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], Jonathan Schildcrout[1]; Ran Tao[1]; Jiangmei Xiong[1]; Megan T. Jones[1]; 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; Aaron Alexander-Bloch[2,3,4,6] and Simon Vandekar[1,6,*]

[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]These authors contributed equally.
*Corresponding authors

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 a meta-analysis of a robust effect size index (RESI) using 63 longitudinal and cross-sectional magnetic resonance imaging studies (75,255 total scans) to demonstrate that optimizing study design is an important way to improve standardized effect sizes in BWAS. Our results of brain volume associations with demographic and cognitive variables indicate that BWAS with larger standard deviation of the independent variable have larger effect size estimates and that longitudinal studies have systematically larger standardized effect sizes than cross-sectional studies by 290%. 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. Using bootstrapping in the Lifespan Brain Chart Consortium we show that modifying study design to increase between-subject standard deviation by 45% increases standardized effect sizes by 42% and adding a second measurement per subject can increase effect sizes by 35%. These findings underscore the importance of considering design features in BWAS and emphasize that increasing sample size is not the only approach to improve the replicability of BWAS.

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

Brain-wide association studies (BWAS) use noninvasive magnetic resonance imaging (MRI) to identify associations of inter-individual differences in behavior, cognition, age, sex, and other biological or clinical measurements with brain structure and function[1,2]. In neuropsychiatric research, the fundamental goal of BWAS is to identify neurological differences that may serve as markers for diagnosis that precede formation of persistent and severe symptoms of a disorder. Robust imaging markers could 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]. Statistical replicability is defined as "obtaining consistent results across studies that had obtained their own data and aimed at answering the same scientific question"[4]. Like power, replicability is a function of the standardized effect sizes (ESs) and the sample size[5,6]. Low replicability in BWAS has been attributed to a combination of small ESs, small sample sizes, and publication bias[1,2,7–11].

The most obvious solution to improve power and replicability of univariate and multivariate BWAS is to increase study sample sizes. Several recent studies showed that thousands of study participants are required to improve the replicability of BWAS because standardized ESs are much smaller than expected based on biased estimates reported in smaller studies[1,2]. To further improve the utility of MRI as a neurological marker and the replicability of BWAS it is critical to understand modifiable features of the study population that can be implemented to increase standardized ESs. In particular, using a longitudinal design can improve power by increasing efficiency and leveraging intra-individual associations between brain and cognitive or biological measurements[1,12–14]. The benefit of considering a longitudinal design 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 over 75,000 scans from the Lifespan Brain Chart Consortium[15] (LBCC) to demonstrate the influence of sample population and longitudinal/cross-sectional study designs on ESs. See reference[15] for details on LBCC datasets, which include multiple publicly available open MRI datasets[16–26]. We use a recently proposed robust ES index (RESI)[27–29] to 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)) among the cognitively normal (CN) participants. As we show, standardized ESs are larger in studies with greater sample heterogeneity (represented by the standard deviation of the independent variable) and ESs for age are dependent on the mean and standard deviation of the age of the population being studied. Longitudinal studies have systematically larger ESs by 290%, which compromises the comparability between cross-sectional and longitudinal studies in meta-analyses. We propose an estimate of 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. We bootstrap data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to show that increasing between-subject variability in the age variable by 45% and adding a second longitudinal measurement per subject increases the GMV-age ES by 42% and 35%, respectively. Finally, we show that increasing between-subject variability of cognitive variables

in the Adolescent Brain Cognitive Development (ABCD) study increases the GMV-cognition ES, but increasing within-subject standard deviation provides no benefit.

Meta-analysis shows ESs depend on study population and design

To fit each study level analysis, we nonlinearly regressed GMV, sGMV and WMV on age and sex in 63 neuroimaging datasets from the LBCC using regression models for the 47 cross-sectional datasets and generalized estimating equation (GEE) models for the 16 longitudinal datasets (See Methods for more details). These analyses encompassed 75,255 MRI scans from 60,913 cognitively normal (CN) participants (sample and demographic information in Tab. S1). We used the robust ES index (RESI) as a standardized measure of ES. 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[27,28] and is equal to ½ Cohen's *d* under some assumptions[27], so Cohen's suggested interpretations for ES are: small (RESI=0.1); medium (RESI=0.25); and large (RESI=0.4). 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 "longitudinal RESI (L-RESI)".

To investigate the effects of study population and design on the RESIs, we performed a meta-analysis across the 63 studies in the LBCC to model the association of study-level population and design features with ESs. Study population and design features were quantified as the sample mean, standard deviation (SD), and skewness of the age variable as nonlinear terms, and a binary variable indicating the design type (cross-sectional/longitudinal). After obtaining estimates of the RESIs of age and sex in each analysis of GMV, sGMV, and WMV, we conducted a meta-analysis of the RESI or L-RESI estimates using a weighted linear regression model with study population and design features as variables (see Methods for more details).

For GMV, the partial regression plots showed the effect of each variable controlling for others in the model and demonstrated a strong cubic-shape relationship between the ES (RESI or L-RESI) of age on GMV and study population mean age, indicating that the strength of the age effect varies with respect to the age of the population being studied, with the largest age ESs on GMV in the human lifespan occurring during early and late adulthood (Tab. 1A; Fig. 1A). There was a strong positive linear effect of the study population SD of age and the GMV-age ESs: for each unit increase in the SD of age (in years), expected RESI increased by about 0.1 (Fig. 1A). Plots of sGMV and WMV showed u-shape changes of the age RESI with respect to study population mean age (Fig. 1B & C), and a similar, but weaker relationship between RESI and study population SD and skewness (Tab. S2 & S3). Finally, the meta-analysis also showed a moderate effect of study design on the RESI of age on GMV, sGMV, and WMV (Fig. 1; Tab. 1; Tab. S2A & S3A). The average RESI of age on GMV in longitudinal studies (RESI=0.355) is 0.26 larger than in cross-sectional studies (RESI=0.091) after controlling for the study population variables, corresponding to a 290% 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 did not improve the RESI for biological sex, because the variable does not vary within-subject. (Tab. S4-S6; Fig. S2).

