Comparing Automated Subcortical Volume Estimation Methods; Amygdala Volumes Estimated by FSL and FreeSurfer Have Poor Consistency

Patrick Sadil[a,*], Martin A. Lindquist[a]

[a] *Johns Hopkins Bloomberg School of Public Health,*

Abstract

Subcortical volumes are a promising source of biomarkers and features in biosignatures, and automated methods facilitate extracting them in large, phenotypically rich datasets. However, while extensive research has verified that the automated methods produce volumes that are similar to those generated by expert annotation, the consistency of methods with each other is understudied. Using data from the UK Biobank, we compare the estimates of subcortical volumes produced by two popular software suites: FSL and FreeSurfer. Although most subcortical volumes exhibit good to excellent consistency across the methods, the tools produce diverging estimates of amygdalar volume. Through simulation, we show that this poor consistency can lead to conflicting results, where one but not the other tool suggests statistical significance, or where both tools suggest a significant relationship but in opposite directions. Considering these issues, we discuss several ways in which care should be taken when reporting on relationships involving amygdalar volume.

Keywords: MRI, Amygdala

1. Introduction

Regional volumes of subcortex have been proposed as biomarkers for several psychopathologies. For example, the volume of the amygdala has been suggested as a biomarker for Alzheimer's, depression symptom severity in young adults, bipolar disorder in youth, migraine frequency, chronic pain, and others (Daftary et al., 2019; Khatri and Kwon, 2022; Pfeifer et al., 2008; Liu et al., 2017; Vachon-Presseau et al., 2016; Rogers et al., 2009; Ruocco et al., 2012; Szeszko et al., 2004). As a biomarker, subcortical volumes are advantageous for being interpretable (given the rich literature linking these structures to many functions), explainable (hypotrophy and hypertrophy are both easily described to healthcare providers and patients), and readily available. The latter

*Corresponding author
Email address: psadil1@jh.edu (Patrick Sadil)

point comes from the fact that it is possible to estimate the regional volumes from any structural image with several automated algorithms.

For estimating regional subcortical volumes, two automated techniques are popular: FMRIB's Integrated Registration and Segmentation Tool (FIRST) from the FMRIB Software Library (FSL) and FreeSurfer's Automated Segmentation (ASEG) (Patenaude, 2007; Patenaude et al., 2011; Fischl, 2012). Both techniques exhibit high consistency with the gold-standard of manual segmentation in healthy adults and some clinical populations (Hsu et al., 2002; Tae et al., 2008; Morey et al., 2009; Pardoe et al., 2009; Dewey et al., 2010; Lehmann et al., 2010; Doring et al., 2011; Nugent et al., 2013; Wenger et al., 2014), although there is variability across regions and between methods. For segmenting the hippocampus, FreeSurfer has been reported as having higher intraclass correlations than FSL (Doring et al., 2011), and neither method appears to have worse reliability (across repeated scans) than manual segmentation (Mulder et al., 2014). For the putamen, FSL has a higher Dice coefficient with manual segmentation, and the methods perform similarly on the caudate (Perlaki et al., 2017). For the amygdala, which method performs better depends on the metric (Morey et al., 2009). However, these comparisons may not generalize to other populations (e.g., pediatric, elderly), given that performance of the automated techniques is not consistently high across populations (Schoemaker et al., 2016; Kim et al., 2012; Sánchez-Benavides et al., 2010; Zhou et al., 2021).

While comparisons to manual segmentation could identify which method produces the best estimates of volume, it remains less clear how the two methods compare to each other. To our knowledge, the two methods have only been compared to each other by Perlaki et al. (2017). In their research, the methods were consistent with each other (exhibiting intraclass correlations that ranged from around 0.7 - 0.9), but the analyses included only the putamen and caudate, and the sample size was relatively small (N=30). We extend the results of Perlaki et al. (2017) to the remaining structures and with a much larger population (tens of thousands). Our investigation should be considered in the context of research that would use estimates of volume as a biomarker by, for example, correlating volume with a health-related outcome. Our primary concern is whether the results of such a study could be expected to depend on the method used for automated segmentation.

2. Methods and Results

First, we looked at the consistency of subcortical volumes between FSL and FreeSurfer using data from the UK Biobank (Alfaro-Almagro et al., 2018). The data were downloaded Jan 2024 and contained 45743 participants with usable anatomical data (Category 190: FreeSurfer ASEG; Category 1102: FSL FIRST). For details on intraclass correlation calculations, see Appendix A.1. Estimates and associated uncertainty are displayed using subscripts, as recommended by Louis and Zeger (2008). For example, an estimate of 0.22 with a 95% confidence interval spanning [0.21, 0.23] will be rendered as ${}_{0.21}0.22_{0.23}$.

