Title

Comparison of approaches to control for intracranial volume in research on the association of brain volumes with cognitive outcomes

Journal

Human Brain Mapping, 45(4)

ISSN

1065-9471

Authors

Wang, Jingxuan
Hill-Jarrett, Tanisha
Buto, Peter
et al.

Publication Date

2024-03-01

DOI

10.1002/hbm.26633

Peer reviewed

Received: 21 August 2023 | Revised: 30 January 2024 | Accepted: 7 February 2024

DOI: 10.1002/hbm.26633

RESEARCH ARTICLE

WILEY

Comparison of approaches to control for intracranial volume in research on the association of brain volumes with cognitive outcomes

Jingxuan Wang[1,2] | Tanisha Hill-Jarrett[2] | Peter Buto[1,2] | Annie Pederson[1,2] | Kendra D. Sims[1,2] | Scott C. Zimmerman[1] | Michelle A. DeVost[1] | Erin Ferguson[1] | Benjamin Lacar[3] | Yulin Yang[1] | Minhyuk Choi[1] | Michelle R. Caunca[4] | Renaud La Joie[4] | Ruijia Chen[1] | M. Maria Glymour[2] | Sarah F. Ackley[2]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, California, USA

[2]Department of Epidemiology, Boston University, Boston, Massachusetts, USA

[3]Bakar Computational Health Sciences Institute, University of California, San Francisco, California, USA

[4]Memory and Aging Center, Department of Neurology, University of California, San Francisco, California, USA

Correspondence
Sarah F. Ackley, Department of Epidemiology, Boston University, Boston, MA, USA.
Email: sackley@bu.edu

Funding information
National Institute on Aging, Grant/Award Numbers: K00AG068431, K99AG073454, T32AG049663, F99AG083306, T32AG078115, F31AG079651, K99AG083121, P30-AG062422, R01AG057869, K99AG065501

Abstract

Most neuroimaging studies linking regional brain volumes with cognition correct for total intracranial volume (ICV), but methods used for this correction differ across studies. It is unknown whether different ICV correction methods yield consistent results. Using a brain-wide association approach in the MRI substudy of UK Biobank ($N = 41{,}964$; mean age = 64.5 years), we used regression models to estimate the associations of 58 regional brain volumetric measures with eight cognitive outcomes, comparing no correction and four ICV correction approaches. Approaches evaluated included: no correction; dividing regional volumes by ICV (proportional approach); including ICV as a covariate in the regression (adjustment approach); and regressing the regional volumes against ICV in different normative samples and using calculated residuals to determine associations (residual approach). We used Spearman-rank correlations and two consistency measures to quantify the extent to which associations were inconsistent across ICV correction approaches for each possible brain region and cognitive outcome pair across 2320 regression models. When the association between brain volume and cognitive performance was close to null, all approaches produced similar estimates close to the null. When associations between a regional volume and cognitive test were not null, the adjustment and residual approaches typically produced similar estimates, but these estimates were inconsistent with results from the crude and proportional approaches. For example, when using the crude

approach, an increase of 0.114 (95% confidence interval [CI]: 0.103–0.125) in fluid intelligence was associated with each unit increase in hippocampal volume. However, when using the adjustment approach, the increase was 0.055 (95% CI: 0.043–0.068), while the proportional approach showed a decrease of −0.025 (95% CI: −0.035 to −0.014). Different commonly used methods to correct for ICV yielded inconsistent results. The proportional method diverges notably from other methods and results were sometimes biologically implausible. A simple regression adjustment for ICV produced biologically plausible associations.

KEYWORDS
brain-wide association studies, cognitive aging, intracranial volume correction, reproducibility, structural magnetic resonance imaging

1 | INTRODUCTION

Structural magnetic resonance imaging (MRI or MR imaging) is widely used to evaluate the relationship between brain volumetric measures and cognitive measures, including memory, attention, and executive function (Aggarwal et al., 2010; Scheltens et al., 2002). In studies examining associations between brain volumetric measures and cognition in aging populations, correcting volumetric measures for intracranial volume (ICV) is often necessary to account for differences in skull size (Barnes et al., 2010; Voevodskaya et al., 2014). This is because skull size is not an independent predictor of dementia (Edland et al., 2002) and is associated with numerous childhood and adulthood socioeconomic status factors that may confound associations between neuroimaging measures and outcomes (Hackman & Farah, 2009). Several different statistical approaches to account for ICV have been adopted in the field. These different approaches may produce inconsistent estimates, but, to date, there is no clear guidance on which approaches are preferable. This presents significant challenges for reproducibility in neuroimaging studies, may account for inconsistent results across studies, and likely contributes to incorrect estimates in some studies. Associations between volumetric measures and treatment or outcomes are used in a wide range of contexts, including drug trials (Chertkow & Black, 2007), and may be used to inform biological understanding of disease, research priorities, and interventions (Veitch et al., 2019). Thus, analysis decisions on ICV correction may ultimately impact individual clinical diagnosis and management (Opfer et al., 2022).

The proportional, adjustment, and residual approaches are the three commonly used approaches to correct for ICV (Box 1). For the proportional approach, each brain volume measure is divided by ICV and this scaled quantity is used to determine associations with a cognitive measure, typically using a regression model (O'Brien et al., 2006, 2011; Voevodskaya et al., 2014). For the adjustment approach, ICV is included along with the brain volume measure as an independent variable in a regression model with cognition as the dependent variable (O'Brien et al., 2011; Voevodskaya et al., 2014). The residual approach uses two regression models, first regressing each regional volume against ICV and calculating the residuals from this model (i.e., variation in brain volume not predicted by ICV), and then using the calculated residuals as the independent variable in a regression with cognitive measures as the predictor. Typically, but not always, the coefficients for the first regression are estimated in a normative sample of healthy controls (O'Brien et al., 2011; Sanfilipo

BOX 1 Summary of commonly used approaches used to correct for intracranial volume (ICV)

Approach	Description
Crude, no correction for ICV	Crude volumes are used in determining associations with cognitive outcomes. ICV is not controlled for in either calculating correlations or performing regressions.
Proportional	Brain volume measures are divided by ICV, and this scaled quantity is used to determine associations with a cognitive measure, typically using a regression model.
Adjustment	ICV is included along with brain volume measures as an independent variable in a regression model with cognition as the dependent variable.
Residual	The residual approach uses two regression models, first regressing regional volumes against ICV and calculating the residuals from this model, and then using the calculated residuals as the independent variable in a regression with cognitive measures as the predictor. Typically, coefficients for the first regression are estimated in a normative sample of healthy controls.

et al., 2004; Voevodskaya et al., 2014). Specific implementations of these methods vary, and sometimes crude volumes are used without correcting for ICV (Van Horn et al., 2014).

Evidence on how different correction approaches modify the estimated associations between MRI volumetric measures and cognition is limited. Some prior studies examine how other associations with MRI volumetric measures, such as sex, gender, and age (Dhamala et al., 2022; Kijonka et al., 2020; Sanchis-Segura et al., 2020), vary with ICV correction strategy. The small number of prior studies evaluating how different ICV correction strategies affect associations between volumetric measures and cognition have important limitations: they do not evaluate all commonly used ICV correction approaches; they use a single or small number of cognitive measures; they are performed in younger samples or small samples; or they are performed in highly select volunteer cohorts (e.g., the Alzheimer's Disease Neuroimaging Initiative cohort) such that results may not generalize to the general aging population (Dhamala et al., 2022; Kijonka et al., 2020; Sanchis-Segura et al., 2020; Voevodskaya et al., 2014).