These results demonstrated that standardized ESs are dependent on study population features (such as mean age and SD of age) and study design features (such as cross-

sectional/longitudinal design) and suggest that modifying study design features, such as increasing variability and conducting longitudinal studies can increase ES in BWAS.

Cross-sectional RESI (CS-RESI) quantifies benefit of longitudinal design

While the systematic difference in the ES between cross-sectional and longitudinal study designs compromised the comparability of the 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 defined and evaluated a new CS-RESI and its confidence interval (see Supplement: Section 1). The CS-RESI represents the RESI in the same study population if the longitudinal study had been conducted cross-sectionally, and allows researchers to compare ESs from studies using different designs (cross-sectional or longitudinal). The difference between the L-RESI and CS-RESI quantifies the benefit of using the longitudinal study design within a single study.

To demonstrate the difference in ES by study design and validate the CS-RESI estimator, we used non-parametric bootstrapping to create longitudinal GMV subsamples by randomly sampling subjects with replacement (1,000 bootstraps per sample size) in the 16 longitudinal neuroimaging datasets. At a given sample size, after obtaining the longitudinal subsample, we created an equivalent cross-sectional sample by randomly selecting one observation from each subject. In each replicate, we fit the same GEE and linear regression models of GMV to the longitudinal and cross-sectional samples to quantify the effects of age and sex on GMV. We obtained the differences of the RESI estimated in the cross-sectional data to the new CS-RESI and L-RESI estimated in the longitudinal data.

When comparing the L-RESI estimated in the longitudinal data to the RESI estimated in cross-sectional data, the longitudinal ESs for age and sex were larger in two large longitudinal datasets in the LBCC (Fig. 2A). The positive differences represent the amount of benefit of using longitudinal designs for these two studies in quantifying the age or sex effect, 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. The bias 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 bias (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 approached zero as the sample size increased (Fig. 2B), indicating that the systematic difference in ESs between longitudinal and cross-sectional datasets was removed. The bootstrap results across 16 studies indicated the benefits of their longitudinal study designs in detecting age and sex effects (Fig. S3). As a further comparison, we repeated the meta-analysis above on the RESIs from all studies using the CS-RESIs in longitudinal studies, demonstrating a reduction in the design effect (Tab. 1B; Tab S2B & S3B).

The benefit of conducting the longitudinal design can be quantified in each study (without bootstrapping), by comparing the CS-RESI estimates with the L-RESI estimates (Fig. 2C-E). Longitudinal designs resulted in a 0.23 increase in the GMV-age ES among the 16 longitudinal neuroimaging datasets, corresponding to a moderate change in ES, using Cohen's recommendations[30]. Importantly, in some studies (e.g., POND and SCZIowa), the L-RESI was smaller than the CS-RESI. This counterintuitive result was due to repeated measurements introducing variance that was not related to the outcome. This interpretation is supported by

theory (Supplement: Section 2) and by our subsequent analyses of cognitive variables in the ABCD study (see below). The benefits of a longitudinal design are substantially reduced for sex relative to the age effect because the benefit is dependent on the variability of the within-subject measurement: age varies across within-subject measurements, while biological sex does not (Fig. S4).

Modifying within- and between-subject sampling can increase ESs

Consistent with statistical principles, the meta-analysis results above suggest that the larger SD of the primary variable produces larger standardized ESs. To investigate the effects of modifying the SD of the age variable on ES in BWAS, we bootstrapped the cross-sectional United Kingdom BioBank (UKB) data under 3 sampling schemes (scheme 0-2; Fig. 3A) that sequentially increase the SD of the age variable in the sample by drawing subjects from the tails of age distribution in the study with higher probability (see Methods). In the UKB, sampling subjects more heavily from the tails of the age distribution increased the SD of age by 25% (scheme 1 vs. scheme 0) and 39% (scheme 2 vs. scheme 0; Fig. 3B; Tab. S7) and with corresponding 21% and 36% increases in the RESI, respectively (Fig. 3C; Tab. S7).

We implemented the same three resampling schemes between- and within-subject in the ADNI dataset to investigate the effect of increasing the SD of the age variable longitudinally (Fig. S5A). Similar to the cross-sectional UKB analysis, increases in the between-subject SD of age by 28% and 45% were associated with increases of 22% and 42% in the CS-RESIs and increases of 10% and 20% in L-RESIs, respectively (Fig. 3D; Tab. S8A). However, increasing the within-subject SD of age (Fig. S5A) did not substantially increase the average observed ESs (Fig. 3D; Tab. S8A). Comparing the CS-RESI to L-RESI suggests that any type of repeated measurement on age increased ESs in a longitudinal design compared to a cross-sectional design, but, when comparing longitudinal designs, increasing within-subject variance of age did not improve age ESs. The reduced influence of within-subject variance may be because it accounted for only 23%, 19%, and 16% of the total variance of the age variable in schemes 0-2, respectively, for fixed within-subject scheme 0.

Finally, we investigated the effect of the number of observations per subject on ESs. The ADNI dataset we used included on average 5.35 observations per subject. We performed the same bootstrap analysis as above, but included only the baseline observation and one random follow-up observation for each subject to reduce the average per-subject observation number to two while keeping the within-subject SD of age at the same level (Fig. S5; Tab. S8B). Reducing the number of observations per subject did not substantially decrease the observed CS- and L-RESI of age on GMV (Fig. 3E; Tab. S8B).

Together, these results suggest having larger spacing in between-subject age measurements increases ES and that conducting a longitudinal study increases ES. Modifications to within-subject study design did not confer a substantial increase in ES for the GMV-age association. Reducing the number of within-subject measurements to two did not substantially reduce ESs. Comparing L-RESI and CS-RESI demonstrated that adding a single longitudinal measurement to a cross-sectional dataset can increase the GMV-age ES by 35% (Tab. S8B).