Across all structures, estimated agreement was lower than estimated consistency (Table A1), reflecting differences in the average volumes estimated by the two methods (Appendix A.2). However, a constant shift across participants would not affect many analyses targeted by our primary concern, analyses related to estimated correlations

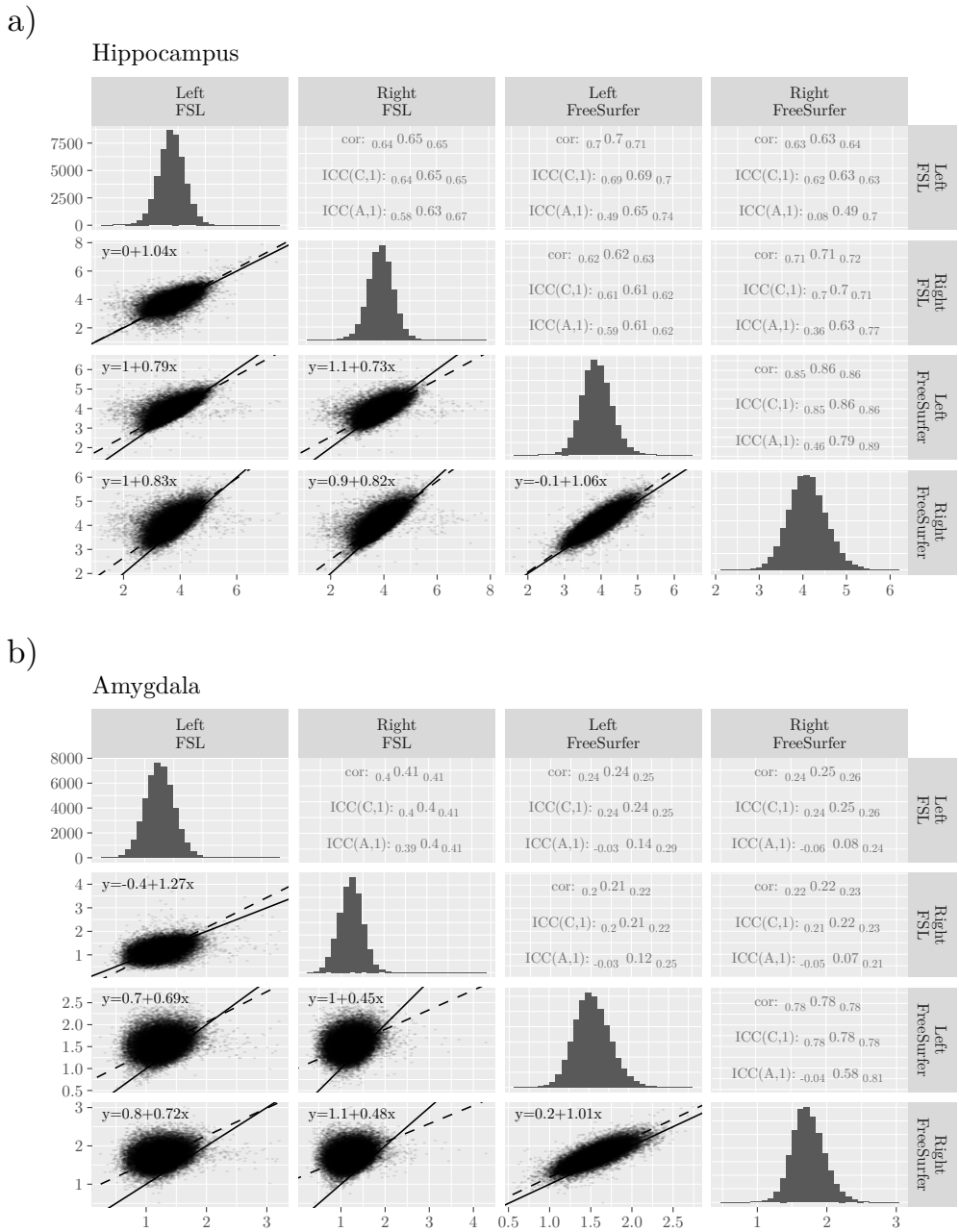

Figure 1: Comparisons of Subcortical Volumes Estimated by FSL and FreeSurfer for two example regions a) Hippocampus and b) Amygdala. For remaining subcortical regions, see Table A1. In the lower triangular panels, the line of equivalence is marked with a solid line, and the dashed lines show the result of orthogonal regression. In the upper panel, the uncertainty estimates span 95% confidence intervals. The histograms along the diagonal display volumes in the full sample.

between regional volumes and some other measure. For that reason, we focus not on agreement but instead on consistency.

Consistency varies by region (Figure 1, Table A1). To interpret the intraclass correlations, consider the categories provided by Cicchetti and Sparrow (1981): <0.4: poor, [0.4, 0.6): fair, [0.6, 0.75): good, [0.75, 1): excellent. Using those categories, the methods exhibit "good" to "excellent" consistency for most regions. However, the consistency of volumes for the amygdala is markedly worse than the others, being around only ${}_{0.24}0.24_{0.25}$ and ${}_{0.21}0.22_{0.23}$ for the left and right hemispheres (ranges of uncertainty span 95% confidence intervals). Compare those values to the values for the hippocampus (Figure 1), which has good consistency (left: ${}_{0.69}0.69_{0.70}$, right: ${}_{0.70}0.70_{0.71}$). For both structures (that is, the amygdala and hippocampus), consistency across hemispheres as reported by FreeSurfer is the highest numerically (Figure 1). For the amygdala, there is a two-way interaction between method and hemisphere (FSL - FreeSurfer: 0.22, $p < 0.001$), with FreeSurfer reporting that the right amygdala is larger than the left (left - right: -0.19, $p < 0.001$) and FSL reporting that the left is larger than the right (left - right: 0.04, $p < 0.001$).

The amygdala has been described as particularly challenging to segment; one small study (N=23) reports a consistency of 0.6 for volumes estimated by FSL across repeated scans of the same individual (Morey et al., 2010). Moreover, automated segmentation algorithms can be affected by experimental factors like site, scanner, participant positioning, and software version (Hedges et al., 2022; Du et al., 2021; McGuire et al., 2017; Yang et al., 2016; Liu et al., 2020; Mulder et al., 2014; Morey et al., 2010; Perlaki et al., 2017), and differential sensitivity to such factors could impact an intraclass correlation. In the UKB, several potentially confounding factors were correlated with estimates of amygdalar volume (Figure A2). However, regressing these factors from the estimates of volume did not improve the consistency between the methods (Appendix A.4).