In this study, we compared estimated associations between MRI volumetric measures and cognitive measures across commonly used ICV correction approaches in the MRI subsample of the UK Biobank. The UK Biobank is a cohort of middle-aged and older adults participating in the National Health Service. As such, Alzheimer's disease and vascular dementia would be expected to be the most common causes of brain atrophy (Brunnström et al., 2009; Goodman et al., 2017). Inspired by brain-wide association studies (BWAS), which evaluate each pairwise association of brain region and outcome (Marek et al., 2022), we evaluated the extent to which different ICV correction approaches give inconsistent associations between a brain volume and cognitive measure for the full-factorial combination of ICV correction approach, brain volumetric measure, and cognitive outcome. Inconsistency across ICV correction approaches may have important implications for reproducibility in neuroimaging research. These inconsistencies may account for conflicting findings across studies and produce spurious associations that would not reach statistical significance with alternative approaches.

2 | MATERIALS AND METHODS

2.1 | Study population

The UK Biobank is a large prospective cohort study of 502,490 UK adults aged 40–69 years at recruitment in 2006–2010. At the baseline visit, participants completed social, physical, and medical assessments. Starting in 2014, participants were invited for MRI neuroimaging at four clinics using identical protocols (Littlejohns et al., 2020). UK Biobank invited eligible participants for second MRIs starting in 2019, although we did not include repeat MRIs in our analysis (Supplementary Material 1). The final target sample for the MRI substudy is 100,000 but at the time of writing, MR imaging data were available for 41,964 participants. Ethical approval was obtained by the UKB study from the National Health Service National Research Ethics Service with all participants providing written informed consent. Analyses were approved by the University of California, San Francisco Institutional Review Board under UK Biobank Resource project no. 74748.

2.2 | MRI volumetric measures

All MRI data were obtained from identical scanners (3T Siemens Skyra, software VD13) equipped with a standard Siemens 32-channel head coil. All image preprocessing was conducted by the UK Biobank neuroimaging team and included non-brain removal, bias-field correction, and tissue segmentation. The total scanning time was 31 min with five additional minutes for subject adjustment, shimming, and so on. Imaging acquisition included T1, resting fMRI, task fMRI, T2 FLAIR, dMRI, and swMR. For T1-weighted scans, 3D MPRAGE protocol with a cubic millimeter isotropic resolution was used. The acquisition may be repeated if significant artifacts were detected when scanning, such as excessive head movement. An automated quality control pipeline was developed to detect issues and quantify data quality. An image processing pipeline based on FSL (the FMRIB Software Library, version 5.0.10) and FreeSurfer (version 6.0) was applied to generate processed data. When processing T1 images with FreeSurfer, T2 FLAIR is utilized in combination with T1 images for enhanced accuracy in cortical modeling, if available. We used imaging data processed by FreeSurfer. Specifically, we included subcortical volumes extracted from the FreeSurfer automatic subcortical segmentation (ASEG) tool and cortical volumes based on the Desikan–Killianny–Tourville (DKT) atlas. This included a total of 27 subcortical and 31 cortical regions (Supplementary Material 2). ICV was also estimated from ASEG. In our primary analysis, we a priori selected eight regions of interest (ROIs) previously linked to cognitive outcomes (Apostolova & Thompson, 2008; Feng et al., 2018; Henneman et al., 2009; Juottonen et al., 1998; Ries et al., 2008). Only data from the first MRI visit was used due to limited follow-up. We combined hemispheres to obtain a single measure for each ROI. Full details on image acquisition, processing, and quality control are available in the UK Biobank Brain Imaging Documentation (Alfaro-Almagro et al., 2018).

2.3 | Cognitive measures

We considered eight cognitive measures: fluid intelligence, numeric memory, prospective memory, pairs matching, Trail Making A, Trail Making B, reaction time, and symbol digit substitution. Detailed descriptions of the eight cognitive tests are included in Supplementary Material 3. All cognitive tests were administered in English via touchscreen interface and designed to be completed without supervision. All cognitive outcomes were measured at the MRI visit. Cognitive scores for pairs matching, Trail Making A, Trail Making B, and reaction time referred to negative one times the score on those cognitive tests; that is, these measures were signed such that higher values indicate better cognition. All continuous cognitive outcomes were z-standardized to produce standardized regression coefficients. Note that very few

individuals had a dementia diagnosis at the MRI visit ($n = 27$) or were diagnosed with incident dementia in follow-up ($n = 44$).

2.4 | ICV correction approaches

We considered no correction and four implementations of widely used approaches to correct for ICV in analyses of associations between regional brain volumes and cognition (in all models, the cognitive measure is the outcome variable of interest) (O'Brien et al., 2006; O'Brien et al., 2011; Voevodskaya et al., 2014). The first approach uses crude, uncorrected volumes, while the other four are ICV correction approaches. The approaches are as follows: (1) *Crude approach*: Crude volumes are used without correcting for ICV to determine associations with cognitive measures in a regression model. (2) *Proportional approach*: Volumetric measures are divided by ICV and this ratio of regional volume to ICV is used to determine associations in a regression model with cognition as the outcome. (3) *Adjustment approach*: Crude volumes are used to determine associations in a regression model that adjusts for ICV as a covariate. (4) *Full-sample residual approach*: Each volumetric measure is first regressed against ICV in the full sample of participants. This regression model is then used to obtain ICV-corrected volumes in the whole sample using the following: $\text{Volume}_{\text{corrected},i} = \text{Volume}_{\text{crude},i} - \widehat{\beta} \times \text{ICV}_i$, where $\widehat{\beta}$ is the slope from the regression model. ICV-corrected volumes (i.e., residuals from the first regression model) are then used as the independent variable in a regression model to determine associations with cognitive measures. (5) *Normative-subsample residual approach*: This approach is identical to the full-sample residual approach, but the first regression is restricted to a "normative" sample of dementia-free participants younger than 60 years.

2.5 | Covariates

All models were adjusted for age, age squared, sex (female and male), race (White, Asian, Black, and other), APOE-ε4 alleles (0, 1, and 2), education (A-levels or above and less than A-levels), and assessment center (Reading, Cheadle, Newcastle, and Bristol) as a proxy for geographic location. Age and age-squared were adjusted for as orthogonal polynomials; the rest of the covariates were adjusted for as categorical variables. At recruitment, age and sex information were obtained from a central registry and subsequently updated by participants. Participants self-reported their race and education through a touchscreen questionnaire at baseline. The number of APOE-ε4 alleles (0, 1, or 2) was determined using the single nucleotide polymorphisms rs7412 and rs429358 (Bekris et al., 2010).

2.6 | Statistical analyses

We summarized demographic characteristics of the subsamples with MRI data who completed each of the following cognitive outcomes: fluid intelligence, numeric memory, prospective memory, pairs matching, Trail Making A, Trail Making B, reaction time, and symbol digit substitution. We calculated Spearman-rank correlation coefficients between ICV and each cognitive outcome. To evaluate the associations between regional volumes and cognitive outcomes, we used linear regression for all cognitive outcomes except for prospective memory, for which we used logistic regression (1 for correct on the first attempt and 0 otherwise). All models were adjusted for age, age squared, sex, race, the number of APOE-ε4 alleles, education, and assessment center as a proxy for geographic location. To facilitate comparison of effect sizes across ICV correction approaches, ROIs, and cognitive outcomes, we *z*-standardized all corrected and crude regional volumes, ICV, and continuous cognitive outcomes by subtracting the sample mean and dividing by the sample standard deviation; this produces standardized regression coefficients.

