to the fact that the same tracer bias would affect both groups equally, which statistically would not influence the detection of group differences.

There are several limitations to the current work. Firstly, on the basis of purely increasing tracer agreement, there are no clear recommendations on the choice of including covariates in ComBat. One caveat to using ComBat is that, unless explicitly accounted for in the covariates, it will assume that any biases due to real biological differences between tracer cohorts are batch differences, which are subsequently removed. Therefore, one should carefully examine the composition of the

data at hand and consider whether it is necessary to model known biological factors via the covariate terms. Secondly, data from the simulation experiment were generated from models trained on a cohort of amyloid-positive subjects from OASIS-3 instead of data from an actual anti-amyloid drug trial. Although simulations were set up to mimic data that would be collected in a successful trial, it remains to be seen whether our hypotheses would hold on real-world clinical trial data. Finally, ComBat assumes that the target features (after residualizing covariate terms) are distributed normally. However, this is often not the case for amyloid imaging data, where the distribution of amyloid burden across subjects often exhibits a bimodal pattern of low amyloid (amyloid-negative) and high amyloid (amyloid-positive) clusters. Recently developed harmonization techniques such as NoDiM [34] and PEACE [29] explicitly model the target as a multi-modal Gaussian mixture. This allows for the batch-specific shifts and scales to be adjusted independently for each cluster. However, neither study focused specifically on regional harmonization of tracer-specific effects. Adapting a cluster-specific harmonization technique for this task is thus an interesting avenue for future investigation, as this framework may be more suitable for amyloid PET imaging data.

Conclusions

Harmonization of amyloid PET radiotracers is imperative for removing tracer-specific biases in amyloid burden measurements for optimal performance of downstream tasks, such as enhancing statistical power and reducing false discoveries in clinical trials. In the current study, we demonstrated that ComBat is effective for harmonizing both global and regional amyloid measurements in an entirely data-driven way. Our experimental results suggest that ComBat not only increases the absolute agreement of measurements made by different tracers, but also provides a significant benefit to the performance of detecting true treatment effects in anti-amyloid

drug trials. ComBat thus presents as a viable technique for harmonizing regional-based analyses of amyloid PET.

List of Abbreviations

PET: positron emission tomography; FBP: [^{18}F]-florbetapir; PiB: [^{11}C]-Pittsburgh Compound-B; OASIS-3: Open Access Series of Imaging Studies 3; SUVR: standardized uptake value ratio; ICC: intraclass correlation coefficient; MsAE: mean standardized absolute error; AD: Alzheimer's disease; ROI: region-of-interest; CL: Centiloid; MRI: magnetic resonance imaging; APOE: apolipoprotein-ε4; GAM: generalized additive model; LME: linear mixed effects; CDR: Clinical Dementia Rating.

Supplementary Information

Additional file 1.docx - Supplementary figures and tables. **Supp. Table 1**: List of FreeSurfer regions-of-interest used in the study, along with their subgroupings. **Supp. Fig. 1**: Regional ICCs plotted on the surface from the tracer head-to-head comparison. **Supp. Fig. 2**: Regional MsAEs plotted on the surface from the tracer head-to-head comparison. **Supp. Fig. 3**: Surface plots of mean statistical power from the simulation experiment. **Supp. Fig. 4**: Subcortical plots of mean statistical power from the simulation experiment. **Supp. Table 2**: Mean statistical power of detecting significant rate-of-change differences between treatment and placebo groups in the simulation experiment for the global summary amyloid estimate. **Supp. Table 3**: Mean statistical power of detecting significant rate-of-change differences between treatment and placebo groups in the simulation experiment for the ROI amyloid measurements.

Declarations

Ethics approval and consent to participate

Ethics approvals were obtained by the OASIS-3 dataset. All participants were consented into Knight ADRC-related projects in accordance with the Declaration of Helsinki and following procedures approved by the Institutional Review Board of Washington University School of Medicine in St. Louis. For more details, we refer the reader to the OASIS-3 reference.

Consent for publication

Not applicable

Availability of data and materials

Data utilized in this study were obtained from the OASIS-3 open access dataset. Data can be requested at https://sites.wustl.edu/oasisbrains.

Code for this study will be made publicly available at https://github.com/sotiraslab. All statistical analyses and simulation experiments were implemented using R version 4.3.2 and python version 3.10.10. The *neuroharmonize* python package (https://github.com/rpomponio/neuroHarmonize) was used to train and apply ComBat and GAM-ComBat models.

Competing Interests

AS reported receiving personal fees from BrightFocus for serving as a grant reviewer and stock from TheraPanacea outside the submitted work. All remaining authors have no conflicting interests to report.

