RESEARCH ARTICLE

A comparison of intracranial volume estimation methods and their cross-sectional and longitudinal associations with age

Stener Nerland[1,2] | Therese S. Stokkan[1,2] | Kjetil N. Jørgensen[2,3] |
Laura A. Wortinger[1,2] | Geneviève Richard[4] | Dani Beck[1,2] |
Dennis van der Meer[5] | Lars T. Westlye[4,6] | Ole A. Andreassen[1,4] |
Ingrid Agartz[1,2,7,8] | Claudia Barth[1,2]

[1]Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway

[2]NORMENT, University of Oslo, Oslo, Norway

[3]Department of Psychiatry, Telemark Hospital, Skien, Norway

[4]NORMENT, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[5]School of Mental Health and Neuroscience, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[6]Department of Psychology, University of Oslo, Oslo, Norway

[7]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

[8]Stockholm Health Care Services, Stockholm Region, Stockholm, Sweden

Correspondence

Stener Nerland and Claudia Barth, Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway.
Email: stener.nerland@medisin.uio.no; claudia.barth@medisin.uio.no

Funding information

Helse Sør-Øst RHF, Grant/Award Numbers: 2017-097, 2019-104, 2020-020; Norges Forskningsråd, Grant/Award Numbers: 213700, 223273, 250358; Stiftelsen Kristian Gerhard Jebsen, Grant/Award Number: SKGJ-MED-008

Abstract

Intracranial volume (ICV) is frequently used in volumetric magnetic resonance imaging (MRI) studies, both as a covariate and as a variable of interest. Findings of associations between ICV and age have varied, potentially due to differences in ICV estimation methods. Here, we compared five commonly used ICV estimation methods and their associations with age. T1-weighted cross-sectional MRI data was included for 651 healthy individuals recruited through the NORMENT Centre (mean age = 46.1 years, range = 12.0–85.8 years) and 2410 healthy individuals recruited through the UK Biobank study (UKB, mean age = 63.2 years, range = 47.0–80.3 years), where longitudinal data was also available. ICV was estimated with FreeSurfer (eTIV and sbTIV), SPM12, CAT12, and FSL. We found overall high correlations across ICV estimation method, with the lowest observed correlations between FSL and eTIV ($r = .87$) and between FSL and CAT12 ($r = .89$). Widespread proportional bias was found, indicating that the agreement between methods varied as a function of head size. Body weight, age, sex, and mean ICV across methods explained the most variance in the differences between ICV estimation methods, indicating possible confounding for some estimation methods. We found both *positive* and *negative* cross-sectional associations with age, depending on dataset and ICV estimation method. Longitudinal ICV reductions were found for all ICV estimation methods, with annual percentage change ranging from −0.293% to −0.416%. This convergence of longitudinal results across ICV estimation methods offers strong evidence for age-related ICV reductions in mid- to late adulthood.

KEYWORDS

aging, brain morphometry, brain segmentation, intracranial volume, magnetic resonance imaging, methods comparison

1 | INTRODUCTION

Intracranial volume (ICV), defined as the volume within the cranium including the brain, meninges, and cerebrospinal fluid (CSF), is an important measure in brain magnetic resonance imaging (MRI) studies. It is frequently used to adjust for individual variation in head size (O'Brien et al., 2011; Voevodskaya et al., 2014) and as a proxy for premorbid brain volume in the study of neurodegenerative diseases (Davis & Wright, 1977). Manual delineation of structural brain MRI scans is considered the most accurate in vivo method for determining ICV (Huo et al., 2016; Klasson et al., 2015; Whitwell et al., 2001). However, this approach is labor intensive and requires training, making it impractical for large datasets. To overcome these limitations, a variety of automated methods for computing ICV using T1-weighted structural MRI have been developed. Previous studies have reported varying consistency and agreement between ICV estimation methods (Malone et al., 2015; Sargolzaei et al., 2015). In one study, associations between hippocampal volume, education and a cognitive measure differed between ICV estimation methods (Nordenskjöld et al., 2013). Some studies have also indicated that the accuracy of automated ICV estimation varies as a function of head size (Klasson et al., 2018). Such findings highlight the importance of assessing ICV estimation methods for potential sources of bias, which may otherwise introduce spurious effects in studies relying on ICV estimates. In the present study, we pursued two related aims: First, we compared ICV estimation methods and assessed potential sources of bias influencing ICV estimation. Secondly, we assessed both cross-sectional and longitudinal associations between age and ICV.

Automated ICV estimation methods typically use T1-weighted MRI images and can be broadly classified as either registration- or segmentation-based. Registration-based methods estimate ICV via an atlas scaling factor given by the determinant of an affine transformation of individual MRI images to a template. The two most common registration-based ICV estimation methods are estimated Total Intracranial Volume (eTIV; Buckner et al., 2004) in FreeSurfer and SIENAX from FSL (FMRIB Software Library; Smith et al., 2002, 2004). With segmentation-based methods, MRI images are first segmented into tissue compartments which are then used to calculate volumetric estimates of the intracranial cavity. One popular segmentation-based method is the Tissue Volumes utility in Statistical Parametric Mapping (SPM; https://www.fil.ion.ucl.ac.uk/spm/). CAT12 is an extension of SPM12 that aims to provide a more robust segmentation algorithm (http://www.neuro.uni-jena.de/cat/). Both SPM12 and CAT12 compute ICV as a probability-weighted sum of grey matter (GM), white matter (WM) and CSF. Sequence Adaptive Multimodal Segmentation (SAMSEG; Puonti et al., 2016) was recently introduced in FreeSurfer version 7 and computes the segmentation-based Total Intracranial Volume (sbTIV).

Throughout the lifespan, ICV increases rapidly from early childhood until early adolescence and is thought to remain relatively stable throughout adulthood (Mills et al., 2016; Pfefferbaum et al., 1994). These findings are consistent with studies on head circumference and computed tomography (Bergerat et al., 2021; Huda et al., 2004;

Neubauer et al., 2009). In adulthood, the presence of ICV change is less well-established, and it remains uncertain if continued changes occur and to what extent. However, past non-MRI studies on ICV and head size have often been limited to childhood and adolescence. A notable exception is the cross-sectional study by Weaver and Christian (1980), reporting no significant association between age and occipitofrontal head circumference in 567 participants (50% female) aged 18 to approximately 67 years.

A number of cross-sectional MRI studies have assessed the association between age and ICV in adulthood, some of which report significant negative associations with age while others find no significant associations. For instance, DeCarli et al. (2005) estimated ICV using in-house software in 2081 participants from 34 to 96 years of age and found subtle negative associations with age (∼0.1% ICV change per year), independent of sex. Similarly, in a study of 147 participants between the ages of 15 and 96, Buckner et al. (2004) reported a slight negative association between manually determined ICV and age (-1.05 cm^3/year). A notable recent example is the study by Ma et al. (2018), where cross-sectional associations with age were analyzed in a large dataset consisting of 7656 scans for 1727 elderly subjects (55–95 years of age) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Three diagnostic groups were included; cognitive normal, mild cognitive impairment, and an Alzheimer's disease group. They observed no significant cross-sectional associations between age and ICV as estimated with three methods: eTIV, SPM12, and multiatlas label fusion (MALF; Huo et al., 2016).

Most studies on putative associations between ICV and age are cross-sectional rather than longitudinal (Good et al., 2001; Kim et al., 2018; Kruggel, 2006). Such study designs are limited in their ability to resolve age trajectories across the lifespan and can be influenced by generational effects such as secular growth rates (Miller & Corsellis, 1977). In Caspi et al. (2020), both longitudinal and cross-sectional age effects on ICV were assessed in participants between the ages of 16 and 55 years. They included 528 participants at baseline, 378 at follow-up, and 309 at the second follow-up, where the mean period between time points was 3.3 years. In the longitudinal analysis, they reported nonlinear associations where the annual percentage change (APC) of ICV was initially positive (0.03% APC at age 20) and then negative in later adulthood (-0.09% APC at age 55). In the cross-sectional analysis, the data indicated a predicted ICV change of 0.21% for males and 0.22% for females at 20 years of age. They attributed the different magnitudes of effects in the longitudinal and cross-sectional analyses to different secular growth rates. We found only one other study on longitudinal ICV changes in the literature, which showed a statistically significant increase in ICV in the youngest group (≤34 years) but no significant longitudinal age changes in mid- (35–54 years) to late (≥54 years) adulthood (Liu et al., 2003) where ICV was estimated with Exbrain (Lemieux et al., 2003). However, this study was limited by a small sample size with only 44 participants in the first group, 37 in the second, and 9 in the third.

In the present study, we compared five of the most frequently used automated ICV estimation methods to test their relative

consistency and absolute agreement, as well as their cross-sectional and longitudinal associations with age. Given the robustness of the ICV measure, we expected that correlation coefficients between estimation methods would exceed 0.9. We further hypothesized that the agreement between ICV estimates would be higher between the two registration-based methods and between the three segmentation-based methods, than across the two groups of ICV estimation methods. To test potential confounding by age, sex, height, body weight, and mean ICV across methods, we computed their explained variance of the pairwise differences between ICV estimation methods. To examine cross-sectional associations with age, we performed exploratory regression analyses for each dataset and ICV estimation method assessing both linear and nonlinear effects of age. The influence of sex, height, and body weight on cross-sectional age trajectories was investigated in separate interaction models. Finally, based on the study by Caspi et al., 2020, we expected lower ICV at follow-up compared with baseline with an APC of $\sim$0.09%.

2 | MATERIALS AND METHODS

2.1 | Participants and MRI

Participants from two MRI datasets were included in this study: (1) Norwegian Centre for Mental Disorders Research (NORMENT), a cross-sectional dataset recruited in the greater Oslo region of Norway, and (2) UK Biobank (UKB), a longitudinal dataset recruited in the United Kingdom. See Table 1 for demographics for each dataset and Figure 1 for the age distribution for each dataset.

2.1.1 | NORMENT dataset

Healthy participants were pooled from three clinical studies at the NORMENT Centre: the Youth-Thematically Organized Psychosis (Y-TOP) study ($n = 61$), the Thematically Organized Psychosis (TOP) study ($n = 275$), and the StrokeMRI study ($n = 315$). Participants with complete data, that is, age, sex, body weight, height, and T1-weighted MRI data, were included. For Y-TOP and TOP, invitations to healthy participants were sent out to a random sample, stratified by age and

region, from the Norwegian National Population Register in the greater Oslo region. For the StrokeMRI study, healthy participants were recruited via advertisement in local newspapers, social media, and word-of-mouth.

Exclusion criteria for Y-TOP and TOP included a history of neurological disorders or moderate to severe head injury, current or previous diagnosis of a psychiatric disorder, a family history of severe mental disorders, IQ below 70 points, and meeting the criteria for alcohol or substance dependency at the time of MR. For TOP, cannabis use within the last 3 months prior to assessment was an additional exclusion criterion. For StrokeMRI, exclusion criteria included serious head trauma, a history of stroke, dementia, or other severe neurological and psychiatric diseases, alcohol- and substance abuse, and medication use thought to affect the nervous system.

Adult participants gave written informed consent to participate. For adolescent participants below 16 years of age, written assent and parental consent was given. Participants were remunerated with a gift card worth 500 NOK. The studies were approved by the Regional Committee for Research Ethics (REK) and the Norwegian Data Protection Authority, and were carried out in accordance with the Helsinki Declaration.

T1-weighted images were acquired at the Oslo University Hospital, Ullevål, on a 3 T General Electric Discovery MR750 scanner, with a 32-channel head coil, between June 2015 and May 2019. An inversion recovery-prepared 3D gradient recalled echo (BRAVO) sequence was employed with the following parameters: repetition time = 8.16 ms, echo time = 3.18 ms, inversion time = 400 ms, field of view = 256 mm, flip angle = $12°$, matrix = 188×256, voxel size = 1 mm isotropic, 188 sagittal slices.

