Revision 2

Title: Study design features that improve effect sizes in cross-sectional and longitudinal brain-wide association studies

Authors: Kaidi Kang[1,*]; Jakob Seidlitz[2,3,4]; Richard A.I. Bethlehem[5]; Jiangmei Xiong[1]; Megan T. Jones[1]; Kahini Mehta[3,4,6]; Arielle S. Keller[3,4,6]; Ran Tao[1]; Anita Randolph[7]; Bart Larsen[7]; Brenden Tervo-Clemmens[8]; Eric Feczko[7]; Oscar Miranda Dominguez[7]; Steve Nelson[7]; Lifespan Brain Chart Consortium; 3R-BRAIN; AIBL; Alzheimer's Disease Neuroimaging Initiative*; Alzheimer's Disease Repository Without Borders Investigators; CALM Team; CCNP; COBRE; cVEDA; Harvard Aging Brain Study; IMAGEN; POND; The PREVENT-AD Research Group; Jonathan Schildcrout[1]; Damien Fair[7]; Theodore D. Satterthwaite[3,4,6]; Aaron Alexander-Bloch[2,3,4,#] and Simon Vandekar[1,#,*]

[1]Department of Biostatistics, Vanderbilt University Medical Center
[2]Department of Child and Adolescent Psychiatry and Behavioral Sciences, The Children's Hospital of Philadelphia
[3]Department of Psychiatry, University of Pennsylvania
[4]Lifespan Brain Institute of The Children's Hospital of Philadelphia and Penn Medicine
[5]Department of Psychology, University of Cambridge
[6]Penn Lifespan Informatics and Neuroimaging Center (PennLINC), Perelman School of Medicine, University of Pennsylvania
[7]Department of Pediatrics, University of Minnesota Medical School
[8]Department of Department of Psychiatry & Behavioral Sciences, University of Minnesota Medical School
[#]These authors contributed equally.
*Corresponding authors

*Data used in the preparation of this article includes data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Please address correspondence to:
simon.vandekar@vumc.org
kaidi.kang@vanderbilt.edu

Abstract:
Brain-wide association studies (BWAS) are a fundamental tool in discovering brain-behavior associations. Several recent studies showed that thousands of study participants are required to improve the replicability of BWAS because actual effect sizes are much smaller than those reported in smaller studies. Here, we perform analyses and meta-analyses of a robust effect size index (RESI) using 63 longitudinal and cross-sectional magnetic resonance imaging studies from the Lifespan Brain Chart Consortium (77,695 total scans) to demonstrate that optimizing study design is critical for improving standardized effect sizes and replicability in BWAS. A meta-analysis of brain volume associations with age indicates that BWAS with larger covariate variance have larger effect size estimates and that the longitudinal studies we examined have systematically larger standardized effect sizes than cross-sectional studies. We propose a cross-sectional RESI to adjust for the systematic difference in effect sizes between cross-sectional and longitudinal studies that allows investigators to quantify the benefit of conducting their study longitudinally. Analyzing age effects on global and regional brain measures from the United Kingdom Biobank and the Alzheimer's Disease Neuroimaging Initiative, we show that modifying longitudinal study design through sampling schemes to increase between-subject variability and adding a single additional longitudinal measurement per subject can improve effect sizes. However, evaluating these longitudinal sampling schemes on cognitive, psychopathology, and demographic associations with structural and functional brain outcome measures in the Adolescent Brain and Cognitive Development dataset shows that commonly used longitudinal models can, counterintuitively, reduce effect sizes. We demonstrate that the benefit of conducting longitudinal studies depends on the strengths of the between- and within-subject associations of the brain and non-brain measures. Explicitly modeling between- and within-subject effects avoids conflating the effects and allows optimizing effect sizes for them separately. These findings underscore the importance of considering study design features to improve the replicability of BWAS.

Brain-wide association studies (BWAS) use noninvasive magnetic resonance imaging (MRI) to identify associations between inter-individual differences in behavior, cognition, biological or clinical measurements and brain structure or function[1,2]. In brain disorder research, the fundamental goal of BWAS is to identify individual differences that improve our understanding of how brain organization and function may be altered in clinical conditions. Identifying robust imaging markers could also be a first step toward the goal of developing earlier and more effective treatments for brain disorders[3].

Recent studies have raised concerns about the replicability of BWAS[1,2,4]. Statistical replicability is defined as "obtaining consistent results across studies that had obtained their own data and aimed at answering the same scientific question"[5]. Like power, replicability is a function of the standardized effect size (ES) and the sample size[6–8]. Standardized ESs (such as Pearson's correlation *r*, Cohen's *d*, and Cohen's *f*) are different from unstandardized effects (such as regression coefficients) because they quantify the strength of an association on a unitless scale. Low replicability in BWAS has been attributed to a combination of small standardized ESs, small sample sizes, *p*-hacking, and publication bias[1,2,9–13]. Throughout this paper, we focus on standardized ESs as these are directly related to power.

The most obvious solution to improve power and replicability of univariate and multivariate BWAS is to increase study sample sizes or scan time[14]. Several recent studies showed that thousands of study participants are required to obtain replicable findings in BWAS because standardized ESs are much smaller than expected based on biased estimates reported in smaller studies[1,2]. However, ESs are a function of modifiable features of the study design that can be adjusted to increase standardized ESs. By increasing ESs through study design, we can better utilize the $1.7 billion the National Institutes of Health (NIH) spends annually on neuroimaging research[15]. In particular, modifications to the sampling scheme and implementing longitudinal designs might improve statistical power for associations of brain and cognitive or biological measurements[1,16–18]. Further, longitudinal BWAS, which utilize repeated measurements from the same individuals, can study intra-individual associations. The benefit of these study design features has not been studied in BWAS and, as yet, there is no established method to quantify the benefit of using a longitudinal versus cross-sectional design.

Here, we analyze and meta-analyze 63 neuroimaging datasets including 77,695 scans from 60,900 cognitively normal (CN) participants from the Lifespan Brain Chart Consortium[19] (LBCC) to demonstrate the influence of sample population and longitudinal/cross-sectional study designs on ESs[19–30]. We quantify the ESs of age and sex on the three main tissue volumes of the cerebrum (total gray matter volume (GMV), total subcortical gray matter volume (sGMV), and total white matter volume (WMV)), mean cortical thickness (CT), as well as regional GMV and CT in CN participants. As we show, standardized ESs are larger in studies with greater sample heterogeneity (represented by the standard deviation of the covariate). The longitudinal studies we analyzed have systematically larger GMV-age ESs compared to cross-sectional studies.

Systematic differences in ESs between cross-sectional and longitudinal studies that compromise their comparability arise because the ES is dependent on study design. Throughout the present study, we use the robust effect size index (RESI)[31–33] as a measure of standardized ES, a recently developed ES index encompassing many types of test statistics which is equal to ½ Cohen's *d* under some assumptions (Methods; Supplement)[31]. Because

standardized ESs are dependent on whether the study design was cross-sectional or longitudinal (Supplement: Section 1.1.1), we refer to the RESI estimate in longitudinal studies as the "longitudinal RESI (L-RESI)". We also propose the "cross-sectional RESI (CS-RESI) for longitudinal datasets" that improves the comparability of ESs between longitudinal and cross-sectional studies and allows us to quantify the benefit of using a longitudinal study design in a single dataset. The CS-RESI represents the RESI in the same study population if the longitudinal study had been conducted cross-sectionally. Comparing L-RESI to the newly developed CS-RESI allows us to quantify the benefit of using a longitudinal study design in a single dataset (see Supplement: Section 1).

We evaluate sampling schemes to increase ESs for the association between structural and functional brain measures and covariates including age, cognitive function, and psychopathology. In the United Kingdom Biobank (UKB) and Alzheimer's Disease Neuroimaging Initiative (ADNI), we show that increasing between-subject variability in the age variable and adding a second measurement per subject increases the total GMV-age ES by up to 60% and 156%, respectively. To generalize these findings, we investigate longitudinal sampling schemes applied to other covariates (NIH toolbox, CBCL, body mass index (BMI), birthweight, and handedness) for structural and functional brain measures in the Adolescent Brain Cognitive Development (ABCD) and demonstrate that, counterintuitively, longitudinal designs can reduce ESs. This reduction in longitudinal designs is due to common model assumptions conflating between- and within-subject effects. We show that modeling these two effects separately allows us to ameliorate the reduction and improve ESs for each by increasing the between- or within-subject variability of the covariate, respectively.

Meta-analyses show ESs depend on study population and design

To fit each study-level analysis, we regress each of the global brain measures (total GMV, total sGMV, total WMV, mean CT) and regional brain measures (regional GMV and CT, based on Desikan-Killiany parcellation[34]) on sex and a nonlinear spline function of age in each of the 63 neuroimaging datasets from the LBCC using linear regression models for the cross-sectional datasets and generalized estimating equations (GEEs) for the longitudinal datasets (Methods). Site effects are removed before the regressions using ComBat (Methods). The analyses for total GMV, total sGMV and total WMV use all 63 neuroimaging datasets (47 cross-sectional and 16 longitudinal) containing 77,695 MRI scans from 60,900 CN participants. Among these, 74,149 MRI scans from 57,538 CN participants who have complete-case regional brain measures across 43 neuroimaging datasets (30 cross-sectional and 13 longitudinal) are used in the analysis for mean CT and regional brain measures(see Methods for more details; sample information in Tab. S1 and S2). To investigate the effects of study design features on the RESIs, we perform meta-analyses for the four global brain measures and two regional brain measures in the LBCC to model the association of study-level design features with ESs. Study design features are quantified as the sample mean, standard deviation (SD), and skewness of the age covariate as nonlinear terms, and a binary variable indicating the design type (cross-sectional or longitudinal). After obtaining estimates of the RESIs of age and sex in each analysis of the global and regional brain measures, we conduct meta-analyses of the RESI or L-RESI estimates using weighted linear regression models with study design features as covariates (see Methods for more details).

For total GMV, the partial regression plots of the effect of each study design feature demonstrate a strong cubic-shape relationship between total GMV-age ES (RESI or L-RESI) and study population mean age, indicating that the strength of the age effect varies with respect to the age of the population being studied. The largest total GMV-age ESs in the human lifespan occur during early and late adulthood (Tab. S3A; Fig. 1a). There is also a strong positive linear effect of the study population SD of age and the total GMV-age ESs: for each unit increase in the SD of age (in years), expected ES increases by about 0.1 (Fig. 1a), based on the well-known relationship between correlation strength and covariate SD. Plots for total sGMV, total WMV, and mean CT show U-shaped changes of the age RESI with respect to study population mean age (Fig. 1b-d), and a similar, but sometimes weaker relationship between expected ES and study population SD and skewness of the age covariate (Fig. 1b-d; Tab. S3-S5). Finally, the meta-analyses also show a moderate effect of study design on the RESI of age on each of the global brain measures (Fig. 1a-d; Tab. S3A; Tab. S4A, Tab. S5A, and Tab. S6A). The average total GMV-age RESI in longitudinal studies (RESI=0.39) is 0.31 larger than in cross-sectional studies (RESI=0.08) after controlling for the study design variables, corresponding to a >380% increase in ES for longitudinal studies. This estimate quantifies the systematic differences in the ESs between the cross-sectional and longitudinal studies among the 63 neuroimaging studies. Notably, longitudinal study design does not improve the ES for biological sex, because the variable does not vary within subjects (Tab. S7-S10; Fig. S2).

For regional GMV and CT, similar effects of study design features also occur across regions (Fig. 1e-f; 34 regions per hemisphere). In most of the regions, the ESs of age on regional GMV and CT are strongly associated with the study population SD of age. Longitudinal study designs generally tend to have a positive effect on the regional GMV-age ESs and a weaker effect on regional CT-age ESs (see supplementary material for the partial regression plots for each regional outcome).

These results demonstrate that standardized ESs are dependent on study design features, such as mean age and SD of age of the sample population and cross-sectional/longitudinal design, and suggest that modifying study design features, such as increasing variability and conducting longitudinal studies can increase ES in BWAS.

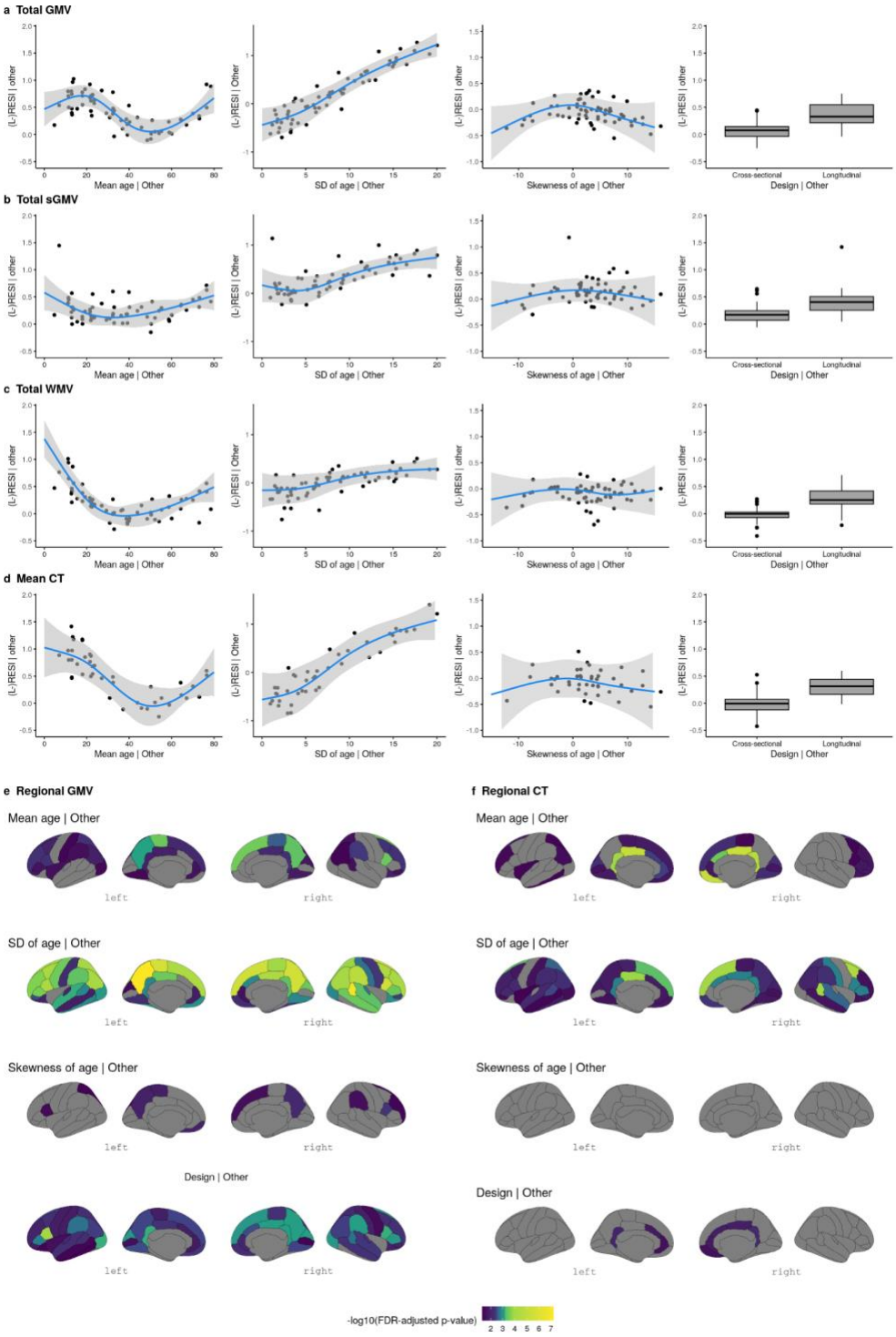

Fig. 1. Meta-analyses reveal study design features that increase effect sizes between age and different brain measures. (**a-d**) Partial regression plots of the meta-analyses of effect sizes (RESI for cross-sectional studies and L-RESI for longitudinal studies) for the association between age and global brain measures (**a**) total gray matter volume (GMV), (**b**) total subcortical gray matter volume (sGMV), (**c**) total white matter volume (WMV), and (**d**) mean cortical thickness (CT) ("| Other" means after fixing the other features at constant levels: design = cross-sectional, mean age = 45 years, sample age SD = 7 and/or skewness of age = 0 (symmetric)) show that effect sizes

vary with respect to the mean and SD of age in each study. The blue curves are the expected effect sizes for age from the locally estimated scatterplot smoothing (LOESS) curves. The gray areas are the 95% confidence bands from the LOESS curves. (**e-f**) The effects of study design features on the effect sizes (RESI and L-RESI) for the associations between age and regional brain measures (regional GMV and CT). Regions with Benjamini-Hochberg adjusted *p*-values<0.05 are shown in color.

Comparison of L-RESI and CS-RESI quantifies benefit of longitudinal design

While the systematic difference in the ES between cross-sectional and longitudinal study designs compromises the comparability of ESs between studies, it is also a feature that we can leverage to quantify the benefit of using a longitudinal design in a single study. With this goal in mind, we define and evaluate a new CS-RESI and its confidence interval (see Supplement: Section 1). The difference between the L-RESI and CS-RESI in a single longitudinal study quantifies the benefit of using the longitudinal study design.

To demonstrate the difference in ES by study design and validate the CS-RESI estimator, we use non-parametric bootstrapping to create longitudinal bootstrap samples by randomly sampling subjects with replacement (1,000 bootstraps per sample size) in each of the 16 longitudinal neuroimaging datasets. In this analysis and subsequent bootstrap analyses, the full dataset represents the unknown population, and bootstrap replicates represent random samples from the population. At a given sample size, after obtaining a longitudinal bootstrap sample, we create an equivalent cross-sectional version of the longitudinal bootstrap sample by randomly selecting one measurement from each subject. In each replicate, we fit the same GEEs and linear regression models of total GMV to the longitudinal and cross-sectional bootstrap samples to quantify the effects of age and sex on total GMV, and obtain the differences of the RESI estimated in the cross-sectional sample to the CS-RESI and L-RESI estimated in the longitudinal sample.

When comparing the L-RESI estimated in the longitudinal samples to the RESI estimated in cross-sectional samples, the longitudinal ESs for age and sex are typically larger (Fig. 2a). The positive differences represent the amount of benefit of using longitudinal designs to quantify the age or sex effect for the two studies shown, with age showing an increase of 0.16 in GUSTO and 0.25 in ADNI, and sex showing an expected smaller increase of 0.02 in GUSTO and ADNI (other studies shown in Fig. S3). The difference we note in longitudinal ESs is not restricted to the L-RESI, as Cohen's *d* for sex computed in a longitudinal study also shows the same slightly positive estimate (Fig. 2a). In contrast, the mean difference between the CS-RESI computed in the longitudinal samples and the RESI computed in the cross-sectional samples approaches zero as the sample size increases (Fig. 2b), indicating that the systematic difference in ESs between longitudinal and cross-sectional datasets is removed. As a further comparison, we repeat the meta-analyses above on the ESs from all studies but now use the CS-RESIs to represent the ESs in longitudinal studies, demonstrating a reduction in the design effect (Tab. S3B, S4B, S5B, and S6B).

The benefit of conducting the longitudinal design can be quantified in each study by comparing the CS-RESI estimates with the L-RESI estimates (Fig. 2c; Fig. S4 and S5). Longitudinal designs result in a 0.31 average increase in the total GMV-age ES among the 16 longitudinal neuroimaging datasets (Tab. S3A), corresponding to a moderate increase in ES[35].

Notably, in some studies (e.g., POND and SCZIowa), the L-RESI is smaller than the CS-RESI, which implies that they lose efficiency by using longitudinal designs. We investigate this counterintuitive result further in our subsequent analyses of cognitive variables in the ABCD study (below).