With poor consistency between measurements of the amygdala, there is concern that reported relationships involving amygdala volumes may depend on which method is used for estimating the volume, a choice that may be considered arbitrary or lab-specific. At least two kinds of issues could arise. First, lower consistency could make it more likely that one but not both methods leads to significant correlations. Second, lower consistency could make it more likely that the two methods produce significant correlations that go in opposite directions.

To investigate how often these two issues could occur, we first simulated experiments with artificial data. In each simulation, datasets with two noisy estimates of volume and a third, outcome, variable were generated such that the two volume estimates had a prespecified intraclass correlation with each other (in expectation), and the true volume had a given product-moment correlation with the outcome variable. The estimated volumes were then tested for a product-moment correlation with the outcome, and the process was repeated for several intraclass correlations and sample sizes. In all simulations, the true product-moment correlation was set to a value that is either typical for neuroimaging research (0.1, Marek et al., 2022), small but non-zero (0.01), or large (0.2). For additional details on the simulation methods, see Appendix A.5.

With lower consistency, the estimated product-moment correlations were more often

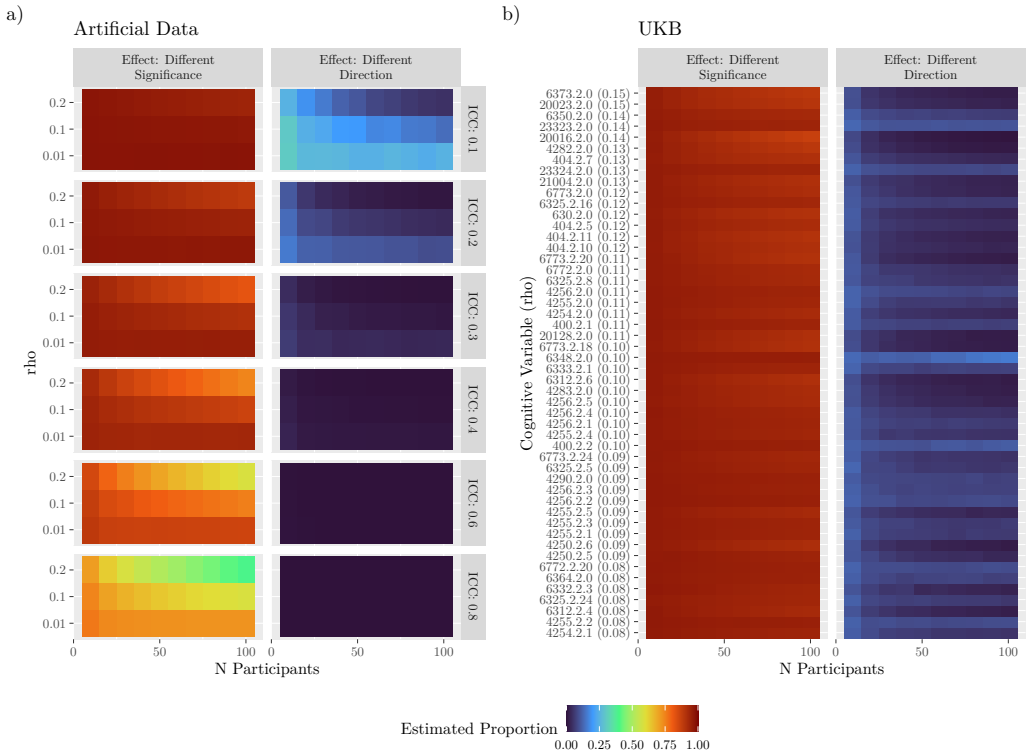

Figure 2: Effects of Low Measurement Consistency. In the subfigures, the left column or panel shows the proportion of simulated experiments where the methods produce correlations on opposite sides of an $\alpha = 0.05$ significance threshold. The right column or panel shows the proportion of simulated experiments where both measures are significantly correlated with an outcome but in opposite directions. Proportions are shown with color (Swihart et al., 2010). a) Simulations with Artificial Data. Rows (rho) indicate pre-specified product-moment correlation with the outcome variable. ICC: pre-specified intraclass correlation between measures of volume. b) Simulations with UKB. Each row corresponds to a non-image derived phenotype from the UKB. The value in parentheses is the absolute correlation of the variable with the average volumes of the left amygdala (average across methods). For a display without color that includes estimates of uncertainty, see Figure A3.

on opposite sides of a significance threshold (Figure 2a, left column). For example, with consistency near the value that was estimated for volumes of the amygdala in the UKB (0.2), a typical effect size (0.1), and experiments with 50 participants, significance differed in around $_{94.9}95.0_{95.1}$ percent of simulations in which one test was significant (for simulations, estimates indicate medians and ranges of uncertainty span 95% equal-tailed intervals, see Appendix A.5). With a sample size of 100, that percentage was only minimally different (to $_{93.7}93.9\%_{94.0}$). For a discussion on this insensitivity to sample size, see Section Appendix A.6.