2.1.2 | UKB dataset

Participants with complete longitudinal MRI data and data on age, sex, weight, and height for both baseline and follow-up imaging sessions were selected from the UKB cohort (www.ukbiobank.ac.uk). Time from baseline to follow-up was 2.3 years on average with a standard deviation of 0.1 years. MRI data was acquired at two different sites and participants were scanned at the same site for baseline and follow-up.

TABLE 1 Sample demographics. Continuous variables reported as mean ± SD. For the UKB dataset, we report age, height, and body weight at baseline imaging session.

|	Sample demographics					
	n	Sex ratio (F/M) (% female)	Age ± SD (years)	Age range (years)	Height ± SD (cm)	Weight ± SD (kg)
UKB	2410	1203/1207 (49.9%)	63.2 ± 7.1	47.0–80.3	170.9 ± 9.6	75.8 ± 14.7
NORMENT	651	366/285 (56.2%)	46.1 ± 18.7	12.0–85.8	173.5 ± 9.3	74.5 ± 14.8
Y-TOP	61	38/23 (62.3%)	17.0 ± 1.9	12.0–20.6	170.8 ± 7.9	62.2 ± 11.3
TOP	275	131/144 (47.6%)	36.9 ± 9.8	17.7–56.7	175.2 ± 9.4	77.7 ± 14.8
StrokeMRI	315	197/118 (62.5%)	59.8 ± 14.5	20.0–85.8	172.5 ± 9.2	74.1 ± 14.1
Combined dataset	3061	1569/1492 (51.3%)	59.6 ± 12.8	12.0–85.8	171.4 ± 9.6	75.5 ± 14.7

Age range across datasets
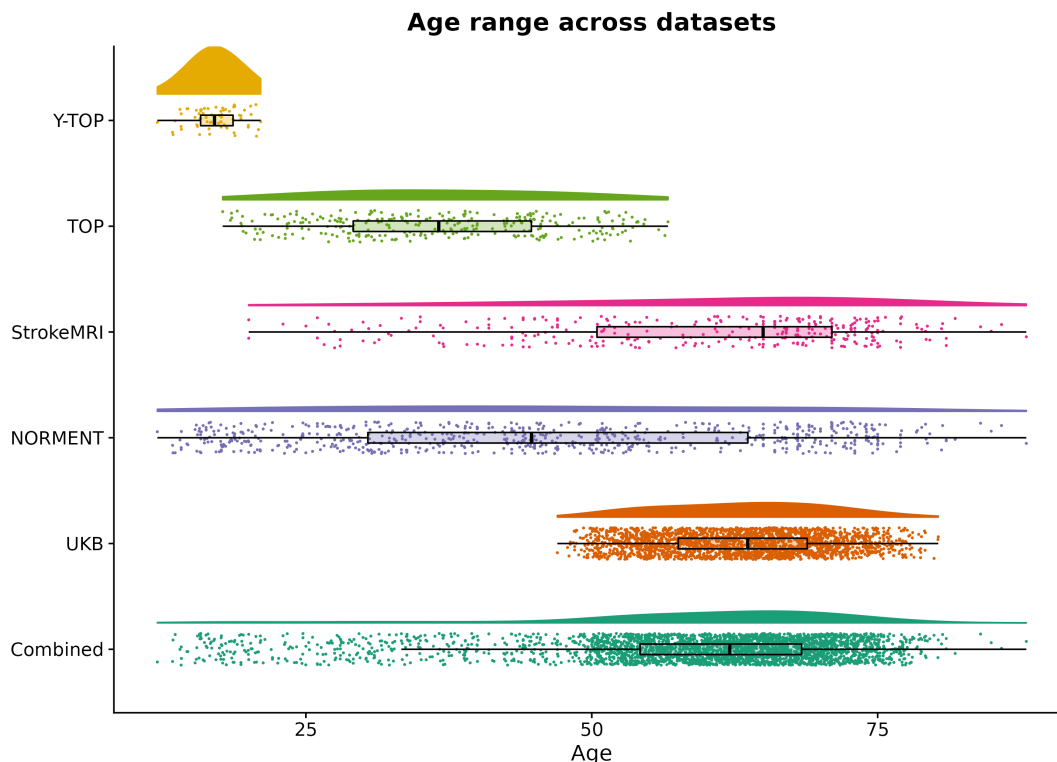
FIGURE 1 Raincloud plot of the age distributions for each subsample. It is composed of flat violin plots, points representing the age of each participant, and horizontal box plots. The lower and upper hinges correspond to the first and third quartiles. The upper and lower whiskers extend to within 1.5 times interquartile range of the hinges. For UKB we report age at baseline imaging session.

We excluded participants with a height difference greater than 5 cm between baseline and follow-up, as we considered this to be indicative of measurement error. Participants were also excluded if they had been diagnosed with disorders known to influence brain structure based on diagnoses from the International Statistical Classification of Diseases and Related Health Problems (ICD-10; World Health Organization, 2004). Diagnostic exclusion criteria included disorders in chapter V and VI, field F; mental and behavioral disorders, including F00–F03 (Alzheimer's disease and dementia), F06.7 (mild cognitive disorder), and field G (diseases of the nervous system), including inflammatory and neurodegenerative diseases (except G55-59; "Nerve, nerve root, and plexus disorders").

An overview of the UK Biobank acquisition protocols is available in Alfaro-Almagro et al. (2018) and Miller et al. (2016). For MRI, a magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence was employed on a Siemens Skyra 3 T scanner with a standard Siemens 32-channel RF receive head coil. The following parameters were used: repetition time = 2000 ms, echo time = 2.01 ms, inversion time = 880 ms, field of view = 256 mm, flip angle = 8°, matrix = 208 × 256, voxel size = 1 mm isotropic.

2.2 | MRI image processing

T1-weighted MRI images were processed to yield two registration-based ICV measures (eTIV, FSL) and three segmentation-based ICV measures (sbTIV, SPM12 and CAT12). For SPM12 and CAT12, we used MATLAB (The MathWorks, Inc., Massachusetts) version R2018b.

2.2.1 | eTIV

We calculated eTIV using the standard processing pipeline, *recon-all*, in FreeSurfer (v5.3.0 for the UKB dataset and v6.0.0 for the NORMENT dataset; https://surfer.nmr.mgh.harvard.edu/). This pipeline performs intensity nonuniformity correction and normalization, skull stripping, and registration to the fsaverage template which is based on the MNI305 template (Evans et al., 1993). The linear scaling factor of this transformation is converted to an ICV estimate by multiplication with the ICV of fsaverage.

2.2.2 | FSL

FSL (v6.0.1; https://fsl.fmrib.ox.ac.uk/fsl/) computes ICV with the SIENAX package by first extracting brain and skull images from a single T1-weighted MRI image which is then affinely registered to the MNI152 template (Grabner et al., 2006). Points along the skull are used when determining the registration scaling factor. The final ICV estimate is calculated by multiplying the scaling factor with the measured ICV of the MNI152 template. We ran the *bet* command with a

fractional intensity threshold of 0.35 and enabled the bias field and neck cleanup flags. The threshold of 0.35 was chosen as it gave the best results in a previous study (Berg et al., 2018), where a systematic test of fractional intensity thresholds was performed. The use of a lower fractional intensity threshold than the default of 0.5 is in line with previous studies (Popescu et al., 2012; Vichianin et al., 2018).

2.2.3 | SPM12

SPM12 (https://www.fil.ion.ucl.ac.uk/spm/) uses a unified segmentation algorithm to perform tissue classification, bias correction and image registration within the same generative model. Based on prior tissue probability maps, it segments the image into tissue classes weighted by the probability of the tissue membership of each tissue type. We used the "Tissue Volumes" utility in SPM12 with the default parameters to calculate ICV as the sum of WM, GM, and CSF.

2.2.4 | CAT12

As with SPM12, CAT12 (v12.7; http://www.neuro.uni-jena.de/cat/) uses tissue probability maps to spatially normalize, skull-strip, and initialize the segmentation. In contrast to SPM12, CAT12 uses an adaptive maximum a posteriori segmentation approach for determining the final segmentation which accounts for local intensity variations in the original image (Tavares et al., 2020). The goal of this procedure is to provide a segmentation algorithm that is less sensitive to differences in image intensity. As with SPM12, ICV is calculated as the partial volume-adjusted sum of WM, GM and CSF. CAT12 processing was performed with the default parameters.

2.2.5 | sbTIV

To compute sbTIV, we used SAMSEG (https://surfer.nmr.mgh.harvard.edu/fswiki/Samseg; Puonti et al., 2016), which creates probability-weighted segmentations of the input image, including skull, non-brain tissue and CSF. sbTIV (https://surfer.nmr.mgh.harvard.edu/fswiki/sbTIV) is computed as a sum of the WM, GM, and CSF volumes. SAMSEG was run with the default parameters using a single T1-weighted image as input.

2.3 | Statistical analyses

In the combined NORMENT and UKB dataset, we computed Pearson correlations between each ICV estimate and performed Bland–Altman and relative importance analysis. To accommodate for the presence of any sample-specific associations, for example, due to generational or scanner differences, cross-sectional associations with age were assessed in the two datasets separately. Longitudinal analyses were limited to the UKB dataset. To ensure the soundness of the fitted

regression models, we visually assessed residuals versus fitted values, scale-location, quantile-quantile, and residuals versus leverage plots. All statistical tests were performed using R Statistical Software (v3.6.3; R Core Team, 2020).

2.3.1 | Outlier correction

To avoid the excess influence of outliers due to measurement errors, we assessed each ICV estimation method for outliers and excluded the corresponding participants in all subsequent analyses. To identify outliers, we used the median absolute deviation method (Leys et al., 2013) implemented in the R package *Routliers* (https://CRAN.R-project.org/package=Routliers). Briefly, this method finds the median of the absolute deviations from the median, which yields the median absolute deviation (MAD). Upper and lower thresholds are determined by taking the median of the original dataset ± b·MAD where b is the specified threshold (unitless). This approach has the advantage of being robust with respect to sample size and the presence of extreme values.

For the cross-sectional datasets, we used a deviation threshold of 3 on the ICV estimates to identify cross-sectional outliers. For the UKB dataset, we also excluded longitudinal outliers based on the pairwise differences between ICV at baseline and follow-up, where we used a less strict deviation threshold of 4. Since outlier removal can affect results, analyses without outlier correction were also performed.

2.3.2 | Pearson correlation analyses

To assess pairwise linear relationships between ICV estimation methods in the combined dataset, we calculated Pearson correlation coefficients (r) with 95% confidence intervals (CI) for each pair of estimation methods using the function *cor.test* from the *stats* R package. To compare the correlations between registration- and segmentation-based methods, we calculated pooled correlations by first applying the Fisher transformation to the correlation coefficients before averaging and back-transforming using the inverse Fisher transformation. We calculated the pooled correlations for each registration-based method with respect to all the segmentation-based methods and compared these pooled correlations to the correlation between the registration-based methods.

2.3.3 | Bland–Altman analysis

For each pair of ICV estimation methods in the combined dataset, we created Bland–Altman plots by plotting differences (ΔICV) between ICV method pairs against their means, which can be seen as a proxy for head size. In the Bland–Altman plots, we used 95% agreement intervals with upper and lower limits of agreement calculated as ΔICV ±1.96 standard deviation (SD) (Altman & Bland, 1983;

Giavarina, 2015). A deviation of ΔICV from zero shows the presence, magnitude, and direction of the difference, or bias, between methods. The bias can be constant or proportional in relation to mean ICV. In the latter case, the difference between methods varies as a function of mean ICV. We quantified these associations by calculating Pearson correlation coefficients between ΔICV and mean ICV and testing their statistical significance.