The factorial combination of brain volume measure, cognitive outcome, and ICV correction approaches leads to 2320 distinct estimates. We evaluated the extent to which different ICV correction approaches yield consistent estimates when applied to the same brain volume ROI measure and cognitive outcome by (1) generating pairwise comparisons for correction approaches across regional volumes and cognitive outcomes; (2) summarizing derived statistics from pairwise comparisons; and (3) creating Manhattan plots to show how rates of statistical significance $\alpha = 0.05$ of associations vary across correction approaches. Consistency of estimates using different ICV corrections was evaluated using Spearman-rank correlation coefficients for each possible pair of ICV correction methods and across all ROI and cognitive outcome combinations. We additionally generated two measures for the consistency of findings using alternative ICV correction methods: For the first measure, we calculated the proportion of pairs of correction methods without significantly different estimates across all ROIs using a conservative *z*-test (DeGroot & Schervish, 2012). For the second measure of consistency, we calculated the proportion of pairs of correction approaches for which associations had consistent signs across all brain regions assessed. Specifically, associations were only considered to have opposite signs if both associations were statistically significantly different from the null.

For the residual approach, we additionally evaluated the impact of varying age thresholds to define the normative sample in associations between the eight selected regional volumes linked to dementia and all cognitive outcomes. Specifically, we defined the normative sample using different age cutoffs (<60, <65, <70, <75, and <80) and assessed whether the analytical choice of cutoffs affected estimated associations.

We conducted four sensitivity analyses: First, to assess whether the estimates of the associations with cognitive outcomes are affected by observations at the extreme ends of the regional volume distributions, we additionally fit models excluding participants with regional volumes beyond extreme percentiles (1st and 99th; Supplementary Material 4). Second, we repeated analyses adjusting only for age and age squared to assess whether results were affected by the adjustment set chosen. Third, we additionally examined incident dementia as an outcome (Supplementary Material 5). Finally, we controlled for total brain volume *in lieu* of ICV.

TABLE 1 Characteristics of analytic sample for each cognitive outcome.

	Fluid intelligence	Numeric memory	Prospective memory	Pairs matching	Trail making A	Trail making B	Reaction time	Symbol digit substitution
	N = 38,607	*N* = 28,793	*N* = 39,330	*N* = 39,349	*N* = 28,861	*N* = 27,417	*N* = 39,100	*N* = 28,150
Female, *N* (%)	20,218 (52.4)	15,073 (52.3)	20,579 (52.3)	20,588 (52.3)	14,692 (52.3)	14,315 (52.2)	20,457 (52.3)	14,731 (52.3)
Age at assessment (years, SD)	64.2 (7.6)	64.9 (7.6)	64.3 (7.6)	64.3 (7.6)	64.9 (7.6)	64.7 (7.6)	64.2 (7.6)	64.9 (7.6)
Race, *N* (%)								
White	37,462 (97.0)	27,939 (97.0)	38,129 (96.9)	38,148 (96.9)	27,272 (97.0)	26,609 (97.1)	37,908 (97.0)	27,313 (97.0)
Asian	527 (1.4)	395 (1.4)	560 (1.4)	560 (1.4)	379 (1.3)	373 (1.4)	554 (1.4)	383 (1.4)
Black	331 (0.9)	247 (0.9)	341 (0.9)	341 (0.9)	243 (0.9)	237 (0.9)	339 (0.9)	245 (0.9)
Other	287 (0.7)	212 (0.7)	300 (0.8)	300 (0.8)	210 (0.7)	198 (0.7)	299 (0.8)	209 (0.7)
Count of APOE-ε4 alleles, *N* (%)								
0	27,951 (72.4)	20,829 (72.3)	28,452 (72.3)	28,466 (72.3)	20,322 (72.3)	19,830 (72.3)	28,277 (72.3)	20,364 (72.3)
1	9802 (25.4)	7321 (25.4)	10,007 (25.4)	10,012 (25.4)	7159 (25.5)	6978 (25.5)	9958 (25.5)	7164 (25.4)
2	854 (2.2)	643 (2.2)	871 (2.2)	871 (2.2)	623 (2.2)	609 (2.2)	865 (2.2)	622 (2.2)
Education, *N* (%)								
A-levels or above	31,149 (80.7)	23,403 (81.3)	31,589 (80.3)	31,601 (80.3)	22,807 (81.2)	22,345 (81.5)	31,417 (80.4)	22,875 (81.3)
Less than A-levels	7458 (19.3)	5390 (18.7)	7741 (19.7)	7748 (19.7)	5297 (18.8)	5072 (18.5)	7683 (19.6)	5275 (18.7)
Assessment center, *N* (%)								
Reading	5788 (15.0)	5794 (20.1)	5837 (14.8)	5837 (14.8)	5794 (20.6)	5676 (20.7)	5820 (14.9)	5795 (20.6)
Cheadle	22,843 (59.2)	13,038 (45.3)	23,450 (59.6)	23,469 (59.6)	12,355 (44.0)	12,057 (44.0)	23,267 (59.5)	12,357 (43.9)
Newcastle	9928 (25.7)	9914 (34.4)	9995 (25.4)	9995 (25.4)	9907 (35.3)	9636 (35.1)	9965 (25.5)	9951 (35.3)
Bristol	48 (0.1)	47 (0.2)	48 (0.1)	48 (0.1)	48 (0.2)	48 (0.2)	48 (0.1)	47 (0.2)

3 | RESULTS

Sample sizes across cognitive outcomes ranged from 27,147 to 39,349 (Table 1) with very similar demographic compositions. Slightly over half the participants were female, average age ranged from 64.2 to 64.9 years, nearly all participants were White (ranging from 96.9% to 97.1%), and a majority reported A-level or above education (ranging from 80.7% to 81.5%). Very few individuals had a dementia diagnosis at the MRI visit ($n = 27$) or were diagnosed with incident dementia in follow-up ($n = 44$). Correlations between cognitive outcomes and ICV were low, ranging from 0.019 for signed pairs matching to 0.167 for fluid intelligence (Supplementary Material 6).

Associations of regional brain volumes with cognitive scores are shown in Figures 1 and S1. The associations varied in both sign and magnitude across ICV correction approaches. Crude estimates were typically farthest from the null, and proportional approaches were typically closest to the null, but sometimes with reversed sign compared with other approaches. For example, when using the crude approach, an increase of 0.114 (95% confidence interval [CI]: 0.103–0.125) in fluid intelligence was associated with each unit increase in hippocampal volume. However, when using the adjustment approach, the increase was 0.055 (95% CI: 0.043–0.068), while the proportional approach showed a decrease of −0.025 (95% CI: −0.035 to −0.014).

Figure 2 shows comparisons across ICV correction approaches for the associations of all brain regions assessed with fluid intelligence, numeric memory, and Trail Making A and B. Specifically, it shows pairwise scatterplots of estimated associations, as well as Spearman-rank correlations and both measures of consistency for the 58 brain regions evaluated. Estimates when ICV corrections are based on the adjustment approach are highly correlated ($\rho \geq 0.89$) with coefficients from the residual approach. These two approaches both show only weak correlations with the proportional approach ($\rho = 0.28$ for the adjustment approach and $\rho = 0.25$ for the residual approach, both for fluid intelligence). Different ICV correction approaches sometimes produce both qualitatively different results: the proportional approach was most likely to produce an association with a reversed sign that