Lower consistency also coincided with a higher proportion of experiments in which the two methods correlate in opposite directions (Figure 2a, right column). Considering the previous example (an intraclass correlation of 0.2, an effect size of 0.1, and 50 participants per experiment), around $_{5.13}5.71_{6.33}$ percent of experiments in which both product-moment correlations were significant resulted in the correlations having opposite signs. As the intraclass correlation increased, the rates of this effect decreased rapidly.

To assess how often these two issues could occur in practice, we repeated the above analyses but with the UKB. From the UKB, we extracted the "Cognitive" variables using the FMRIB UKBiobank Normalisation, Parsing And Cleaning Kit (McCarthy, 2023), restricting analyses to variables with instance 2 and that exhibited one of the largest (in absolute magnitude) 50 rank correlations (between the cognitive variables and the average of the two estimates of the volume of the left amygdala), and further to only those participants that had values for all 50 of those variables (23486 participants). Each simulation resulted in a rank correlation between the two volume estimates for the left amygdala and each of the 50 variables (results were comparable with the right amygdala). For additional details, see Appendix A.5.

Considering differing significance levels, the rates across experiments using the UKB resembled the rates from experiments with artificial data. In simulated experiments with 50 UKB participants in which at least one correlation was significant, one correlation was not significant in between $_{90.7}90.8_{91.0}$ and $_{95.1}95.2_{95.3}$ percent of simulations (the two estimates cover the 50 cognitive variables; Figure 2b, left panel). Proportions were generally lower for variables that were more strongly correlated with the measure of volume (for illustration, compare the variables that are higher versus lower in the left panel of Figure 2b).

Considering significant correlations with differing signs, the rates across experiments with the UKB bracketed the rates with artificial data. In experiments simulated with 50 UKB participants, the rates ranged from $_{1.28}1.45_{1.64}$ to $_{7.64}8.32_{9.03}$. For most variables, increasing the number of participants decreased the proportion of experiments in which the two correlations exhibited opposite signs, but for some the proportion increased. Across the 50 variables, 9 had a higher proportion at N=100 than N=50, 2 of which had non-overlapping 95% equal-tailed intervals (6348: duration to complete numeric path; 400: time to complete round of pairs matching game; see also Figure A3).

3. Discussion

We examined the consistency of subcortical volumes within the UK Biobank (Alfaro-Almagro et al., 2018), observing that two common methods of estimating the volume of

the amygdala, one from FSL and one from FreeSurfer, have poor consistency with each other. The main concern in this report is that consistency this poor can lead to conflicting results. Two kinds of conflict were explored: the methods producing correlations that are on opposite sides of significance thresholds, and the methods producing volumes with significant correlations that have differing signs. The prevalence of these occurrences was estimated with artificial data and data from the UKB. Based on the observed rates, we make the following recommendations.

3.1. Recommendations

3.1.1. When testing for new biomarkers, report relationships with multiple automated methods (e.g., both FSL and FreeSurfer).

Researchers may have idiosyncratic reasons for selecting a method, particularly when the choice is viewed as arbitrary. If the choice between methods is arbitrary, then reporting the outcome across selections clarifies the fragility or robustness of a result (for a general discussion, see Steegen et al., 2016). The choice may not always be arbitrary, as there are metrics along which and study populations for which one method may perform better (Zhou et al., 2021; Morey et al., 2009; Huizinga et al., 2021). Note that one effect of low consistency could be a downward bias on the magnitude of estimated correlations (Section Appendix A.7), and so there may be advantages to not only reporting but also combining the estimates across methods when predicting health-related outcomes (e.g., by averaging, or including both as independent variables in predictive models). Reporting estimates from multiple reasonable tools will help move conclusions beyond "there exists a correlation with the volume of the amygdala as estimated by method M (version x)" to simply "there exists a correlation with the volume of the amygdala".

3.1.2. When reviewing or conducting meta-analyses of relationships with amygdala volume, consider the method that was used to estimate volume.

As mentioned in the Introduction, the amygdala has received substantial attention due to being predictive of an array of health-related outcomes. As presented in this report, the volume that is estimated by one automated method may only weakly correspond to the volume estimated by another, and so it may be misleading to conduct a meta-analysis without accounting for the algorithms used by the individual studies. In the UKB, there are differences in the strength of the correlations between the measures of volume and the cognitive variables; in nearly all of the variables considered in this report, the magnitude of the correlations involving FreeSurfer's method were numerically larger (Figure A5). Larger correlations do not imply higher veracity, but they further indicate that the methods track aspects of amygdalar anatomy differently.

3.1.3. When replicating or extending research on a relationship that involves the volume of the amygdala, use the method reported in the original publications.

This recommendation follows standard practice for a replication study. We highlight it here in consideration of both extension studies that aim to apply a biomarker or biosignature that includes amygdala volume (such as when testing a putative relationship in a new population), and also in consideration of the ongoing evolution of methods for automatically estimating subcortical volumes. Although FSL and FreeSurfer are two of the most popular methods, others exist (e.g., Akhondi-Asl and Warfield, 2013), including

newer techniques based on deep-learning approaches (e.g., Billot et al., 2023; for review, see Singh and Singh, 2021). Newer methods may have better correspondence with manual segmentation, warranting their use in replication or extension studies. But as this report shows, two methods can perform well while exhibiting poor consistency with each other. So when building on prior findings, it remains important to use the methods of those prior findings, even when they are superseded.

References

Akhondi-Asl, A., Warfield, S.K., 2013. Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. IEEE Transactions on Medical Imaging 32, 1840–1852. doi:10.1109/TMI.2013.2266258.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. NeuroImage 166, 400–424. doi:10.1016/j.neuroimage.2017.10.034.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J.L., Bastiani, M., Miller, K.L., Nichols, T.E., Smith, S.M., 2021. Confound modelling in uk biobank brain imaging. NeuroImage 224, 117002. doi:10.1016/j.neuroimage.2020.117002.