2.3.4 | Relative importance analysis

To assess the influence of age, sex, height, body weight, and mean ICV across the compared estimation methods on the differences between ICV estimates (i.e., ΔICV) in the combined dataset, we performed relative importance analyses with the *relaimpo* package in R (https://CRAN.R-project.org/package=relaimpo) using the Lindeman, Merenda, and Gold (LMG) metric. This method performs an averaging over the orderings of the explanatory variables to decompose the explained variance of the full model, $\Delta ICV \sim Age + Sex + Weight + Height + Mean$, into the non-negative contributions of each of the variables. The decomposition is constrained so that its sum equals the R^2 of the full model. The advantage of this method over sequential or nested approaches is that the internal correlational structure of the regressors is taken into account, and the resulting variance decomposition is unbiased (Lindeman et al., 1980).

If neither measure is confounded by age, sex, height, body weight, or mean ICV across methods, we would expect them to explain a negligible proportion of the nonshared variation between them, that is, the variation of the difference. Thus, we interpreted the degree to which ICV estimate differences were explained by these variables as an indication of possible bias. Note that wherever these variables explain a large proportion of the variance in the difference between two methods, it is not possible to say which of the methods are confounded.

2.3.5 | Cross-sectional associations between age and ICV

To test for cross-sectional associations between age and ICV, we first fitted linear regression models for each method with ICV as the outcome variable and age and sex as independent variables. These analyses were conducted separately for the NORMENT and UKB datasets. In additional models, we also included age^2 to account for possible quadratic associations with age. To visualize the relationship between age and ICV, we created partial regression plots (Velleman & Welsch, 1981) where both age and ICV were residualized with respect to sex and plotted against each other. Cross-sectional Annual Percentage Change (CS-APC) was calculated by expressing the estimated coefficient of the age term in the fitted model as a percentage relative to the mean ICV for each dataset.

In addition to the linear models, we tested for nonlinear relationships between ICV and age using a generalized additive model (GAM).

This method models possible nonlinear relationships using smooth functions (Hastie & Tibshirani, 1986) which can account for higher order nonlinearities. We used cubic regression splines to model age while adjusting for sex. The restricted maximum likelihood method was used as in Sørensen et al. (2021) with the R package *mgcv* (https://CRAN.R-project.org/package=mgcv).

To test the influence of height and body weight on cross-sectional age associations, we fitted additional models including age, sex, height, and body weight as independent variables, as well as separate models with age-by-height and age-by-body weight interactions. To test the influence of sex on age associations, we included models with age, sex, and an age-by-sex interaction term.

To compare the relative qualities of the linear, quadratic and regression spline models, we used the Akaike Information Criterion (AIC; Akaike, 1974). We calculated and compared AIC scores within each ICV estimation method. A low score indicates less information loss in the model, which is a trade-off between goodness of fit and the simplicity of the model. A difference in AIC scores greater than 2 is considered to indicate a significantly better relative quality for the model with the lower score.

Since the age range differed between the NORMENT and the UKB datasets, we performed additional age-matched analyses where we only included a subset of 289 participants in the NORMENT dataset with an age above 47 years.

In post-hoc analyses, we ran additional linear regression models where we examined if the age effect differed in the two datasets. Here, ICV was used as the outcome variable and age, sex, and the age-by-cohort interaction as independent variables. The analyses were run both with the complete NORMENT dataset, as well as with the age-matched subset of the NORMENT dataset as described above.

2.3.6 | Longitudinal associations between age and ICV

To test for longitudinal associations between age and ICV, we fitted linear mixed-effects (LME) models for each ICV estimation method. We entered baseline age, baseline age^2, scanner, and sex as fixed effects. The primary variable of interest was time, which was defined as 0 for measurements at baseline and as time in years since baseline for follow-up measurements. We used individual intercepts as a random effect, in order to allow for between-participant variation in ICV estimates.

To fit the models we used the *lmer* function from the *lmerTest* package (https://CRAN.R-project.org/package=lmerTest) using the restricted maximum likelihood criterion (REML). The *P*-values were estimated using Satterthwaite's approximation. This approach to evaluating significance in LME models compared favorably to other approximations in Luke (2017). Annual Percentage Change (APC) was estimated by taking the estimated coefficient of the time term in the fitted model and expressing it as a percentage relative to mean ICV at baseline.

We fitted additional models to examine the influence of sex, height, and body weight on longitudinal age associations, including a sex-by-time, height-by-time, or weight-by-time interaction term in addition to the terms above. To test the influence of baseline age on longitudinal age associations, we included models with an age-by-time interaction term.

3 | RESULTS

3.1 | Outlier analysis

We identified and excluded 7 outliers in the NORMENT dataset (1.1% of sample) and 12 outliers in the cross-sectional UKB dataset (0.5% of sample). We also identified and excluded 63 longitudinal outliers for the UKB dataset (2.6% of sample). Two participants were marked both as cross-sectional and longitudinal outliers in the UKB dataset.

We found that eTIV was the largest single contributor of both cross-sectional outliers in the NORMENT dataset and longitudinal outliers in the UKB dataset. See Table S1 for demographic information on participants identified as outliers and Figure S1 for an UpSet plot depicting the ICV estimation methods for which outliers were identified.

The results of analyses without outlier correction were similar to those of the outlier-corrected analyses, with a slight decrease in the correlations between ICV estimation methods, especially between eTIV and SPM12. Specifically, the estimated cross-sectional and longitudinal associations with age were similar for with and without outlier correction.

3.2 | Pearson correlation analyses

Pearson correlations between ICV estimation methods were overall high. The lowest correlation coefficient was observed between eTIV and FSL ($r = .873$, CI = [0.865, 0.882]), followed by FSL and CAT12 ($r = .895$, CI = [0.887, 0.902]). The highest correlation coefficients were observed between sbTIV and SPM12 ($r = .969$, CI = [0.967, 0.971]) and sbTIV and CAT12 ($r = .961$, CI = [0.958, 0.963]). See Figure 2 for a correlogram showing Pearson correlations with 95% confidence intervals.

The Pearson correlation between the registration-based methods (eTIV and FSL) was lower ($r = .873$) than the pooled correlation between eTIV and the segmentation-based methods (pooled $r = .936$) and the pooled correlation between FSL and the segmentation-based methods (pooled $r = .925$). The pooled correlation between the segmentation-based methods (SPM12, CAT12, and sbTIV) was .959.

3.3 | Bland–Altman analysis

See Figure 3 for Bland–Altman plots. We found that, on average, FSL systematically estimated lower ICV (63–143 ml negative bias) and

sbTIV higher ICV (56–143 ml positive bias) compared with the other methods. We also observed statistically significant proportional bias, that is, correlations between ΔICV and mean ICV for most pairwise comparisons, as indicated by the regression lines in Figure 3. The presence of proportional bias indicates that the agreement between methods differs as a function of the mean ICV across estimation methods. The strongest proportional bias was seen in comparisons of sbTIV with the other ICV estimation methods, where the magnitude of correlations ranged from 0.07 to 0.30 ($P < 10^{-4}$). Weaker proportional bias was seen for SPM12 compared with the other methods (excluding sbTIV), here the correlations between ΔICV and mean ICV ranged from 0.10 to 0.05 ($P < 10^{-4}$). We also found weak proportional bias for CAT12 compared with eTIV ($r = -.05$, $P < .01$). Inspection of diagnostic plots revealed no unexplained nonlinearity in the residuals, indications of heteroscedasticity, or influential cases exceeding a Cook's distance of 1. See Figures S2 and S3 for diagnostic plots.

3.4 | Relative importance analysis

When including age, sex, height, body weight, and mean ICV across the ICV estimation methods in the comparison as explanatory variables, the explained variance of the total model ranged from 4.35% for the FSL–sbTIV difference to 30.05% for the SPM12–sbTIV difference. For the SPM12–sbTIV difference, a large proportion of explained variance was due to sex (9.81%) and body weight (10.05%). For the SPM12–CAT12 difference, a large proportion of explained variance was due to body weight (12.37%) and age (9.11%). Notably, mean ICV across methods explained 10.48% of the variance of the eTIV–sbTIV difference and 7.38% of the eTIV–SPM12 difference. Across pairwise differences, body weight, age, sex, and mean ICV across methods were the best explanatory variables for the differences in ICV estimates. Height explained a small proportion of the variance in most ICV differences, except for the SPM12-sbTIV difference (4.42%). See Figure 4 for bar plots showing the variance decomposition for each explanatory variable.

3.5 | Cross-sectional associations between age and ICV

Linear models provided the best fit for both datasets and all ICV estimates, except for sbTIV in the UKB dataset where the age^2 term was significant ($P < .05$) and the difference in AIC of the two models exceeded -2, that is, markedly lower for the quadratic model. Diagnostic plots revealed no violations of the standard assumptions of linear regression. See Figures S4 and S5 for diagnostic plots for each of the main linear regression models and Table S2 for the AIC scores for each model and ICV estimation method for both the NORMENT and the UKB datasets.

For the NORMENT dataset, we found significant *positive* cross-sectional associations with age (ranging from 12.0 to 85.8 years) for
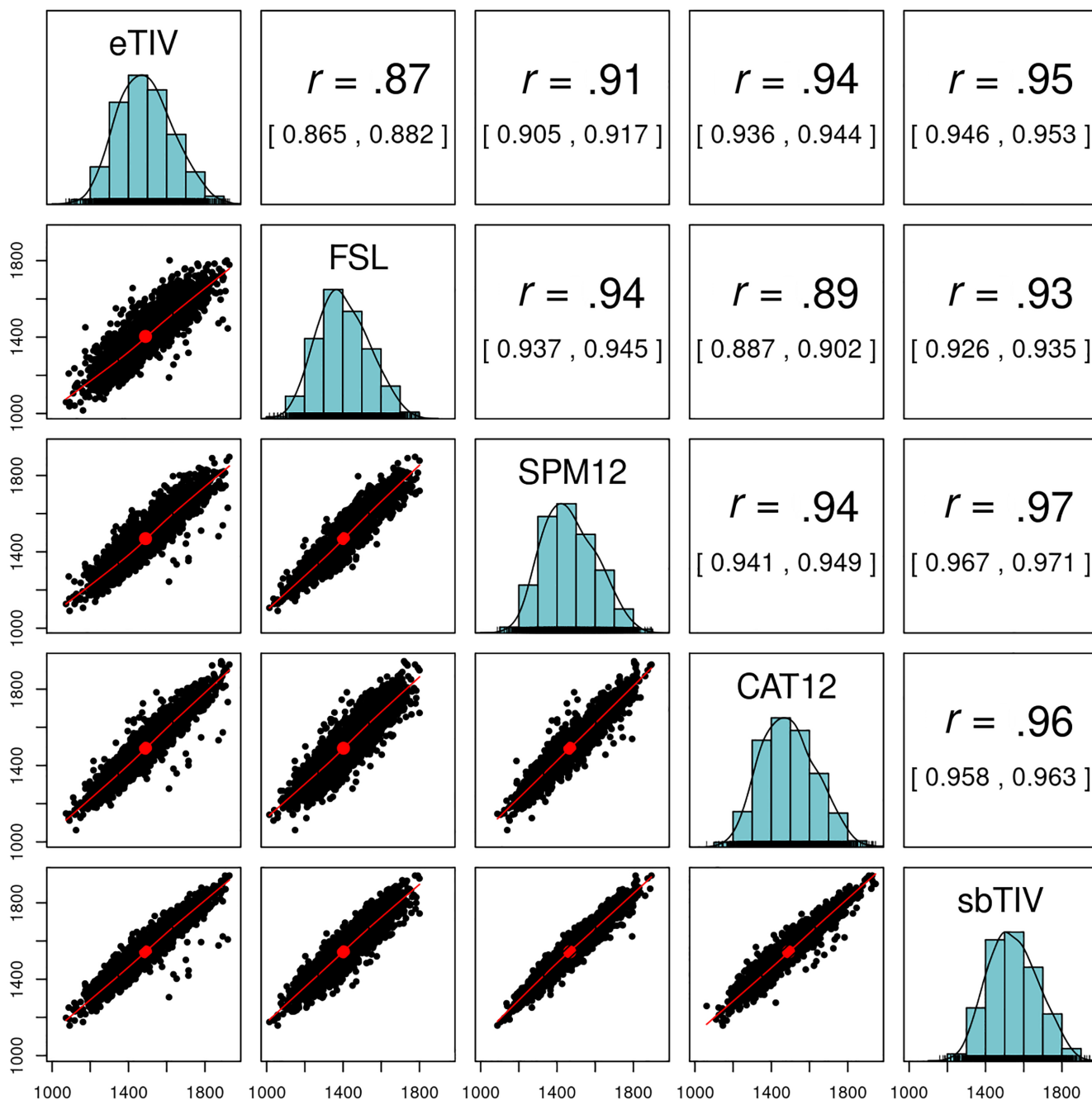
FIGURE 2 Correlations between each pair of ICV estimation methods in the combined dataset ($n = 2981$). The main diagonal shows the distribution of each ICV estimate, the lower diagonal shows scatter plots for each pair of ICV estimates, and the top diagonal shows Pearson correlations with 95% confidence intervals in brackets.