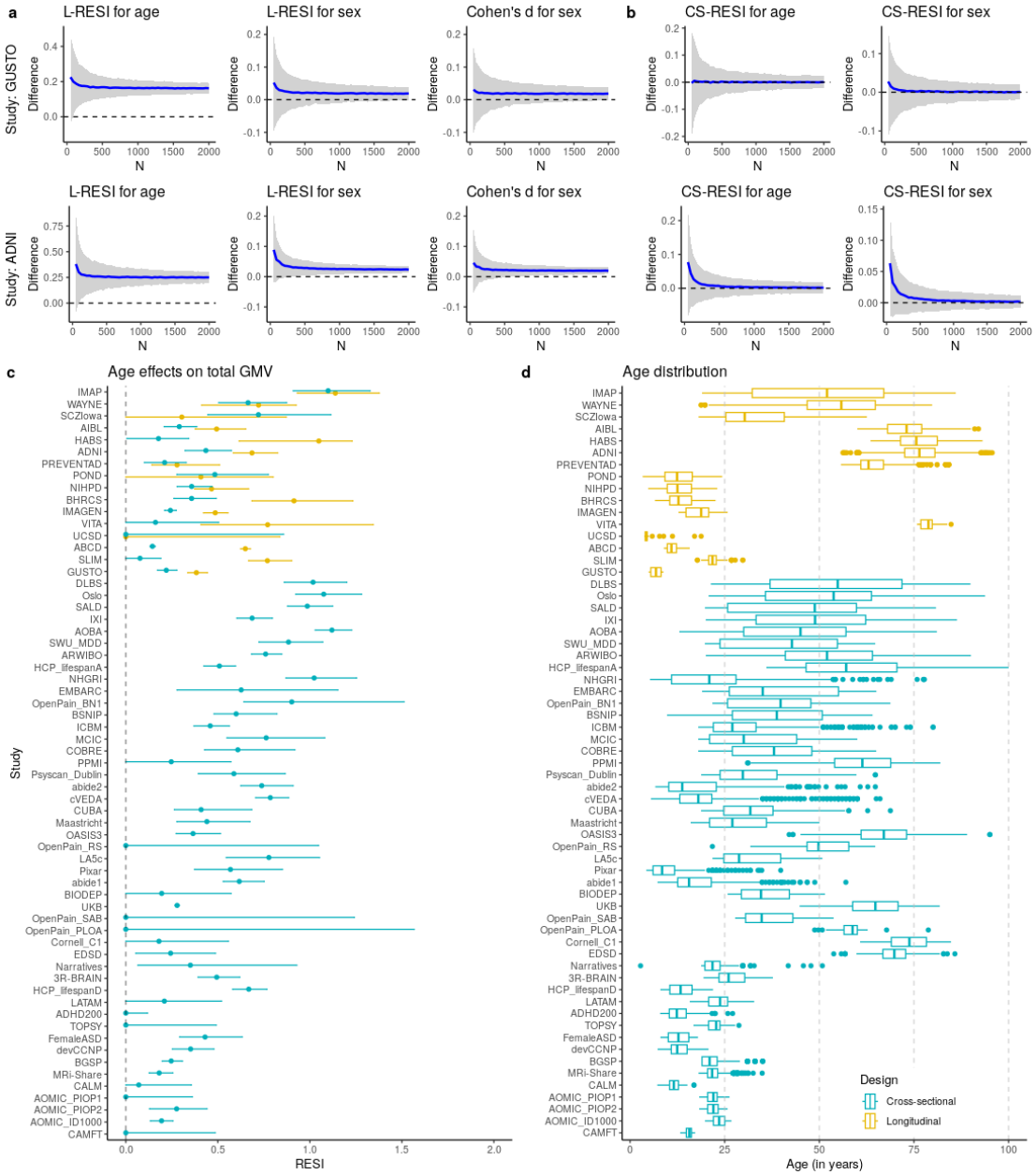

Fig. 2. (**a**) Bootstrap analysis shows that effect sizes of age and sex on total gray matter volume (GMV) in longitudinal studies are systematically larger than effect sizes in cross-sectional studies (Difference = L-RESI - RESI) and (**b**) the CS-RESI for longitudinal analysis estimates and removes the difference (Difference = CS-RESI - RESI). We conduct 1,000 bootstraps per sample size (indicated by N on the x-axes). The blue curves represent the mean difference curves and the gray areas are the 95% confidence bands from the bootstraps. (**c**) Forest plot of the effect sizes for total GMV-age association shows that the CS-RESIs in longitudinal studies are more comparable to the regular RESIs in cross-sectional studies. Regular RESIs (blue) are estimated for cross-sectional

studies, CS-RESIs (blue) and L-RESIs (yellow) are both estimated for longitudinal studies. The dots and lines represent RESI estimates and their 95% confidence intervals (CIs). The studies are sorted by study design and decreasing standard deviation (SD) of age in each study. (**d**) The age distribution in each study; studies with smaller SD of age tend to have smaller age effects.

Improved sampling schemes can increase ESs

Consistent with statistical principles, the meta-analysis results above suggest that the larger variability of the primary covariate produces larger standardized ESs. To investigate the effects of modifying the variability of the age covariate on ESs in BWAS, we implement three sampling schemes that produce different sample SD of the age covariate, treating the large-scale cross-sectional United Kingdom Biobank (UKB) data as the population and drawing samples whose age distributions follow a pre-specified shape (bell-shaped, uniform, and U-shaped; Fig. S6; also see Methods). In UKB, the U-shaped sampling scheme on age increases the total GMV-age ES by 60% compared to bell-shaped and by 27% compared to uniform (Fig. 3a), with an associated increase in power (Fig. 3b). To achieve 80% power to detect the total GMV-age association, <100 subjects is sufficient if using the U-shaped sampling scheme, whereas about 200 subjects are needed if the bell-shaped sampling scheme is used (Fig. 3b). A similar pattern can be seen for the regional outcomes of GMV and CT (Fig. 3c-f). The U-shaped sampling scheme typically provides the largest ESs of age and the highest power, followed by the uniform and bell-shaped schemes. The region-specific improvement in GMV-age ES and power by using the U-shaped sampling scheme compared with the bell-shaped can be found in Fig. S7.

To investigate the effect of increasing the variability of the age covariate longitudinally, we implement sampling schemes to adjust the between- and within-subject variability of age in the bootstrap samples from the longitudinal ADNI dataset. In the bootstrap samples, each subject has two measurements (baseline and a follow-up). To imitate the true operation of a study, we select each subject's two measurements based on baseline age and the follow-up age by targeting specific distributions for the baseline age and the age change at the follow-up time point (Fig. S8; Methods). We evaluate the effect on the L-RESI because it is directly related to power in longitudinal studies (see Methods: Sample size calculation for a target power with a given effect size). Increasing between-subject and within-subject variability of age increases the average observed ESs, with corresponding increases in power (Fig. 4a-b). U-shaped between-subject sampling scheme on age increases the total GMV-age ES by 23.6% compared to bell-shaped and by 12.1% compared to uniform, when using uniform within-subject sampling scheme (Fig. 4a).

Additionally, we investigate the effect of the number of measurements per subject on ES and power in longitudinal data using the ADNI dataset. Adding a single additional measurement after the baseline provides the biggest improvement in total GMV-age ES by 156% and power by 112%, the benefit of additional measurements is minimal (Fig. 4c-d). Finally, we also evaluate the effects of the longitudinal sampling schemes on regional GMV and CT in ADNI (Fig. 4e-h). When sampling two measurements per subject, the between- and within-subject sampling schemes producing larger age variability increase the ES and power across all regions.

Together, these results suggest having larger spacing in between- and within-subject age measurements increases ES and power. Most of the benefit of the number of within-subject measurements in a longitudinal study is due to the first additional measurement after baseline.

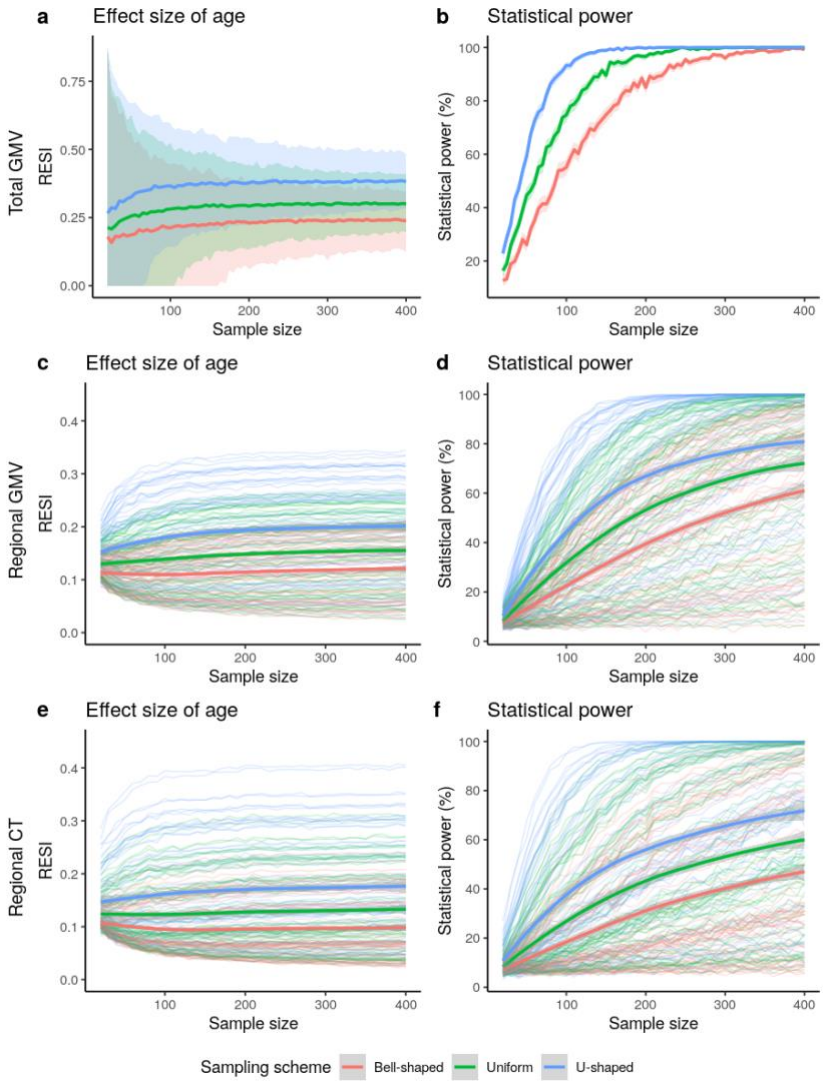

Fig. 3. Improved effect sizes and statistical power for age associations with different brain measures: total gray matter volume (GMV) (a-b), regional GMV (c-d), and regional cortical thickness (CT) (e-f), under three sampling schemes in the UKB study (*N* = 29,031 for total GMV and *N* = 29,030 for regional GMV and CT). The sampling schemes target different age distributions to increase the variance of age: Bell-shaped<Uniform<U-shaped (Fig. S6). Using the resampling schemes that increase age variance increases (**a**) the total GMV-age effect sizes, and (**b**) statistical power (at significance level of 0.05). The same result holds for (**c-d**) regional GMV and (**e-f**) regional CT. The curves represent the average effect size or estimated power at a given sample size and sampling scheme. The shaded areas represent the corresponding 95% confidence bands. (**c-f**) The bold curves are the average effect size or power across all regions with significant uncorrected effect sizes using the full UKB data.

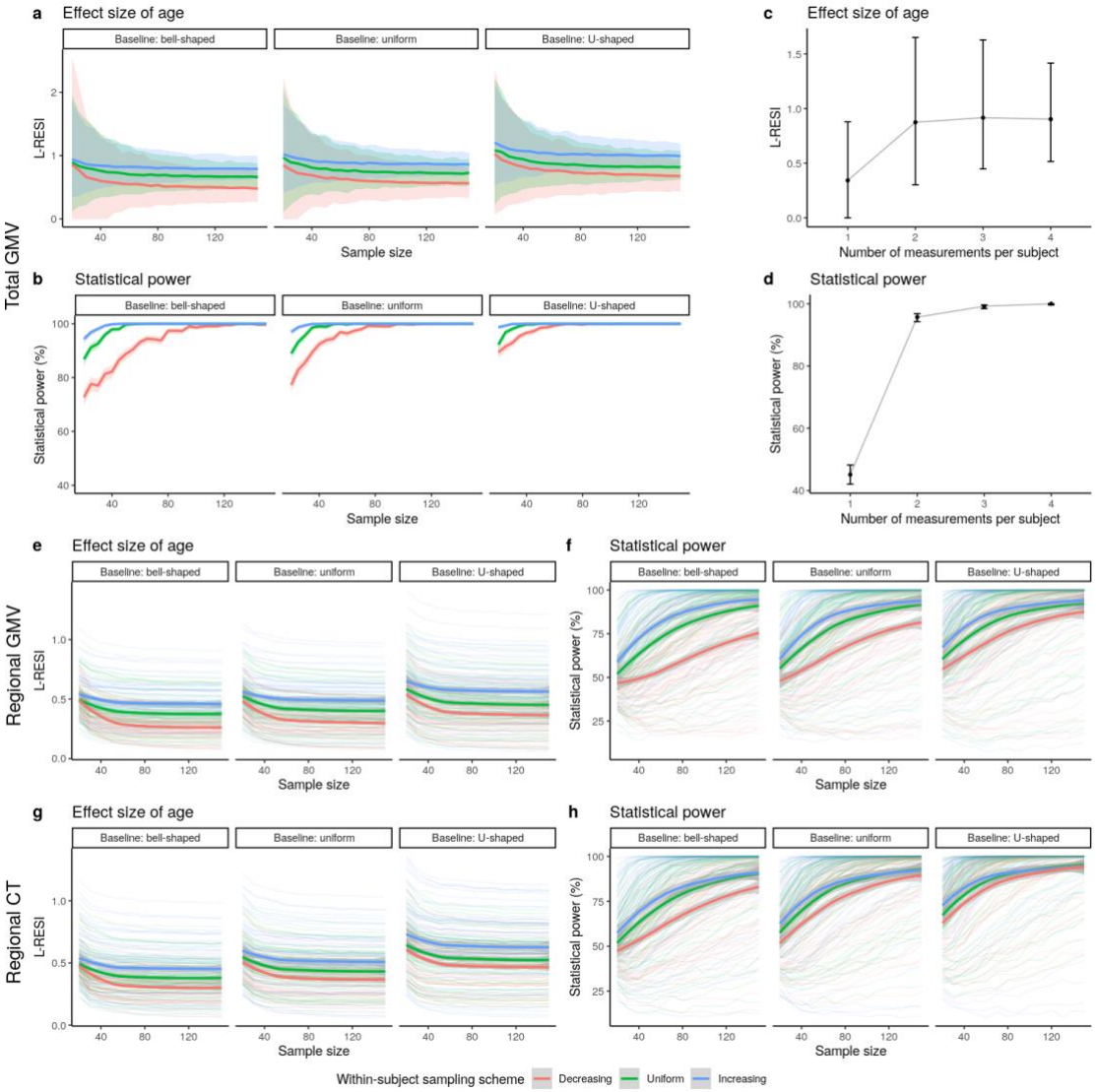

Fig. 4. Improved effect sizes and statistical power for age associations with structural brain measures under different longitudinal sampling schemes in the ADNI data. Three different sampling schemes (Fig. S8) are implemented in bootstrap analyses to modify the distributions of baseline age and the age change from baseline and therefore to modify the between- and within-subject variability of age. (**a**) Implementing the sampling schemes results in higher (between- and/or within-subject) variability and increases the total gray matter volume (GMV)-age effect size and (**b**) statistical power at significance level of 0.05. The curves represent the average effect size or estimated power and the gray areas are the 95% confidence bands across the 1,000 bootstraps. (**c-d**) Increasing the number of measurements from one to two per subject provides the most benefit on (**c**) the total GMV-age effect size and (**d**) statistical power when using uniform between- and within-subject sampling schemes and N = 30. The points represent the mean effect size or estimated power and the whiskers are the 95% confidence intervals. (**e-h**) Improved effect sizes and power for age associations with regional (**e-f**) GMV and (**g-h**) cortical thickness (CT), across all brain regions under different sampling schemes. The bold curves are the average effect size or estimated power across all regions with significant uncorrected effect sizes using the full ADNI data. The gray areas are the corresponding 95% confidence bands. Increasing the between- and within-subject variability of age by implementing different sampling schemes can increase the effect size of age, and the associated statistical power, on regional GMV and regional CT.

Preferred sampling schemes and longitudinal designs depend on state versus trait associations between brain and non-brain measures

To explore if the proposed sampling schemes are effective on a variety of non-brain covariates and their associations with structural and functional brain measures, we implement the sampling schemes on subjects (cognitively normal and atypical) with cross-sectional and longitudinal measurements from the ABCD dataset. The non-brain covariates include the NIH toolbox[36], CBCL, body mass index (BMI), birthweight, and handedness (See Tab. S11 and Tab. S12 for the descriptive summary; Methods). Functional connectivity (FC) is used as a functional brain measure and is computed for all pairs of regions in the Gordon atlas[37] (Methods). We use the bell- and U-shaped target sampling distributions to control the between- and within-subject variability of each non-brain covariate (Methods). For each non-brain covariate, we show the results for the four combinations of between- and within-subject sampling schemes. Overall, there is a consistent benefit to increasing between-subject variability of the covariate (Fig. 5), which leads to >1.8 factor reduction in sample size needed to scan and >1.4 factor increase in ES for over 50% of associations. Moreover, 72.1% of covariate-outcome associations had increased ESs by increasing between-subject sampling (Fig. S9).

Surprisingly, increasing within-subject variability decreases ESs for many structural associations (Fig. 5a-f), suggesting that conducting longitudinal analyses can result in decreased power compared to cross-sectional analyses. To confirm this, we compare the ESs of the non-brain covariates on each type of brain measure estimated using only the baseline measurements to the ESs estimated using the full longitudinal data (Fig. 6a-d). Consistent with the reduction in ES by increasing within-subject variability, for most structural associations (GMV and CT), conducting cross-sectional analyses using only the baseline measures results in larger ESs (and therefore, higher power and replicability) than conducting longitudinal analyses using the full longitudinal data (Fig. 6a-c). For the FC outcomes, there is a slight positive effect of conducting longitudinal analyses (Fig. 6d). Identical results are found using linear mixed models (LMMs) with individual-specific random intercepts, which are commonly used in BWAS (Fig. S10). Together, these results suggest that the benefit of conducting longitudinal studies and adjusting the within-subject variability is highly dependent on the brain-behavior association and, counterintuitively, longitudinal designs can reduce ESs.

To investigate why increasing within-subject variance or using longitudinal designs is not beneficial for some associations, we examined an assumption common to widely used LMMs and GEEs in BWAS: there is consistent association strength between the brain measure and non-brain covariate across between- and within-subject changes in the non-brain covariate. These effects are independent of the random effects parameters and can differ if there are different state or trait associations between the brain measure and non-brain covariate. For example, measures of crystallized intelligence are subject to within-individual differences in state more than total GMV, which has low state variance so there is no within-subject association between these variables (Tab. S13). In contrast, FC measures are also subject to within-individual differences in state so can have stronger within-subject associations with crystallized intelligence. To demonstrate this, we fit models that estimate distinct effects for between- and within-subject associations in the ABCD (Methods) and find that there are large between-subject parameter estimates and small within-subject parameter estimates in total and

regional GMV (Supplement: Section 3.2; Tab. S13; Fig. 6e), whereas the FC associations are distributed more evenly across between- and within-subject parameters (Fig. 6f). If the between- and within-subject associations are different, fitting these associations separately avoids averaging the larger effect with the smaller effect and can inform our understanding of brain-behavior associations (Supplement: Section 3.2). This approach ameliorates the reduction in ESs caused by longitudinal designs for structural brain measures in the ABCD (Fig. S11). See Supplemental Section 3 for technical details.