Effects of within- and between-subject sampling on cognitive variable RESIs

in the ABCD cohort

To explore whether increasing the SD of cognitive measurements affects the RESI of GMV, sGMV, and WMV in a similar way to age, we bootstrapped 11,811 subjects with cross-sectional or longitudinal measurements (either cognitively normal or atypical) from the ABCD dataset to analyze 4 cross-sectional and 6 longitudinal standardized cognitive measurements from the NIH toolbox obtained at baseline and 1-year follow-up[31] (Tab. S9; Methods). We used sampling schemes 0 and 2 applied above, to increase the between-subject SD of each cognitive measurement (Methods) and used two resampling schemes to increase the SD of the two cognitive measurements within-subject (schemes 0 and 1; Fig. 4A; Methods). Results showed small to medium ESs for all GMV-cognition associations (Fig. 4; Tab. 2). As with the GMV-age association above, increasing the between-subject SD of the cognitive measurement increased the GMV-cognition RESI in both the cross-sectional (Fig. 4B) and longitudinal analyses (Fig. 4C) unless the RESI was close to zero.

For the 6 cognitive measurements analyzed longitudinally, increasing the within-subject SD of the cognitive measurements did not increase CS-RESI or L-RESI, and in some cases it decreased (Fig. 4C). In addition, the L-RESI was equal to or smaller than CS-RESI across all cognitive variables. These results indicate that there is no efficiency improvement in conducting a longitudinal study design to study GMV-cognition associations in the ABCD study population. Similar patterns were observed for sGMV and WMV (Fig. S6 and S7).

Discussion

BWAS are a fundamental tool in discovering brain-behavior associations that may inform our understanding of neuropsychiatric disorders and their treatments. As the evidence of small ESs increases, optimizing study design to improve standardized ESs is critical to identify markers that might be indicative of psychiatric disorders. We used the RESI to show the influence of study design characteristics on the ESs of age, sex, and cognitive measures with GMV, sGMV, and WMV using 63 neuroimaging datasets. Our results indicated that longitudinal studies have larger standardized ESs than cross-sectional studies and that larger between-subject SD of the variable of interest systematically increases standardized ESs. In contrast, larger within-subject SD was not associated with increases in ESs, indicating that there may be little benefit to optimizing this within-subject design feature in longitudinal studies. 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 study design differences when comparing longitudinal and cross-sectional studies. While large sample sizes remain a significant consideration, our findings emphasize the importance of considering other design features in BWAS to improve standardized ESs.

Study design considerations

Increasing the independent variable SD corresponds to a similar factor increase in the ES. Thus, increasing ESs can be as simple as sampling subjects with small and large values of the variable of interest with higher probability. We showed that increasing a variable's SD improved ESs, even when there was significant nonlinearity in the association, such as with age and GMV (Fig. S1). Sampling schemes that increase independent variable SD are easily implemented to variables that have low measurement cost and are stable over short time periods. For cognitive or other behavioral measurements that can vary substantially over short time periods, more

complex designs may be necessary to increase standardized ESs because investigators might need to collect these measurements contemporaneously with the imaging data.

In our meta-analysis, longitudinal studies on average had 290% larger standardized ESs than cross-sectional studies. Statistically, the benefit of conducting a longitudinal design is a complicated function of the within-subject correlation of the brain measurement, the within-subject correlation of the independent variable, and the association of within-subject changes in GMV with within-subject changes in the independent variable. The apparent lack of benefit in the GMV-cognitive measurement association is due to a lack of association between state variability in the cognitive measurements with within-subject changes in GMV (Fig. S8). This reduces the ES by averaging the stronger between-subject effect with the weaker within-subject association. In contrast, there was a benefit of longitudinal design in the ADNI as there was correlation between the within-subject change in age and the change in GMV (Fig. S9).

Because standardized ESs are directly related to study efficiency, design considerations that optimize study efficiency will increase ESs and study replicability. Two-phase and outcome-dependent sampling designs can inform what subjects should be selected for imaging from a larger sampling in order to increase the efficiency and ESs of imaging-behavior associations[32–34].

Ranges of ES in BWAS

We found that standardized ESs depend on the mean age of the study population and vary substantially across human growth: for the GMV-age association ESs were largest in young and late adulthood when age-related changes in GMV are strongest. These are the times when age-related differences account for the largest proportion of variance in GMV and may be due to a reduced influence of other factors at these periods of the lifespan.

Effect sizes in our analyses ranged from 0-1.14 across datasets (in the GMV-age association) and 0-0.16 in the cognitive variables corresponding to absent to very large and absent to small effects, respectively, using Cohen's recommendations[30]. These ranges are more encouraging than previous reports[1,2], and may differ for several reasons: our focus on age effects and broad cerebral tissue classes is likely to find stronger ESs; we considered datasets with a larger age SD which will have larger ESs, compared to the primary focus on the ABCD sample in previous papers; lastly, previous reports collapsed across multiple scales of imaging data and different behavioral measures, many of these associations are likely to be at or near null values and the distribution of ESs in those papers will be a mixture of null and non-null findings.

Between- and within-subject effects

Our analyses assume the between- and within-subject associations are equal. Under this assumption, longitudinal studies guarantee larger ESs when the within-subject correlation of the outcome and the within-subject correlation of the independent variable are both positive[35]. This is likely to be common in imaging and behavioral data collected for BWAS. However, these benefits do not hold when the cross-sectional and longitudinal effects are different (see Supplement: Section 2 for a more technical discussion). Cross-sectional and longitudinal effects can be different when, for example, there is an association between the trait of a cognitive measurement and the outcome, but no association between cognitive measurement state. For

these types of associations, cross-sectional and longitudinal effects can be modeled as separate fixed effects[35–37] and modeling them separately might yield different conclusions with respect to the detrimental effects of longitudinal design we found in the ABCD cognitive variables.

Univariate versus Multivariate BWAS

It is important to distinguish univariate and multivariate BWAS[3,38]. Univariate methods view the imaging variable as the outcome and the models seek to identify brain regions that are correlated with cognitive outcome. In contrast, multivariate methods treat the cognitive measurement as the outcome and use machine learning methods to predict the outcome variable as a function of the imaging variable. Multivariate methods have larger ESs[39] than univariate methods because the signal in neuroimaging data is likely to be indirect and dispersed across networks of brain regions. Though we focused on univariate analyses with a small number of brain measurements, our findings directly generalize to mass-univariate methods because ESs do not depend on the number of comparisons. Our findings may generalize to the multivariate setting, where the imaging variable is used as the covariate, based on previous research that optimizes power using outcome-dependent sampling and two-phase study design[34].