FSL ($b = 1.22$ ml/year, $P < 10^{-6}$) and SPM12 ($b = 1.01$ ml/year, $P < 10^{-5}$). These changes corresponded to a CS-APC of 0.086% for FSL and 0.069% for SPM12. Age was not significantly associated with any other ICV measure in the NORMENT dataset. For the UKB dataset, we found significant *negative* cross-sectional associations with age (ranging from 47.0 to 80.3 years) for all ICV estimation methods. The greatest CS-APC were seen for CAT12 (−0.107% CS-APC) and eTIV (−0.107% CS-APC). The smallest effect size was seen for FSL with an estimated −0.049% CS-APC.

Where the CS-APC estimates are percentages relative to the mean ICV of each dataset.

In the full model, including body weight and height as independent variables in addition to sex and age, we found significant contributions of height for all methods. In the NORMENT dataset, body weight was also significantly associated with ICV for FSL ($b = 1.27$ ml/kg, $P < .001$), SPM12 ($b = 1.38$ ml/kg, $P < 10^{-4}$), and sbTIV ($b = 0.87$ ml/kg, $P < .05$). In the UKB dataset, body weight was associated with ICV for FSL ($b = 0.51$ ml/kg, $P < .01$) and SPM12
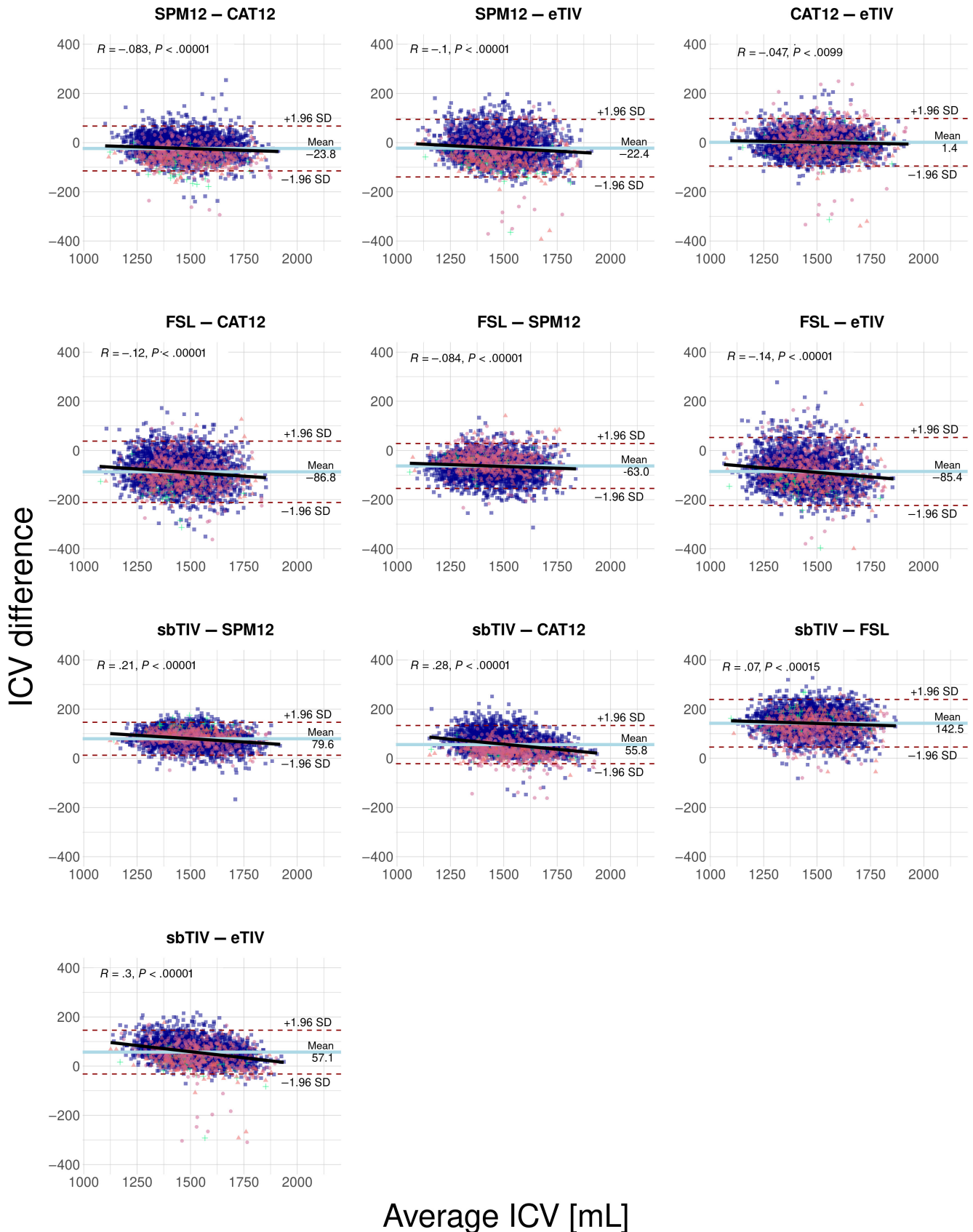
Bland Altman plots

FIGURE 3 Bland–Altman plots for each pair of ICV estimates in the combined dataset ($n = 2981$). The means of each pair of estimates (x-axis) are plotted against the percentage differences of the estimates, ΔICV (y-axis). The Pearson correlation coefficient between mean ICV and ΔICV is shown on the top of each plot along with its *P*-value.
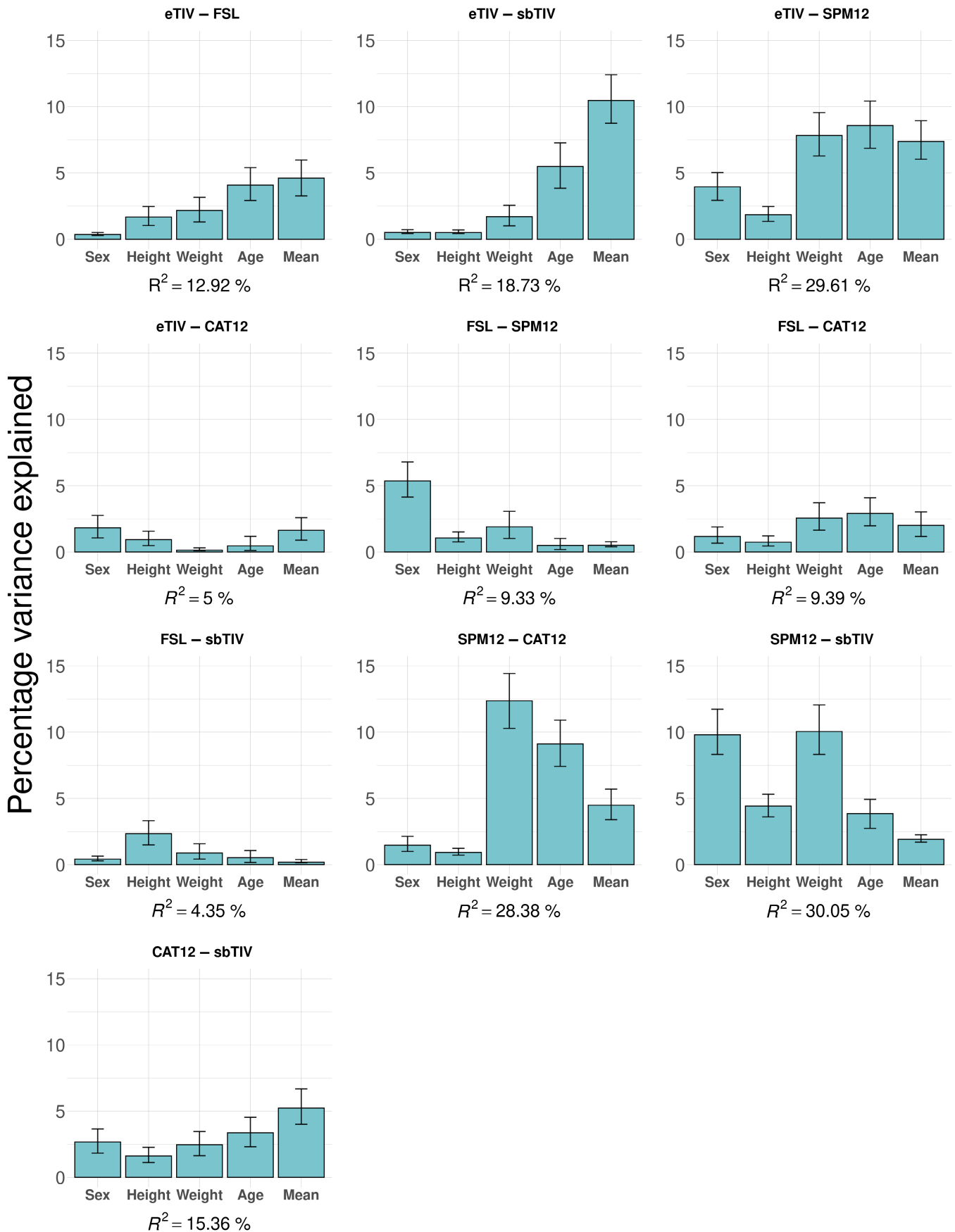
Relative importance plots

FIGURE 4 Relative importance of sex, height, body weight, and age, and mean ICV across methods on the differences between each ICV measure in the combined dataset ($n = 2981$).

($b = 1.04$ ml/kg, $P < 10^{-9}$), but not for eTIV, CAT12, or sbTIV. In the NORMENT dataset, the significant associations between age and ICV remained after additionally adjusting for body weight and height for FSL ($b = 1.06$ ml/years, $P < 10^{-5}$) and SPM12 ($b = 0.84$ ml/years, $P < 10^{-3}$). However, in the UKB dataset the cross-sectional associations with age were no longer significant for FSL, SPM12, and sbTIV when adjusting for body weight and height.

In the NORMENT dataset, we observed *positive* interactions between age and sex for SPM12 ($b = 0.98$ ml/years for males, $P \leq .05$) and CAT12 ($b = 1.07$ ml/years for males, $P < .05$), indicating a more positive slope in males. In the UKB dataset, age-by-sex interactions were found only for FSL ($b = -1.52$ ml/years for males, $P < .05$), and in contrast to the findings in the NORMENT dataset, this showed a *negative* interaction between age and sex, indicating a more negative slope in males. We found no significant age-by-height interactions for any ICV estimation method. A significant age-by-weight association was observed for eTIV in the UKB dataset ($b = 0.05$ ml/years-by-kg, $P < .05$), but not for any of the other ICV estimation methods or in the NORMENT dataset.

In the analyses including a subset of 289 participants from the NORMENT dataset that were age-matched with the UKB dataset, we observed no significant effects of age for any of the ICV estimation methods in the linear model covarying for sex only. When we also adjusted for height and weight, we saw a significant effect of age only for SPM12 ($b = 1.26$ ml/years, $P < .05$).