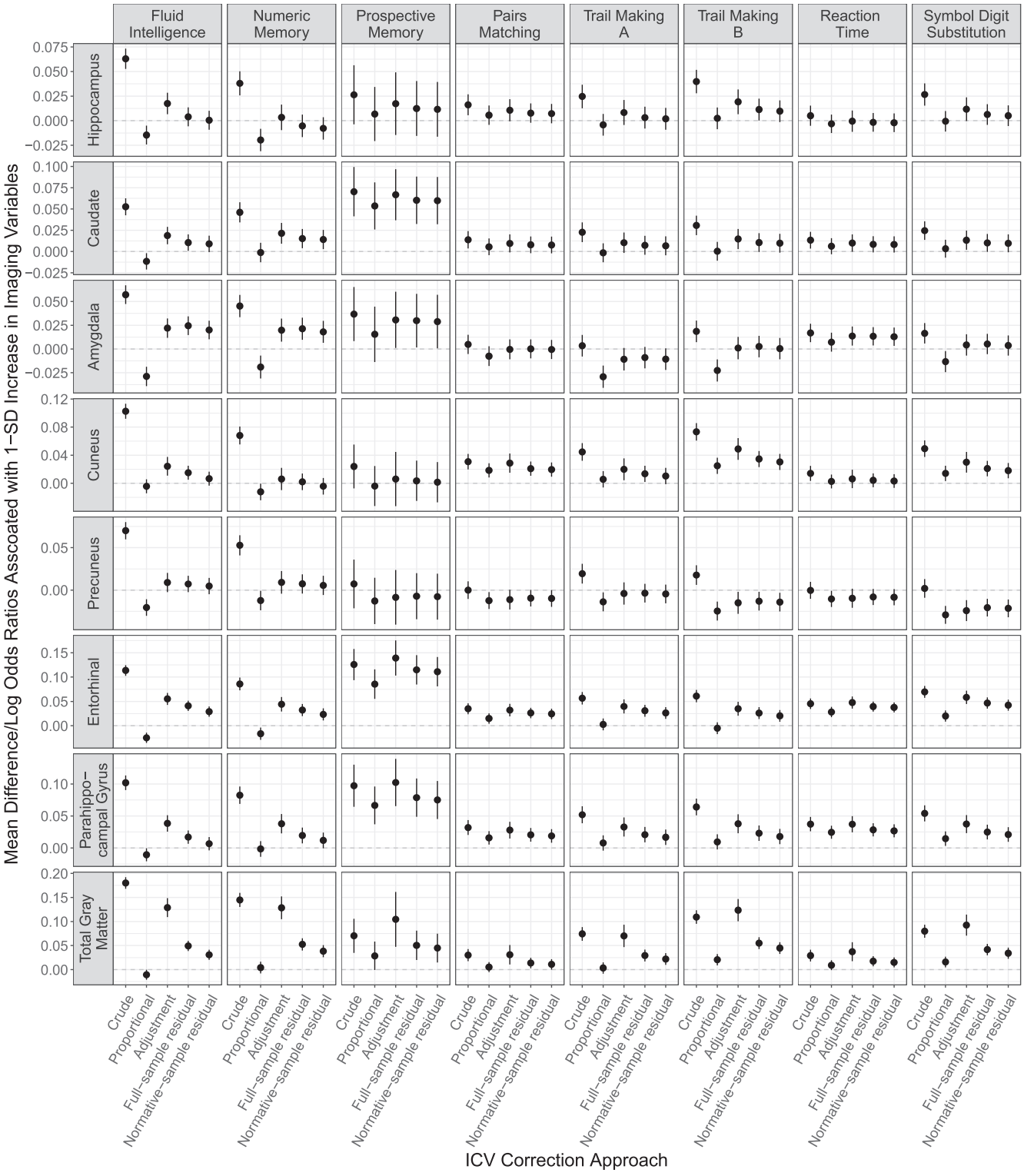

FIGURE 1 Associations of volumes of the cuneus, entorhinal cortex, parahippocampal gyrus, precuneus, caudate nucleus, hippocampus, amygdala, and total grey matter with each cognitive outcome. Cognitive scores are signed such that higher values indicate better cognition. Specifically, cognitive scores for pairs matching, Trail Making A, Trail Making B, and reaction time refer to negative one times the score on those cognitive tests. Models adjusted for age, age squared, sex, race, the number of APOE-ε4 alleles, education, and assessment center.

reaches statistical significance. Even when point estimates are not of the opposite side of the null, estimates may be statistically significantly inconsistent. Figure S2 is the same plot for the remaining cognitive outcomes.

Discrepancies between ICV correction approaches were prominent for fluid intelligence, numeric memory, and Trail Making A and B. Numeric memory and Trail Making A and B are often used in clinical settings as indicators of simple attention/working