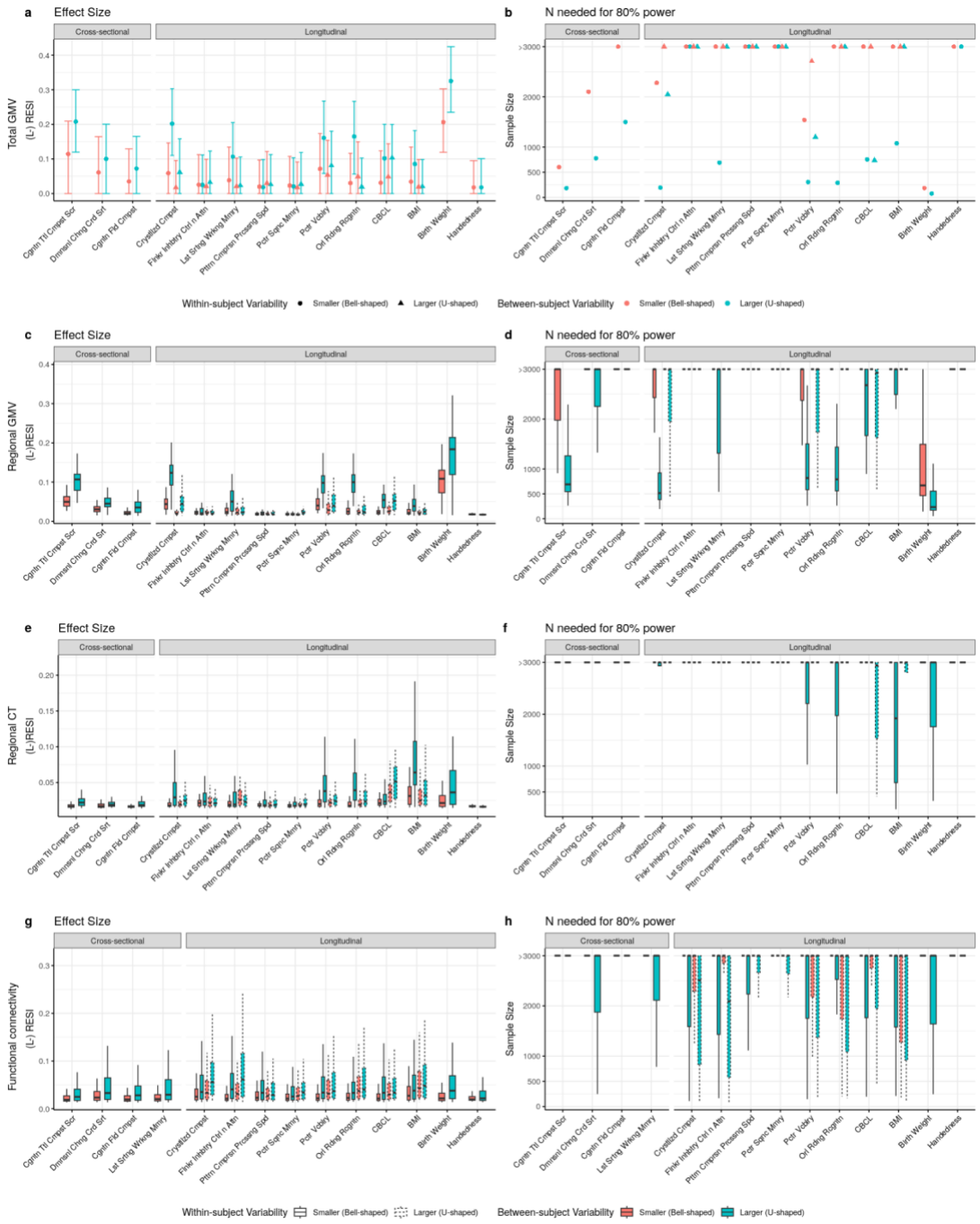

Fig. 5. Heterogeneous improvement of effect sizes for cognitive, mental health, and demographic associations with structural and functional brain measures in the ABCD study at *N* = 500. (**a**) U-shaped between-subject sampling scheme (blue) that increases between-subject variability of the non-brain covariate produces larger effect sizes and (**b**) reduces the number of participants scanned to obtain 80% power in total gray matter volume (GMV). The points and triangles are the average effect sizes across bootstraps and the whiskers are the 95% confidence intervals. Increasing within-subject sampling (triangles) can reduce effect sizes. A similar pattern holds in (**c-d**) regional GMV and (**e-f**) regional cortical thickness (CT); boxplots show the distributions of the effect sizes across regions. In contrast, (**g**) regional pairwise functional connectivity (FC) effect sizes increase by increasing between- (blue) and within-subject variability (dashed borders) with a corresponding reduction in the (**h**) number of participants scanned for 80% power.

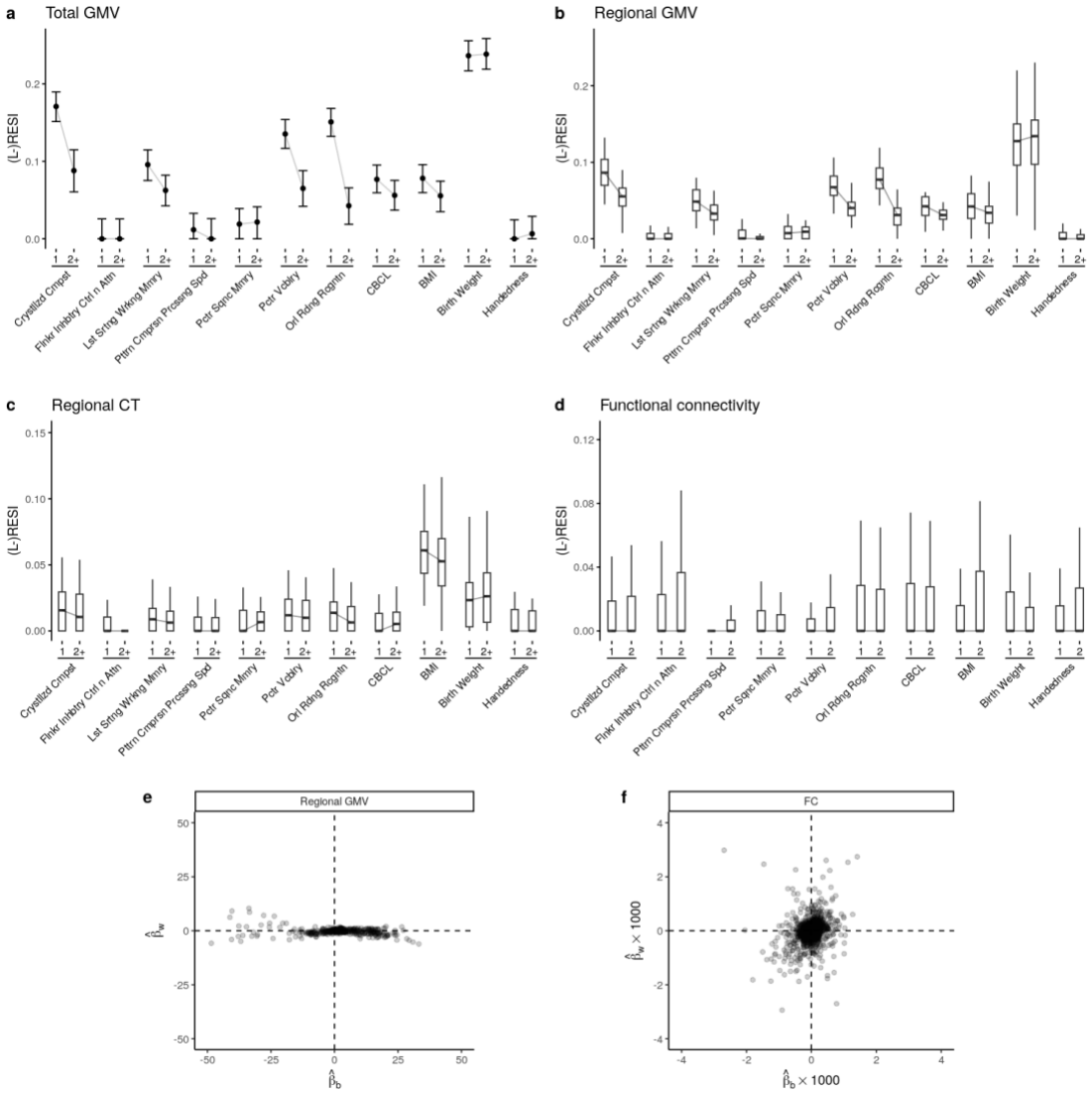

Fig. 6. Longitudinal study designs can reduce effect sizes and replicability due to heterogeneous between- and within-subject associations between brain and behavior

measures. Boxplots show the distributions of the effect sizes across regions. (**a-c**) Cross-sectional analyses (indicated by "1"s on the x-axes) can have larger effect sizes than the same longitudinal analyses (indicated by "2" or "2+") for structural brain measures in ABCD. (**d**) The functional connectivity (FC) measures have a slight benefit of longitudinal modeling. The cross-sectional analyses only use the baseline measures; the longitudinal analyses use the full longitudinal data. (**e**) Most regional gray matter volume (GMV) associations (Fig. 5c) have larger between-subject parameter estimates (β_b, x-axis) than within-subject parameter estimates (β_w, y-axis; see Supplement: Eqn (13)), whereas (**f**) FC associations (Fig. 5g) show less heterogeneous relationships between the two parameters.

Discussion

BWAS are a fundamental tool to discover brain-behavior associations that may inform our understanding of brain disorders and their treatments. As the evidence of small ESs in BWAS increases, optimizing study design to improve standardized ESs is critical to understanding brain-behavior associations that might inform research on brain health across the lifespan[4,38,39]. We use the RESI to show the influence of study design features on ESs of age, sex, and behavioral measures with structural and functional brain measures using diverse neuroimaging datasets from LBCC. Our results indicate that increasing between-subject SD of the covariate of interest systematically increases standardized ESs and that the benefit of performing a longitudinal study strongly depends on the nature of the association. We use statistical principles to show that typical longitudinal models used in BWAS conflate between- and within-subject effects. When we model between- and within-subject effects separately we can increase power for both the between- or within-subject effects. We propose a CS-RESI for longitudinal analysis that can be used to estimate the benefit in ES of a longitudinal design and to correct for ES differences between study designs when comparing longitudinal and cross-sectional studies in the same study population. While large sample sizes and increased scan duration remain useful considerations, our findings emphasize the importance of considering other study design features in BWAS as well to improve standardized ESs and replicability[14,38].

Increasing between-subject variability to increase effect sizes

Our results demonstrate that ESs can be increased by sampling subjects with small and large values of the covariate of interest with higher probability. This is well-known in linear models where the ES is explicitly a function of the SD of the covariate[40]. We show that increasing a variable's SD improves ESs by a median factor of 1.4, even when there is nonlinearity in the association, such as with age and GMV (Fig. S1). When the association is very non-monotonic, e.g., if there is a U-shape relationship between covariate and outcome, sampling the tails more heavily could decrease power, and it also may decrease our ability to detect nonlinearities in the center of the study population. In this case, sampling to obtain a uniform distribution of the covariate balances power across the range of the covariate and can still increase power relative to random sampling when the covariate has a bell-shaped or normal distribution in the population. Increasing between-subject variability is beneficial in more than 72% of the association pairs we studied, despite potential nonlinearities (Figure S9).

It is important to note that, in this study, we endeavor to increase ES *parameters*, which are unknown values quantifying the strength of associations, by intentionally modifying design features, such as the covariate variability to increase the efficiency of the unstandardized effect

estimate (see Supplement: Section 1). Larger ES parameters are always preferable in hypothesis testing as they have higher replicability. In contrast, larger ES *estimates* do not necessarily imply greater generalizability or replicability, as they can be positively biased due to a small sample size or by the analysis procedures (such as by multiple testing, biased reporting, or *p*-hacking). Unstandardized effects (such as regression coefficients), which are often related to the biological interpretation of the population, are not affected by the sampling modifications used here and can be unbiasedly estimated using correctly specified models[41].

Because standardized ESs are directly related to study efficiency, design considerations that optimize study efficiency will increase ESs and study replicability. Two-phase, extreme group, and outcome-dependent sampling designs can inform what subjects should be selected for imaging from a larger sample in order to increase the efficiency and ESs of brain-behavior associations[42–48]. When there are multiple covariates of interest, multivariate optimal sampling designs can be used to increase ESs[49]. Multivariate designs are also needed to stratify sampling to avoid confounding by sociological variables. As we see here, the use of optimal designs increases power, but positively biases the standardized ES parameters relative to their population values due to manipulation of the sample variance[46]. Weighted regression or other statistical methods can be used to compute unbiased ES estimates that account for sampling design[49–51]. When there are concerns about implementing optimal design strategies, choosing optimal psychometric measurements or interventions (e.g., medications or neuromodulation within a clinical trial) may also be effective for increasing ESs for a target association. The decision to pursue a particular design will depend on the costs and complexity of acquiring the non-brain covariates of interest and translational goals that may require population representative ES estimates.

Longitudinal design considerations

In the meta-analysis, longitudinal studies of the total GMV-age effect have, on average, >380% larger standardized ESs than cross-sectional studies. However, we see in subsequent analyses that the benefit of conducting a longitudinal design is highly dependent on the between- and within-subject effects. When the between- and within-subject effect parameters are equal and the within-subject brain measure error is low, longitudinal studies offer larger ESs than cross-sectional studies (Supplement: Section 3.1)[52]. This is the reason there is a benefit of longitudinal design in the ADNI in the total GMV-age association (Fig. S12). While the ES in these analyses did not account for cost per measurement; comparing efficiency per measurement analytically (Supplement: Section 3.1) supports the approach of collecting two measurements per subject in this scenario.

When the between- and within-subject effects differ but we still fit them with a single effect, the coefficient for the covariate does not have the same interpretation (as when the effects are equal) because it is a weighted average of the between- and within-subject effects whose weights are determined by the study design features (see Supplement: Section 3). This difference in interpretation affects both the unstandardized effects and standardized ESs. The apparent lack of benefit in the GMV-psychometric measurement association in the ABCD is due to a lack of association between within-subject changes in the psychometric measurements with within-subject changes in GMV (Fig. 6e; Tab. S13). The smaller ESs we find in longitudinal analyses are due to the larger weight on the smaller within-subject effect for the weighted

average effect in the longitudinal model (see Supplement: Section 3 Eqn. (14)). As shown in our analyses, fitting the between- and within-subject effects separately prevents conflating the two effects.

Longitudinal models offer the unique ability to separately estimate between- and within-subject associations. When estimating these effects separately, investigators can use a GEE or LMM with any working covariance structure which will yield unbiased estimates of the between- and within-subject effects if the model is correctly specified. Using sampling strategies to increase between- and within-subject variance will separately increase ESs for the between- and within-subject associations, respectively. Estimates of between- and within-subject associations will not be directly comparable to estimates obtained from a cross-sectional model because they have a different interpretation (See Supplement: Section 3.2)[53–55]. We discuss several other reasonable modeling approaches for longitudinal BWAS in the Supplement (Section 3).

Conclusions

It is difficult to provide general recommendations for design and analysis that apply to all studies, and study design must consider other features in addition to optimizing ESs, such as practical challenges in collecting longitudinal data. Nevertheless, the present study provides guidelines for designing and analyzing BWAS for optimal ESs (Fig. S13 and S14), based on our empirical and theoretical results. Although the decision for a particular design or analysis strategy will depend on unknown features of the brain and non-brain measures and their association, these characteristics can be evaluated in pilot data or the analysis dataset (Fig. S12; Supplement: Section 3.1). Broadly, increasing the covariate SD in the study population has a deterministic increase in standardized ESs for most associations so is a practical approach, but the benefit of a longitudinal design is dependent on the form of the association between the covariate and outcome. Our findings highlight the importance of rigorous evaluation of statistical designs and analysis methods to improve power and replicability of BWAS.

Reference

1. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
2. Owens, M. M. *et al.* Recalibrating expectations about effect size: A multi-method survey of effect sizes in the ABCD study. *PLOS ONE* **16**, e0257535 (2021).
3. Nour, M. M., Liu, Y. & Dolan, R. J. Functional neuroimaging in psychiatry and the case for failing better. *Neuron* **110**, 2524–2544 (2022).
4. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate

sample sizes. *Nature* **615**, E4–E7 (2023).

5. National Academies of Sciences, E. *et al. Understanding Reproducibility and Replicability*. *Reproducibility and Replicability in Science* (National Academies Press (US), 2019).
6. Nosek, B. A. *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
7. Patil, P., Peng, R. D. & Leek, J. T. What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspect. Psychol. Sci.* **11**, 539–544 (2016).
8. Liu, S., Abdellaoui, A., Verweij, K. J. H. & van Wingen, G. A. Replicable brain–phenotype associations require large-scale neuroimaging data. *Nat. Hum. Behav.* **7**, 1344–1356 (2023).
9. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
10. Szucs, D. & Ioannidis, J. P. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. *NeuroImage* **221**, 117164 (2020).
11. Reddan, M. C., Lindquist, M. A. & Wager, T. D. Effect Size Estimation in Neuroimaging. *JAMA Psychiatry* **74**, 207–208 (2017).
12. Vul, E. & Pashler, H. Voodoo and circularity errors. *NeuroImage* **62**, 945–948 (2012).
13. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect. Psychol. Sci.* **4**, 274–290 (2009).
14. Ooi, L. Q. R. *et al.* MRI economics: Balancing sample size and scan duration in brain wide association studies. *BioRxiv Prepr. Serv. Biol.* 2024.02.16.580448 (2024) doi:10.1101/2024.02.16.580448.
15. NIH RePORTER.

https://reporter.nih.gov/search/3UVwEyrEZ0y6i8_NCs_ucg/projects/charts.

16. Nee, D. E. fMRI replicability depends upon sufficient individual-level data. *Commun. Biol.* **2**, 1–4 (2019).

17. Smith, P. L. & Little, D. R. Small is beautiful: In defense of the small-N design. *Psychon. Bull. Rev.* **25**, 2083–2101 (2018).

18. Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M. & Mills, K. L. Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Dev. Cogn. Neurosci.* **47**, 100902 (2021).

19. Bethlehem, R. a. I. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).

20. The ADHD-200 Consortium. The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. *Front. Syst. Neurosci.* **6**, 62 (2012).

21. Di Martino, A. *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).

22. Lukas Snoek *et al.* AOMIC-PIOP1. *OpenNeuro* (2020) doi:doi:10.18112/openneuro.ds002785.v2.0.0.

23. Bilder, R *et al.* UCLA Consortium for Neuropsychiatric Phenomics LA5c Study. *OpenNeuro* (2020) doi:doi:10.18112/openneuro.ds000030.v1.0.0.

24. Samuel A. Nastase; *et al.* Narratives. *OpenNeuro* (2020) doi:doi:10.18112/openneuro.ds002345.v1.1.4.

25. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* **4**, 170181 (2017).

26. Lukas Snoek; *et al.* AOMIC-PIOP2. *OpenNeuro* (2020) doi:doi:10.18112/openneuro.ds002790.v2.0.0.

27. Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. Development of the social brain from age three to twelve years. *Nat. Commun.* **9**, 1027 (2018).

28. Kuklisova-Murgasova, M. *et al.* A dynamic 4D probabilistic atlas of the developing brain. *NeuroImage* **54**, 2750–2763 (2011).

29. Lukas Snoek; *et al.* AOMIC-ID1000. *OpenNeuro* (2021) doi:doi:10.18112/openneuro.ds003097.v1.2.1.

30. Reynolds, J. E., Long, X., Paniukov, D., Bagshawe, M. & Lebel, C. Calgary Preschool magnetic resonance imaging (MRI) dataset. *Data Brief* **29**, 105224 (2020).