Funding

BY was supported by the Imaging Science Pathways NIH T32 EB014855. AS was supported by NIH award R01 AG067103 and BrightFocus Foundation grant ADR A2021042S.

Computations were performed using the facilities of the Washington University Research Computing and Informatics Facility, which were partially funded by NIH grants S10OD025200, 1S10RR022984-01A1 and 1S10OD018091-01. Additional support is provided by The McDonnell Center for Systems Neuroscience.

Authors' contributions

All authors contributed to the conceptualization and design of the study. BY implemented all data analyses and experiments and wrote the first draft of the manuscript. AS, BG and TB contributed to the interpretation of data. TE, SK and DK provided technical support. All authors were involved with manuscript revision, and all approved of the final draft.

Acknowledgements

Acknowledgement is made to the donors of the ADR A2021042S, a program of the BrightFocus Foundation, for support of this research. Data were provided by OASIS-3: Longitudinal Multimodal Neuroimaging (Principal Investigators: T. Benzinger, D. Marcus, J. Morris). OASIS-3 was supported by the following funding sources: NIH P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

References

1. Chapleau M, Iaccarino L, Soleimani-Meigooni D, Rabinovici GD. The Role of Amyloid PET in Imaging Neurodegenerative Disorders: A Review. J Nucl Med. 2022;63:13S-19S.

2. Swanson CJ, Zhang Y, Dhadda S, Wang J, Kaplow J, Lai RYK, et al. A randomized, double-blind, phase 2b proof-of-concept clinical trial in early Alzheimer's disease with lecanemab, an anti-Aβ protofibril antibody. Alz Res Therapy. 2021;13:80.

3. Shcherbinin S, Evans CD, Lu M, Andersen SW, Pontecorvo MJ, Willis BA, et al. Association of Amyloid Reduction After Donanemab Treatment With Tau Pathology and Clinical Outcomes: The TRAILBLAZER-ALZ Randomized Clinical Trial. JAMA Neurology. 2022;79:1015–24.

4. Mathotaarachchi S, Pascoal TA, Shin M, Benedet AL, Kang MS, Beaudry T, et al. Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiology of Aging. 2017;59:80–90.

5. Choi H, Jin KH. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behavioural Brain Research. 2018;344:103–9.

6. Pfeil J, Hoenig MC, Doering E, van Eimeren T, Drzezga A, Bischof GN. Unique regional patterns of amyloid burden predict progression to prodromal and clinical stages of Alzheimer's disease. Neurobiology of Aging. 2021;106:119–29.

7. Pascoal TA, Therriault J, Mathotaarachchi S, Kang MS, Shin M, Benedet AL, et al. Topographical distribution of Aβ predicts progression to dementia in Aβ positive mild cognitive impairment. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2020;12:e12037.

8. Ezzati A, Abdulkadir A, Jack CR, Thompson PM, Harvey DJ, Truelove-Hill M, et al. Predictive value of ATN biomarker profiles in estimating disease progression in Alzheimer's disease dementia. Alzheimer's & Dementia. 2021;17:1855–67.

9. Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Annals of Neurology. 2004;55:306–19.

10. Su Y, Flores S, Wang G, Hornbeck RC, Speidel B, Joseph-Mathurin N, et al. Comparison of Pittsburgh compound B and florbetapir in cross-sectional and longitudinal studies. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2019;11:180–90.

11. Wolk DA, Zhang Z, Boudhar S, Clark CM, Pontecorvo MJ, Arnold SE. Amyloid imaging in Alzheimer's disease: comparison of florbetapir and Pittsburgh compound-B positron emission tomography. J Neurol Neurosurg Psychiatry. 2012;83:923–6.

12. Landau SM, Breault C, Joshi AD, Pontecorvo M, Mathis CA, Jagust WJ, et al. Amyloid-β Imaging with Pittsburgh Compound B and Florbetapir: Comparing Radiotracers and Quantification Methods. Journal of Nuclear Medicine. 2013;54:70–7.

13. Villemagne VL, Mulligan RS, Pejoska S, Ong K, Jones G, O'Keefe G, et al. Comparison of 11C-PiB and 18F-florbetaben for Aβ imaging in ageing and Alzheimer's disease. Eur J Nucl Med Mol Imaging. 2012;39:983–9.

14. Adamczuk K, Schaeverbeke J, Nelissen N, Neyens V, Vandenbulcke M, Goffin K, et al. Amyloid imaging in cognitively normal older adults: comparison between 18F-flutemetamol and 11C-Pittsburgh compound B. Eur J Nucl Med Mol Imaging. 2016;43:142–51.