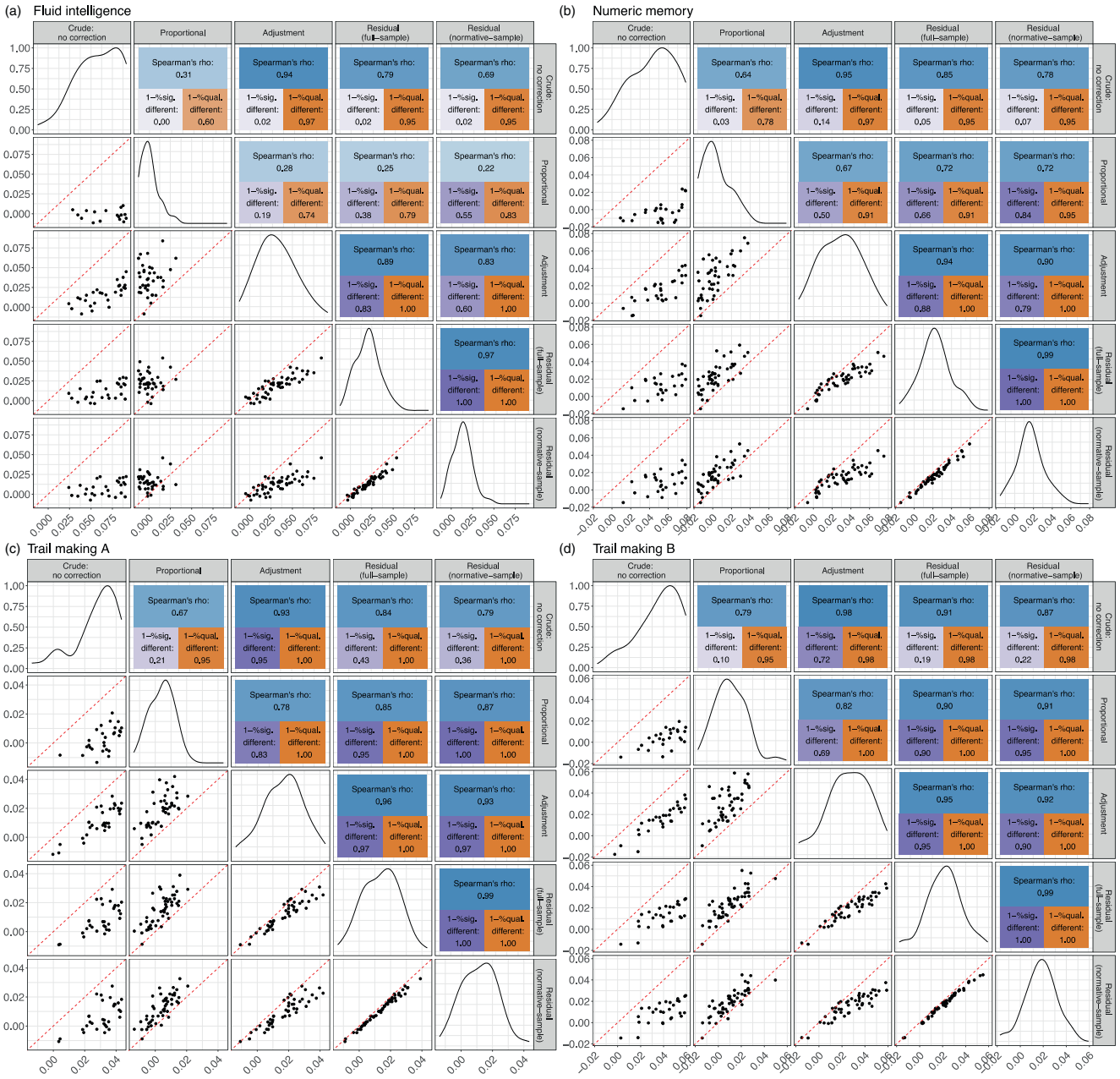

FIGURE 2 Comparison of coefficients based on alternative ICV corrections for estimated associations of 58 brain region volume measures with (a) fluid intelligence, (b) numeric memory, (c) Trail Making A, and (d) Trail Making B. Pairwise comparisons for estimated associations between volumetric measures and fluid intelligence among six total intracranial volume (ICV) correction approaches. Diagonal panels are smoothed density estimates for the distribution of associations across all 58 brain regions assessed under the corresponding ICV correction approach. Lower diagonal panels are scatterplots for the associations under each pair of ICV correction approaches across all 58 brain regions assessed. Upper diagonal panels show the correlation and two consistency measures between each pair. Spearman's rho (range from −1 to 1) stands for Spearman-rank correlation coefficient; 1 – %sig. different (range from 0 to 1) is one minus the proportion of pairs with significantly different estimates across all ROIs using a conservative *z*-test; and 1 – %qual. different (range from 0 to 1) is the one minus proportion of pairs with significantly opposite signs. Color intensity represents the value, with fully saturated colors indicating a 1.

memory and executive function, respectively, are known to be affected in more advanced stages of Alzheimer's disease and are prominent areas of cognition affected in other forms of dementia (e.g., frontotemporal dementia) (Ashendorf et al., 2008; Castel et al., 2009; Peña-Casanova et al., 2012). Associations with prospective memory were less appreciably affected by the ICV correction approach, but this may reflect a single limited assessment task in the UK Biobank.

Figure 3 shows brain-wide associations in a Manhattan plot (Ehret, 2010) for all cognitive outcomes and across all brain regions evaluated, comparing ICV correction approaches. Which and how many associations are statistically significant varies with ICV

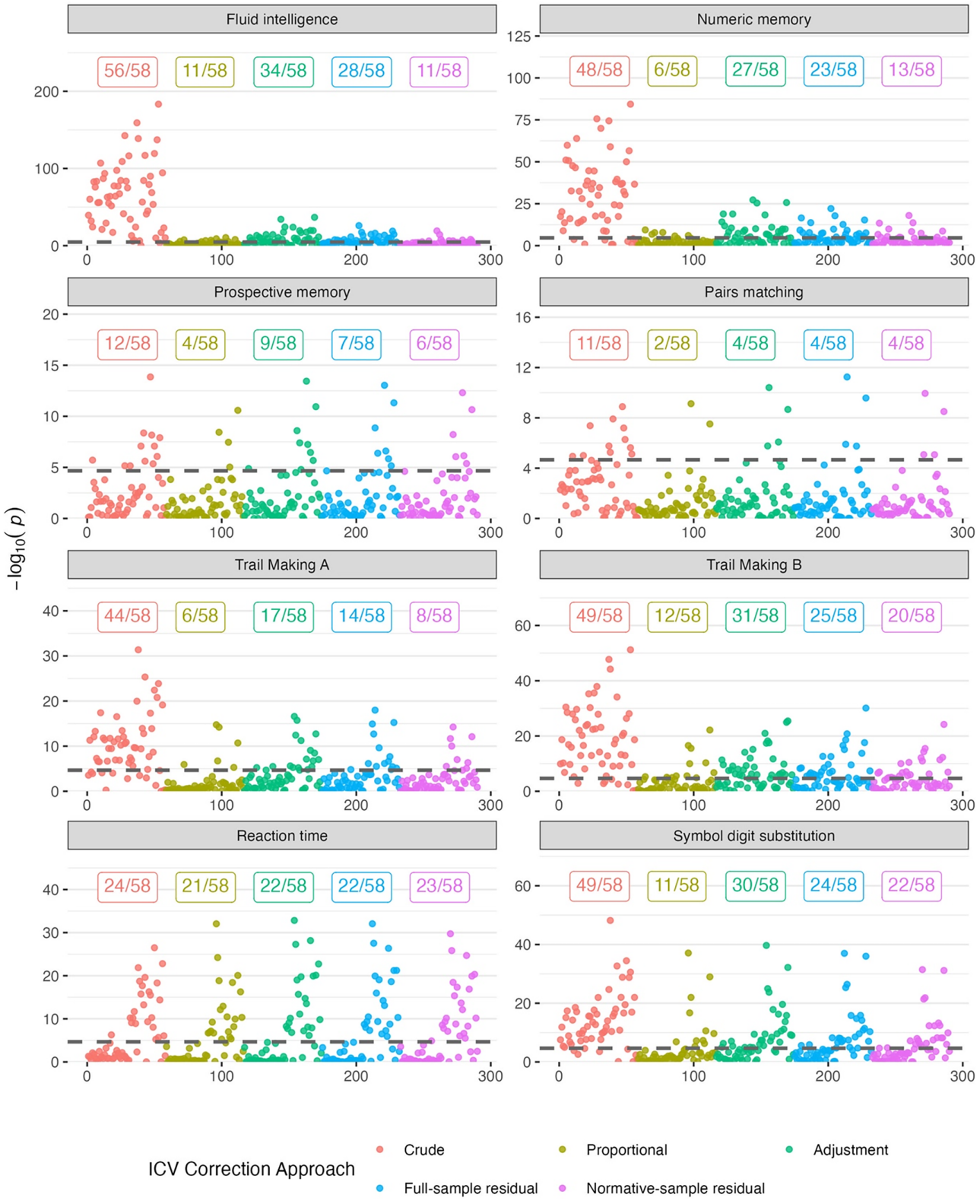

FIGURE 3 Manhattan plot for the association of cognitive outcomes with 58 regional brain volumes, comparing ICV correction approaches. Dashed lines indicate the Bonferroni-corrected threshold for brain-wide significance ($p < 0.05/(58*8*5)=2.2*10\text{-}\hat{}5$). Number annotations indicate the number of brain-wide significant regional associations across all 58 brain regions assessed.

correction approach and cognitive outcome assessed. The crude approach produces the smallest *p*-values, consistent with the premise that failing to correct for ICV inflates associations and statistical significance. Figures S3 and S4 give the same Manhattan plot but use different Bonferroni-corrected statistical significance thresholds ($\alpha = 0.01/(58*8*5)$ and $\alpha = 0.001/(58*8*5)$).

The full-sample and normative-sample (age <60) residual approaches produced highly correlated associations (Figure 2), with comparable *p*-values (Figure 3). However, estimated associations can vary, as shown in Figure 1, with estimates with the normative-sample residual approach typically attenuated relative to those of approach the full-sample residual approach. Figure 4 extends this, showing how using an increasingly older, and presumably less healthy sample with more age-associated neurodegeneration, delivers larger estimated associations between hippocampal volume and fluid intelligence. Consistent with results in Figure 1, Figure 4 shows an increasingly younger sample produces an attenuated association between selected regions and fluid intelligence, numeric memory, and Trail Making B. This highlights that the definition of the normative sample can affect estimated associations. In sensitivity analyses, we find that results are comparable excluding outliers in ICV and with different adjustment sets (see Figures S5–S8). In addition, we did not observe a similar inconsistency in time-to-event analyses (Figure S9), but the number of incident dementia cases was small. Figures S10 and S11 show similar inconsistecies between correction approaches as the main analysis using total brain volume *in lieu* of intracranial volume.

4 | DISCUSSION

We evaluated whether estimated associations between regional brain volumetric measures and cognitive outcomes differed when using crude volumes and four approaches to correct for ICV using the large UK Biobank MRI subsample. Although estimates based on the adjustment and residual approaches were similar, estimates based on the proportional and crude approaches were inconsistent. Inconsistencies were largest when estimates from the adjustment and residual approaches were further from the null.