Barnes, J., Ridgway, G.R., Bartlett, J., Henley, S.M., Lehmann, M., Hobbs, N., Clarkson, M.J., MacManus, D.G., Ourselin, S., Fox, N.C., 2010. Head size, age and gender adjustment in mri studies: a necessary nuisance? NeuroImage 53, 1244–1255. doi:10.1016/j.neuroimage.2010.06.025.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48. doi:10.18637/jss.v067.i01.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2023. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical Image Analysis 86, 102789. doi:10.1016/j.media.2023.102789.

Cicchetti, D.V., Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. American Journal of Mental Deficiency 86, 127–137. PMID: 7315877.

Daftary, S., Van Enkevort, E., Kulikova, A., Legacy, M., Brown, E.S., 2019. Relationship between depressive symptom severity and amygdala volume in a large community-based sample. Psychiatry Research: Neuroimaging 283, 77–82. doi:10.1016/j.pscychresns.2018.12.005.

Dewey, J., Hana, G., Russell, T., Price, J., McCaffrey, D., Harezlak, J., Sem, E., Anyanwu, J.C., Guttmann, C.R., Navia, B., Cohen, R., Tate, D.F., 2010. Reliability and validity of mri-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in hiv-infected patients from a multisite study. NeuroImage 51, 1334–1344. doi:10.1016/j.neuroimage.2010.03.033.

Doring, T.M., Kubo, T.T., Cruz, L.C.H., Juruena, M.F., Fainberg, J., Domingues, R.C., Gasparetto, E.L., 2011. Evaluation of hippocampal volume based on mr imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. Journal of Magnetic Resonance Imaging 33, 565–572. doi:10.1002/jmri.22473.

Du, J., Liang, P., He, H., Tong, Q., Gong, T., Qian, T., Sun, Y., Zhong, J., Li, K., 2021. Reproducibility of volume and asymmetry measurements of hippocampus, amygdala, and entorhinal cortex on traveling volunteers: a multisite mp2rage prospective study. Acta Radiologica 62, 1381–1390. doi:10.1177/0284185120963919.

Fischl, B., 2012. Freesurfer. NeuroImage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Frost, C., Thompson, S.G., 2000. Correcting for regression dilution bias: Comparison of methods for a single predictor variable. Journal of the Royal Statistical Society: Series A (Statistics in Society) 163, 173–189. doi:10.1111/1467-985X.00164.

Gamer, M., Lemon, J., <puspendra.pusp22@gmail.com>, I.F.P.S., 2019. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1.

Gomez-Ramirez, J., Quilis-Sancho, J., Fernandez-Blazquez, M.A., 2022. A comparative analysis of mri automated segmentation of subcortical brain volumes in a large dataset of elderly subjects. Neuroinformatics 20, 63–72. doi:10.1007/s12021-021-09520-z.

Hedges, E.P., Dimitrov, M., Zahid, U., Brito Vega, B., Si, S., Dickson, H., McGuire, P., Williams, S., Barker, G.J., Kempton, M.J., 2022. Reliability of structural mri measurements: The effects of scan

session, head tilt, inter-scan interval, acquisition sequence, freesurfer version and processing stream. NeuroImage 246, 118751. doi:10.1016/j.neuroimage.2021.118751.

Hsu, Y., Schuff, N., Du, A., Mark, K., Zhu, X., Hardin, D., Weiner, M.W., 2002. Comparison of automated and manual mri volumetry of hippocampus in normal aging and dementia. Journal of Magnetic Resonance Imaging 16, 305–310. doi:10.1002/jmri.10163.

Huizinga, W., Poot, D.H.J., Vinke, E.J., Wenzel, F., Bron, E.E., Toussaint, N., Ledig, C., Vrooman, H., Ikram, M.A., Niessen, W.J., Vernooij, M.W., Klein, S., 2021. Differences Between MR Brain Region Segmentation Methods: Impact on Single-Subject Analysis. Frontiers in Big Data 4, 577164. doi:10.3389/fdata.2021.577164.

Khatri, U., Kwon, G.R., 2022. Alzheimer's disease diagnosis and biomarker analysis using resting-state functional mri functional brain network with multi-measures features and hippocampal subfield and amygdala volume of structural mri. Frontiers in Aging Neuroscience 14, 818871. doi:10.3389/fnagi.2022.818871.

Kim, H., Chupin, M., Colliot, O., Bernhardt, B.C., Bernasconi, N., Bernasconi, A., 2012. Automatic hippocampal segmentation in temporal lobe epilepsy: Impact of developmental abnormalities. NeuroImage 59, 3178–3186. doi:10.1016/j.neuroimage.2011.11.040.

Lehmann, M., Douiri, A., Kim, L.G., Modat, M., Chan, D., Ourselin, S., Barnes, J., Fox, N.C., 2010. Atrophy patterns in alzheimer's disease and semantic dementia: A comparison of freesurfer and manual volumetric measurements. NeuroImage 49, 2264–2274. doi:10.1016/j.neuroimage.2009.10.056.

Liu, H.Y., Chou, K.H., Lee, P.L., Fuh, J.L., Niddam, D.M., Lai, K.L., Hsiao, F.J., Lin, Y.Y., Chen, W.T., Wang, S.J., et al., 2017. Hippocampus and amygdala volume in relation to migraine frequency and prognosis. Cephalalgia 37, 1329–1336. doi:10.1177/0333102416678624.