Conclusions

While the principles for improving standardized ESs we illustrate are generally applicable, the range of ESs and the benefits of conducting longitudinal research are outcome and variable dependent. Increasing the independent variable SD in the study population has a deterministic increase in standardized ESs, but the benefit of a longitudinal design is dependent on the form of the association between the variable and outcome. We have considered design features to improve ESs and replicability but analysis methods to improve power can also improve replicability, such as by using more efficient statistical methods[40,41]. Together, these points highlight the importance of rigorous evaluation of statistical design and analysis methods to reduce costs and 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. National Academies of Sciences, E. *et al. Understanding Reproducibility and Replicability*.

Reproducibility and Replicability in Science (National Academies Press (US), 2019).

5. Nosek, B. A. *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
6. 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).
7. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
8. 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).
9. Reddan, M. C., Lindquist, M. A. & Wager, T. D. Effect Size Estimation in Neuroimaging. *JAMA Psychiatry* **74**, 207–208 (2017).
10. Vul, E. & Pashler, H. Voodoo and circularity errors. *NeuroImage* **62**, 945–948 (2012).
11. 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).
12. Nee, D. E. fMRI replicability depends upon sufficient individual-level data. *Commun. Biol.* **2**, 1–4 (2019).
13. Smith, P. L. & Little, D. R. Small is beautiful: In defense of the small-N design. *Psychon. Bull. Rev.* **25**, 2083–2101 (2018).
14. 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).
15. Bethlehem, R. a. I. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).
16. 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).

17. 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).
18. Lukas Snoek *et al.* AOMIC-PIOP1. *OpenNeuro* (2020) doi:10.18112/openneuro.ds002785.v2.0.0.
19. Bilder, R *et al.* UCLA Consortium for Neuropsychiatric Phenomics LA5c Study. *OpenNeuro* (2020) doi:10.18112/openneuro.ds000030.v1.0.0.
20. Samuel A. Nastase; *et al.* Narratives. *OpenNeuro* (2020) doi:10.18112/openneuro.ds002345.v1.1.4.
21. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* **4**, 170181 (2017).
22. Lukas Snoek; *et al.* AOMIC-PIOP2. *OpenNeuro* (2020) doi:10.18112/openneuro.ds002790.v2.0.0.
23. 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).
24. Kuklisova-Murgasova, M. *et al.* A dynamic 4D probabilistic atlas of the developing brain. *NeuroImage* **54**, 2750–2763 (2011).
25. Lukas Snoek; *et al.* AOMIC-ID1000. *OpenNeuro* (2021) doi:10.18112/openneuro.ds003097.v1.2.1.
26. Reynolds, J. E., Long, X., Paniukov, D., Bagshawe, M. & Lebel, C. Calgary Preschool magnetic resonance imaging (MRI) dataset. *Data Brief* **29**, 105224 (2020).
27. Vandekar, S., Tao, R. & Blume, J. A Robust Effect Size Index. *Psychometrika* **85**, 232–246 (2020).
28. 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.

29. Jones, M., Kang, K. & Vandekar, S. RESI: Robust Effect Size Index (RESI) Estimation. (2022).

30. Cohen, J. *Statistical power analysis for the behavioral sciences*. (Erlbaum Associates, 1988).

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

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

33. 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).

34. 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).

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

36. Neuhaus, J. M. & Kalbfleisch, J. D. Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics* **54**, 638–645 (1998).

37. Guillaume, B., Hua, X., Thompson, P. M., Waldorp, L. & Nichols, T. E. Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. *NeuroImage* **94**, 287–302 (2014).

38. 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).

39. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* **615**, E4–E7 (2023).

40. Noble, S., Scheinost, D. & Constable, R. T. Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. *NeuroImage* **209**, 116468 (2020).

41. 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).

42. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).

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

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

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

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

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

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 the prior work[15] (see Supplemental Information and Tab. S1 from prior study). We filtered to a subset of 61,091 cognitively normal 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. The final dataset includes 75,255 MRI scans from 60,913 cognitively normal participants across 63 studies (Tab. S1).

Details of data processing are described in the prior work[15]. Briefly, 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).

Cognitive measurements

The ABCD dataset is a large-scale repository aiming to track the brain and psychological development of over 10,000 children from ages 9 to 10 by measuring hundreds of variables including demographic, physical, cognitive and mental health variables[42]. The NIH Toolbox is a set of assessment tools designed to measure cognitive, emotional, sensory, and motor functions. In the current study, we examined the effects of the 10 cognitive measures from the NIH Toolbox (fully corrected *T*-scores of the individual subscales and composite scores)[31] on the three main tissue volumes of the cerebrum (i.e., GMV, sGMV and WMV) after controlling for age and sex using 11,811 cognitively normal or atypical participants in the ABCD study.

There are 4 cognitive measurements with less than 10% of the 1-year follow-ups (i.e., List Sorting Working Memory Test, Dimensional Change Card Sort Test, Cognition Fluid Composite, and Cognition Total Composite Score); only their baseline cognitive measurements were included in the analyses. For the remaining 6 variables (i.e., Picture Vocabulary Test, Flanker Inhibitory Control and Attention Test, Pattern Comparison Processing Speed Test, Picture Sequence Memory Test, Oral Reading Recognition Test, and Crystallized Composite), both the baseline and 1-year follow-up observations were used.

Statistical analysis

Removal of site effects

For multi-site or multi-study neuroimaging studies, we need to control the potential heterogeneities between sites to obtain unconfounded and generalizable results. Before

estimating the main effects of age and sex on the main tissue volumes of the cerebrum (GMV, WMV, and sGMV), we applied ComBat[43] and LongComBat[44] 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 GMV:

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

where ns denotes natural cubic splines on two degrees of freedom. 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 random intercepts:

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

When estimating the effects of cognitive measurements in the ABCD dataset, ComBat was used to control the site effects respectively for each of the 4 cross-sectional cognitive measurements. The ComBat model is specified as illustrated below using GMV:

GMV ~ ns(age, df = 2) + sex + cogn + site.

LongComBat was used for each of the 6 longitudinal cognitive measurements with a linear mixed effects model with random intercept only:

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

The Combat and LongComBat were implemented using the neuroCombat[45] and longCombat[46] R packages. Site effects were removed before all subsequent analyses including the bootstrap analyses described below.