For the direct comparisons of the age effects between the two datasets, we found significant age-by-cohort interactions for each ICV estimation method when using the complete NORMENT dataset. These interactions indicated a more positive effect of age in the NORMENT dataset compared with the UKB dataset. We also saw a main effect of cohort, indicating higher ICV estimates in the NORMENT dataset compared with the UKB dataset for all ICV estimation methods except sbTIV. When we used the age-matched NORMENT dataset, we found similar significant age-by-cohort interactions for SPM12 ($b = 1.71$ ml/years for NORMENT, $P < .05$), CAT12 ($b = 2.21$ ml/years for NORMENT, $P < .01$), and sbTIV ($b = 1.63$ ml/years for NORMENT, $P < .05$).

See Figures 5 and 6 for partial regression plots for each dataset where age and ICV are residualized with respect to sex and plotted against each other. See Tables S3 and S4 for further details concerning the main linear regression models used to test cross-sectional associations between ICV and age, and Tables S5 and S6 for the full models including body weight and height. See Tables S7 and S8 for the results of the age-matched analyses and Tables S9 and S10 for the results of the direct comparisons of age effects across datasets.

3.6 | Longitudinal associations between age and ICV

In the main longitudinal analyses, with fixed factors sex, baseline age, baseline age^2, scanner, and time (years since baseline), in the UKB dataset (age range = 47.0–80.3 years), we found significant associations with time since baseline for all ICV estimates. Effect sizes ranged from an APC of −0.293% for sbTIV to −0.416% for CAT12. The longitudinal effects of time on ICV remained for all ICV estimation methods when we included height and weight as fixed factors. Where the APC estimates are percentages relative to the mean ICV at baseline.

We found significant interactions with time for sex and height for eTIV, but not for any of the other ICV estimation methods. For sex, the interaction indicated lower longitudinal reduction for male participants ($b = 2.53$ ml/years for males, $P < 10^{-4}$). For height, the interaction showed a lower longitudinal reduction for taller participants ($b = 0.11$ ml/years-by-cm, $P < .001$). There were no statistically significant interactions between longitudinal ICV change and baseline age.

See Table S11 for details on the main LME model used to investigate longitudinal associations with age. See Figure 7 for spaghetti plots depicting longitudinal ICV change in the UKB dataset stratified by sex. See Figure S6 for histograms depicting the raw annual percentage change between baseline and follow-up for each ICV estimation method. Diagnostic plots are included as Figure S7.

4 | DISCUSSION

We compared five commonly used ICV estimation methods, eTIV, FSL, SPM12, CAT12, and sbTIV, and tested their cross-sectional and longitudinal associations with age. Correlations were overall high, but we found notable differences between the estimation methods. In particular, age, body weight, sex, and mean ICV across methods explained a large proportion for the nonshared variation between several ICV estimation methods. Different cross-sectional effects of age were observed in the two datasets. In the UKB dataset (age range = 47.0–80.3 years), all ICV estimation methods indicated *negative* cross-sectional associations with age, whereas for the NORMENT dataset (age range = 12.0–85.8 years), we observed significant effects only for two of the ICV estimation methods indicating *positive* cross-sectional associations with age. Finally, we observed a striking convergence of results across ICV estimation methods in the longitudinal analyses in the UKB dataset, with an average of 2.3 years from baseline to follow-up, supporting past reports of longitudinal ICV reduction in mid- to late adulthood.

4.1 | Outlier analysis

In the outlier analyses, we found that eTIV was the largest single contributor (i.e., outliers detected for eTIV, but no other ICV estimates) both for cross-sectional outliers in the NORMENT dataset (86% of outliers) and for longitudinal outliers in the UKB dataset (25% of outliers). In comparison, no participants were marked as outliers for sbTIV alone. This could indicate that eTIV is especially prone to measurement errors and requires a more comprehensive quality control procedure than the other methods, including assessment of the Talairach registration that is performed in the *recon-all* processing stream in FreeSurfer.

Partial regression plots - NORMENT
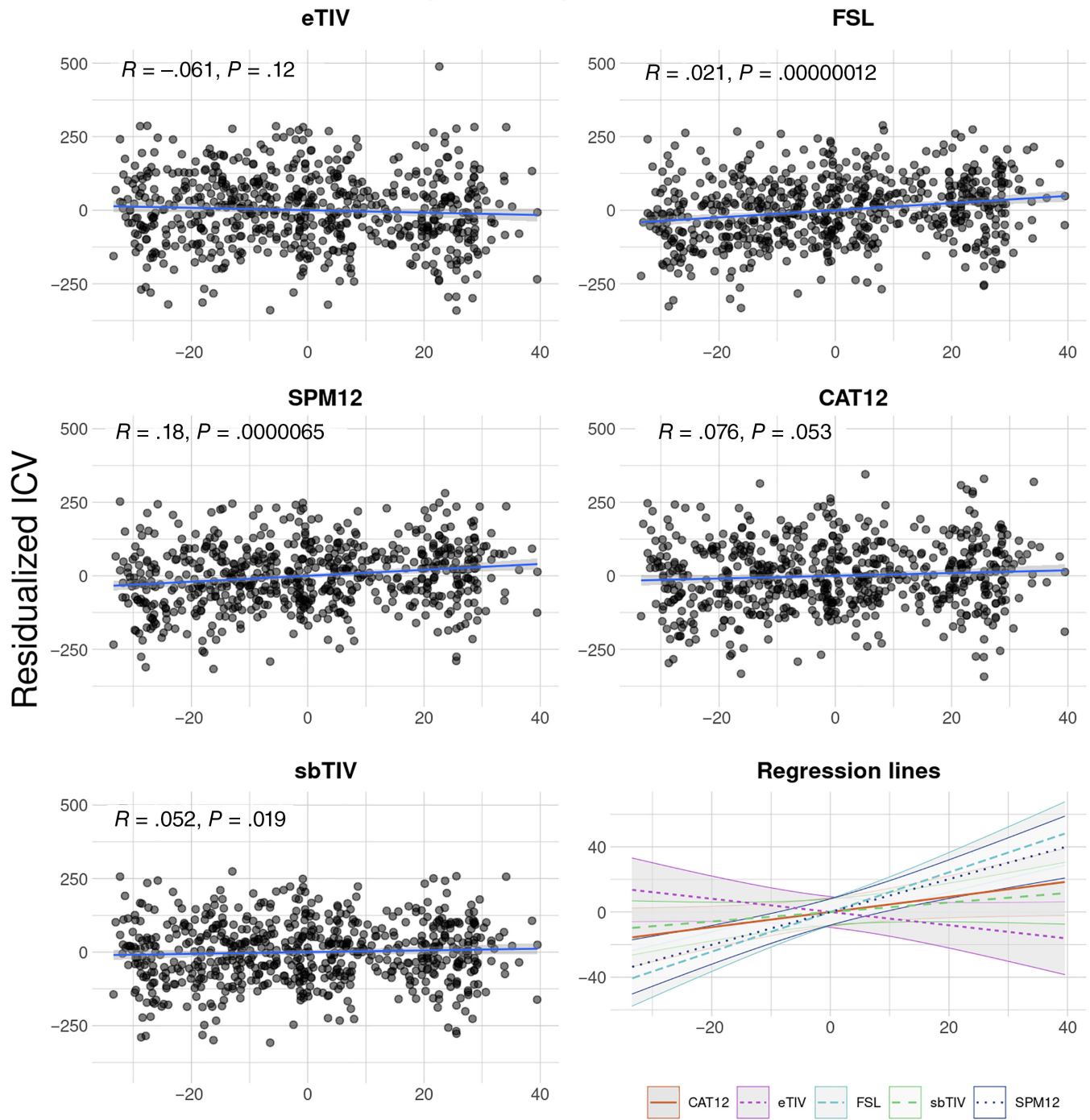
FIGURE 5 Partial regression plots in the NORMENT dataset ($n = 644$) for the associations between age and ICV for each estimation method. The effect of sex has been regressed out for both age and ICV and the regression lines show the residual effect of age on ICV. As age has been residualized with respect to sex, the x-axis is centered at the sex-adjusted mean.

4.2 | Correlations between ICV estimation methods

Contrary to our hypothesis, we found higher correlations for each of the registration-based methods, eTIV and FSL, with the segmentation-based methods (eTIV pooled $r = .936$; FSL pooled $r = .925$) than between eTIV and FSL ($r = .873$). The pooled correlation between the segmentation-based methods (SPM12, CAT12, and sbTIV) was .959. This points to greater internal consistency within the segmentation-based compared with the registration-
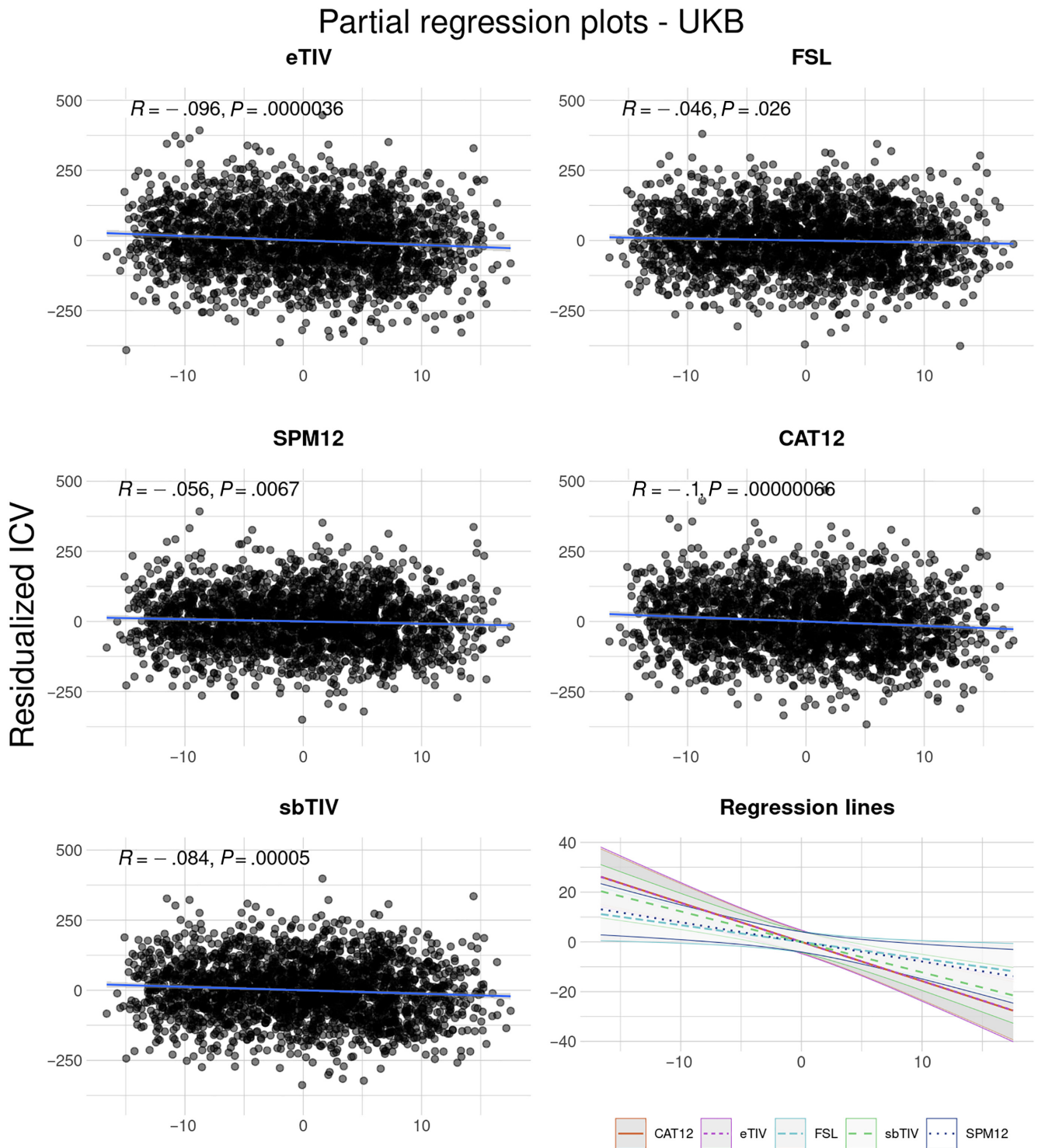
FIGURE 6 Partial regression plots in the cross-sectional UKB dataset ($n = 2337$) for the associations between age and ICV for each estimation method. The effect of sex has been regressed out from both age and ICV and the regression lines show the residual effect of age on ICV. As age has been residualized with respect to sex, the x-axis is centered at the sex-adjusted mean.