15. Mountz JM, Laymon CM, Cohen AD, Zhang Z, Price JC, Boudhar S, et al. Comparison of qualitative and quantitative imaging characteristics of [11C]PiB and [18F]flutemetamol in normal control and Alzheimer's subjects. NeuroImage: Clinical. 2015;9:592–8.

16. Landau SM, Thomas BA, Thurfjell L, Schmidt M, Margolin R, Mintun M, et al. Amyloid PET imaging in Alzheimer's disease: a comparison of three radiotracers. Eur J Nucl Med Mol Imaging. 2014;41:1398–407.

17. Chen CD, McCullough A, Gordon B, Joseph-Mathurin N, Flores S, McKay NS, et al. Longitudinal head-to-head comparison of 11C-PiB and 18F-florbetapir PET in a Phase 2/3 clinical trial of anti-amyloid-β monoclonal antibodies in dominantly inherited Alzheimer's disease. Eur J Nucl Med Mol Imaging. 2023;50:2669–82.

18. Klunk WE, Koeppe RA, Price JC, Benzinger TL, Devous Sr. MD, Jagust WJ, et al. The Centiloid Project: Standardizing quantitative amyloid plaque estimation by PET. Alzheimer's & Dementia. 2015;11:1-15.e4.

19. Pegueroles J, Montal V, Bejanin A, Vilaplana E, Aranha M, Santos-Santos MA, et al. AMYQ: An index to standardize quantitative amyloid load across PET tracers. Alzheimer's & Dementia. 2021;17:1499–508.

20. Bourgeat P, Doré V, Doecke J, Ames D, Masters CL, Rowe CC, et al. Non-negative matrix factorisation improves Centiloid robustness in longitudinal studies. NeuroImage. 2021;226:117593.

21. Chen K, Ghisays V, Luo J, Chen Y, Lee W, Wu T, et al. Harmonizing florbetapir and PiB PET measurements of cortical Aβ plaque burden using multiple regions-of-interest and machine learning techniques: An alternative to the Centiloid approach. Alzheimer's & Dementia. 2024;20:2165–72.

22. Liu H, Nai Y-H, Saridin F, Tanaka T, O' Doherty J, Hilal S, et al. Improved amyloid burden quantification with nonspecific estimates using deep learning. Eur J Nucl Med Mol Imaging. 2021;48:1842–53.

23. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.

512 24. Richter S, Winzeck S, Correia MM, Kornaropoulos EN, Manktelow A, Outtrim J, et al.
513 Validation of cross-sectional and longitudinal ComBat harmonization methods for magnetic
514 resonance imaging data on a travelling subject cohort. Neuroimage: Reports. 2022;2:100136.

515 25. Sun D, Rakesh G, Haswell CC, Logue M, Baird CL, O'Leary EN, et al. A comparison of
516 methods to harmonize cortical thickness measurements across scanners and sites.
517 NeuroImage. 2022;261:119509.

518 26. Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, et al. Harmonization
519 of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan.
520 NeuroImage. 2020;208:116450.

521 27. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of
522 cortical thickness measurements across scanners and sites. NeuroImage. 2018;167:104–20.

523 28. Leithner D, Schöder H, Haug A, Vargas HA, Gibbs P, Häggström I, et al. Impact of ComBat
524 Harmonization on PET Radiomics-Based Tissue Classification: A Dual-Center PET/MRI and
525 PET/CT Study. Journal of Nuclear Medicine. 2022;63:1611–6.

526 29. Bilgel M. Probabilistic estimation for across-batch compatibility enhancement for amyloid
527 PET. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2023;15:e12436.

528 30. LaMontagne PJ, Benzinger TLS, Morris JC, Keefe S, Hornbeck R, Xiong C, et al. OASIS-3:
529 Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer
530 Disease. medRxiv. 2019;2019.12.13.19014902.

531 31. Su Y, D'Angelo GM, Vlassenko AG, Zhou G, Snyder AZ, Marcus DS, et al. Quantitative
532 Analysis of PiB-PET with FreeSurfer ROIs. PLOS ONE. 2013;8:e73377.

32. Su Y, Flores S, Hornbeck RC, Speidel B, Vlassenko AG, Gordon BA, et al. Utilizing the Centiloid scale in cross-sectional and longitudinal PiB PET studies. NeuroImage: Clinical. 2018;19:406–16.

33. Su Y, Blazey TM, Snyder AZ, Raichle ME, Marcus DS, Ances BM, et al. Partial volume correction in quantitative amyloid imaging. NeuroImage. 2015;107:55–64.

34. Properzi MJ, Buckley RF, Chhatwal JP, Donohue MC, Lois C, Mormino EC, et al. Nonlinear Distributional Mapping (NoDiM) for harmonization across amyloid-PET radiotracers. NeuroImage. 2019;186:446–54.