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 the three main tissue volumes of the cerebrum using GEE and linear regression models in the longitudinal datasets and cross-sectional datasets, respectively. The mean model was specified as illustrated below using GMV after ComBat/LongComBat:

$$y_i \sim ns(age_i, df = 2) + sex_i ,$$

Where y_i was taken to be GMV, WMV, or sGMV and the age effect was estimated with natural cubic splines with 2 knots. For GEE models, an exchangeable correlation structure was specified as the working structure. GEE models were fit with the geepack package[47] in R. We used the robust ES index (RESI) as a standardized measure of ES. RESIs and confidence intervals were computed using the RESI R package (version 1.0.1)[28,29]. 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[27,28]. The RESI is equal to ½ Cohen's *d* under some assumptions[27], so Cohen's suggested interpretations for ES are: small (RESI=0.1); medium (RESI=0.25); and large (RESI=0.4).

Meta-analysis of the age and sex effects

The weighted linear regression model for the meta-analysis of age effects across the 63 studies was specified as:

$$\hat{S}_{age,i} = design_i + ns[mean(age)_i, 3] + ns[sd(age)_i, 3] + ns[skew(age)_i, 3] + \epsilon_i,$$

where $\hat{S}_{age,i}$ denotes the estimated RESI or L-RESI for study *i,* and the weights are the inverse of the standard error of each RESI estimate. The sample mean, standard deviation (SD), and skewness of the age were included as nonlinear terms estimated using natural splines with 3 knots, and a binary variable indicating the design type (cross-sectional/longitudinal) was also included. The weights were the inverse of the standard error of each RESI or L-RESI estimate. These analyses were performed for GMV, sGMV, and WMV.

The weighted linear regression model for the meta-analysis of sex effects across the 63 studies was specified as

$$\hat{S}_{sex,i} = design_i + ns[mean(age)_i, 3] + ns[sd(age)_i, 3] + ns[pr(male), 3] + \epsilon_i,$$

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

In the partial regression plots, the expected RESIs for the age effects in the partial regression plots were estimated from the meta-analysis model after fixing study design at cross-sectional, mean age at 45, SD of age at 7, and/or skewness at 0; the expected RESIs for the sex effects were estimated from the meta-analysis after fixing study design at cross-sectional, 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. Each bootstrap sample was created by sampling subjects with replacement and then a cross-sectional dataset was created by uniformly sampling a single measurement within each of the subjects selected in the bootstrap sample. L-RESI and CS-RESI were computed in each of the longitudinal bootstrap samples and compared to the RESI computed in the cross-sectional bootstrap sample. This analysis was repeated for all of the 16 longitudinal studies included in our LBCC subset. All site effects were removed using LongComBat prior to performing the bootstrap analysis.

Bootstrap analysis of sampling strategies for age in the UKB and ADNI

We used bootstrapping to evaluate different sampling strategies in the cross-sectional UKB and longitudinal ADNI datasets. To increase the between-subject SD of the age variable we used the following three sampling schemes – scheme 0: random sampling with equal probability; scheme 1: linear increase in the resampling probabilities of the observations closer to the range limits of age in the study; scheme 2: quadratically increasing the sampling probability near the range limits of age (Fig. 3A). In the ADNI dataset, after selecting each subject with replacement based on the baseline age percentile using the between-subject sampling scheme, we then sampled within-subject measurements with replacement from each subject using the same three sampling schemes weighting the probabilities based on the age percentiles within the subject. Sampling from the tails of the baseline age across subjects was used to increase the between-subject SD of age while maintaining the within-subject SD of age at approximately the same level (Fig. S5A). All subjects still have the same number of observations after the

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

To investigate the effect of modifying the number of observations per subject while controlling the within-subject SD of age at the same level, after having the above bootstrap data using the between-subject and within-subject resampling schemes, we only kept the baseline observation and a random follow-up observation for each subject.

Bootstrap analysis of resampling strategies for cognitive variables in ABCD

We used bootstrapping to study how different sampling strategies affect the RESI in the ABCD dataset. To increase the between-subject variability of each cognitive measurement, we applied sampling schemes 0 and 2, described above, in each bootstrap to the baseline measurements. Because there were only two measurements per subject in the ABCD dataset we could not use the same within-subject bootstrap used for the ADNI dataset. Instead, we applied two sampling schemes to sample subjects based on the percentile of the absolute difference between the cognitive measurements (Fig. 4A) to increase the within-subject variability of the cognitive measurement.

Linear regression models and GEEs were fitted to estimate the effect of each cognitive variable on GMV, sGMV and WMV after controlling for age and sex for the 4 cross-sectional and 6 longitudinal cognitive variables,

$$\text{GMV_combat} \sim \text{ns}(\text{age}, \text{df} = 2) + \text{cogn} + \text{sex}.$$

GEE models assumed an exchangeable working covariance structure.

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.

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. We acknowledge the usage of several openly shared MRI datasets: 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.

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.

TABLES:

Tab. 1: Meta-analysis results for the GMV-age association RESI. The residual degrees of freedom are 52. The regression model was weighted by the inversed standard errors of the RESIs.

A. Outcome: RESI or L-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.26	0.08	9.39	1	<0.01	0.36	(0.10, 0.66)
Mean age			8.03	3	<0.01	0.57	(0.22, 1.06)
SD of age			25.94	3	<0.01	1.07	(0.57, 1.77)
Skewness of age			2.47	3	0.07	0.26	(0.00, 0.74)

B. Outcome: RESI or CS-RESI

	Estimate	S.E.	*F*	d.f.	*p*-value	Meta-RESI	95% CI
Design: longitudinal	0.02	0.08	0.07	1	0.80	0.00	(0.00, 0.29)
Mean age			3.90	3	0.01	0.36	(0.18, 0.98)
SD of age			20.67	3	<0.01	0.95	(0.60, 1.71)
Skewness of age			0.80	3	0.50	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 effect size index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; SD: standard deviation

Tab. 2: The estimated effect size for each GMV-cognition association after controlling for age and sex. Only the baseline measurements were available for the first 4 cognitive measurements due to severe missingness in the 1-year follow-up (Tab. S9). For the remaining 6 measurements, both the baselines and 1-year follow-ups were used.