The proportional, adjustment, and residual approaches are all commonly used strategies (O'Brien et al., 2006) to correct for ICV, and crude estimates are frequently presented as well (Van Horn et al., 2014). However, we found that the proportional approach frequently produced estimates that were inconsistent with our understanding of neurobiology and even had the opposite sign of estimates derived with adjustment and residual approaches. As an example, the hippocampus plays a vital role in encoding and consolidation of new memories (Reber, 2013). Numerous studies have demonstrated an association between hippocampal atrophy and poorer neuropsychological test performance, particularly with regard to memory tasks (Golomb et al., 1993; Peng et al., 2015; Rusinek et al., 2003; Schuff et al., 2009). This relationship is further corroborated by existing experimental studies and clinical case studies involving direct damage to the hippocampus (Fortin et al., n.d.; Clark et al., 2005; Dickerson & Eichenbaum, 2010). However, estimates for the association between hippocampal volume and cognitive outcomes were inconsistent across ICV correction approaches: the proportional approach indicated that larger hippocampal volume was associated with worse cognition, which contravenes extant understanding of neurobiology and neurodegenerative diseases (see Figure 1) (Hardcastle et al., 2020; Peng et al., 2015). While more often consistent, even adjustment and residual approaches do not always produce consistent estimated associations, particularly since the associations produced with the residual approach depended on the normative sample adopted (see Figure 4). Specifically, we find using increasingly younger normative samples attenuated estimated associations between hippocampal volume and fluid intelligence.

These results have important implications for reproducibility in neuroimaging studies, our understanding of disease biology, and intervention evaluation. We recognize that these ICV correction methods represent different biological constructs and that method employed may depend on the nature of the research question. For example, in some studies, we may not want to account for ICV to account for perinatal and childhood factors that influence cranium size and regional volumes. However, findings from different studies may diverge merely because of the selected ICV correction approach—ostensibly a minor statistical decision. Estimated associations from crude volumes tended to be further from the null than findings after ICV correction, suggesting that ICV captures confounding by lifetime peak brain size or childhood growth. If atrophy-related neurodegeneration is the construct we intend to capture, ICV correction will often be necessary. Moreover, it is important to note that ICV correction is also crucial for accurately identifying and interpreting volumetric measures that are associated with cognitive functioning or dementia symptoms (Edland et al., 2002). Our findings suggest the proportional approach may be misleading and, if used, should be interpreted in conjunction with results from other approaches. The adjustment and residual approaches tend to produce comparable estimates and it may be reasonable to favor these two approaches. However, these two approaches represent different principles. The adjustment approach evaluates the relationship between regional volumes and cognition, keeping ICV constant. In contrast, the two-step residual approach removes the influence of ICV on regional volumes by regressing regional volumes on ICV and obtaining the residuals. This adjustment and residual approaches are analogous to Type III sums of squares in ANOVA which simultaneously consider all variables, and the sequential approach of Type I sums of squares, respectively. Thus, selecting between adjustment and residual approaches may depend on the research question and context. In addition, estimated associations with residual approaches can vary with reference sample used (Figure 4). Even more concerning, as the age of the normative sample is decreased, biologically plausible effects are increasingly attenuated, approaching the null as the sample becomes younger and healthier. In addition, the residual approach, when not applied to a separate normative sample, ideally would include a standard error correction to account for the fact that the two-stages of estimation are performed on overlapping samples. Thus, if ICV is to be corrected for, we would tend to favor adjustment over the other methods since it is the simplest to implement while producing biologically plausible associations.

Our results are consistent with small prior studies indicating that ICV correction approach can affect study results. Previous studies have examined whether associations of sex and age with volumetric measures persisted across ICV correction approaches (Kijonka

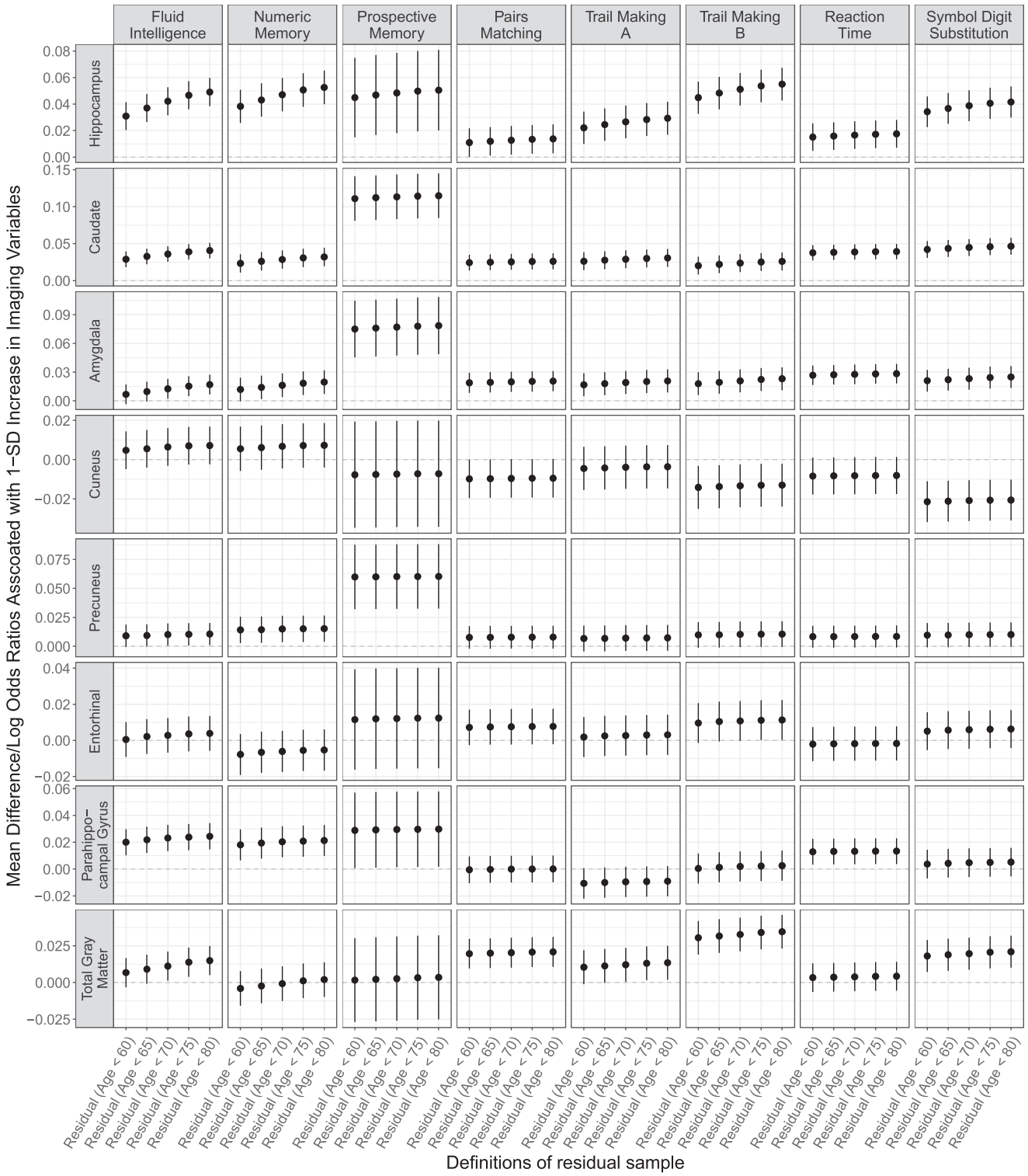

FIGURE 4 Effect of the definition of residual sample on associations between volumes of the cuneus, entorhinal cortex, parahippocampal gyrus, precuneus, caudate nucleus, hippocampus, amygdala, and total gray matter and all cognitive outcomes.

et al., 2020; Nordenskjöld et al., 2015; Sanchis-Segura et al., 2020). Prior work has also examined associations between brain volumetric measures and cognition (Voevodskaya et al., 2014), finding that ICV correction approach altered associations, with a flipped direction of association for the proportional method. This study was performed in two smaller and select samples ($N = 406$ and 724) and examined only the first item of word recall from the ADAS-Cog. The small samples have left substantial uncertainty in whether their conclusions hold for larger and less select samples. For example, this previous work found that the association between hippocampal volumes and cognition was

not statistically significantly different across ICV correction approaches, possibly due to imprecise estimates. Our sample, with 25–35 times as many participants as the previous study and a much broader range of volumetric measures and cognitive outcomes assessed (Voevodskaya et al., 2014), provides far more conclusive findings on the importance of ICV correction approach.