Liu, S., Hou, B., Zhang, Y., Lin, T., Fan, X., You, H., Feng, F., 2020. Inter-scanner reproducibility of brain volumetry: influence of automated brain segmentation software. BMC Neuroscience 21, 35. doi:10.1186/s12868-020-00585-1.

Louis, T.A., Zeger, S.L., 2008. Effective communication of standard errors and confidence intervals. Biostatistics 10, 1–2. doi:10.1093/biostatistics/kxn014.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., et al., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660. doi:10.1038/s41586-022-04492-9.

Mathalon, D.H., Sullivan, E.V., Rawles, J.M., Pfefferbaum, A., 1993. Correction for head size in brain-imaging measurements. Psychiatry Research: Neuroimaging 50, 121–139. doi:10.1016/0925-4927(93)90016-B.

McCarthy, P., 2023. funpack. Zenodo. doi:10.5281/ZENODO.1997626.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46. doi:10.1037/1082-989X.1.1.30.

McGuire, S.A., Wijtenburg, S.A., Sherman, P.M., Rowland, L.M., Ryan, M., Sladky, J.H., Kochunov, P.V., 2017. Reproducibility of quantitative structural and physiological mri measurements. Brain and Behavior 7, e00759. doi:10.1002/brb3.759.

Morey, R.A., Petty, C.M., Xu, Y., Pannu Hayes, J., Wagner, H.R., Lewis, D.V., LaBar, K.S., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. NeuroImage 45, 855–866. doi:10.1016/j.neuroimage.2008.12.033.

Morey, R.A., Selgrade, E.S., Wagner, H.R., Huettel, S.A., Wang, L., McCarthy, G., 2010. Scan–rescan reliability of subcortical brain volumes derived from automated segmentation. Human Brain Mapping 31, 1751–1762. doi:10.1002/hbm.20973.

Mulder, E.R., De Jong, R.A., Knol, D.L., Van Schijndel, R.A., Cover, K.S., Visser, P.J., Barkhof, F., Vrenken, H., 2014. Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. NeuroImage 92, 169–181. doi:10.1016/j.neuroimage.2014.01.058.

Nugent, A.C., Luckenbaugh, D.A., Wood, S.E., Bogers, W., Zarate, C.A., Drevets, W.C., 2013. Automated subcortical segmentation using FIRST: Test-retest reliability, interscanner reliability, and comparison to manual segmentation: Reliability of Automated Segmentation Using FIRST. Human Brain Mapping 34, 2313–2329. doi:10.1002/hbm.22068.

Pardoe, H.R., Pell, G.S., Abbott, D.F., Jackson, G.D., 2009. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? Epilepsia 50, 2586–2592. doi:10.1111/j.1528-1167.2009.02243.x.

Patenaude, B., 2007. Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation. Ph.D. thesis. University of Oxford.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922. doi:10.1016/j.neuroimage.2011.02.046.

Perlaki, G., Horvath, R., Nagy, S.A., Bogner, P., Doczi, T., Janszky, J., Orsi, G., 2017. Comparison of accuracy between fsl's first and freesurfer for caudate nucleus and putamen segmentation. Scientific Reports 7, 2418. doi:10.1038/s41598-017-02584-5.

Pfeifer, J.C., Welge, J., Strakowski, S.M., Adler, C., Delbello, M.P., 2008. Meta-analysis of amygdala volumes in children and adolescents with bipolar disorder. Journal of the American Academy of Child & Adolescent Psychiatry 47, 1289–1298. doi:10.1097/CHI.0b013e318185d299.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rogers, M.A., Yamasue, H., Abe, O., Yamada, H., Ohtani, T., Iwanami, A., Aoki, S., Kato, N., Kasai, K., 2009. Smaller amygdala volume and reduced anterior cingulate gray matter density associated with history of post-traumatic stress disorder. Psychiatry Research: Neuroimaging 174, 210–216. doi:10.1016/j.pscychresns.2009.06.001.

Ruocco, A.C., Amirthavasagam, S., Zakzanis, K.K., 2012. Amygdala and hippocampal volume reductions as candidate endophenotypes for borderline personality disorder: A meta-analysis of magnetic resonance imaging studies. Psychiatry Research: Neuroimaging 201, 245–252. doi:10.1016/j.pscychresns.2012.02.012.

Schoemaker, D., Buss, C., Head, K., Sandman, C.A., Davis, E.P., Chakravarty, M.M., Gauthier, S., Pruessner, J.C., 2016. Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of freesurfer and fsl against manual segmentation. NeuroImage 129, 1–14. doi:10.1016/j.neuroimage.2016.01.038.

Singh, M.K., Singh, K.K., 2021. A review of publicly available automatic brain segmentation methodologies, machine learning models, recent advancements, and their comparison. Annals of Neurosciences 28, 82–93. doi:10.1177/0972753121990175.

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W., 2016. Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science 11, 702–712. doi:10.1177/1745691616658637.

Swihart, B.J., Caffo, B., James, B.D., Strand, M., Schwartz, B.S., Punjabi, N.M., 2010. Lasagna Plots: A Saucy Alternative to Spaghetti Plots. Epidemiology 21, 621–625. doi:10.1097/EDE.0b013e3181e5b06a.