31. Vandekar, S., Tao, R. & Blume, J. A Robust Effect Size Index. *Psychometrika* **85**, 232–246 (2020).

32. Kang, K. *et al.* Accurate Confidence and Bayesian Interval Estimation for Non-centrality Parameters and Effect Size Indices. *Psychometrika* (2023) doi:10.1007/s11336-022-09899-x.

33. Jones, M., Kang, K. & Vandekar, S. RESI: An R Package for Robust Effect Sizes. Preprint at https://doi.org/10.48550/arXiv.2302.12345 (2023).

34. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).

35. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Erlbaum Associates, Hillsdale, NJ, 1988).

36. Carlozzi, N. E. *et al.* Construct Validity of the NIH Toolbox Cognition Battery in Individuals With Stroke. *Rehabil. Psychol.* **62**, 443–454 (2017).

37. Gordon, E. M. *et al.* Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cereb. Cortex N. Y. NY* **26**, 288–303 (2016).

38. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).

39. Noble, S., Mejia, A. F., Zalesky, A. & Scheinost, D. Improving power in functional

magnetic resonance imaging by moving beyond cluster-level inference. *Proc. Natl. Acad. Sci.* **119**, e2203020119 (2022).

40. Boos, D. D. & Stefanski, L. A. *Essential Statistical Inference: Theory and Methods*. (Springer-Verlag, New York, 2013).

41. Lawless, J. F., Kalbfleisch, J. D. & Wild, C. J. Semiparametric Methods for Response-Selective and Missing Data Problems in Regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61**, 413–438 (1999).

42. Tao, R., Zeng, D. & Lin, D.-Y. Optimal Designs of Two-Phase Studies. *J. Am. Stat. Assoc.* **115**, 1946–1959 (2020).

43. Schildcrout, J. S., Garbett, S. P. & Heagerty, P. J. Outcome Vector Dependent Sampling with Longitudinal Continuous Response Data: Stratified Sampling Based on Summary Statistics. *Biometrics* **69**, 405–416 (2013).

44. Tao, R., Zeng, D. & Lin, D.-Y. Efficient Semiparametric Inference Under Two-Phase Sampling, With Applications to Genetic Association Studies. *J. Am. Stat. Assoc.* **112**, 1468–1476 (2017).

45. Fisher, J. E., Guha, A., Heller, W. & Miller, G. A. Extreme-groups designs in studies of dimensional phenomena: Advantages, caveats, and recommendations. *J. Abnorm. Psychol.* **129**, 14–20 (2020).

46. Preacher, K. J., Rucker, D. D., MacCallum, R. C. & Nicewander, W. A. Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychol. Methods* **10**, 178–192 (2005).

47. Amanat, S., Requena, T. & Lopez-Escamez, J. A. A Systematic Review of Extreme Phenotype Strategies to Search for Rare Variants in Genetic Studies of Complex Disorders. *Genes* **11**, 987 (2020).

48. Lotspeich, S. C., Amorim, G. G. C., Shaw, P. A., Tao, R. & Shepherd, B. E. Optimal

multiwave validation of secondary use data with outcome and exposure misclassification. *Can. J. Stat.* **n/a**, (2023).

49. Tao, R. *et al.* Analysis of Sequence Data Under Multivariate Trait-Dependent Sampling. *J. Am. Stat. Assoc.* **110**, 560–572 (2015).

50. Lin, H. *et al.* Strategies to Design and Analyze Targeted Sequencing Data. *Circ. Cardiovasc. Genet.* **7**, 335–343 (2014).

51. Tao, R., Lotspeich, S. C., Amorim, G., Shaw, P. A. & Shepherd, B. E. Efficient semiparametric inference for two-phase studies with outcome and covariate measurement errors. *Stat. Med.* **40**, 725–738 (2021).

52. Peter Diggle, Patrick Heagerty, Kung-Yee Liang, & Scott Zeger. *Analysis of Longitudinal Data*. vol. Second edition (OUP Oxford, Oxford, 2013).

53. Pepe, M. S. & Anderson, G. L. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Commun. Stat. - Simul. Comput.* (1994) doi:10.1080/03610919408813210.

54. Begg, M. D. & Parides, M. K. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat. Med.* **22**, 2591–2602 (2003).

55. Curran, P. J. & Bauer, D. J. The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change. *Annu. Rev. Psychol.* **62**, 583–619 (2011).

56. Nielsen, A. N., Barch, D. M., Petersen, S. E., Schlaggar, B. L. & Greene, D. J. Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 791–798 (2020).

57. Feczko, E. *et al.* Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities. 2021.07.09.451638 Preprint at https://doi.org/10.1101/2021.07.09.451638 (2021).

58. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging

acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).

59. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

60. Beer, J. C. *et al.* Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage* **220**, 117129 (2020).

61. Jean-Philippe Fortin. neuroCombat: Harmonization of multi-site imaging data with ComBat. *R package version 1.0.13* (2023).

62. Joanne Beer. longCombat: Longitudinal ComBat for Harmonizing Multi-batch Longitudinal Data. *R package version 0.0.0.90000* (2020).

63. Højsgaard, S., Halekoh, U., Yan, J. & Ekstrøm, C. T. geepack: Generalized Estimating Equation Package. (2022).

64. Long, J. S. & Ervin, L. H. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *Am. Stat.* **54**, 217–224 (2000).

65. Agresti, A. & Coull, B. A. Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions. *Am. Stat.* **52**, 119–126 (1998).

Methods

LBCC dataset and processing

The original LBCC dataset includes 123,984 MRI scans from 101,457 human participants across more than 100 studies and is described in our prior work[19] (see Supplemental Information and Tab. S1 from prior study). We filtered to the subset of cognitively normal (CN) participants whose data were processed using FreeSurfer version 6.1. Studies were curated for the analysis by excluding duplicated observations and studies with less than 4 unique age points, sample size less than 20, and/or only participants of one sex. If there were fewer than three participants having longitudinal observations, only the baseline observations were included and the study was considered cross-sectional. If a subject had changing demographic information during the longitudinal follow-up (e.g., changing biological sex), only the most recent observation was included. We updated the LBCC dataset with the ABCD release 5 resulting in a final dataset that includes 77,695 MRI scans from 60,900 CN participants who have available total GMV, sGMV and GMV measures across 63 studies (Tab. S1), among which, 74,148 MRI scans from 57,538 participants across 43 studies have complete-case regional brain measures

(regional GMV, regional surface area and regional cortical thickness (CT), based on Desikan-Killiany parcellation[34]) (Tab. S2). The global brain measure mean CT was derived using the regional brain measures (see below).

The LBCC dataset used in the preparation of this article includes data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Structural brain measures

Details of data processing are described in our prior work[19]. Briefly, total GMV, sGMV, and WMV were estimated from T1- and T2-weight (when available) MRIs using the "aseg" output from Freesurfer 6.0.1. All three cerebrum tissue volumes were extracted from the aseg.stats files output by the recon-all process: 'Total cortical gray matter volume' for GMV; 'Total cortical/cerebral (FreeSurfer version dependent) white matter volume' for WMV; and 'Subcortical gray matter volume' for sGMV (inclusive of thalamus, caudate nucleus, putamen, pallidum, hippocampus, amygdala, and nucleus accumbens area; https://freesurfer.net/fswiki/SubcorticalSegmentation). Regional GMV and CT across 68 regions (34 per hemisphere, based on Desikan-Killiany parcellation[34]) were obtained from the aparc.stats files output by the recon-all process. Mean CT across the whole brain is the weighted average of the regional CT weighted by the corresponding regional surface areas.

Preprocessing specific to ABCD

Functional connectivity measures

Longitudinal functional connectivity (FC) measures were obtained from the ABCD-BIDS community collection, which houses a community-shared and continually updated ABCD neuroimaging dataset available under Brain Imaging Data Structure (BIDS) standards. The data used in these analyses were processed using the abcd-hcp-pipeline version 0.1.3, an updated version of The Human Connectome Project MRI pipeline[57]. Briefly, resting state fMRI time series were demeaned and detrended and a GLM was used to regress out mean WM, CSF, and global signal as well as motion variables and then band-pass filtered. High motion frames (filtered-FD > 0.2mm) were censored during the demeaning and detrending. After preprocessing the time series were parcellated using the 352 regions of the Gordon atlas (including 19 subcortical structures) pairwise Pearson correlations were computed among the regions. FC measures were estimated from resting state fMRI time series using a minimum of 5-minutes of data. After Fisher z-transformation, the connectivities were averaged across the 24 canonical functional networks[37], forming 276 inter-network connectivities and 24 intra-network connectivities.

Cognitive and other covariates

The ABCD dataset is a large-scale repository aiming to track the brain and psychological development of over 10,000 children ages 9 to 16 by measuring hundreds of variables including

demographic, physical, cognitive and mental health variables[58]. We used the release 5 of ABCD study to examine the effect of the sampling schemes on other types of covariates including cognition (fully corrected *T*-scores of the individual subscales and total composite scores of NIH Toolbox[36]), mental health (total problem CBCL syndrome scale), and other common demographic variables (body mass index (BMI), birth weight, and handedness). For each of the covariates, we evaluated the effect of the sampling schemes on their associations with the global and regional structural brain measures and FC after controlling for nonlinear age and sex (and, for FC outcomes only, mean frame displacement (FD)).

For structural brain measures, there were 3 non-brain covariates with less than 5% non-missing follow-ups at both 2-year and 4-year follow-ups (i.e., Dimensional Change Card Sort Test, Cognition Fluid Composite, and Cognition Total Composite Score; Tab. S11), and only their baseline cognitive measurements were included in the analyses. For the remaining 11 variables (i.e., Picture Vocabulary Test, Flanker Inhibitory Control and Attention Test, List Sorting Working Memory Test, Pattern Comparison Processing Speed Test, Picture Sequence Memory Test, Oral Reading Recognition Test, Crystallized Composite, birth weight, BMI and handedness), all of the available baseline, 2-year and 4-year follow-up observations were used. For the FC, only the baseline observations for List Sorting Working Memory Test were used due to missingness (Tab. S12).

The records with BMI lying outside the lower and upper 1% quantiles (i.e., BMI < 13.5 or BMI > 36.9) were considered misinput and replaced with missing values. The variable handedness was imputed using last observation carried forward method.

Statistical analysis

Removal of site effects

For multi-site or multi-study neuroimaging studies, we need to control for potential heterogeneity between sites to obtain unconfounded and generalizable results. Before estimating the main effects of age and sex on the global and regional brain measures (total GMV, total WMV, total sGMV, mean CT, regional GMV and regional CT), we applied ComBat[59] and LongComBat[60] in cross-sectional datasets and longitudinal datasets, respectively, to remove the potential site effects. The ComBat algorithm involves several steps including data standardization, site effect estimation, empirical Bayesian adjustment, removing estimated site effects, and data rescaling. In the analysis of cross-sectional datasets, the model for ComBat was specified as a linear regression model illustrated below using total GMV:

$$\text{GMV} \sim \text{ns}(\text{age}, \text{df} = 2) + \text{sex} + \text{site},$$

where ns denotes natural cubic splines on two degrees of freedom, which means that there were two boundary knots and one interval knot placed at the median of the covariate age. Splines were used to accommodate nonlinearity in the age effect. For the longitudinal datasets, the model for LongComBat used a linear mixed effects model with subject-specific random intercepts:

$$\text{GMV} \sim (1 \mid \text{subject}) + \text{ns}(\text{age}, \text{df} = 2) + \text{sex} + \text{site}.$$

When estimating the effects of other non-brain covariates in the ABCD dataset, ComBat was used to control the site effects respectively for each of the cross-sectional covariates. The ComBat model is specified as illustrated below using GMV:

$$\text{GMV} \sim \text{ns}(\text{age}, \text{df} = 2) + \text{sex} + \text{x} + \text{site},$$

where x denotes the non-brain covariate. LongComBat was used for each of the longitudinal covariates with a linear mixed effects model with subject-specific random intercepts only:

GMV ~ (1 | subject) + ns(age, df = 2) + sex + x + site.

When estimating the effects of other covariates on the FC in the ABCD data, we additionally controlled for the mean frame displacement (FD) of the frames remaining after scrubbing. The longComBat model is specified as:

FC ~ (1 | subject) + ns(age, df = 2) + sex + ns(mean_FD, df = 3) + x + site,

The Combat and LongComBat were implemented using the neuroCombat[61] and longCombat[62] R packages. Site effects were removed before all subsequent analyses including the bootstrap analyses described below.

Robust effect size index (RESI) for association strength

The RESI is a recently developed ES index that has consistent interpretation across many model types, encompassing all types of test statistics in most regression models[31,32]. Briefly, the RESI is a standardized parameter describing the deviation of the true parameter value(s) β from the reference value(s) β_0 from the statistical null hypothesis $H_0: \beta = \beta_0$,

$$S = \sqrt{(\beta - \beta_0)^T \Sigma_\beta{}^{-1} (\beta - \beta_0)}\ ,$$

where *S* denotes the parameter RESI, β and β_0 can be vectors, Σ_β is the covariance matrix for $\hat{\beta}$ (the estimator for β). More details about the definition can be found in Supplement.

In previous work, we defined a consistent estimator for RESI[31]

$$\hat{S} = \left(\max\left\{0, \frac{T^2 - m}{N}\right\}\right)^{1/2},$$

where T^2 is the chi-square test statistics $T^2 = N(\beta - \beta_0)^T \Sigma_\beta{}^{-1} (\beta - \beta_0)$ for testing the null hypothesis $H_0: \beta = \beta_0$, m is the number of parameters being tested (i.e., the length of β) and N is the number of subjects.

As RESI is generally applicable across different models and data types, it's also applicable to the situation where Cohen's *d* was defined. In this scenario, the RESI is equal to ½ Cohen's *d*[31], so Cohen's suggested thresholds for ES can be adopted for RESI: small (RESI=0.1); medium (RESI=0.25); and large (RESI=0.4). Because RESI is robust, when the assumptions of Cohen's *d* are not satisfied, such as when the variances between the groups are not equal, RESI is still a consistent estimator, but Cohen's *d* is not. The confidence intervals (CIs) for RESI were constructed using 1,000 non-parametric bootstraps[32].

Study-level models

After removing the site effects using ComBat or LongComBat in the multi-site data, we estimated the effects of age and sex on each of the global/regional brain measures using GEEs and linear regression models in the longitudinal datasets and cross-sectional datasets, respectively. The mean model was specified as below after ComBat/LongComBat:

$$y_{ij} \sim ns(age_{ij}, df = 2) + sex_i\ ,$$

where y_{ij} was taken to be a global brain measure (i.e., total GMV, WMV, sGMV or mean CT) or regional brain measures (i.e., regional GMV or CT) at the *j*-th visit from the subject *i*, and *j* = 1 for cross-sectional datasets. The age effect was estimated with natural cubic splines with two

degrees of freedom, which means that there were two boundary knots and one interval knot placed at the median of the covariate age. For the GEEs, we used an exchangeable correlation structure as the working structure and identity linkage function. The model assumes the mean is correctly specified, but makes no assumption about the error distribution. The GEEs were fit with the geepack package[63] in R. We used the RESI as a standardized ES measure. RESIs and confidence intervals (CIs) were computed using the RESI R package (version 1.2.0)[33].

Meta-analysis of the age and sex effects

The weighted linear regression model for the meta-analysis of age effects across the studies was specified as:

$$\hat{S}_{age,k} = design_k + ns[mean(age)_k, 3] + ns[sd(age)_k, 3] + ns[skew(age)_k, 3] + \epsilon_k,$$

where $\hat{S}_{age,k}$ denotes the estimated RESI or L-RESI for study *k,* and the weights are the inverse of the standard error of each RESI or L-RESI estimate. The sample mean, standard deviation (SD), and skewness of the age were included as nonlinear terms estimated using natural splines with 3 degrees of freedom (i.e., two boundary knots plus two interval knots at the 33th and 66th percentiles of the covariates), and a binary variable indicating the design type (cross-sectional/longitudinal) was also included.

The weighted linear regression model for the meta-analysis of sex effects across the studies was specified as

$$\hat{S}_{sex,k} = design_k + ns[mean(age)_k, 3] + ns[sd(age)_k, 3] + ns[pr(male)_k, 3] + \epsilon_k,$$

where $\hat{S}_{sex,k}$ denotes the estimated RESI or L-RESI of sex for study *k,* and the weights were the inverse of the standard error of each RESI or L-RESI estimate. The sample mean, SD of the age covariate, and the proportion of males in each study were included as nonlinear terms estimated using natural splines with 3 degrees of freedom, and a binary variable indicating the design type (cross-sectional/longitudinal) was also included.

These meta-analyses were performed for each of the global and regional brain measures. Inference was performed using robust standard errors[64]. In the partial regression plots, the expected ESs for the age effects were estimated from the meta-analysis model after fixing mean age at 45, SD of age at 7 and/or skewness at 0; the expected ESs for the sex effects were estimated from the meta-analysis model after fixing mean age at 45, SD of age at 7 and/or proportion of males at 0.5.

Bootstrap analysis evaluating proposed CS-RESI

Bootstrapping was used to estimate the bias of the L-RESI and CS-RESI in estimating the RESI computed from cross-sectional studies. Longitudinal bootstrap samples were created by cluster sampling subjects (including all measurements for the selected subject) with replacement. Then the equivalent cross-sectional bootstrap sample (i.e., the study dataset from the same study population that was collected cross-sectionally) was created by randomly sampling a single measurement within each of the subjects selected in the longitudinal bootstrap sample. L-RESI and CS-RESI were computed in the longitudinal bootstrap samples and compared to the RESI computed in the cross-sectional bootstrap samples. 1,000 bootstrap samples were drawn per sample size and the mean differences were calculated across the 1,000 bootstraps. This analysis was repeated for all of the 16 longitudinal studies included in our LBCC subset, with a

focus on the outcome total GMV. All site effects were removed using LongComBat prior to performing the bootstrap analysis. Simulation studies were also conducted to evaluate the proposed CS-RESI (see Supplement: Section 1).

Sampling schemes for age in the UKB and ADNI

We used bootstrapping to evaluate the effect of different sampling schemes with different target sample covariate distributions on the effect size (ES) and power in the cross-sectional UKB and longitudinal ADNI datasets. For a given sample size and sampling schemes, 1,000 bootstrap replicates were conducted. The ES was estimated as the mean ES across the bootstrap replicates. The 95% confidence intervals (CIs) for ES were estimated using the lower and upper 2.5% quantiles across the 1,000 ES estimates in the bootstrap replicates. Power was calculated as the proportion of bootstrap replicates producing *p*-values less than and equal to 5% for those associations that were significant at 0.05 in the full sample. In the UKB, only one region was not significant for age in each of GMV and CT, and in ADNI, only 1 and 4 regions were not significant for age in GMV and CT, respectively. The 95% CIs for power were derived using Wilson's method[65].

In the UKB dataset, to modify the (between-subject) variability of the age variable we used the following three target sampling distributions (Fig. S6) – bell-shaped: the target distribution has most of the subjects distributed in the middle age range; uniform; and U-shaped: the distribution has most of the subjects distributed closer to the range limits of the age in the study. The sample with U-shaped age distribution has the largest sample variance of age, followed by the one with uniform age distribution and the one with bell-shaped age distribution. The bell-shaped and U-shaped functions were proportional to a quadratic function. To sample according to these distributions, each record was first inversely weighted by the frequency of the records with age falling in the range of +/- 0.5 years of the age for that record to achieve the uniform sampling distribution, then rescaled to derive the weights for bell- and U-shaped sampling distributions. The records with age <50 or >78 years were winsorized at 50 or 78 years when assigning weights, respectively, to limit the effects of outliers on the weight assignment, but the actual ages were used when analyzing each bootstrapped data.