based methods. While both eTIV and FSL estimate ICV via an atlas scaling factor, the target template for eTIV is the MNI305-derived fsaverage template (Evans et al., 1993; Fischl et al., 1999), whereas

FSL uses the MNI152 template as its target (Mazziotta et al., 1995). Furthermore, FSL uses skull points to constrain the transformation to the template (Smith, 2002; Smith et al., 2001), which may give
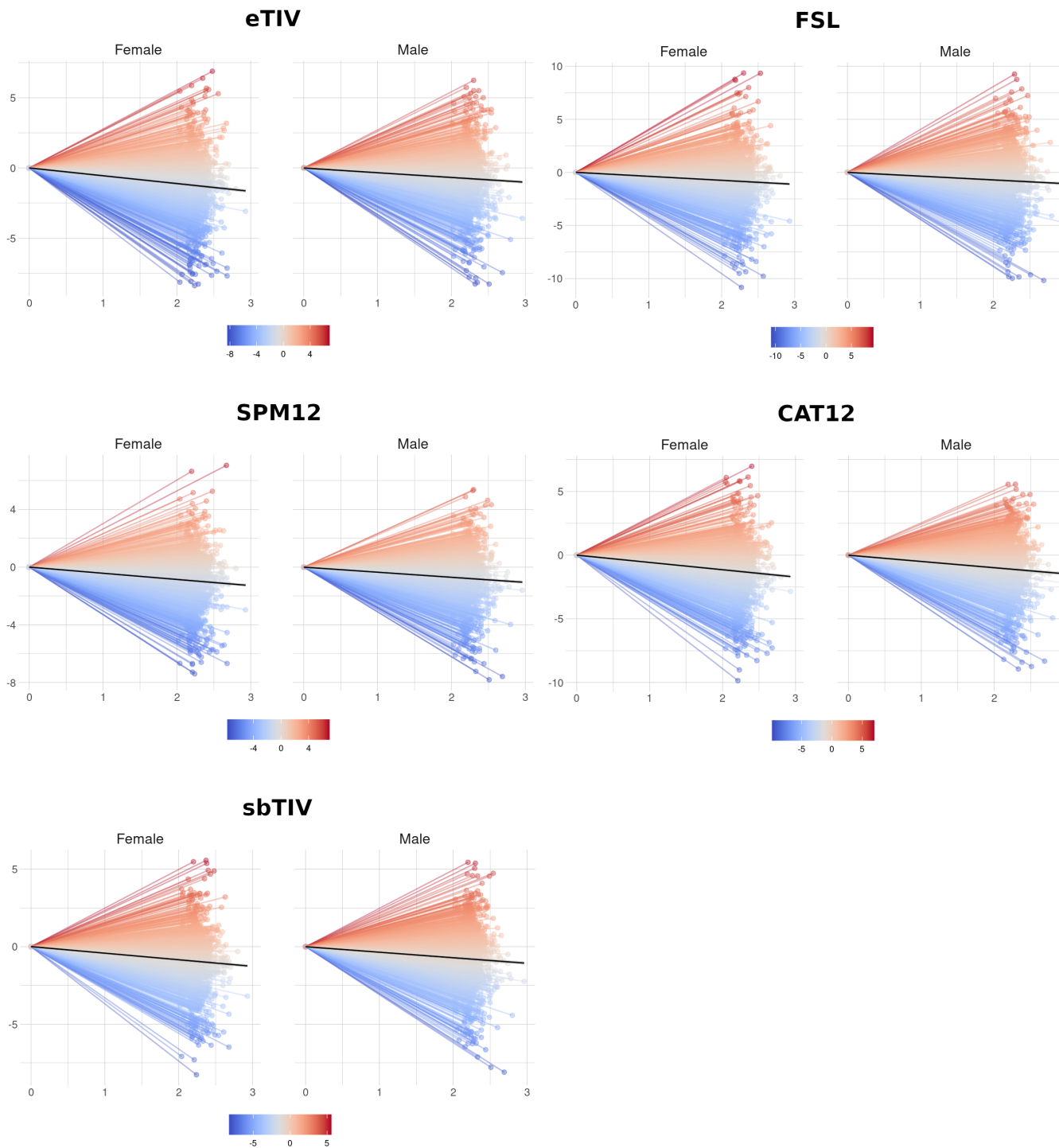
FIGURE 7 Normalized spaghetti plots depicting longitudinal change for each ICV estimate stratified by sex. The x-axis depicts the interscan interval normalized by time of first imaging session, the y-axis represents ICV change from baseline to follow-up as a percentage of the total ICV across both time points, and the color indicates positive ICV change in red and negative ICV change in blue. The black line depicts the linear trend of longitudinal ICV change.

better registration results compared with eTIV (Buckner et al., 2004). Together, these differences could explain the relatively low correlation between eTIV and FSL, and the differences in their pairwise correlations with the segmentation-based estimation methods.

4.3 | Quantitative differences between estimated ICV

Consistent with the previous literature, we found systematic quantitative differences between ICV estimates. In particular, FSL estimated

lower ICV, while sbTIV estimated higher ICV than the other methods. Since registration-based methods calculate ICV as the product of an atlas scaling factor and the predetermined ICV of the template, under- or over-estimation of ICV can occur due to the estimated ICV of the template. This could explain the tendency of FSL to underestimate ICV, and a similar overestimation for eTIV relative to manually determined ICV has been reported previously (Klasson et al., 2018). As such, one should be cautious to interpret registration-based methods as absolute measures of ICV.

It is unclear why the segmentation-based method sbTIV appeared to overestimate ICV, but it is worth noting that the anatomical extent of the cranial cavity inferiorly along the spinal cord is not clearly defined, and systematic differences in the cut-off in the different ICV estimation methods may contribute to systematic volumetric differences. Therefore, while segmentation-based methods estimate ICV using a more direct approach, that is, probability-weighted voxel counts, these methods can also be subject to systematic biases and it remains important to assess their agreement with manually determined ICV if the goal is quantitative ICV estimation.

4.4 | Associations between mean ICV and ICV differences

Bland–Altman analyses revealed widespread associations between mean ICV across methods and the differences between them. In other words, the agreement between ICV estimates differed as a function of the mean ICV across methods. This finding was supported by the relative importance analyses where mean ICV across methods explained a substantial proportion of the variance in the eTIV-sbTIV, eTIV-SPM12, SPM12-CAT12, and CAT12-sbTIV differences. Klasson et al. (2018) reported findings consistent with a bias due to head size on eTIV. If mean ICV is considered a proxy for head size, our results lend support to these findings and indicate that similar bias is also present for other ICV estimation methods. It should be noted that while our methods revealed systematic associations in the differences between ICV estimation methods and mean ICV, they cannot say which of the ICV estimation methods have a higher accuracy. We encourage future methodological studies on the validity of automated ICV estimation methods to also include manually determined ICV.

4.5 | Relative importance of explanatory variables on ICV differences

In the relative importance analyses, we explored the explained variance of age, sex, body weight, height, and mean ICV across methods on the pairwise differences between ICV estimates. If ICV estimation is unbiased by these variables, we would expect to see low explained variance of the estimated ICV differences. Contrary to this expectation, we found a high proportion of explained variance for the total

model in the comparison of SPM12 with sbTIV (30.05%), CAT12 (28.38%), and eTIV (29.61%), and in the comparison of sbTIV with eTIV (18.73%) and CAT12 (15.36%).

Body weight explained 10.05% of the variance in the SPM12–sbTIV difference and 12.37% in SPM12–CAT12 difference. This suggests a systematic impact of body weight on ICV differences between these estimation methods. Such bias can be grounds for concern in studies where the variable of interest is associated with weight differences. For example, weight gain is a known side effect of antipsychotic medication (Dayabandara et al., 2017).

We found that sex explained 9.81% of the SPM12–sbTIV difference. There are contradictory results in the literature on the presence and strength of sexual dimorphism in brain volumes and age-related associations with brain volumes (Fjell et al., 2009; Inano et al., 2013; Ritchie et al., 2018). It remains an open question whether sex-dependent volumetric differences remain after correction for ICV. Past studies have also shown that the statistical correction method can affect sex differences in brain volumes (Pintzka et al., 2015; Sanchis-Segura et al., 2020). The choice of ICV estimation method may explain some of the discrepancy in the past findings. We encourage future studies on sex dimorphism of brain volumes to carefully assess the accuracy of the chosen ICV estimation method, given that confounding by sex can affect results.

4.6 | Cross-sectional associations with age

We found negative cross-sectional associations between ICV and age across all ICV estimation methods in the UKB dataset. Notably, associations with age did not reproduce across all ICV estimation methods in the NORMENT dataset. For this dataset, we only observed statistically significant associations between ICV and age for two (FSL and SPM12) out of five ICV estimation methods. This supports our hypothesis that the choice of ICV estimation method can, at least in some cases, cause discrepancies in reported associations between ICV and age. In post hoc analyses, we found statistically significant differences in the age effects in the two datasets for all ICV estimation methods indicating more negative age trajectories in the UKB dataset.

Unlike the study by Ma et al. (2018), we found statistically significant negative associations with age for all ICV estimation methods in the UKB dataset, which had a similar age range. While this study included a large number of MRI scans (7656 scans), the number of participants was less than that of the UKB dataset (1727 participants). The data was also acquired on both 1.5 T and 3 T scanners which could reduce the sensitivity of the analyses. Furthermore, their statistical models may not have been optimal for distinguishing between within-subject longitudinal associations and between-subject effects. It should also be noted that differences in the secular growth rates in the ADNI dataset compared with the UKB dataset may have contributed to the discrepancy with our results.

Since the mean age and the age range differed between the UKB and the NORMENT datasets, we performed additional analyses on a subset of 289 participants from the NORMENT dataset that were

age-matched to the UKB dataset. In these analyses, none of the main models showed significant associations between age and ICV. However, in the full models, corrected for age, sex, height, and body weight, we saw a significant *positive* association between age and ICV for SPM12—in line with our findings in the complete dataset. Given the loss of statistical power in these analyses it is difficult to draw firm conclusions. However, the replication of the positive association with age for SPM12 in the full model, and the positive signs of the coefficients for the age term in the main models for FSL and SPM12, suggest that the age associations we observed were not entirely explained by differences in the age ranges. Furthermore, we observed significant age-by-cohort differences in the direct comparisons between the UKB dataset and the age-matched NORMENT dataset for SPM12, CAT12, and sbTIV, but not for eTIV and FSL.

We did not see evidence for nonlinear effects of age on ICV in the NORMENT dataset, and the quadratic age term was statistically significant only for sbTIV in the UKB dataset. Past studies on ICV and head size, including computed tomography and head circumference studies (Bergerat et al., 2021; Huda et al., 2004; Neubauer et al., 2009; Weaver & Christian, 1980), lend some support to our findings of linear age trajectories across most ICV estimation methods.

In the NORMENT dataset, we observed sex-by-age interactions for SPM12 and CAT12, indicating a greater cross-sectional ICV increase in males for SPM12 and CAT12. In the UKB dataset, on the other hand, we found a negative sex-by-age interaction for FSL, indicating that male sex is associated with greater cross-sectional ICV decrease for ICV estimated with FSL. Since the sex-by-age interactions did not replicate with the other ICV estimation methods, they may point toward sex-dependent bias to ICV estimation with these methods. These findings show that sex can, for some ICV estimation methods, affect the cross-sectional associations between age and ICV. This can be a particular concern in clinical group comparisons where sex distributions may vary between groups.