Our study has several strengths in addition to the large sample: these include a comprehensive evaluation of ICV correction approaches and a wide range of cognitive outcomes evaluated. Employing a BWAS-inspired approach, we examined a large combination of associations between cognitive tests and regional brain volumes for a total of 2320 regressions with several measures of consistency to further support the robustness of our findings. Our study has several limitations. First, our results only pertained to cross-sectional evaluation of MRI volumetric measures and cognitive outcomes. Other neuroimaging measures—including cortical thicknesses, diffusion tensor imaging, and longitudinal change in volumetric measures (Schwarz et al., 2016)—may warrant further investigation (Barnes et al., 2010). UK Biobank participants are known to be healthier on average than the general UK population, have higher socioeconomic status, and are predominately White (Fry et al., 2017). Patterns may differ in more diverse populations (Mukadam et al., 2022). Second, we did not evaluate alternative ICV estimation methods (Buckner et al., 2004), and we did not examine less commonly used methods (e.g., weighted ICV matching (Dhamala et al., 2022)). However, due to potential limitations of the FreeSurfer-ASEG ICV-estimation approach, we performed a sensitivity analysis using total brain volume, and results are comparable. Finally, it has been previously suggested that the proportional approach may be biased precisely because of how the error in the ratio behaves (Sanfilipo et al., 2004), which would necessarily be non-differential with respect to the quantity itself due to less error in the ratio for larger volumes. However, we did not consider how measurement error could lead to or exacerbate the biases in estimated associations, and this is an area that merits further exploration.

5 | CONCLUSION

In conclusion, different ICV correction approaches can produce substantively different estimated associations between MRI-derived measures of brain volume and cognitive outcomes. These differences are largest when the associations are large. The proportional approach is most likely to produce estimates that are inconsistent with adjustment or residual control approaches and biologically implausible. Residual and adjustment approaches are more plausible but since they may produce different results, results based on only one of the approaches should be interpreted with caution.

ACKNOWLEDGMENTS

JW, THJ, KDS, MAD, EF, RLJ, RC, MMG, and SFA receive research support from the NIH. This project was supported by NIH/NIA [F99AG083306, T32AG078115, K99AG083121, F31AG079651, T32AG049663, K99AG065501, P30AG062422, K00AG068431, R01AG057869, and K99AG073454].

FUNDING INFORMATION

This work was supported by the National Institutes of Health (NIH), the National Institute on Aging (K00AG068431, T32AG049663, F99AG083306, T32AG078115, F31AG079651, K99AG083121, P30AG062422, K99AG065501, R01AG057869, and K99AG073454). This content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NIA.

CONFLICT OF INTEREST STATEMENT

Benjamin Lacar was supported by Innovate for Health Data Science Fellowship from Johnson & Johnson and is currently employed by Seer, Inc.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from UK Biobank. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access with the permission of the UK Biobank.

ORCID

Jingxuan Wang https://orcid.org/0000-0002-9387-0480

REFERENCES

Aggarwal, N. T., Wilson, R. S., Bienias, J. L., de Jager, P. L., Bennett, D. A., Evans, D. A., & DeCarli, C. (2010). The association of magnetic resonance imaging measures with cognitive function in a biracial population sample. *Archives of Neurology*, *67*, 475–482.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., … Smith, S. M. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *NeuroImage*, *166*, 400–424.

Apostolova, L. G., & Thompson, P. M. (2008). Mapping progressive brain structural changes in early Alzheimer's disease and mild cognitive impairment. *Neuropsychologia*, *46*, 1597–1612.

Ashendorf, L., Jefferson, A. L., O'Connor, M. K., Chaisson, C., Green, R. C., & Stern, R. A. (2008). Trail making test errors in normal aging, mild cognitive impairment, and dementia. *Archives of Clinical Neuropsychology*, *23*, 129–137.

Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M. D., Lehmann, M., Hobbs, N., Clarkson, M. J., MacManus, D. G., Ourselin, S., & Fox, N. C. (2010). Head size, age and gender adjustment in MRI studies: A necessary nuisance? *NeuroImage*, *53*, 1244–1255.

Bekris, L. M., Yu, C.-E., Bird, T. D., & Tsuang, D. W. (2010). Review article: Genetics of Alzheimer disease. *Journal of Geriatric Psychiatry and Neurology*, *23*, 213–227.

Brunnström, H., Gustafson, L., Passant, U., & Englund, E. (2009). Prevalence of dementia subtypes: A 30-year retrospective survey of neuropathological reports. *Archives of Gerontology and Geriatrics*, *49*, 146–149.

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and validation against manual measurement of total intracranial volume. *NeuroImage*, *23*, 724–738.

Castel, A. D., Balota, D. A., & McCabe, D. P. (2009). Memory efficiency and the strategic control of attention at encoding: Impairments of value-directed remembering in Alzheimer's disease. *Neuropsychology*, *23*, 297–306.

Chertkow, H., & Black, S. (2007). Imaging biomarkers and their role in dementia clinical trials. *Canadian Journal of Neurological Sciences*, *34*, S77–S83.

Clark, R. E., Broadbent, N. J., & Squire, L. R. (2005). Impaired remote spatial memory after hippocampal lesions despite extensive training beginning early in life. *Hippocampus*, *15*(3), 340–346.

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics*. Pearson Education.

Dhamala, E., Ooi, L. Q. R., Chen, J., Kong, R., Anderson, K. M., Chin, R., Yeo, B. T. T., & Holmes, A. J. (2022). Proportional intracranial volume correction differentially biases behavioral predictions across neuroanatomical features and populations. *Neuroimage*, *260*, 119485. https://doi.org/10.1101/2022.03.15.483970

Dickerson, B. C., & Eichenbaum, H. (2010). The episodic memory system: Neurocircuitry and disorders. *Neuropsychopharmacology*, *35*, 86–104.

Edland, S. D., Xu, Y., Plevak, M., O'Brien, P., Tangalos, E. G., Petersen, R. C., & Jack, C. R. (2002). Total intracranial volume: Normative values and lack of association with Alzheimer's disease. *Neurology*, *59*, 272–274.

Ehret, G. B. (2010). Genome-wide association studies: Contribution of genomics to understanding blood pressure and essential hypertension. *Current Hypertension Reports*, *12*, 17–25.

Feng, F., Wang, P., Zhao, K., Zhou, B., Yao, H., Meng, Q., Wang, L., Zhang, Z., Ding, Y., Wang, L., An, N., Zhang, X., & Liu, Y. (2018). Radiomic features of hippocampal subregions in Alzheimer's disease and amnestic mild cognitive impairment. *Frontiers in Aging Neuroscience*, *10*, 290.

Fortin, N. J., Agster, K. L., & Eichenbaum, H. B. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, *5*, 458–462.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., & Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *American Journal of Epidemiology*, *186*, 1026–1034.