Szeszko, P.R., MacMillan, S., McMeniman, M., Lorch, E., Madden, R., Ivey, J., Banerjee, S.P., Moore, G.J., Rosenberg, D.R., 2004. Amygdala Volume Reductions in Pediatric Patients with Obsessive–Compulsive Disorder Treated with Paroxetine: Preliminary Findings. Neuropsychopharmacology 29, 826–832. doi:10.1038/sj.npp.1300399.

Sánchez-Benavides, G., Gómez-Ansón, B., Sainz, A., Vives, Y., Delfino, M., Peña-Casanova, J., 2010. Manual validation of freesurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and alzheimer disease subjects. Psychiatry Research: Neuroimaging 181, 219–225. doi:10.1016/j.pscychresns.2009.10.011.

Tae, W.S., Kim, S.S., Lee, K.U., Nam, E.C., Kim, K.W., 2008. Validation of hippocampal volumes measured using a manual method and two automated methods (freesurfer and ibaspm) in chronic major depressive disorder. Neuroradiology 50, 569–581. doi:10.1007/s00234-008-0383-9.

Vachon-Presseau, E., Tétreault, P., Petre, B., Huang, L., Berger, S.E., Torbey, S., Baria, A.T., Mansour, A.R., Hashmi, J.A., Griffith, J.W., et al., 2016. Corticolimbic anatomical characteristics predetermine risk for chronic pain. Brain 139, 1958–1970. doi:10.1093/brain/aww100.

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., Wahlund, L.O., Larsson, E.M., Westman, E., Initiative, A.D.N., 2014. The effects of intracranial volume adjustment approaches on multiple regional mri volumes in healthy aging and alzheimer's disease. Frontiers in Aging Neuroscience 6. doi:10.3389/fnagi.2014.00264.

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N.C., Kühn, S., Schaefer, S., Heinze, H.J., Düzel, E., Bäckman, L., Lindenberger, U., Lövdén, M., 2014. Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. Human Brain Mapping 35, 4236–4248. doi:10.1002/hbm.22473.

Yang, C.Y., Liu, H.M., Chen, S.K., Chen, Y.F., Lee, C.W., Yeh, L.R., 2016. Reproducibility of brain morphometry from short-term repeat clinical mri examinations: A retrospective study. PLOS ONE 11, e0146913. doi:10.1371/journal.pone.0146913.

Zhou, Q., Liu, S., Jiang, C., He, Y., Zuo, X.N., 2021. Charting the human amygdala development across childhood and adolescence: Manual and automatic segmentation. Developmental Cognitive Neuroscience 52, 101028. doi:10.1016/j.dcn.2021.101028.

Appendix A. Supplementary Materials

Appendix A.1. Intraclass Correlation

The intraclass correlation was based on a two-way mixed linear model (McGraw and Wong, 1996). In the model, the volume x for the region of participant i as measured by method k was treated as equal to the sum of an intercept, ν, the "true" volume, λ_i, a method bias, c_k, and an error term, ϵ_{ik}

$$
\begin{aligned}
x_{ik} &= \nu + \lambda_i + c_k + \epsilon_{ik} \\
\sum_k c_k &= 0 \\
\lambda_i &\sim N(0, \sigma_\lambda^2) \\
\epsilon_{ik} &\sim N(0, \sigma_\epsilon^2)
\end{aligned}
$$

Note that the c_k terms are assumed fixed, with variance given by $\sigma_c^2 = \sum_k c_k^2/(k-1)$.

An important assumption of this model is that the two methods are expected to have the same mean-squared error. The model does not include features that would allow for the errors in the two methods to be correlated (e.g., participants are not distinguished by characteristics that coincide with the methods performing better or worse).

The consistency version of the intraclass correlation, $ICC(C,1)$, and the absolute agreement, $ICC(A,1)$ were given as fractions of the variance components

$$
\begin{aligned}
ICC(C,1) &= \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2} \\
ICC(A,1) &= \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2 + \sigma_c^2}
\end{aligned}
$$

Variance components and associated confidence intervals estimated using the `R` package `irr` (R Core Team, 2023; Gamer et al., 2019), which uses the mean square approach described by McGraw and Wong (1996).

Appendix A.2. Differences in Average Volumes

Although we primarily focused on the consistency of the methods, we also observed shifts in the average volumes reported by the two methods (Figure A1). Differences in averages across methods have been reported previously (Gomez-Ramirez et al., 2022; Perlaki et al., 2017; Dewey et al., 2010; Huizinga et al., 2021). FSL tends to report volumes that are larger than those from FreeSurfer for the Accumbens, Amygdala, Hippocampus, and Pallidum, whereas for the Caudate, Putamen, and Thalamus FSL tends to report values that are smaller than those from FreeSurfer (all $p < 0.0001$ for two-sided t-test). Across structures, the variability in the difference appears to covary with the average of the two volume estimates, increasing as the average volume estimate decreases (Figure A1 b).