4.7 | Longitudinal associations with age

In the longitudinal analyses, we observed statistically significant ICV reduction at follow-up compared with baseline across all estimation methods. The consistency of the longitudinal findings stands in contrast to the cross-sectional analyses in the NORMENT dataset where statistically significant associations with age were found for two ICV estimation methods, but not for the other three methods. This may be due to the superior ability of longitudinal data to resolve age trajectories (Sørensen et al., 2021). In particular, the estimation of longitudinal change has the advantage of being based on within-subject variation where potential biases from variables such as sex, height, and birth year may have a lesser influence on the results.

Our results indicated an APC ranging from −0.29% to −0.42%, which was considerably higher than the APC estimated by Caspi et al. (2020) where they reported an APC of −0.09% at age 55. However, their age range was from 16 to 55 years of age, and thus only partially

overlapped with ours. Notably, they also included people with a diagnosis of schizophrenia and relatives of these participants in the analysis. It is known that ICV has a significant genetic component (Adams et al., 2016), and is affected by genes that confer risk for developing a psychotic disorder (Smeland et al., 2018). Furthermore, cross-sectional studies have found associations between a diagnosis of schizophrenia and ICV (Gurholt et al., 2020; van Erp et al., 2016). Thus, the sample used in Caspi et al. (2020) may be confounded by genetic risk of psychotic disorder, both directly in patients, and indirectly in relatives.

We observed significant sex-by-time interactions for eTIV, indicating less ICV decrease between baseline and follow-up in males, but no such interactions for the other ICV estimation methods. While this relationship was not replicated with the other ICV estimation methods, it is in line with a semi-longitudinal study of elderly participants (age range = 71.1–74.3 years) by Royle et al. (2013) where two sets of manual ICV segmentations were compared in a dataset of elderly participants. The first segmentation measured current ICV, whereas the other was derived from expert manual segmentation where inner table skull thickening (i.e., thickening of the inner bony structure of the skull) was taken into account. The segmentations were used to estimate the longitudinal effect of inner skull thickening on ICV change, yielding an estimated median ICV decrease of 6.2% in males and 8.3% in females across the lifespan, in line with the literature on physiological skull changes with age (Harding, 1949). However, given that eTIV was the only ICV estimation method that showed this sex-by-time interaction it may be due to biases inherent to eTIV. It is also unclear whether skull thickening occurs at a constant rate across the lifespan and consequently how much of this effect would be detectable in the age range of the UKB dataset. Furthermore, for eTIV alone we also found a significant height-by-time interaction indicating less longitudinal change in taller participants. It is possible that the sex-by-time interaction is driven by height rather than other physiological differences between males and females in this dataset.

We did not find statistically significant baseline age-by-time interactions for any ICV estimation methods, suggesting that the rate of longitudinal ICV change does not differ with age in mid- to late-adulthood. This finding should, however, be interpreted with caution, given the narrow age range of this sample. In line with the past literature on age-related ICV change, particularly in adolescence, it would be expected that the rate of change differs if longitudinal change is estimated for a greater age range.

4.8 | Strengths and limitations

Strengths of our study include the use of large, well-characterized datasets composed of participants with no known neurological or psychiatric disorders thought to affect head size. In particular, the UKB dataset is the largest dataset used for assessing longitudinal ICV change to date. In the cross-sectional analyses, we had a broad age range suitable to assess the presence of nonlinear associations between age and ICV. The data was processed in a

harmonized framework, facilitating the direct comparison between ICV estimation methods. For the NORMENT dataset, MRI acquisition was conducted on the same scanner system with no major upgrades.

Some limitations also apply. We processed a total of 5471 T1-weighted images (651 NORMENT scans + 4820 UKB baseline and follow-up scans) with five different ICV estimation methods. The resulting dataset of 27,355 ICV estimates precluded detailed quality control of each output. It also prevented us from performing manual segmentation of images, given its time-consuming nature (e.g., ~30 min per subject in Ambarki et al., 2012). Without manually determined ICV estimates, we cannot draw firm conclusions about which estimation methods are more accurate with respect to manual delineation. As with all longitudinal studies, systematic time-dependent bias (e.g., due to scanner drift) can affect the longitudinal results. Finally, our longitudinal dataset only covered a narrow age range, and we did not investigate longitudinal ICV change in adolescence and early adulthood.

5 | CONCLUSIONS

Correlations between ICV estimation methods were lower between the registration-based methods than between segmentation-based methods, suggesting greater internal consistency for the segmentation-based methods. We found that sex, body weight, age, and mean ICV across methods explained a large proportion of the nonshared variation for some pairs of ICV estimation methods. This may represent bias that can be grounds for concern in clinical studies. We observed significant proportional bias for most pairwise comparisons, that is, varying agreement as a function of mean ICV across methods.

In the NORMENT dataset, spanning adolescence to old age, two ICV estimation methods were *positively* associated with age. In the UKB dataset, spanning mid- to late adulthood, all ICV estimates were *negatively* associated with age. The discovery of age-related effects only for two ICV estimation methods in the NORMENT dataset illustrate how the choice of ICV estimation method can affect findings of age-related associations with ICV. Relationships between ICV and age were significantly different between the two datasets for all estimation methods in the complete dataset, and for three methods in age-matched analyses. This may be due to different secular growth rates in the two cohorts. The convergence of longitudinal results across ICV estimation methods, in the largest dataset to date, offers strong evidence for longitudinal age-related ICV reductions in mid- to late adulthood.

In conclusion, the choice of ICV estimation method is a possible source of bias, not only in studies investigating ICV as a variable of interest, but also in studies that use ICV as an adjustment factor. We encourage future studies to investigate the validity of automated ICV estimation methods and implement quality control procedures to assess the accuracy of the ICV estimation method as with other morphometric brain measures.

AUTHOR CONTRIBUTIONS
Stener Nerland: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Therese S. Stokkan:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Kjetil N. Jørgensen:** Methodology, Writing - original draft, Writing - review & editing. **Laura A. Wortinger:** Writing - original draft, Writing - review & editing. **Geneviève Richard:** Data curation, Writing - original draft, Writing - review & editing. **Dani Beck:** Writing - original draft, Writing - review & editing. **Dennis van der Meer:** Data curation, Writing - original draft, Writing - review & editing. **Lars T. Westlye:** Resources, Writing - original draft, Writing - review & editing, Funding acquisition. **Ole A. Andreassen:** Resources, Writing - original draft, Writing - review & editing, Funding acquisition. **Ingrid Agartz:** Resources, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition, Project administration. **Claudia Barth:** Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration.

CONFLICT OF INTEREST
OAA has received a speaker honorarium from Lundbeck, and is a consultant for HealthLytix. The other authors report no financial relationships with commercial interest.

DATA AVAILABILITY STATEMENT
The data incorporated from the UK Biobank study is publicly available pending a standard registration procedure (https://www.ukbiobank.ac.uk/). Derived data for this dataset, i.e., estimated ICV for each estimation method employed in this study, will be made publicly available through the UK Biobank Returns Catalogue https://biobank.ndph.ox.ac.uk/showcase/docs.cgi?id=1). Given a formal data sharing agreement and approval from the local ethics committees, access to the NORMENT dataset can be provided.

ORCID
Stener Nerland https://orcid.org/0000-0001-8771-8856
Therese S. Stokkan https://orcid.org/0000-0003-3609-1058
Kjetil N. Jørgensen https://orcid.org/0000-0002-5751-4809

Laura A. Wortinger https://orcid.org/0000-0003-2721-3172
Geneviève Richard https://orcid.org/0000-0001-6475-2576
Dani Beck https://orcid.org/0000-0002-0974-9304
Dennis van der Meer https://orcid.org/0000-0002-0466-386X
Lars T. Westlye https://orcid.org/0000-0001-8644-956X
Ole A. Andreassen https://orcid.org/0000-0002-4461-3568
Ingrid Agartz https://orcid.org/0000-0002-9839-5391
Claudia Barth https://orcid.org/0000-0001-6544-0945

REFERENCES

Adams, H. H., Hibar, D. P., Chouraki, V., Stein, J. L., Nyquist, P. A., Rentería, M. E., Trompet, S., Arias-Vasquez, A., Seshadri, S., Desrivières, S., Beecham, A. H., Jahanshad, N., Wittfeld, K., Van der Lee, S. J., Abramovic, L., Alhusaini, S., Amin, N., Andersson, M., Arfanakis, K., ... Thompson, P. M. (2016). Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nature Neuroscience*, *19*(12), 1569–1582. https://doi.org/10.1038/nn.4398

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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. https://doi.org/10.1016/j.neuroimage.2017.10.034

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *32*(3), 307–317. https://doi.org/10.2307/2987937

Ambarki, K., Lindqvist, T., Wåhlin, A., Petterson, E., Warntjes, M. J. B., Birgander, R., Malm, J., & Eklund, A. (2012). Evaluation of automatic measurement of the intracranial volume based on quantitative MR imaging. *American Journal of Neuroradiology*, *33*(10), 1951–1956. https://doi.org/10.3174/ajnr.A3067

Berg, A. O., Jørgensen, K. N., Nerhus, M., Athanasiu, L., Popejoy, A. B., Bettella, F., Norbom, L. C. B., Gurholt, T. P., Dahl, S. R., Andreassen, O. A., Djurovic, S., Agartz, I., & Melle, I. (2018). Vitamin D levels, brain volume, and genetic architecture in patients with psychosis. *PLoS One*, *13*(8), e0200250. https://doi.org/10.1371/journal.pone.0200250

Bergerat, M., Heude, B., Taine, M., Nguyen The Tich, S., Werner, A., Frandji, B., Blauwblomme, T., Sumanaru, D., Charles, M. A., Chalumeau, M., & Scherdel, P. (2021). Head circumference from birth to five years in France: New national reference charts and comparison to WHO standards. *The Lancet Regional Health-Europe*, *5*, 100114. https://doi.org/10.1016/j.lanepe.2021.100114

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*(2), 724–738. https://doi.org/10.1016/j.neuroimage.2004.06.018

Caspi, Y., Brouwer, R. M., Schnack, H. G., van de Nieuwenhuijzen, M. E., Cahn, W., Kahn, R. S., Niessen, W. J., van der Lugt, A., & Pol, H. H. (2020). Changes in the intracranial volume from early adulthood to the sixth decade of life: A longitudinal study. *NeuroImage*, *220*, 116842. https://doi.org/10.1016/j.neuroimage.2020.116842

Davis, P. J. M., & Wright, E. A. (1977). A new method for measuring cranial cavity volume and its application to the assessment of cerebral atrophy at autopsy. *Neuropathology and Applied Neurobiology*, *3*(5), 341–358. https://doi.org/10.1111/j.1365-2990.1977.tb00595.x

Dayabandara, M., Hanwella, R., Ratnatunga, S., Seneviratne, S., Suraweera, C., & de Silva, V. A. (2017). Antipsychotic-associated weight gain: Management strategies and impact on treatment adherence. *Neuropsychiatric Disease and Treatment*, *13*, 2231–2241. https://doi.org/10.2147/NDT.S113099

DeCarli, C., Massaro, J., Harvey, D., Hald, J., Tullberg, M., Au, R., Beiser, A., D'Agostino, R., & Wolf, P. A. (2005). Measures of brain morphology and infarction in the Framingham heart study: Establishing what is normal. *Neurobiology of Aging*, *26*(4), 491–510. https://doi.org/10.1016/j.neurobiolaging.2004.05.004

Evans, A. C., Collins, D., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In 1993 IEEE conference record nuclear science symposium and medical imaging conference. https://doi.org/10.1109/NSSMIC.1993.373602

Fischl, B., Sereno, M. I., Tootell, R. B. H., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*(4), 272–284. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4

Fjell, A. M., Westlye, L. T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Agartz, I., Salat, D. H., Greve, D. N., Fischl, B., Dale, A. M., & Walhovd, K. B. (2009). Minute effects of sex on the aging brain: A multisample magnetic resonance imaging study of healthy aging and Alzheimer's disease. *The Journal of Neuroscience*, *29*(27), 8774–8783. https://doi.org/10.1523/JNEUROSCI.0115-09.2009