In each bootstrap from the ADNI dataset, each subject was sampled to have two records. We modified the between- and within-subject variability of age, respectively, by making the baseline age follow one of the three target sampling distributions used for the UKB dataset and the age change independently follow one of three new distributions: decreasing; uniform; and increasing (Fig. S8). The increasing and decreasing functions were proportional to an exponential function. The sample with "increasing" age change distribution has the largest within-subject variability of age, followed by the one with the uniform age change distribution and the one with "decreasing" distribution of age change.

To modify the baseline age and the age change from baseline independently, we first created all combinations of the baseline record and one follow-up from each participant, and calculated the baseline age and age change for each combination. Each combination was then inversely weighted by the "bivariate" frequency of the combinations with baseline age and age change falling into the range of +/- 0.5 years of the baseline age and age change of this combination to achieve uniform bivariate distribution of baseline age and age change. The weight for each combination was then rescaled to make the baseline age and age change follow

different sampling distributions independently. The combinations with baseline age <65 or >85 years were winsorized at 65 or 85 years, and the ones with age change greater than 5 years were winsorized at 5 years when assigning weights to limit the effects of outliers on the weight assignment, but the actual ages were used when analyzing each bootstrapped data.

The sampling methods can be easily extended to the scenario where each subject has three records (and more than three) in the bootstrap data by making combinations of the baseline and two follow-ups. Each combination was inversely weighted to achieve uniform baseline age and age change distributions, respectively, by the "trivariate" frequency of the combinations with baseline age, and the two age changes from baseline for the two follow-ups falling into the range of +/- 0.5 years of the corresponding values for this combination. Since we only investigated the effect of modifying the number of measurements per subject under uniform between- and within-subject sampling schemes (Fig. 4c-d), we didn't need to consider rescaling the weights here to achieve other sampling distributions. For the scenario where each subject only has one measurement (Fig. 4c-d), the ESs and power were estimated only using the baseline measurements.

All site effects were removed using ComBat or LongComBat prior to performing the bootstrap analysis.

Sampling schemes for other non-brain covariates in ABCD

We used bootstrapping to study how different sampling strategies affect the RESI in the ABCD dataset. Each subject in the bootstrap data has two measurements. We applied the same weight assignment method described above for the ADNI dataset to modify the between- and within-subject variability of a covariate. We made the baseline covariate and the change in covariate follow bell-shaped and/or U-shaped distributions, to let the sample have larger or smaller between- and/or within-subject variability of the covariate, respectively. The baseline covariate and change of covariate were winsorized at the upper and lower 5% quantiles to limit the effect of outliers on sampling. For each cognitive variable, only the subjects with non-missing baseline measurements and at least one non-missing follow-up were included.

GLMs and GEEs were fitted to estimate the effect of each non-brain covariate on the structural brain measures after controlling for age and sex,

$$\text{GMV} \sim \text{ns}(\text{age}, \text{df} = 2) + \text{sex} + x,$$

where x denotes one of the non-brain covariates. GEEs assumed an exchangeable working covariance structure.

For the non-brain covariates that were stable over time (e.g., birth weight and handedness), only the between-subject sampling schemes were applied (i.e., the subjects were sampled based on their baseline covariate values, and then a follow-up was selected randomly for each subject). The sampling schemes to increase the between-subject variability in the covariate "handedness", which is a binary variable (right-handed or not), was specified differently. The expected proportion of right-handed subjects in the bootstrap samples is 50% under the sampling scheme with "larger" between-subject variability and 10% under the sampling scheme with "smaller" between-subject variability.

For given between- and/or within-subject sampling schemes, 1,000 bootstrap replicates were conducted. The ES was estimated as the mean ES across the bootstrap replicates. The 95% CIs for ES were estimated using the lower and upper 2.5% quantiles across the 1,000 ES

estimates in the bootstrap replicates. The 95% CIs for power were derived using Wilson's method[65]. The sample sizes needed for 80% power were estimated based on the (mean) estimated ES and *F*-distribution (see below).

Analysis of FC in ABCD

In a subset of the ABCD where we have preprocessed longitudinal FC data at two timepoints (baseline and 2-year follow-up), we only restricted to the subjects with non-missing measurements at both of the two times. In the GEEs used to estimate the effects of non-brain covariates on FC, the mean model was specified as below after LongComBat:

$$y_{ij} \sim ns(age_{ij}, df = 2) + sex_i + ns(mean_FD, df = 3) + x_{ij},$$

where y_{ij} was taken to be an FC outcome and *x* denotes a non-brain covariate. Mean frame displacement (FD) was also included as a covariate with natural cubic splines with 3 degrees of freedom. We used exchangeable correlation structure as the working structure in the GEEs, and used the frame count of each scan as the weights.

When evaluating the effect of different sampling schemes on the ESs, for given between- and/or within-subject sampling schemes, 1,000 bootstrap replicates were conducted. The ES was estimated as the mean effect size across the bootstrap replicates. CIs were computed as described above. The sample sizes needed for 80% power were estimated based on the (mean) estimated ESs and *F*-distribution (see below).

Sample size calculation for a target power with a given effect size

After estimating the ES for an association, the calculation of the corresponding sample size N needed for detecting this association with $\gamma \times 100\%$ power at significance level of α was based on an *F*-distribution. Let *df* denote the total degree of freedom of the analysis model, $F(x;\ \lambda)$ denote the cumulative density function for the random variable *X,* which follows the (non-central) *F*-distribution with degrees of freedom being 1 and *N - df* and non-centrality parameter λ. The corresponding sample size N is:

$$N = \{N: F(F^{-1}(1-\alpha);\ \lambda = N\hat{S}^2) = \gamma\},$$

where $\hat{S}$ is the estimated RESI for the ES. Power curves for the RESI are given in Figure 3 of Vandekar et al. (2002)[31].

Estimation of the cross-sectional and longitudinal effects

For the non-brain covariates that were analyzed longitudinally in ABCD dataset, GEEs with exchangeable correlation structures were fitted to estimate their cross-sectional and longitudinal effects on structural and functional brain measures after controlling for age and sex, respectively. The mean model was specified as below illustrated with GMV:

GMV ~ ns(age, df = 2) + sex + X_bl + X_change,

where X_bl denotes the subject-specific baseline covariate values, and the X_change denotes the difference of the covariate value at each visit to the subject-specific baseline covariate value (see Supplement: Section 3.2). The subjects without baseline measures were not included in the modeling. The model coefficients for the terms "X_bl" and "X_change" represent the cross-sectional and longitudinal effects of this non-brain covariate on total GMV, respectively. For the FC data, the same covariates and weighting were used as described above.

Data availability

Participant-level data from many datasets are available according to study-level data access rules. Study-level model parameters are available at https://github.com/KaidiK/RESI_BWAS.

We acknowledge the usage of several openly shared MRI datasets, which are available at the respective consortia websites and are subject to the sharing policies of each consortium: OpenNeuro (https://openneuro.org/), UK BioBank (https://www.ukbiobank.ac.uk/), ABCD (https://abcdstudy.org/), the Laboratory of NeuroImaging (https://loni.usc.edu/), data made available through the Open Science Framework (https://osf.io/), the Human Connectome Project (http://www.humanconnectomeproject.org/), and the OpenPain project (https://www.openpain.org). The ABCD data repository grows and changes over time. The ABCD data used in this report came from DOI 10.15154/1503209. Data used in this article were provided by the brain consortium for reliability, reproducibility and replicability (3R-BRAIN) (https://github.com/zuoxinian/3R-BRAIN). Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at https://www.aibl.csiro.au. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. More information on the ARWIBO consortium can be found at https://www.arwibo.it/. More information on CALM team members can be found at https://calm.mrc-cbu.cam.ac.uk/team/ and in the Supplementary Information. Data used in this article were obtained from the developmental component 'Growing Up in China' of the Chinese Color Nest Project (http://deepneuro.bnu.edu.cn/?p=163). Data used in the preparation of this article were obtained from the IConsortium on Vulnerability to Externalizing Disorders and Addictions (c-VEDA), India (https://cveda-project.org/). Data used in the preparation of this article were obtained from the Harvard Aging Brain Study (HABS P01AG036694) (https://habs.mgh.harvard.edu). Data used in the preparation of this article were obtained from the IMAGEN consortium (https://imagen-europe.com/). The POND network (https://pond-network.ca/) is a Canadian translational network in neurodevelopmental disorders, primarily funded by the Ontario Brain Institute.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health

(www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Code availability
All code used to produce the analyses presented in this paper is available at https://github.com/KaidiK/RESI_BWAS.

Acknowledgments
S.N.V was supported by R01MH123563. A.A.B. and J.S. were partially supported by 5K08MH120564-05. Data processing done at the University of Cambridge is supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and NIHR Applied Research Collaboration East of England. Any views expressed are those of the author(s) and not necessarily those of the funders, IHU-JU2, the NIHR or the Department of Health and Social Care.

Disclosures
J.S., R.B., and A.A.B. are directors and hold equity in Centile Bioscience. A.A.B. receives consulting income from Octave Bioscience.

Supplementary Figures:

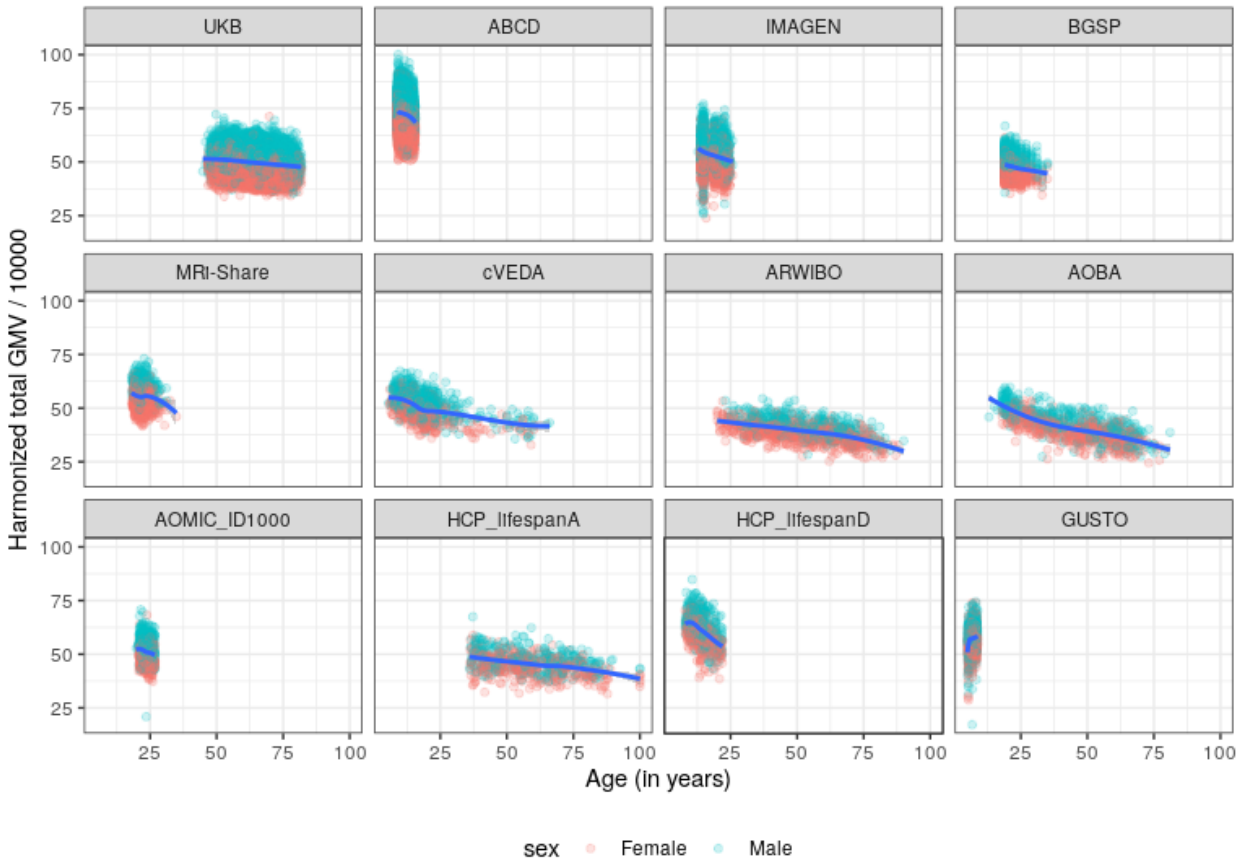

Fig. S1. Scatterplots of the harmonized total gray matter volume (GMV) measures (GMV after removing site effect using ComBat or longCombat) and age (in years) in the 12 largest studies among the 63 studies. The blue curves are the locally estimated scatterplot smoothing (LOESS) curves. The shaded areas are the 95% confidence bands. MRI datasets containing 77,695 scans from 63 studies with different age sample distributions were used. Nonlinear models are used to fit total GMV, white matter volume (WMV), subcortical gray matter volume (sGMV), and mean cortical thickness (CT) as a function of age (in years) and sex in each study.

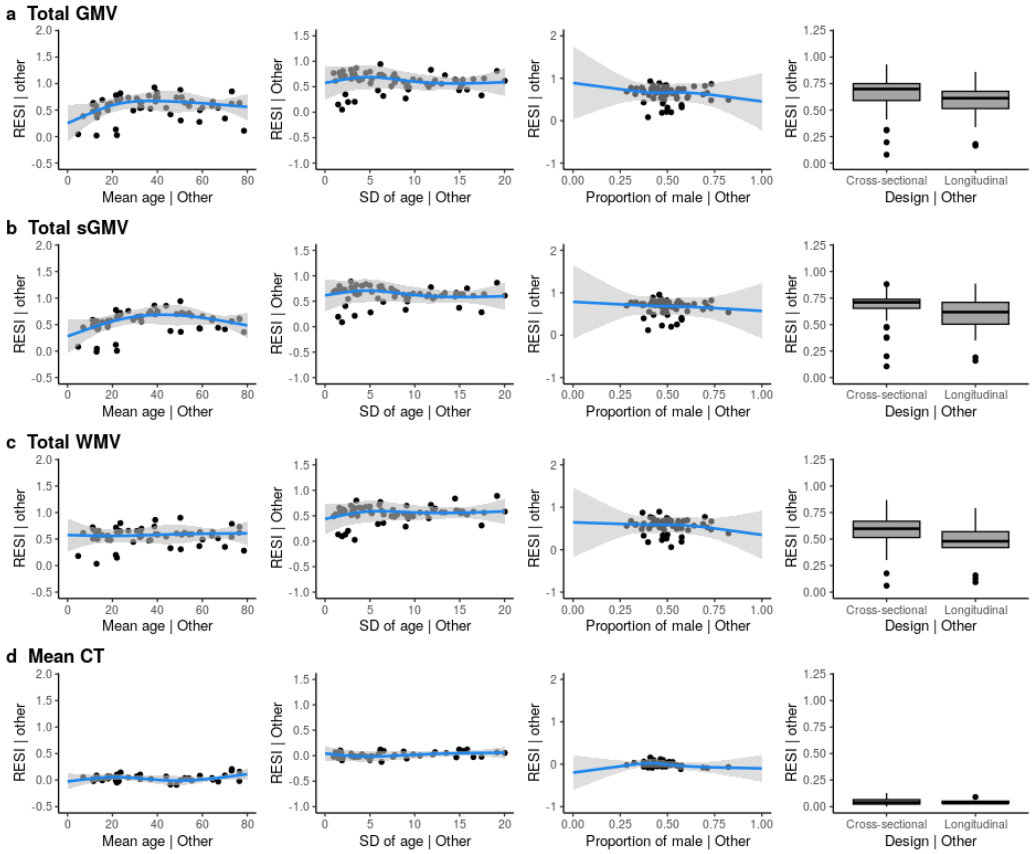

Fig. S2. Meta-analysis results regarding the effects of study features on the associations between sex and different brain measures. (**a-d**) Partial regression plots of the meta-analyses of effect sizes (RESI for cross-sectional studies and L-RESI for longitudinal studies) for the association between sex and global brain measures (**a**) total gray matter volume (GMV), (**b**) total subcortical gray matter volume (sGMV), (**c**) total white matter volume (WMV), and (**d**) mean cortical thickness (CT) ("| Other" means after fixing the other features at constant levels: design = cross-sectional, mean age = 45 years, sample age standard deviation (SD) = 7 and/or proportion of males = 0.5). The blue curves are the expected effect size for sex. The gray areas are the 95% confidence bands.

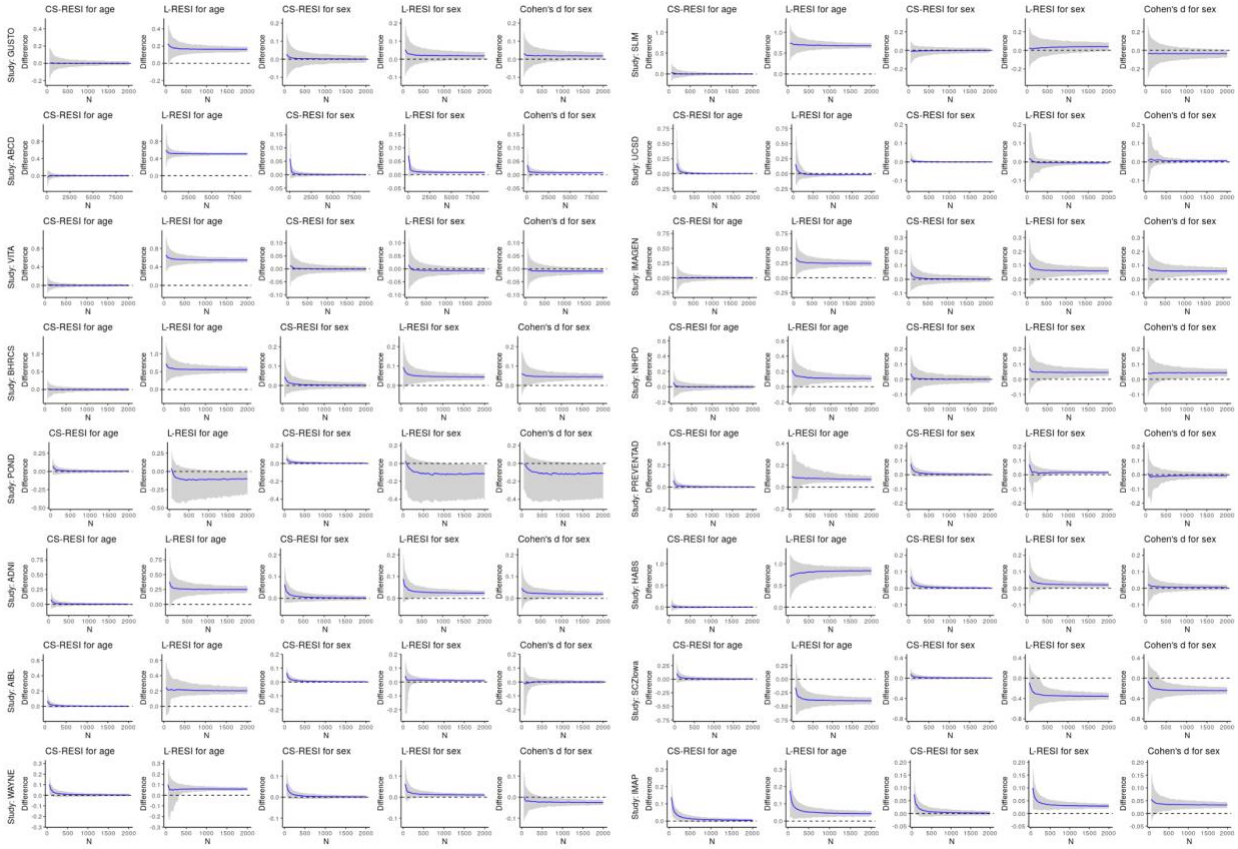

Fig. S3. The systematic difference in effect size by study designs and the comparability of CS-RESI (cross-sectional RESI) in longitudinal studies to RESI in cross-sectional studies. We conduct 1,000 bootstraps per sample size N. The blue curves represent the mean difference curves and the gray areas are the 95% confidence bands from the bootstraps. For most of the studies, the differences of L-RESI (longitudinal RESI) in longitudinal datasets to regular RESI in cross-sectional datasets approach non-zero constants as sample size increases. However, the differences of CS-RESI in longitudinal datasets to the regular RESI in cross-sectional datasets always approach zero at large sample sizes, indicating the systematic difference in effect sizes between longitudinal and cross-sectional datasets is removed by using CS-RESI.