cognitive measurement	Design	RESI or CS-RESI (95% CI)	L-RESI (95% CI)
List Sorting Working Memory	CS	0.10 (0.08, 0.12)	
Dimensional Change Card Sort	CS	0.06 (0.05, 0.08)	
Cognition Fluid Composite	CS	0.06 (0.04, 0.08)	
Cognition Total Composite Score	CS	0.15 (0.13, 0.17)	
Picture Vocabulary	L	0.16 (0.13, 0.17)	0.06 (0.03, 0.08)
Flanker Inhibitory Control and Attention	L	0.03 (0.00, 0.05)	0.00 (0.00, 0.03)
Pattern Comparison Processing Speed	L	0.00 (0.00, 0.03)	0.00 (0.00, 0.03)
Picture Sequence Memory	L	0.03 (0.00, 0.05)	0.02 (0.00, 0.04)
Oral Reading Recognition	L	0.16 (0.14, 0.18)	0.04 (0.02, 0.07)
Crystallized Composite	L	0.19 (0.17, 0.22)	0.06 (0.04, 0.09)

* RESI: robust ES index; L-RESI: longitudinal RESI; CS-RESI: cross-sectional RESI; CI: confidence interval; CS: cross-sectional; L: longitudinal.

FIGURES:

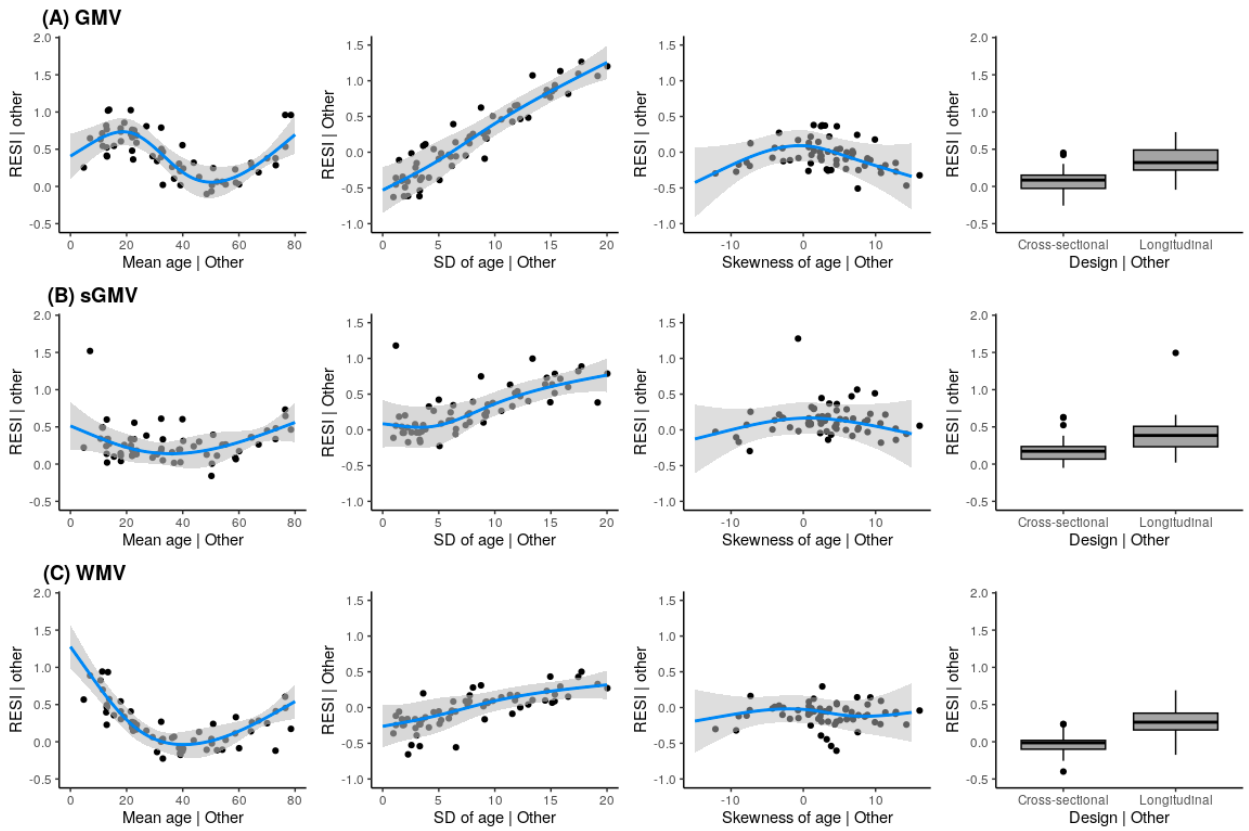

Fig. 1: Partial regression plots of a meta-analysis of RESI and L-RESI for the GMV-age (**A**), sGMV-age (**B**), and WMV-age (**C**) association with respect to the sample mean, SD, skewness of age and study design after fixing the other features at constant levels (i.e., design = cross-sectional, mean 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. Sixty-three MRI datasets containing 75,255 MRI scans from 60,913 cognitively normal participants were modeled as a function of age and sex separate study-level models. The gray areas are the 95% confidence bands from the LOESS smooth.