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N. A., Friston, K. J., & Frackowiak, R. S. J. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, *14*(1), 21–36. https://doi.org/10.1006/nimg.2001.0786

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., & Collins, D. L. (2006). Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults. In R. Larsen, M. Nielsen, & J. Sporring (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2006 (pp. 58–66)*. Springer. https://doi.org/10.1007/11866763_8

Gurholt, T. P., Lonning, V., Nerland, S., Jørgensen, K. N., Haukvik, U. K., Alloza, C., Arango, C., Barth, C., Bearden, C. E., Berk, M., Bohman, H., Dandash, O., Díaz-Caneja, C. M., Edbom, C. T., Erp, T. G. M., Fett, A. J., Frangou, S., Goldstein, B. I., Grigorian, A., ... For the ENIGMA-EOP Working Group. (2020). Intracranial and subcortical volumes in adolescents with early-onset psychosis: A multisite mega-analysis from the ENIGMA consortium. *Human Brain Mapping*, *43*, 373–384. https://doi.org/10.1002/hbm.25212

Harding, F. E. (1949). Endocrinopathies associated with hyperostosis frontalis interna. *The American Journal of Medicine*, *6*(3), 329–335. https://doi.org/10.1016/0002-9343(49)90169-2

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310. https://doi.org/10.1214/ss/1177013604

Huda, W., Lieberman, K. A., Chang, J., & Roskopf, M. L. (2004). Patient size and x-ray technique factors in head computed tomography examinations I. Radiation doses. *Medical Physics*, *31*(3), 588–594. https://doi.org/10.1118/1.1646232

Huo, Y., Asman, A. J., Plassard, A. J., & Landman, B. A. (2016). Simultaneous total intracranial volume and posterior fossa volume estimation using multi-atlas label fusion. *Human Brain Mapping*, *38*(2), 599–616. https://doi.org/10.1002/hbm.23432

Inano, S., Takao, H., Hayashi, N., Yoshioka, N., Mori, H., Kunimatsu, A., Abe, O., & Ohtomo, K. (2013). Effects of age and gender on

neuroanatomical volumes. *Journal of Magnetic Resonance Imaging*, *37*(5), 1072–1076. https://doi.org/10.1002/jmri.23910

Kim, Y. S., Park, I. S., Kim, H. J., Kim, D., Lee, N. J., & Rhyu, I. J. (2018). Changes in intracranial volume and cranial shape in modern Koreans over four decades. *American Journal of Physical Anthropology*, *166*(3), 753–759. https://doi.org/10.1002/ajpa.23464

Klasson, N., Olsson, E., Eckerström, C., Malmgren, H., & Wallin, A. (2018). Estimated intracranial volume from FreeSurfer is biased by total brain volume. *European Radiology Experimental*, *2*(1), 24. https://doi.org/10.1186/s41747-018-0055-4

Klasson, N., Olsson, E., Rudemo, M., Eckerström, C., Malmgren, H., & Wallin, A. (2015). Valid and efficient manual estimates of intracranial volume from magnetic resonance images. *BMC Medical Imaging*, *15*(1), 5. https://doi.org/10.1186/s12880-015-0045-4

Kruggel, F. (2006). MRI-based volumetry of head compartments: Normative values of healthy adults. *NeuroImage*, *30*(1), 1–11. https://doi.org/10.1016/j.neuroimage.2005.09.063

Lemieux, L., Hammers, A., Mackinnon, T., & Liu, R. S. N. (2003). Automatic segmentation of the brain and intracranial cerebrospinal fluid in T1-weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. *Magnetic Resonance in Medicine*, *49*(5), 872–884. https://doi.org/10.1002/mrm.10436

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott Foresman & Co. http://catalog.hathitrust.org/api/volumes/oclc/5310754.html

Liu, R. S. N., Lemieux, L., Bell, G. S., Sisodiya, S. M., Shorvon, S. D., Sander, J. W. A. S., & Duncan, J. S. (2003). A longitudinal study of brain morphometrics using quantitative magnetic resonance imaging and difference image analysis. *NeuroImage*, *20*(1), 22–33. https://doi.org/10.1016/S1053-8119(03)00219-2

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Ma, D., Popuri, K., Bhalla, M., Sangha, O., Lu, D., Cao, J., Jacova, C., Wang, L., & Beg, M. F. (2018). Quantitative assessment of field strength, total intracranial volume, sex, and age effects on the goodness of harmonization for volumetric analysis on the ADNI database. *Human Brain Mapping*, *40*(5), 1507–1527. https://doi.org/10.1002/hbm.24463

Malone, I. B., Leung, K. K., Clegg, S., Barnes, J., Whitwell, J. L., Ashburner, J., Fox, N. C., & Ridgway, G. R. (2015). Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance. *NeuroImage*, *104*, 366–372. https://doi.org/10.1016/j.neuroimage.2014.09.034

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (ICBM). *NeuroImage*, *2*, 89–101. https://doi.org/10.1006/nimg.1995.1012

Miller, A. K. H., & Corsellis, J. A. N. (1977). Evidence for a secular increase in human brain weight during the past century. *Annals of Human Biology*, *4*(3), 253–257. https://doi.org/10.1080/03014467700007142

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Mills, K. L., Goddings, A.-L., Herting, M. M., Meuwese, R., Blakemore, S.-J., Crone, E. A., Dahl, R. E., Güroğlu, B., Raznahan, A., Sowell, E. R., & Tamnes, C. K. (2016). Structural brain development between childhood and adulthood: Convergence across four longitudinal samples. *NeuroImage*, *141*, 273–281. https://doi.org/10.1016/j.neuroimage.2016.07.044

Neubauer, S., Gunz, P., & Hublin, J.-J. (2009). The pattern of endocranial ontogenetic shape changes in humans. *Journal of Anatomy*, *215*(3), 240–255. https://doi.org/10.1111/j.1469-7580.2009.01106.x

Nordenskjöld, R., Malmberg, F., Larsson, E.-M., Simmons, A., Brooks, S. J., Lind, L., Ahlström, H., Johansson, L., & Kullberg, J. (2013). Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements. *NeuroImage*, *83*, 355–360. https://doi.org/10.1016/j.neuroimage.2013.06.068

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*, *193*(2), 113–122. https://doi.org/10.1016/j.pscychresns.2011.01.007

Pfefferbaum, A., Mathalon, D. H., Sullivan, E. V., Rawles, J. M., Zipursky, R. B., & Lim, K. O. (1994). A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Archives of Neurology*, *51*(9), 874–887. https://doi.org/10.1001/archneur.1994.00540210046012

Pintzka, C. W. S., Hansen, T. I., Evensmoen, H. R., & Håberg, A. K. (2015). Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: A HUNT MRI study. *Frontiers in Neuroscience*, *9*, 238. https://doi.org/10.3389/fnins.2015.00238

Popescu, V., Battaglini, M., Hoogstrate, W. S., Verfaillie, S. C. J., Sluimer, I. C., van Schijndel, R. A., van Dijk, B. W., Cover, K. S., Knol, D. L., Jenkinson, M., Barkhof, F., de Stefano, N., & Vrenken, H. (2012). Optimizing parameter choice for FSL-brain extraction tool (BET) on 3D T1 images in multiple sclerosis. *NeuroImage*, *61*(4), 1484–1494. https://doi.org/10.1016/j.neuroimage.2012.03.074

Puonti, O., Iglesias, J. E., & Van Leemput, K. (2016). Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. *NeuroImage*, *143*, 235–249. https://doi.org/10.1016/j.neuroimage.2016.09.011

R Core Team. (2020). *R: A language and environment for statistical computing (v3.6.3) [computer software]*. R Foundation for Statistical Computing. https://www.R-project.org/

Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., Harris, M. A., Alderson, H. L., Hunter, S., Neilson, E., Liewald, D. C. M., Auyeung, B., Whalley, H. C., Lawrie, S. M., Gale, C. R., Bastin, M. E., McIntosh, A. M., & Deary, I. J. (2018). Sex differences in the adult human brain: Evidence from 5216 UK biobank participants. *Cerebral Cortex*, *28*(8), 2959–2975. https://doi.org/10.1093/cercor/bhy109

Royle, N. A., Hernández, M. C. V., Maniega, S. M., Arabisala, B. S., Bastin, M. E., Deary, I. J., & Wardlaw, J. M. (2013). Influence of thickening of the inner skull table on intracranial volume measurement in older people. *Magnetic Resonance Imaging*, *31*(6), 918–922. https://doi.org/10.1016/j.mri.2013.01.012

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*(1), 12953. https://doi.org/10.1038/s41598-020-69361-9

Sargolzaei, S., Sargolzaei, A., Cabrerizo, M., Chen, G., Goryawala, M., Pinzon-Ardila, A., Gonzalez-Arias, S., & Adjouadi, M. (2015). Estimating intracranial volume in brain research: An evaluation of methods. *Neuroinformatics*, *13*, 427–441. https://doi.org/10.1007/s12021-015-9266-5

Smeland, O. B., Wang, Y., Frei, O., Li, W., Hibar, D. P., Franke, B., Bettella, F., Witoelar, A., Djurovic, S., Chen, C.-H., Thompson, P. M., Dale, A. M., & Andreassen, O. A. (2018). Genetic overlap between schizophrenia and volumes of hippocampus, putamen, and intracranial volume indicates shared molecular genetic mechanisms. *Schizophrenia Bulletin*, *44*(4), 854–864. https://doi.org/10.1093/schbul/sbx148

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Smith, S. M., De Stefano, N., Jenkinson, M., & Matthews, P. M. (2001). Normalized accurate measurement of longitudinal brain change. *Journal of Computer Assisted Tomography*, *25*(3), 466–475.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Smith, S. M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P. M., Federico, A., & De Stefano, N. (2002). Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, *17*(1), 479–489. https://doi.org/10.1006/nimg.2002.1040

Sørensen, Ø., Walhovd, K. B., & Fjell, A. M. (2021). A recipe for accurate estimation of lifespan brain trajectories, distinguishing longitudinal and cohort effects. *NeuroImage*, *226*, 117596. https://doi.org/10.1016/j.neuroimage.2020.117596

Tavares, V., Prata, D., & Ferreira, H. A. (2020). Comparing SPM12 and CAT12 segmentation pipelines: A brain tissue volume-based age and Alzheimer's disease study. *Journal of Neuroscience Methods*, *334*, 108565. https://doi.org/10.1016/j.jneumeth.2019.108565

van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., Agartz, I., Westlye, L. T., Haukvik, U. K., Dale, A. M., Melle, I., Hartberg, C. B., Gruber, O., Kraemer, B., Zilles, D., Donohoe, G., Kelly, S., McDonald, C., Morris, D. W., … Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547–553. https://doi.org/10.1038/mp.2015.63

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, *35*(4), 234–242. https://doi.org/10.2307/2683296

Vichianin, Y., Kareesaw, A., Chawalparit, O., & Ohki, M. (2018). Optimizing fractional intensity threshold for FSL-brain extraction tool (BET) and comparing with freeSurfer on 3D T1W MR images. *Siriraj Medical Journal*, *70*(5), 391–396. https://doi.org/10.14456/smj.2018.62

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.

Weaver, D. D., & Christian, J. C. (1980). Familial variation of head size and adjustment for parental head circumference. *The Journal of Pediatrics*, *96*(6), 990–994. https://doi.org/10.1016/S0022-3476(80)80623-8

Whitwell, J. L., Crum, W. R., Watt, H. C., & Fox, N. C. (2001). Normalization of cerebral volumes by use of intracranial volume: Implications for longitudinal quantitative MR imaging. *AJNR: American Journal of Neuroradiology*, *22*(8), 1483–1489.

World Health Organization. (2004). *ICD-10: International statistical classification of diseases and related health problems: Tenth revision* (2nd ed.). World Health Organization.

SUPPORTING INFORMATION

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