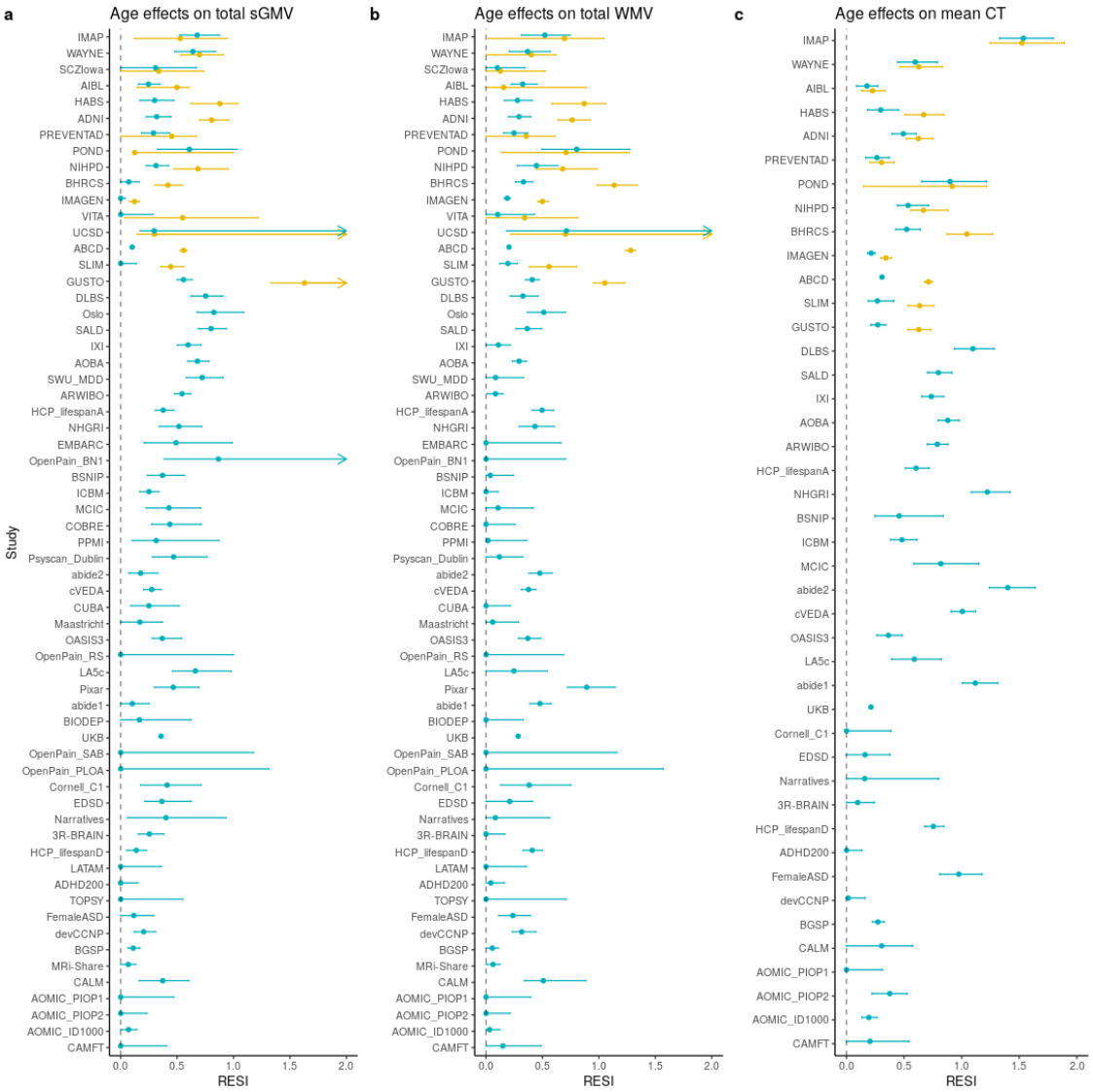

Fig. S4. Forest plots of age effects on total subcortical gray matter volume (sGMV), total white matter volume (WMV) and mean cortical thickness (CT), respectively. The age effects on total sGMV and WMV are estimated in 63 studies and the age effects on mean CT are estimated in 43 studies where mean CT is available. RESIs (blue) are estimated for cross-sectional studies, both CS-RESIs (blue) and L-RESIs (yellow) are estimated for longitudinal studies. The dots and lines represent the point estimates of the RESIs and their 95% confidence intervals (CIs). The blue dots and lines represent the CS-RESIs in longitudinal and RESIs in cross-sectional studies. The yellow dots and lines are for the L-RESIs in longitudinal studies. Arrows mean the upper limits of the CIs are greater than the axis boundary.

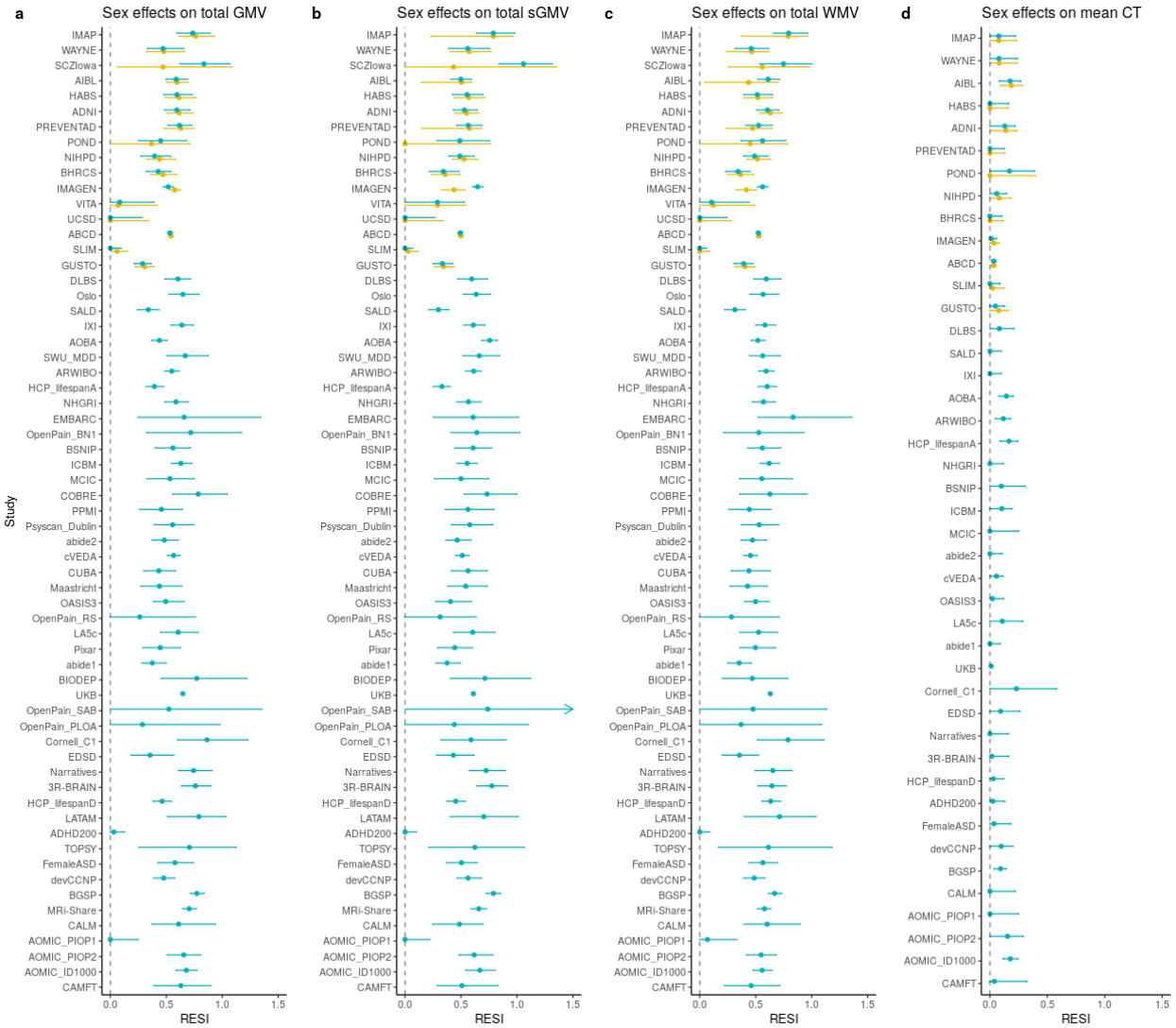

Fig. S5. Forest plots of sex effects on total gray matter volume (GMV), total subcortical gray matter volume (sGMV), total white matter volume (WMV) and mean cortical thickness (CT), respectively. The sex effects on total GMV, sGMV and WMV are estimated in 63 studies and the sex effects on mean CT are estimated in 43 studies where mean CT is available. Both CS-RESI (blue) and L-RESI (yellow) are estimated for longitudinal studies. The dots and lines represent the point estimates of the RESIs and their 95% confidence intervals (CIs). The blue dots and lines represent the CS-RESIs in longitudinal and RESIs in cross-sectional studies. The yellow dots and lines are for the L-RESIs in longitudinal studies. Arrow means the upper limit of the CI is greater than the axis boundary.

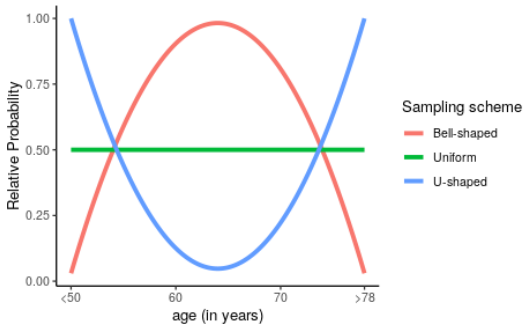

Fig. S6. The sampling scheme implemented in UKB. The sampling schemes adjust the variance of age in the samples by assigning heavier or lighter weights to the subjects with age at the two tails of the population. The U-shaped scheme produces the largest variability of age in the samples, followed by uniform and bell-shaped sampling schemes.

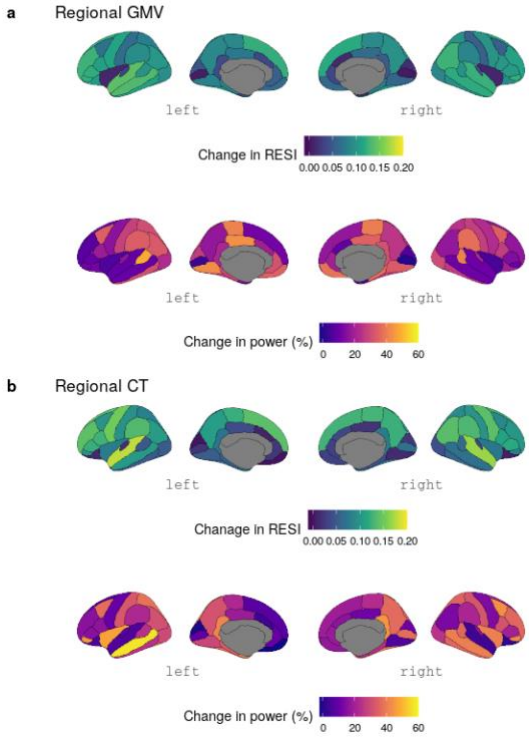

Fig. S7. Region-specific improvement in the RESI and power for the association between age and regional gray matter volume (GMV) (**a**) and regional cortical thickness (CT) (**b**), respectively, by using U-shaped sampling scheme compared with bell-shaped sampling scheme, when N = 300.

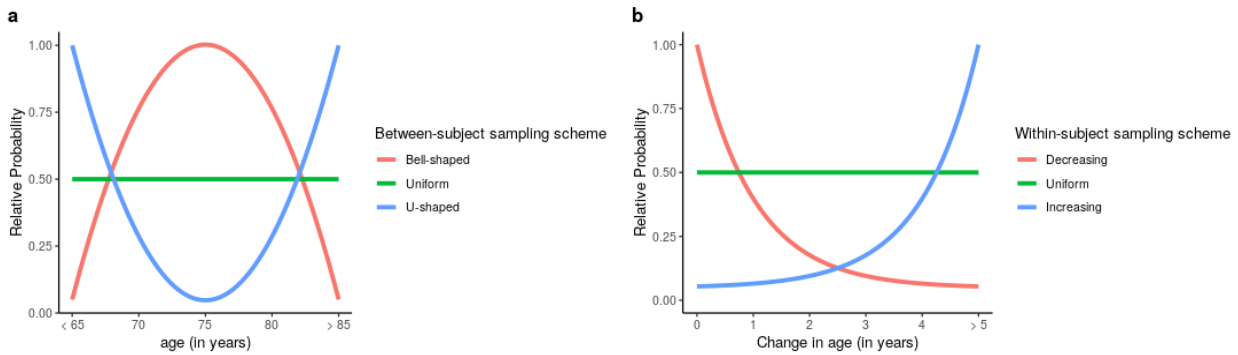

Fig. S8. Between- and within-subject sampling schemes implemented in ADNI. (**a**) The between-subject variability of age is adjusted by assigning heavier or lighter weights to the subjects with baseline age closer to the two tails of the population baseline age distribution. (**b**) The within-subject variability in age is adjusted by increasing or decreasing the probability of selecting the follow-up observation(s) with a larger change in age since baseline.

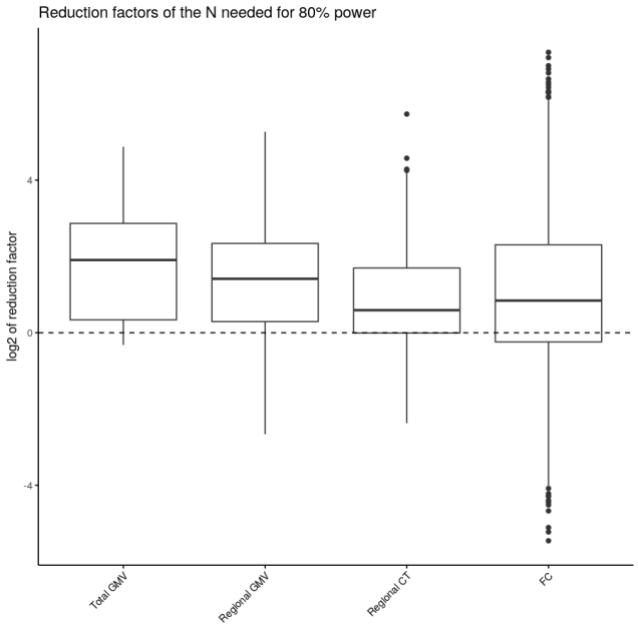

Fig. S9. The boxplots showing the distributions of (log2 of) reduction factors of the sample size N needed for 80% power by increasing between-subject variability of the covariates across all the associations with each of the outcomes in Fig. 5. The reduction factors were derived by comparing the sample sizes needed for 80% power with U-shaped to the one with bell-shaped between-subject sampling scheme when the within-subject sampling scheme is bell-shaped (Fig. 5).

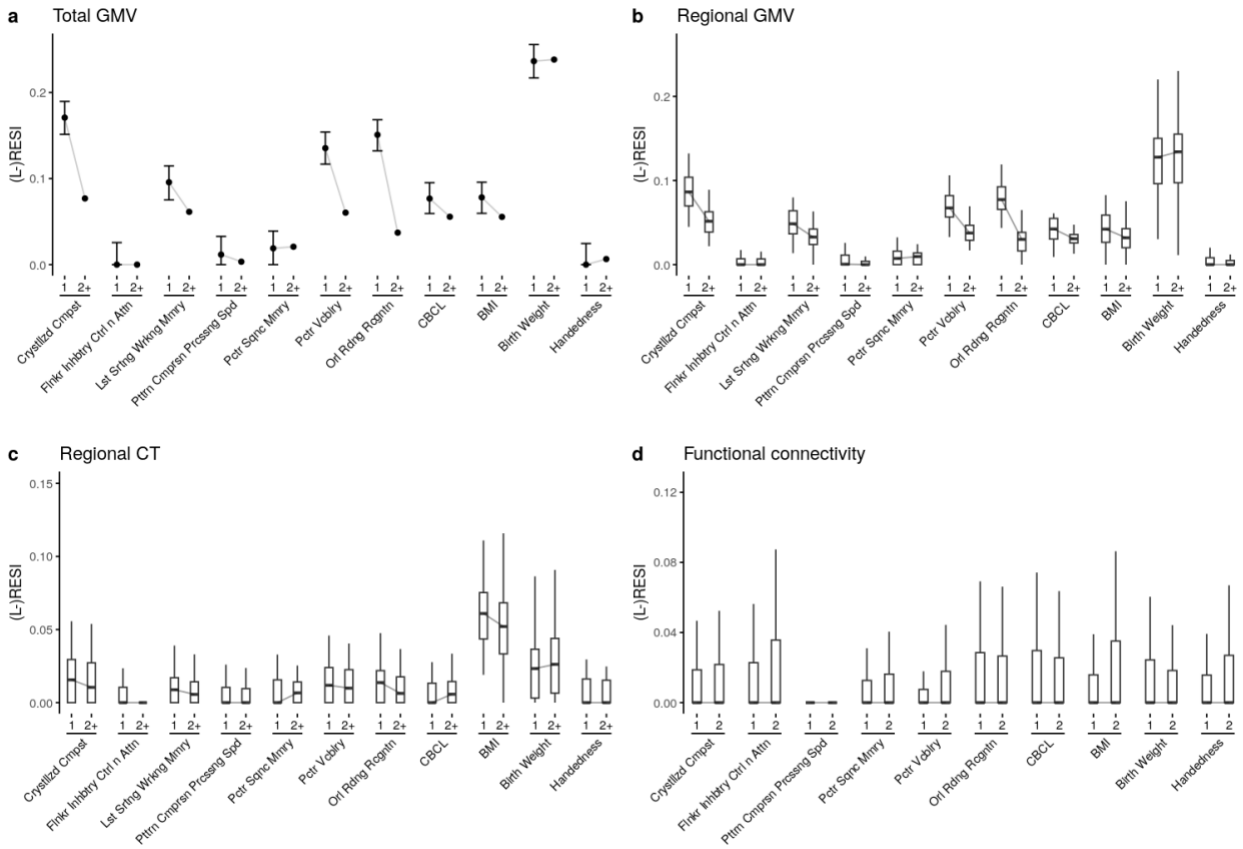

Fig. S10. Replication of Fig. 6a-d using linear mixed models (LMMs) with individual-specific random intercepts. Boxplots show the distributions of the effect sizes across regions. The effect size estimates (RESI) from cross-sectional analyses that only use the baseline measures are indicated by “1”s on the x-axes; the effect size estimates (L-RESI) from longitudinal analyses that use the full longitudinal data are indicated by “2“s or “2+”s on the x-axes. The results are similar using generalized estimating equations (GEEs) and LMMs and show that longitudinal study designs can reduce effect sizes for most structural brain associations. Only the point estimates of L-RESI are obtained as the bootstrap confidence intervals (CIs) in LMMs have not been evaluated.