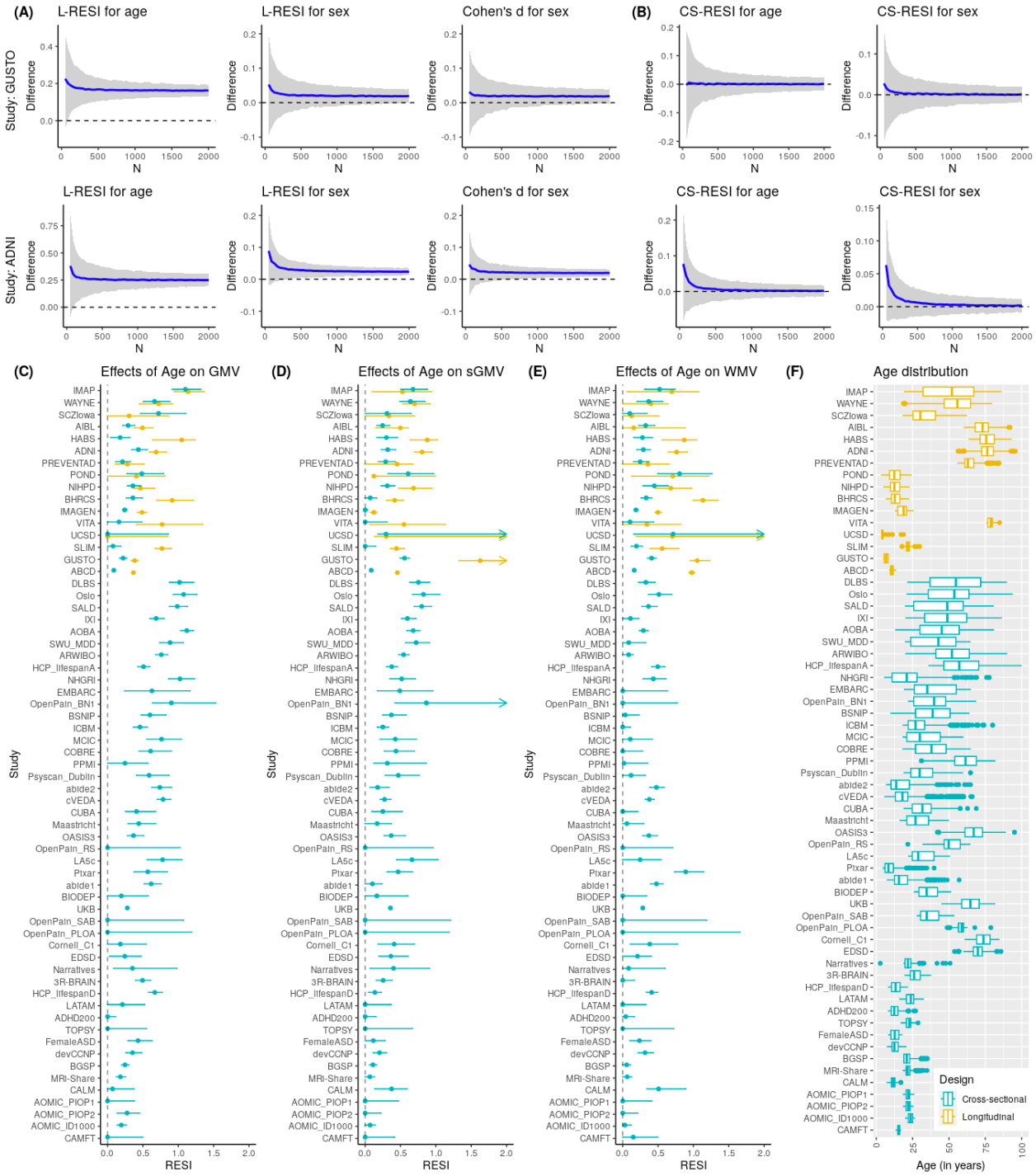

Fig. 2. Bootstrap analysis shows that (**A**) effect sizes in longitudinal studies are systematically larger than effect sizes in cross-sectional studies (Difference = L-RESI - RESI) and (**B**) the modified CS-RESI for longitudinal analysis estimates and removes the difference (Difference = CS-RESI - RESI). Forest plots of the RESI (for cross-sectional studies), and CS-RESI (blue) and L-RESI (yellow) (for longitudinal studies), for age on GMV (**C**), sGMV (**D**), and WMV (**E**) show that the CS-RESI is more comparable to the cross-sectional studies. The dots and lines represent RESI estimates and their 95% CIs. Arrows indicate that the upper limit is greater than

the axis boundary. **(F)** The age distribution in each study, sorted by study design and decreasing standard deviation of age in each study.