Golomb, J., de Leon, M. J., Kluger, A., George, A. E., Tarshish, C., & Ferris, S. H. (1993). Hippocampal atrophy in normal aging. An association with recent memory impairment. *Archives of Neurology*, *50*, 967–973.

Goodman, R. A., Lochner, K. A., Thambisetty, M., Wingo, T. S., Posner, S. F., & Ling, S. M. (2017). Prevalence of dementia subtypes in United States Medicare fee-for-service beneficiaries, 2011–2013. *Alzheimer's & Dementia*, *13*, 28–37.

Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Sciences*, *13*, 65–73.

Hardcastle, C., O'Shea, A., Kraft, J. N., Albizu, A., Evangelista, N. D., Hausman, H. K., Boutzoukas, E. M., van Etten, E. J., Bharadwaj, P. K., Song, H., Smith, S. G., Porges, E. C., Dekosky, S., Hishaw, G. A., Wu, S. S., Marsiske, M., Cohen, R., Alexander, G. E., & Woods, A. J. (2020). Contributions of hippocampal volume to cognition in healthy older adults. *Frontiers in Aging Neuroscience*, *12*, 593833.

Henneman, W. J. P., Sluimer, J. D., Barnes, J., van der Flier, W. M., Sluimer, I. C., Fox, N. C., Scheltens, P., Vrenken, H., & Barkhof, F. (2009). Hippocampal atrophy rates in Alzheimer disease. *Neurology*, *72*, 999–1007.

Juottonen, K., Lehtovirta, M., Helisalmi, S., Riekkinen, P. J., & Soininen, H. (1998). Major decrease in the volume of the entorhinal cortex in patients with Alzheimer's disease carrying the apolipoprotein E ε4 allele. *Journal of Neurology, Neurosurgery & Psychiatry*, *65*, 322–327.

Kijonka, M., Borys, D., Psiuk-Maksymowicz, K., Gorczewski, K., Wojcieszek, P., Kossowski, B., Marchewka, A., Swierniak, A., Sokol, M., & Bobek-Billewicz, B. (2020). Whole brain and cranial size adjustments in volumetric brain analyses of sex- and age-related trends. *Frontiers in Neuroscience*, *14*, 278.

Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J. D., Boultwood, C., Collins, R., Conroy, M. C., Crabtree, N., Doherty, N., Frangi, A. F., Harvey, N. C., Leeson, P., Miller, K. L., Neubauer, S., Petersen, S. E., Sellors, J., … Allen, N. E. (2020). The UK biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions. *Nature Communications*, *11*, 2624.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*, 654–660.

Mukadam, N., Marston, L., Lewis, G., & Livingston, G. (2022). Risk factors, ethnicity and dementia: A UK biobank prospective cohort study of White, South Asian and Black participants. *PLoS One*, *17*, e0275309.

Nordenskjöld, R., Malmberg, F., Larsson, E. M., Simmons, A., Ahlström, H., Johansson, L., & Kullberg, J. (2015). Intracranial volume normalization methods: Considerations when investigating gender differences in regional brain volume. *Psychiatry Research: Neuroimaging*, *231*, 227–235.

O'Brien, L. M., Ziegler, D. A., Deutsch, C. K., Frazier, J. A., Herbert, M. R., & Locascio, J. J. (2011). Statistical adjustments for brain size in volumetric neuroimaging studies: Some practical implications in methods. *Psychiatry Research: Neuroimaging*, *193*, 113–122.

O'Brien, L. M., Ziegler, D. A., Deutsch, C. K., Kennedy, D. N., Goldstein, J. M., Seidman, L. J., Hodge, S., Makris, N., Caviness, V., Frazier, J. A., & Herbert, M. R. (2006). Adjustment for whole brain and cranial size in volumetric brain studies: A review of common adjustment factors and statistical methods. *Harvard Review of Psychiatry*, *14*, 141–151.

Opfer, R., Krüger, J., Spies, L., Kitzler, H. H., Schippling, S., & Buchert, R. (2022). Single-subject analysis of regional brain volumetric measures can be strongly influenced by the method for head size adjustment. *Neuroradiology*, *64*, 2001–2009.

Peña-Casanova, J., Sánchez-Benavides, G., de Sola, S., Manero-Borrás, R. M., & Casals-Coll, M. (2012). Neuropsychology of Alzheimer's disease. *Archives of Medical Research*, *43*, 686–693.

Peng, G.-P., Feng, Z., He, F. P., Chen, Z. Q., Liu, X. Y., Liu, P., & Luo, B. Y. (2015). Correlation of hippocampal volume and cognitive performances in patients with either mild cognitive impairment or Alzheimer's disease. *CNS Neuroscience & Therapeutics*, *21*, 15–22.

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*, 2026–2042.

Ries, M. L., Carlsson, C. M., Rowley, H. A., Sager, M. A., Gleason, C. E., Asthana, S., & Johnson, S. C. (2008). Magnetic resonance imaging characterization of brain structure and function in mild cognitive impairment: A review. *Journal of the American Geriatrics Society*, *56*, 920–934.

Rusinek, H., de Santi, S., Frid, D., Tsui, W. H., Tarshish, C. Y., Convit, A., & de Leon, M. J. (2003). Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal MR imaging study of normal aging. *Radiology*, *229*, 691–696.

Sanchis-Segura, C., Ibañez-Gual, M. V., Aguirre, N., Cruz-Gómez, Á. J., & Forn, C. (2020). Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. *Scientific Reports*, *10*, 12953.

Sanfilipo, M. P., Benedict, R. H. B., Zivadinov, R., & Bakshi, R. (2004). Correction for intracranial volume in analysis of whole brain atrophy in multiple sclerosis: The proportion vs. residual method. *NeuroImage*, *22*, 1732–1743.

Scheltens, P., Fox, N., Barkhof, F., & De Carli, C. (2002). Structural magnetic resonance imaging in the practical assessment of dementia: Beyond exclusion. *The Lancet Neurology*, *1*, 13–21.

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L. M., Trojanowski, J. Q., Thompson, P. M., Jack, C. R., Weiner, M. W., & Alzheimer's Disease Neuroimaging Initiative. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*, *132*, 1067–1077.

Schwarz, C. G., Gunter, J. L., Wiste, H. J., Przybelski, S. A., Weigand, S. D., Ward, C. P., Senjem, M. L., Vemuri, P., Murray, M. E., Dickson, D. W., Parisi, J. E., Kantarci, K., Weiner, M. W., Petersen, R. C., Jack, C. R., & Alzheimer's Disease Neuroimaging Initiative. (2016). A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. *NeuroImage: Clinical*, *11*, 802–812.

Van Horn, J. D., Bowman, I., Joshi, S. H., & Greer, V. (2014). Graphical neuroimaging informatics: Application to Alzheimer's disease. *Brain Imaging and Behavior*, *8*, 300–310.

Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jr., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Alzheimer's Disease Neuroimaging Initiative. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *15*, 106–152.

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., Wahlund, L. O., Larsson, E. M., Westman, E., & Alzheimer's Disease Neuroimaging Initiative. (2014). The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Frontiers in Aging Neuroscience*, *6*, 264.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang, J., Hill-Jarrett, T., Buto, P., Pederson, A., Sims, K. D., Zimmerman, S. C., DeVost, M. A., Ferguson, E., Lacar, B., Yang, Y., Choi, M., Caunca, M. R., La Joie, R., Chen, R., Glymour, M. M., & Ackley, S. F. (2024). Comparison of approaches to control for intracranial volume in research on the association of brain volumes with cognitive outcomes. *Human Brain Mapping*, *45*(4), e26633. https://doi.org/10.1002/hbm.26633