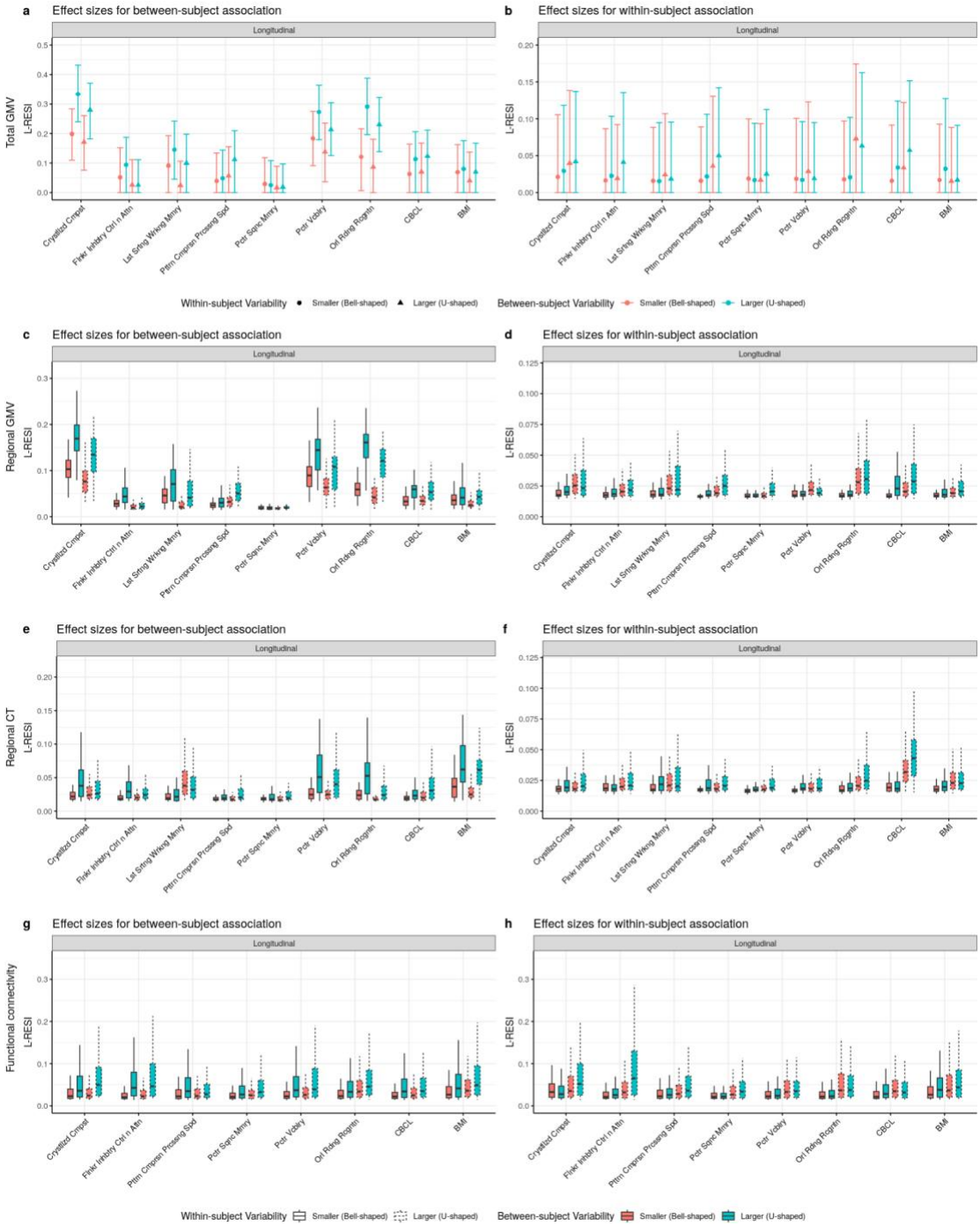

Fig. S11. The influence of sampling schemes on the effect sizes for between- and within-subject associations, respectively, of cognition, mental health, and demographic covariates with different brain measures in the ABCD study at *N* = 500. Boxplots show the distribution of the effect sizes across regions. Between-subject effect sizes are predominantly affected by the between-subject variance, whereas within-subject effect sizes are predominantly affected by the within-subject variance. The results for covariates birthweight and handedness, which do not vary within subjects, are not included as the within-subject sampling schemes do not apply to them.

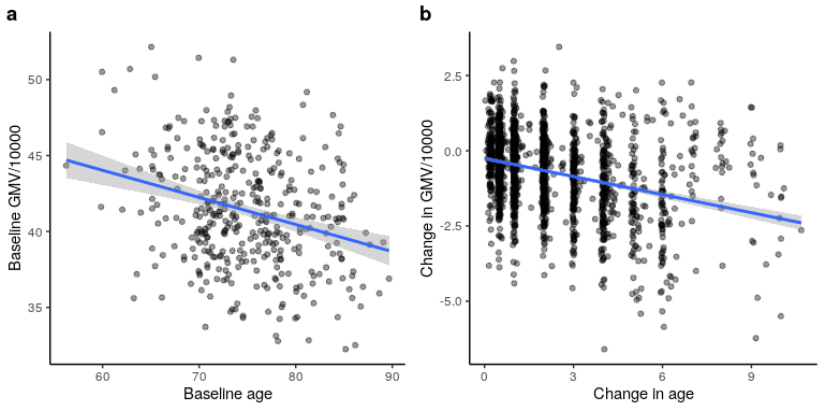

Fig. S12. Visualization of the crude between- and within-subject associations between age and total gray matter volume (GMV) in ADNI. There are strong associations (**a**) between baseline total GMV and baseline age and (**b**) between change in total GMV and change in age since baseline, indicating there exist strong between- and within-subject effects of age on total GMV in ADNI study. The blue lines are the simple linear regression lines and the gray areas are the 95% confidence bands.

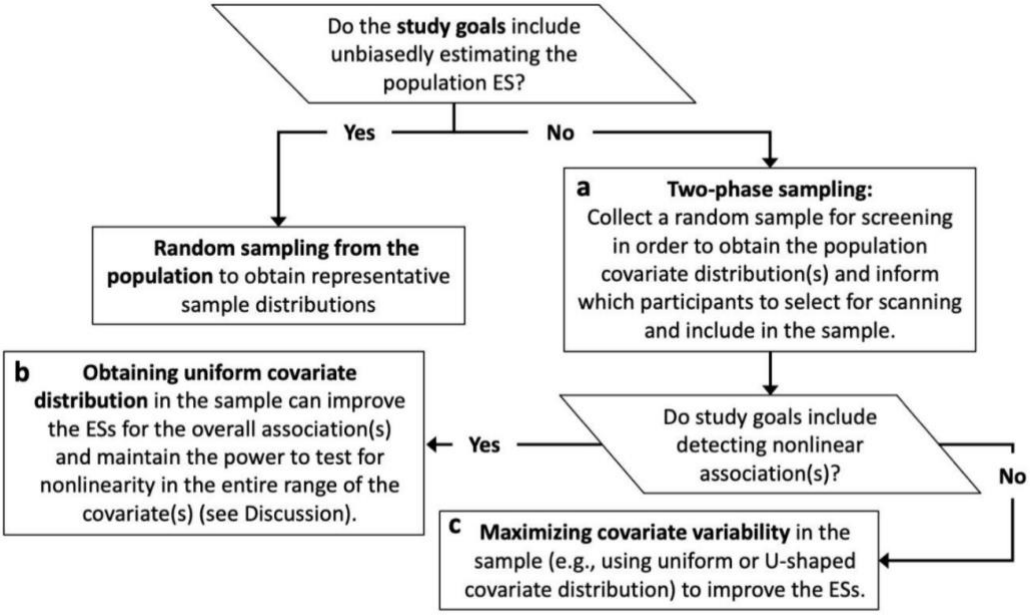

Fig. S13: Decision tree for modified sampling strategy for a single primary covariate. Random/representative sampling is needed to unbiasedly estimate the variance of the covariate distribution in the population in order to obtain unbiased effect size (ES) estimates for the population. (**a**) A two-phase design is needed to modify the covariate distribution(s) in the sample to improve ES and power, where random sampling is performed first in a larger dataset to collect covariate values and biased sampling a subset based on collected covariates values to optimize ESs and power; unbiased population ES estimates still can be obtained using weighted estimation (see Discussion: Increasing between-subject variability to increase effect sizes). (**b**) If the distribution(s) of the covariate(s) in the population is bell-shaped, a uniform covariate distribution in the sample can still increase the ES and power to detect the overall association. (**c**) The particular target distribution will depend on the difficulty of collecting participants in the tail of the distributions (see Supplement: Section 2.1).

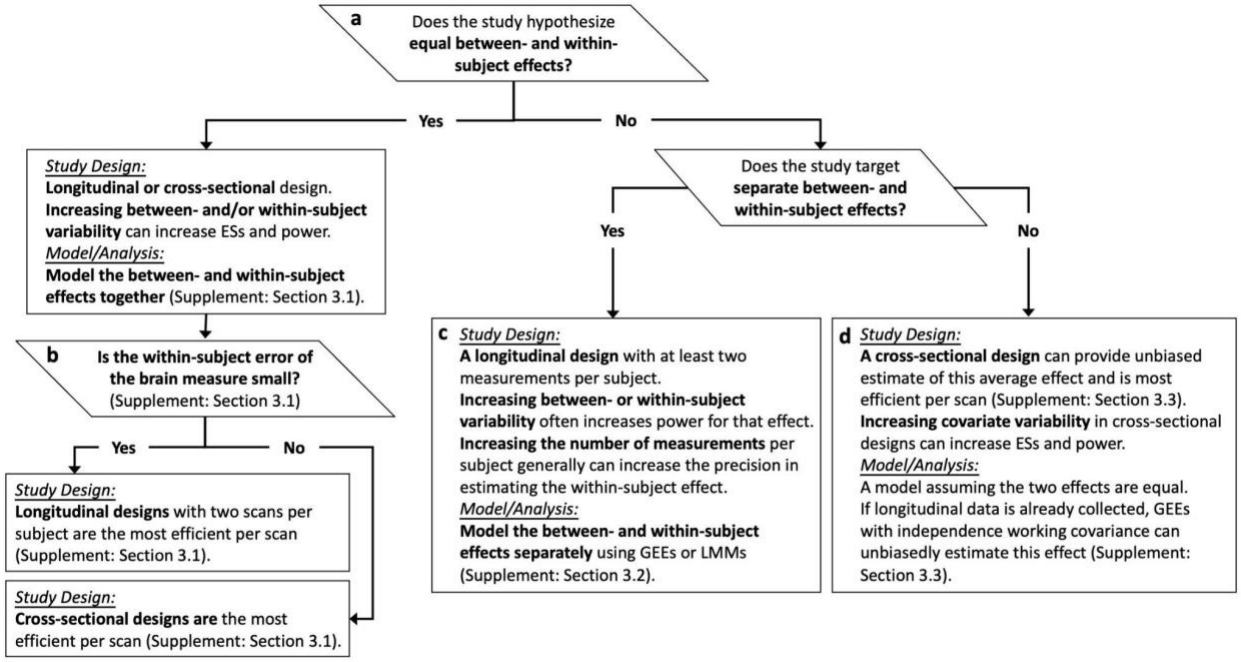

Fig. S14: Optimal study design and analysis depends on characteristics of the hypothesized association(s). (**a**) Visualization can be performed in pilot or study data to evaluate this assumption as in Fig. S12. See Supplement: Section 3 for mathematical details. (**b**) If the between- and within-subject effects are hypothesized to be equal, either a cross-sectional or longitudinal design can be applied, but the efficiency per scan depends on the size of the within-subject error of the brain measure; pilot/study data can be used to evaluate this question (Supplement: Section 3.1). (**c**) If estimating the between- and within-subject effects separately, a longitudinal design is required and common longitudinal data analysis tools such as generalized estimating equations (GEEs) and linear mixed models (LMMs) with separate between- and within-subject effects are required to unbiasedly estimate these effects (see Supplement: Section 3.2). (**d**) If targeting the average effect (i.e., a weighted average of the underlying between- and within-subject effects) is still of interest, despite the different between- and within-subject effects, longitudinal designs are biased as the weights depend on the within-subject variability of the covariate. Only GEEs with independence covariance structure target the same average effect as the cross-sectional models, but are still less efficient in estimation than the cross-sectional model (see Supplement: Section 3.3). These recommendations are based on the empirical findings in the paper and the theory for exchangeable covariance longitudinal linear models in the Supplement.

Supplementary Tables:

Tab. S1. The descriptive summary of the 63 neuroimaging datasets used in the meta-analyses for the total gray matter volume (GMV), subcortical gray matter volume (sGMV) and white matter volume (WMV).

Study	*N*	Total number of observations	Design	Mean age	SD of age	Skewness of age	Proportion of males
UKB	29031	29031	Cross-sectional	64.64	7.58	-3.96	0.50
ABCD	8874	17210	Longitudinal	11.18	1.60	1.35	0.50
IMAGEN	2107	4446	Longitudinal	17.98	3.35	2.34	0.48
BGSP	1570	1570	Cross-sectional	21.56	2.89	3.40	0.42
MRi-Share	1341	1341	Cross-sectional	22.01	2.21	2.31	0.31
cVEDA	1193	1193	Cross-sectional	19.19	9.82	12.78	0.58
ARWIBO	1077	1077	Cross-sectional	52.19	15.33	4.37	0.38
AOBA	964	964	Cross-sectional	43.90	15.83	-3.36	0.47
AOMIC_ID1000	925	925	Cross-sectional	23.63	1.71	0.76	0.48
HCP_lifespanA	689	689	Cross-sectional	58.92	14.94	10.75	0.43
HCP_lifespanD	655	655	Cross-sectional	13.79	3.81	3.00	0.49
GUSTO	642	1230	Longitudinal	6.98	1.18	-0.73	0.51
OASIS3	600	600	Cross-sectional	67.01	9.18	-4.27	0.40
ICBM	588	588	Cross-sectional	30.84	12.27	14.27	0.52
SLIM	562	981	Longitudinal	21.48	1.46	1.44	0.47
abide1	560	560	Cross-sectional	17.88	7.76	8.90	0.82
IXI	538	538	Cross-sectional	48.52	16.54	4.59	0.42
ADHD200	481	481	Cross-sectional	12.92	3.31	3.24	0.52
SALD	473	473	Cross-sectional	45.82	17.44	4.61	0.37
AIBL	470	822	Longitudinal	73.02	6.54	4.59	0.46
devCCNP	423	423	Cross-sectional	12.94	2.93	2.08	0.47
NIHPD	423	882	Longitudinal	12.94	4.12	2.52	0.47
ADNI	417	2232	Longitudinal	76.57	6.06	1.48	0.51
NHGRI	410	410	Cross-sectional	22.71	14.63	16.09	0.57
abide2	391	391	Cross-sectional	17.94	10.58	12.82	0.71
PREVENTAD	383	1678	Longitudinal	64.28	5.21	5.09	0.28
3R-BRAIN	330	330	Cross-sectional	27.05	4.55	3.60	0.56
BHRCS	319	642	Longitudinal	13.35	3.61	2.63	0.55
DLBS	313	313	Cross-sectional	55.26	20.03	-7.84	0.37
Oslo	280	280	Cross-sectional	50.83	17.72	-7.31	0.41
HABS	277	670	Longitudinal	76.48	6.29	4.64	0.42
FemaleASD	229	229	Cross-sectional	12.75	3.03	1.02	0.48
AOMIC_PIOP2	224	224	Cross-sectional	21.97	1.79	0.77	0.43
IMAP	214	418	Longitudinal	50.21	19.16	-7.41	0.46
CUBA	201	201	Cross-sectional	32.26	9.36	8.50	0.73
SWU_MDD	200	200	Cross-sectional	40.17	15.35	6.00	0.34
WAYNE	199	291	Longitudinal	54.10	15.11	-12.16	0.33
BSNIP	197	197	Cross-sectional	39.18	12.98	-1.84	0.44
Narratives	181	181	Cross-sectional	22.75	5.00	6.68	0.45
EDSD	162	162	Cross-sectional	70.05	5.90	-2.88	0.48

Pixar	155	155	Cross-sectional	11.33	8.05	9.28	0.46
Psyscan_Dublin	154	154	Cross-sectional	32.93	11.36	11.36	0.47
Maastricht	143	143	Cross-sectional	29.26	9.27	7.63	0.48
LA5c	125	125	Cross-sectional	32.32	8.77	7.45	0.53
SCZIowa	106	223	Longitudinal	32.91	9.09	7.54	0.61
PPMI	105	105	Cross-sectional	60.06	11.56	-9.29	0.61
POND	97	109	Longitudinal	13.05	5.07	3.43	0.55
MCIC	90	90	Cross-sectional	32.41	12.01	8.87	0.69
COBRE	86	86	Cross-sectional	38.69	11.80	6.89	0.73
UCSD	81	87	Longitudinal	4.74	2.27	3.85	0.57
CALM	77	77	Cross-sectional	11.51	1.99	1.13	0.49
AOMIC_PIOP1	73	73	Cross-sectional	21.97	1.90	1.04	0.40
LATAM	73	73	Cross-sectional	23.52	3.49	2.51	0.70
CAMFT	68	68	Cross-sectional	15.50	0.96	-0.67	0.49
Cornell_C1	66	66	Cross-sectional	73.12	6.17	-2.77	0.41
BIODEP	55	55	Cross-sectional	36.21	7.74	5.86	0.36
VITA	54	99	Longitudinal	78.53	2.57	2.45	0.49
OpenPain_BN1	39	39	Cross-sectional	39.89	13.36	9.91	0.64
EMBARC	38	38	Cross-sectional	39.04	14.49	10.67	0.37
OpenPain_RS	33	33	Cross-sectional	50.38	8.98	-8.94	0.58
TOPSY	29	29	Cross-sectional	22.34	3.28	0.71	0.59
OpenPain_PLOA	20	20	Cross-sectional	58.68	6.49	6.80	0.50
OpenPain_SAB	20	20	Cross-sectional	36.91	7.40	5.91	0.55

Tab. S2. Descriptive summary table for the 43 studies used in the meta-analyses for mean cortical thickness (CT), regional gray matter volume (GMV), and regional CT.