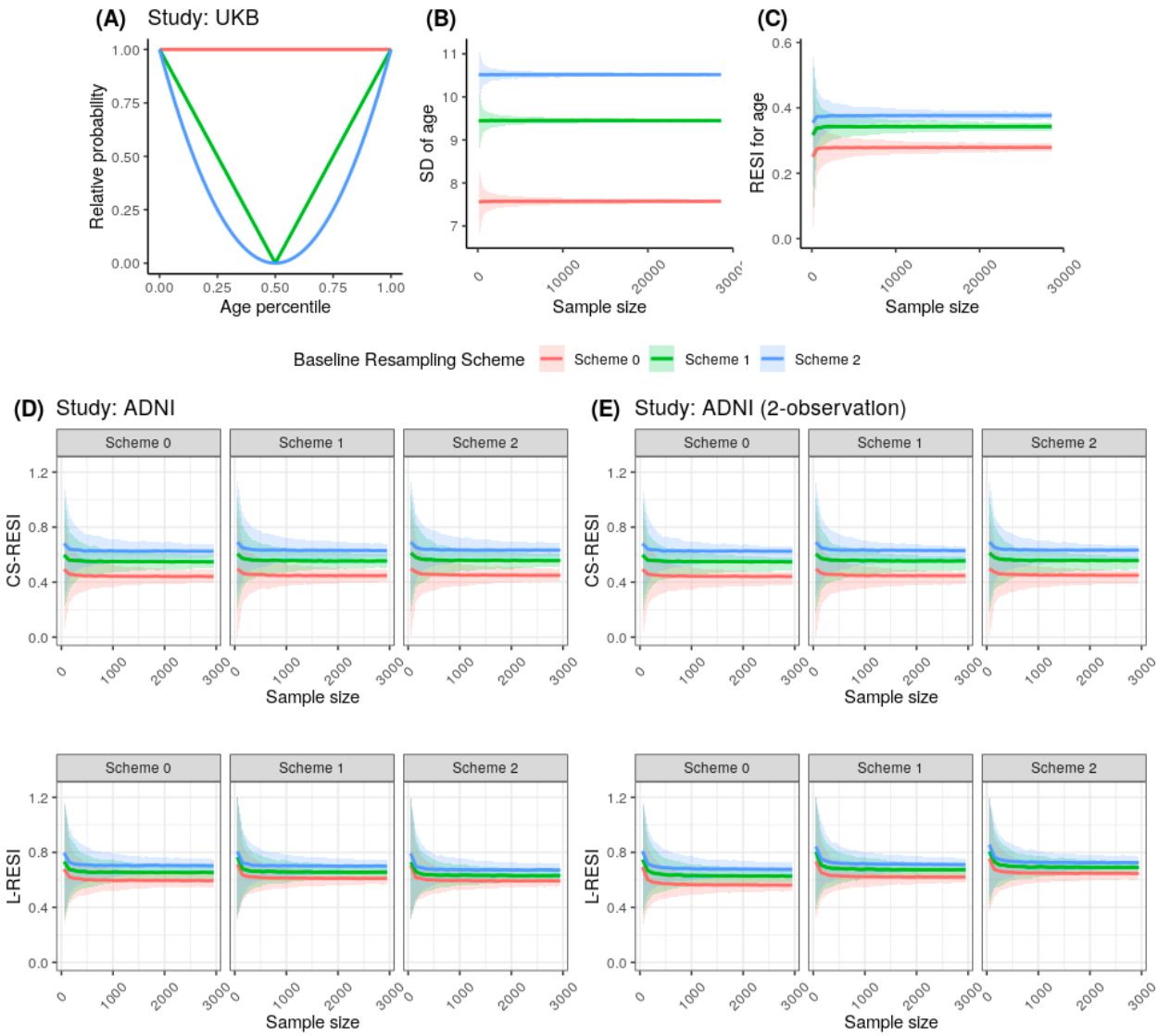

Fig 3. (**A**) Resampling schemes that weighted the tails (**B**) increased the standard deviation of age and (**C**) increased GMV-age robust effect sizes (RESI); The same resampling schemes were implemented for between- and within-subjects in the ADNI data. (**D**) Increasing the between-subject SD of age increased the CS-RESI and L-RESI, however, increasing the within-subject SD of age did not substantially increase ESs. (**E**) Reducing the number of observations to two per subject while maintaining the between- and within-subject SD of age (Fig. S5B) did not substantially change the CS-RESI and L-RESI.

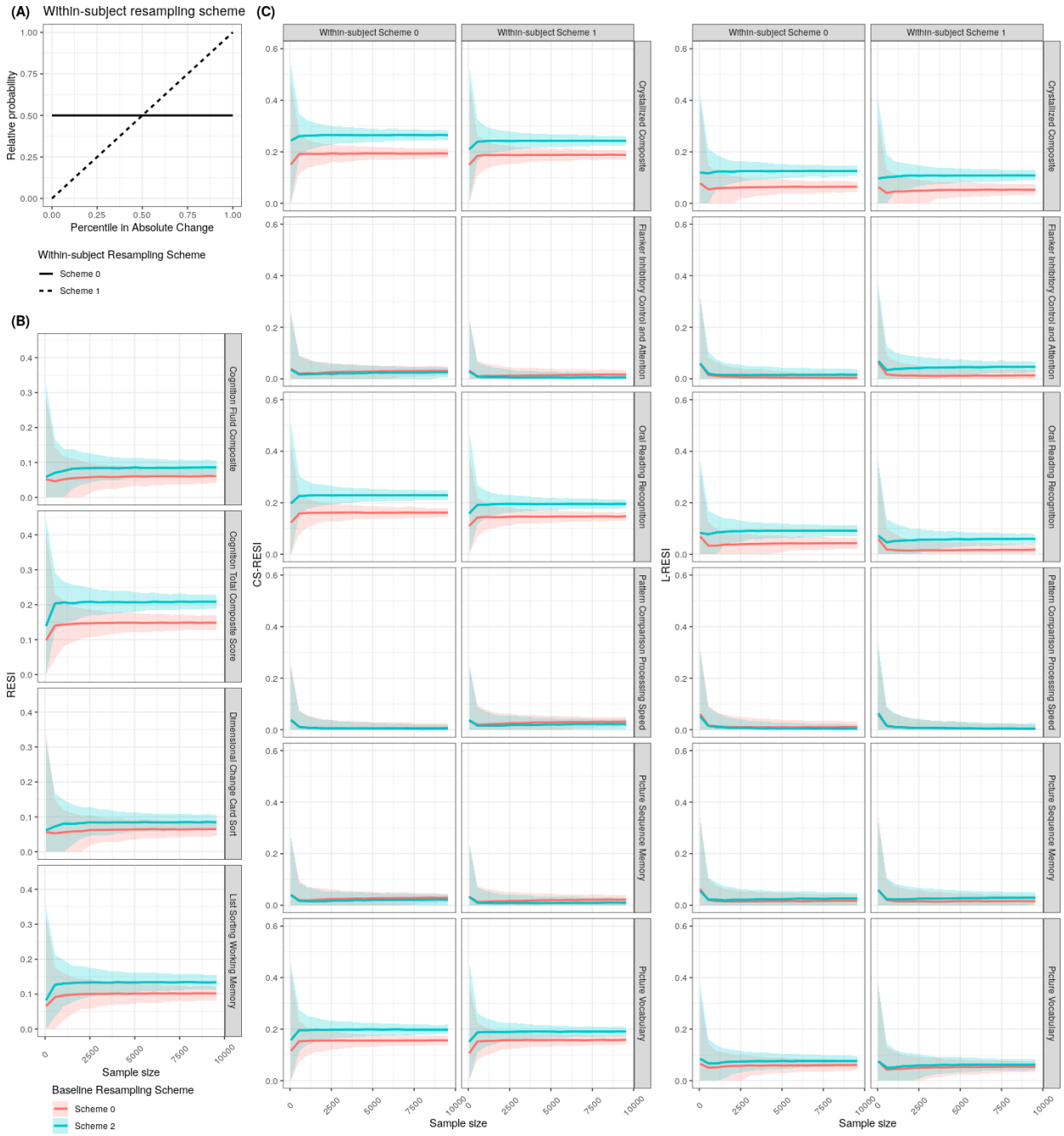

Fig 4. (**A**) Relative probability within-subject resampling schemes to adjust the within-subject variability by increasing the probability of sampling subjects with a larger absolute change in the cognitive measurements. Increasing the between-subject standard deviation of the cognitive measurements (using resampling scheme 2 shown in Fig. 3A) increased the RESIs of some cognitive measurements on GMV in both (**B**) cross-sectional and (**C**) longitudinal analyses. For the cognitive measurements analyzed longitudinally, increasing the within-subject standard deviation in the cognitive measurements did not increase the average observed CS-RESI or L-RESI. (**C**) Longitudinal effect sizes were smaller than cross-sectional effect sizes.