Study	*N*	Total number of observations	Design	Mean age	SD of age	Skewness of age	Proportion of males
UKB	29030	29030	Cross-sectional	64.64	7.58	-3.96	0.50
ABCD	8874	17210	Longitudinal	11.18	1.60	1.35	0.50
IMAGEN	2105	4437	Longitudinal	17.99	3.35	2.33	0.48
BGSP	1569	1569	Cross-sectional	21.56	2.89	3.40	0.42
cVEDA	1193	1193	Cross-sectional	19.19	9.82	12.78	0.58
ARWIBO	1077	1077	Cross-sectional	52.19	15.33	4.37	0.38
AOBA	964	964	Cross-sectional	43.90	15.83	-3.36	0.47
AOMIC_ID1000	924	924	Cross-sectional	23.63	1.71	0.76	0.48
HCP_lifespanA	689	689	Cross-sectional	58.92	14.94	10.75	0.43
HCP_lifespanD	655	655	Cross-sectional	13.79	3.81	3.00	0.49
GUSTO	642	1228	Longitudinal	6.98	1.18	-0.73	0.51
OASIS3	598	598	Cross-sectional	67.08	9.11	-3.94	0.40
ICBM	588	588	Cross-sectional	30.84	12.27	14.27	0.52
SLIM	562	981	Longitudinal	21.48	1.46	1.44	0.47
abide1	560	560	Cross-sectional	17.88	7.76	8.90	0.82
IXI	538	538	Cross-sectional	48.52	16.54	4.59	0.42
ADHD200	481	481	Cross-sectional	12.92	3.31	3.24	0.52
SALD	473	473	Cross-sectional	45.82	17.44	4.61	0.37
AIBL	469	821	Longitudinal	73.02	6.54	4.59	0.46
devCCNP	422	422	Cross-sectional	12.93	2.93	2.09	0.47
NIHPD	422	881	Longitudinal	12.94	4.12	2.51	0.47
ADNI	417	2232	Longitudinal	76.57	6.06	1.48	0.51
NHGRI	410	410	Cross-sectional	22.71	14.63	16.09	0.57
abide2	391	391	Cross-sectional	17.94	10.58	12.82	0.71
PREVENTAD	383	1677	Longitudinal	64.28	5.21	5.09	0.28
BHRCS	319	638	Longitudinal	13.37	3.61	2.62	0.55
DLBS	313	313	Cross-sectional	55.26	20.03	-7.84	0.37
3R-BRAIN	291	291	Cross-sectional	27.05	4.58	3.66	0.55
HABS	277	670	Longitudinal	76.48	6.29	4.64	0.42
FemaleASD	229	229	Cross-sectional	12.75	3.03	1.02	0.48
AOMIC_PIOP2	224	224	Cross-sectional	21.97	1.79	0.77	0.43
IMAP	214	418	Longitudinal	50.21	19.16	-7.41	0.46
WAYNE	197	287	Longitudinal	54.02	15.20	-12.10	0.32
Narratives	181	181	Cross-sectional	22.75	5.00	6.68	0.45
EDSD	162	162	Cross-sectional	70.05	5.90	-2.88	0.48
LA5c	125	125	Cross-sectional	32.32	8.77	7.45	0.53
BSNIP	99	99	Cross-sectional	37.16	13.51	8.40	0.39
POND	97	109	Longitudinal	13.05	5.07	3.43	0.55
MCIC	90	90	Cross-sectional	32.41	12.01	8.87	0.69
CALM	77	77	Cross-sectional	11.51	1.99	1.13	0.49
AOMIC_PIOP1	73	73	Cross-sectional	21.97	1.90	1.04	0.40
CAMFT	68	68	Cross-sectional	15.50	0.96	-0.67	0.49
Cornell_C1	66	66	Cross-sectional	73.12	6.17	-2.77	0.41

Tab. S3. Meta-analysis results for the RESI for age on total gray matter volume (GMV). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.31	0.10	9.80	1	<0.01	0.37	(0.10, 0.70)
Mean age			5.62	3	<0.01	0.46	(0.09, 0.97)
SD of age			18.90	3	<0.01	0.90	(0.50, 1.67)
Skewness of age			2.34	3	0.08	0.24	(0.00, 0.72)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.02	0.08	0.06	1	0.81	0.00	(0.00, 0.26)
Mean age			3.91	3	0.01	0.36	(0.18, 1.02)
SD of age			20.23	3	<0.01	0.94	(0.58, 1.72)
Skewness of age			0.77	3	0.52	0.00	(0.00, 0.61)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust effect size index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S4. Meta-analysis results for the RESI for age on total subcortical gray matter volume (sGMV). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.23	0.11	4.92	1	0.03	0.24	(0.00, 0.78)
Mean age			1.61	3	0.20	0.16	(0.00, 0.67)
SD of age			6.98	3	<0.01	0.52	(0.00, 1.18)
Skewness of age			1.27	3	0.29	0.10	(0.00, 0.63)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.03	0.07	0.20	1	0.65	0.00	(0.00, 0.30)
Mean age			1.10	3	0.36	0.05	(0.00, 0.77)
SD of age			7.09	3	<0.01	0.53	(0.06, 1.11)
Skewness of age			0.17	3	0.91	0.00	(0.00, 0.47)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S5. Meta-analysis results for the RESI for age on total white matter volume (WMV). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.42	0.13	11.11	1	<0.01	0.39	(0.21, 0.88)
Mean age			7.50	3	<0.01	0.54	(0.33, 1.39)
SD of age			1.19	3	0.32	0.08	(0.00, 0.75)
Skewness of age			0.30	3	0.83	0.00	(0.00, 0.5)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.00	0.06	0.00	1	0.95	0.00	(0.00, 0.28)
Mean age			4.88	3	<0.01	0.42	(0.21, 1.23)
SD of age			3.36	3	0.03	0.33	(0.00, 0.92)
Skewness of age			0.46	3	0.71	0.00	(0.00, 0.60)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S6. Meta-analysis results for the RESI for age on mean cortical thickness (CT). The residual degrees of freedom are 32. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.25	0.19	1.79	1	0.19	0.13	(0.00, 0.82)
Mean age			2.30	3	0.10	0.28	(0.00, 1.26)
SD of age			4.27	3	0.01	0.46	(0.10, 1.55)
Skewness of age			0.15	3	0.93	0.00	(0.00, 0.67)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.05	0.17	0.08	1	0.79	0.00	(0.00, 0.50)
Mean age			1.84	3	0.16	0.22	(0.00, 1.27)
SD of age			4.11	3	0.01	0.45	(0.00, 1.44)
Skewness of age			0.05	3	0.99	0.00	(0.00, 0.69)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S7. Meta-analysis results for the RESI for sex on total gray matter volume (GMV). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.03	0.11	0.06	1	0.80	0.00	(0.00, 0.31)
Mean age			0.45	3	0.72	0.00	(0.00, 0.71)
SD of age			0.40	3	0.75	0.00	(0.00, 0.50)
Proportion of male			0.18	3	0.91	0.00	(0.00, 0.49)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.05	0.12	0.19	1	0.66	0.00	(0.00, 0.29)
Mean age			0.26	3	0.85	0.00	(0.00, 0.68)
SD of age			0.41	3	0.75	0.00	(0.00, 0.48)
Proportion of male			0.22	3	0.88	0.00	(0.00, 0.47)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S8: The meta-analysis results for the RESI for sex on total subcortical gray matter volume (sGMV). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.07	0.12	0.27	1	0.60	0	(0, 0.32)
Mean age			0.60	3	0.62	0	(0, 0.69)
SD of age			0.59	3	0.62	0	(0, 0.49)
Proportion of male			0.12	3	0.95	0	(0, 0.39)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.03	0.15	0.05	1	0.82	0.00	(0, 0.34)
Mean age			1.34	3	0.27	0.12	(0, 0.86)
SD of age			1.26	3	0.30	0.10	(0, 0.53)
Proportion of male			0.15	3	0.93	0.00	(0, 0.41)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S9. The meta-analysis results for the RESI for sex on total white matter volume (WMV). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.10	0.12	0.70	1	0.41	0.00	(0.00, 0.42)
Mean age			0.10	3	0.96	0.00	(0.00, 0.48)
SD of age			0.13	3	0.94	0.00	(0.00, 0.43)
Proportion of male			0.28	3	0.84	0.00	(0.00, 0.48)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.07	0.13	0.28	1	0.60	0.00	(0.00, 0.31)
Mean age			0.01	3	1.00	0.00	(0.00, 0.46)
SD of age			0.33	3	0.80	0.00	(0.00, 0.44)
Proportion of male			0.34	3	0.79	0.00	(0.00, 0.46)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S10. The meta-analysis results for the RESI for sex on mean cortical thickness (CT). The residual degrees of freedom are 52. The regression model was weighted by the inverse standard errors of the RESIs. For cross-sectional studies the outcome is RESI, for longitudinal studies the outcome is (**A**) L-RESI or (**B**) CS-RESI.

C. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.01	0.03	0.05	1	0.82	0.00	(0.00, 0.30)
Mean age			0.27	3	0.85	0.00	(0.00, 0.77)
SD of age			0.48	3	0.70	0.00	(0.00, 0.63)
Proportion of male			0.77	3	0.52	0.00	(0.00, 0.86)

D. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	-0.02	0.03	0.26	1	0.61	0.00	(0.00, 0.36)
Mean age			0.21	3	0.89	0.00	(0.00, 0.69)
SD of age			0.25	3	0.86	0.00	(0.00, 0.54)
Proportion of male			0.73	3	0.54	0.00	(0.00, 0.72)

*S.E.: robust (sandwich) standard error; F: F statistic; d.f.: the first degree(s) of freedom for the F statistics; RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. S11. The descriptive summary of the non-brain measures in ABCD study used for the structural brain outcome analyses. The records with missing global and regional brain measures and the subjects with missing baseline records were removed.

	Baseline (*N*=11725)	**2-year follow-up (*N*=8023)**	**4-year follow-up (*N*=3000)**	**Overall (*N*=22748)**
Picture Vocabulary Test				
Mean (SD)	52.3 (11.0)	49.7 (10.2)	50.1 (10.3)	51.1 (10.7)
Median [Min, Max]	51.0 [-12.0, 131]	49.0 [19.0, 88.0]	49.0 [19.0, 81.0]	50.0 [-12.0, 131]
Missing	748 (6.4%)	463 (5.8%)	135 (4.5%)	1346 (5.9%)
Flanker Inhibitory Control and Attention Test				
Mean (SD)	46.1 (9.13)	46.9 (9.64)	48.2 (10.6)	46.6 (9.54)
Median [Min, Max]	46.0 [10.0, 113]	46.0 [19.0, 81.0]	47.0 [19.0, 81.0]	46.0 [10.0, 113]
Missing	754 (6.4%)	989 (12.3%)	224 (7.5%)	1967 (8.6%)
List Sorting Working Memory Test				
Mean (SD)	49.5 (9.90)	48.2 (10.5)	51.1 (10.1)	49.8 (9.97)
Median [Min, Max]	48.0 [9.00, 108]	46.0 [19.0, 68.0]	51.0 [19.0, 81.0]	49.0 [9.00, 108]
Missing	790 (6.7%)	7966 (99.3%)	146 (4.9%)	8902 (39.1%)
Dimensional Change Card Sort Test				
Mean (SD)	47.5 (9.65)	47.6 (10.6)	47.5 (11.6)	47.5 (9.67)
Median [Min, Max]	46.0 [24.0, 132]	45.0 [26.0, 76.0]	45.0 [26.0, 81.0]	46.0 [24.0, 132]
Missing	751 (6.4%)	7974 (99.4%)	2923 (97.4%)	11648 (51.2%)
Pattern Comparison Processing Speed Test				
Mean (SD)	45.3 (14.4)	54.4 (13.7)	60.4 (13.8)	50.4 (15.2)
Median [Min, Max]	45.0 [-5.00, 100]	54.0 [19.0, 84.0]	61.0 [19.0, 81.0]	50.0 [-5.00, 100]
Missing	769 (6.6%)	1021 (12.7%)	229 (7.6%)	2019 (8.9%)
Picture Sequence Memory Test				
Mean (SD)	49.5 (11.0)	53.1 (11.4)	54.9 (13.6)	51.5 (11.7)
Median [Min, Max]	49.0 [10.0, 96.0]	52.0 [19.0, 81.0]	54.0 [19.0, 81.0]	50.0 [10.0, 96.0]
Missing	757 (6.5%)	396 (4.9%)	152 (5.1%)	1305 (5.7%)

Oral Reading Recognition Test				
Mean (SD)	49.4 (11.6)	48.7 (10.6)	51.4 (10.8)	49.4 (11.2)
Median [Min, Max]	47.0 [26.0, 142]	48.0 [24.0, 97.0]	51.0 [23.0, 81.0]	48.0 [23.0, 142]
Missing	759 (6.5%)	487 (6.1%)	153 (5.1%)	1399 (6.2%)
Cognition Fluid Composite				
Mean (SD)	45.8 (11.1)	44.9 (12.1)	47.4 (12.0)	45.8 (11.2)
Median [Min, Max]	45.0 [10.0, 128]	41.5 [24.0, 79.0]	47.5 [23.0, 73.0]	45.0 [10.0, 128]
Missing	961 (8.2%)	7971 (99.4%)	2950 (98.3%)	11882 (52.2%)
Crystallized Composite				
Mean (SD)	51.0 (11.2)	49.2 (10.7)	50.0 (11.3)	50.3 (11.1)
Median [Min, Max]	49.0 [8.00, 147]	48.0 [21.0, 100]	49.0 [22.0, 81.0]	49.0 [8.00, 147]
Missing	910 (7.8%)	1652 (20.6%)	2493 (83.1%)	5055 (22.2%)
Cognition Total Composite Score				
Mean (SD)	47.8 (11.2)	46.2 (10.4)	48.8 (12.3)	47.8 (11.2)
Median [Min, Max]	47.0 [5.00, 155]	45.5 [29.0, 71.0]	48.5 [24.0, 78.0]	47.0 [5.00, 155]
Missing	965 (8.2%)	7971 (99.4%)	2950 (98.3%)	11886 (52.3%)
Total Problem CBCL Syndrome Scale				
Mean (SD)	45.8 (11.4)	44.9 (11.3)	44.4 (11.5)	45.3 (11.4)
Median [Min, Max]	45.0 [24.0, 83.0]	45.0 [24.0, 88.0]	45.0 [24.0, 84.0]	45.0 [24.0, 88.0]
Missing	7 (0.1%)	1345 (16.8%)	23 (0.8%)	1375 (6.0%)
Birth Weight (in oz)				
Mean (SD)	112 (23.5)	113 (23.6)	112 (24.0)	112 (23.6)
Median [Min, Max]	115 [34.0, 236]	115 [34.0, 236]	115 [41.0, 200]	115 [34.0, 236]
Missing	1245 (10.6%)	762 (9.5%)	238 (7.9%)	2245 (9.9%)
BMI				
Mean (SD)	18.9 (3.96)	20.6 (4.53)	22.4 (4.81)	19.9 (4.46)
Median [Min, Max]	17.7 [13.5, 36.8]	19.4 [13.5, 36.9]	21.3 [13.5, 36.8]	18.8 [13.5, 36.9]
Missing	248 (2.1%)	140 (1.7%)	107 (3.6%)	495 (2.2%)

Handedness				
Right-handed	9312 (79.4%)	6368 (79.4%)	2448 (81.6%)	18128 (79.7%)
Not right-handed	2413 (20.6%)	1655 (20.6%)	552 (18.4%)	4620 (20.3%)

Tab. S12. The descriptive summary of the non-brain measures in the subset of ABCD study used for the functional connectivity (FC) outcome analyses. Only the subjects with both baseline and 2-year follow-up are included.

	Baseline (*N*=1192)	**2-year follow-up (*N*=1192)**	**Overall (*N*=2384)**
Picture Vocabulary Test			
Mean (SD)	52.8 (11.2)	50.2 (10.4)	51.5 (10.9)
Median [Min, Max]	52.0 [14.0, 112]	49.0 [19.0, 88.0]	50.0 [14.0, 112]
Missing	62 (5.2%)	75 (6.3%)	137 (5.7%)
Flanker Inhibitory Control and Attention Test			
Mean (SD)	46.0 (9.17)	46.9 (9.40)	46.5 (9.29)
Median [Min, Max]	45.5 [10.0, 113]	46.0 [19.0, 81.0]	46.0 [10.0, 113]
Missing	62 (5.2%)	130 (10.9%)	192 (8.1%)
List Sorting Working Memory Test			
Mean (SD)	50.1 (9.55)	52.5 (12.0)	50.1 (9.58)
Median [Min, Max]	49.0 [26.0, 91.0]	48.0 [34.0, 68.0]	49.0 [26.0, 91.0]
Missing	67 (5.6%)	1181 (99.1%)	1248 (52.3%)
Dimensional Change Card Sort Test			
Mean (SD)	47.9 (9.78)	47.1 (14.4)	47.9 (9.81)
Median [Min, Max]	46.0 [26.0, 98.0]	41.0 [33.0, 71.0]	46.0 [26.0, 98.0]
Missing	63 (5.3%)	1184 (99.3%)	1247 (52.3%)
Pattern Comparison Processing Speed Test			
Mean (SD)	45.9 (13.9)	54.4 (13.3)	50.0 (14.3)
Median [Min, Max]	46.0 [-2.00, 88.0]	54.0 [19.0, 81.0]	50.0 [-2.00, 88.0]
Missing	64 (5.4%)	137 (11.5%)	201 (8.4%)
Picture Sequence Memory Test			
Mean (SD)	49.8 (11.2)	53.6 (11.4)	51.7 (11.4)
Median [Min, Max]	49.0 [15.0, 96.0]	52.0 [19.0, 81.0]	51.0 [15.0, 96.0]
Missing	63 (5.3%)	66 (5.5%)	129 (5.4%)

Oral Reading Recognition Test			
Mean (SD)	49.6 (11.3)	49.4 (10.9)	49.5 (11.1)
Median [Min, Max]	48.0 [28.0, 118]	48.0 [24.0, 97.0]	48.0 [24.0, 118]
Missing	64 (5.4%)	78 (6.5%)	142 (6.0%)
Cognition Fluid Composite			
Mean (SD)	46.4 (11.2)	49.8 (17.9)	46.5 (11.3)
Median [Min, Max]	46.0 [20.0, 113]	47.0 [29.0, 79.0]	46.0 [20.0, 113]
Missing	81 (6.8%)	1182 (99.2%)	1263 (53.0%)
Crystallized Composite			
Mean (SD)	51.3 (11.2)	49.9 (10.9)	50.7 (11.1)
Median [Min, Max]	50.0 [22.0, 110]	49.0 [23.0, 100]	49.0 [22.0, 110]
Missing	80 (6.7%)	201 (16.9%)	281 (11.8%)
Cognition Total Composite Score			
Mean (SD)	48.4 (11.2)	50.7 (12.8)	48.4 (11.2)
Median [Min, Max]	48.0 [18.0, 127]	53.5 [29.0, 70.0]	48.0 [18.0, 127]
Missing	84 (7.0%)	1182 (99.2%)	1266 (53.1%)
Total Problem CBCL Syndrome Scale			
Mean (SD)	45.4 (11.1)	44.3 (11.0)	44.9 (11.1)
Median [Min, Max]	45.0 [24.0, 81.0]	44.5 [24.0, 78.0]	45.0 [24.0, 81.0]
Missing	0 (0%)	160 (13.4%)	160 (6.7%)
Birth Weight (in oz)			
Mean (SD)	112 (23.8)	112 (23.8)	112 (23.8)
Median [Min, Max]	114 [35.0, 208]	114 [35.0, 208]	114 [35.0, 208]
Missing	121 (10.2%)	116 (9.7%)	237 (9.9%)
BMI			
Mean (SD)	18.0 (3.27)	19.5 (3.67)	18.7 (3.56)
Median [Min, Max]	17.1 [13.2, 30.7]	18.7 [13.2, 32.2]	17.9 [13.2, 32.2]
Missing	26 (2.2%)	21 (1.8%)	47 (2.0%)

Handedness			
Right-handed	968 (81.2%)	968 (81.2%)	1936 (81.2%)
Not right-handed	224 (18.8%)	224 (18.8%)	448 (18.8%)

Tab. S13. The estimated cross-sectional and longitudinal parameters of the non-brain covariates on total gray matter volume (GMV), respectively, in ABCD study. See Methods for more details about the estimation of the cross-sectional and longitudinal parameters. The parameter estimates that are significantly different from zero are in bold. The intra-subject correlation of the outcome total GMV is estimated from the GEEs (generalized estimating equations) with an exchangeable correlation structure used for the modeling.

Outcome	Non-brain covariate	Between-subject parameter (95% CI)	Within-subject parameter (95% CI)	Intra-subject correlation
Total GMV	Crystallized Composite	**0.09** (0.08, 0.10)	-0.01 (-0.01, 0.00)	0.90
	Flanker Inhibitory Control and Attention	0.01 (-0.01, 0.02)	0.00 (0.00, 0.01)	0.91
	List Sorting Working Memory	0.06 (0.05, 0.07)	0.00 (-0.02, 0.01)	0.80
	Pattern Comparison Processing Speed	**0.01** (0.00, 0.01)	0.00 (-0.01, 0.00)	0.91
	Picture Sequence Memory	**0.01** (0.00, 0.02)	0.00 (0.00, 0.01)	0.92
	Picture Vocabulary	**0.07** (0.06, 0.08)	0.00 (0.00, 0.01)	0.91
	Oral Reading Recognition	**0.07** (0.06, 0.08)	**-0.01** (-0.01, 0.00)	0.91
	CBCL	**-0.04** (-0.05, -0.03)	-0.01 (-0.01, 0.00)	0.92
	BMI	**-0.13** (-0.16, -0.11)	-0.01 (-0.03, 0.01)	0.92

*CI: confidence interval; GMV: gray matter volume.