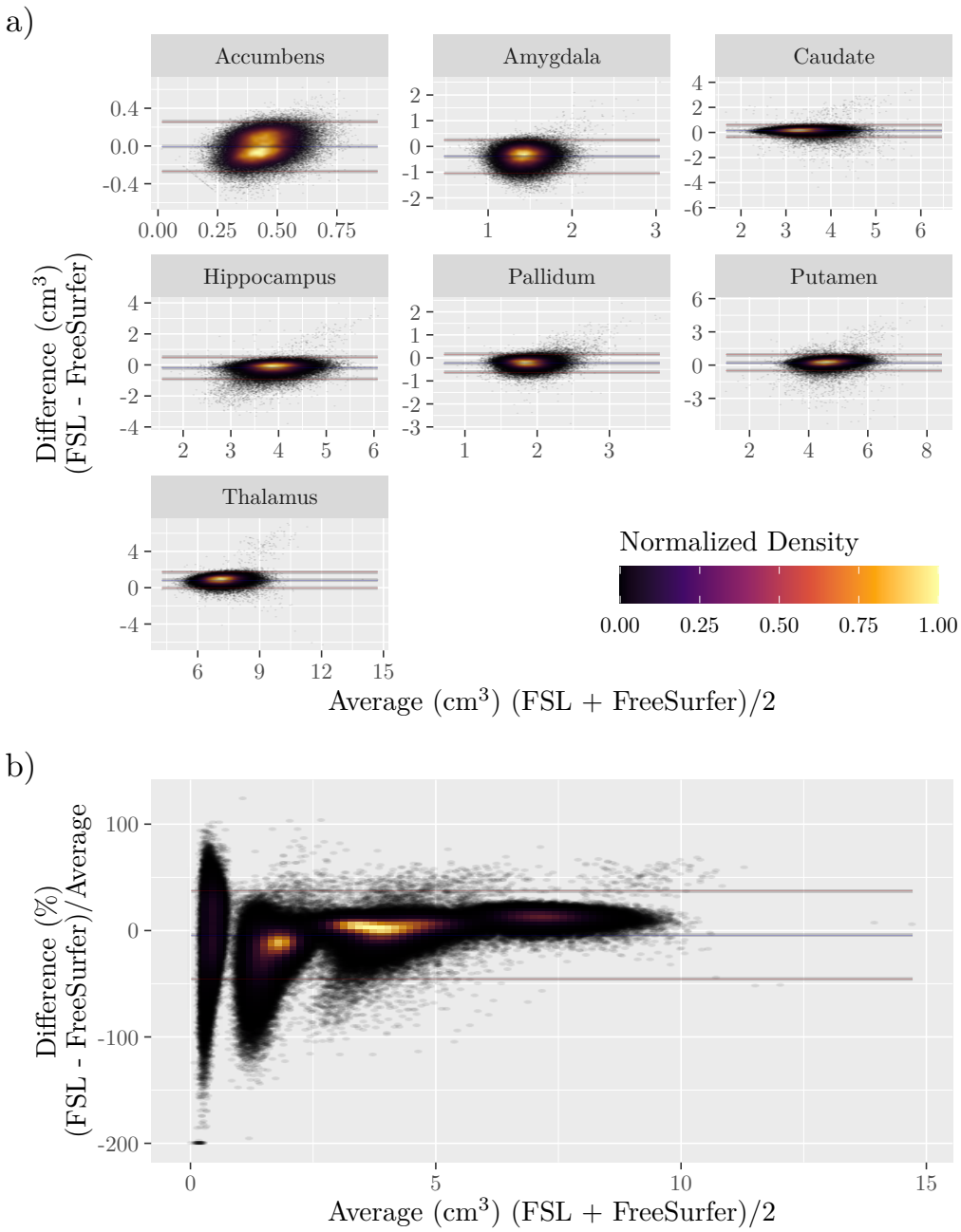

Figure A1: Bland-Altman Plots for Subcortical Volume. Horizontal lines show average difference and limits of agreement (1.96 standard deviations), with ribbons marking 95% confidence intervals. Panels correspond to subcortical structures. Left and right structures are plotted together. Shifts in the average estimate correspond to the central ribbon excluding zero. The color overlay indicates the degree of overplotting. a) Raw Differences. Note that axes are across panels independently. b) Differences by Percent Average.

Appendix A.3. Intraclass Correlations Across All Subcortical Regions

Structure	Hemisphere	ICC(C,1)	ICC(A,1)
Accumbens	Left	$_{0.57}0.58_{0.58}$	$_{0.05}0.44_{0.65}$
Accumbens	Right	$_{0.56}0.57_{0.57}$	$_{-0.03}0.38_{0.63}$
Amygdala	Left	$_{0.24}0.24_{0.25}$	$_{-0.03}0.14_{0.29}$
Amygdala	Right	$_{0.21}0.22_{0.23}$	$_{-0.05}0.07_{0.21}$
Caudate	Left	$_{0.85}0.85_{0.86}$	$_{0.75}0.83_{0.88}$
Caudate	Right	$_{0.86}0.86_{0.87}$	$_{0.62}0.82_{0.90}$
Hippocampus	Left	$_{0.69}0.69_{0.70}$	$_{0.49}0.65_{0.74}$
Hippocampus	Right	$_{0.70}0.70_{0.71}$	$_{0.36}0.63_{0.77}$
Pallidum	Left	$_{0.68}0.68_{0.69}$	$_{-0.09}0.41_{0.71}$
Pallidum	Right	$_{0.66}0.67_{0.67}$	$_{0.04}0.51_{0.73}$
Putamen	Left	$_{0.79}0.79_{0.80}$	$_{0.54}0.74_{0.84}$
Putamen	Right	$_{0.82}0.83_{0.83}$	$_{0.56}0.77_{0.87}$
Thalamus	Left	$_{0.82}0.82_{0.83}$	$_{-0.08}0.49_{0.79}$
Thalamus	Right	$_{0.83}0.83_{0.83}$	$_{-0.09}0.51_{0.81}$

Table A1: Intraclass Correlation for Subcortical Structures. Subscripts indicate 95% confidence intervals.

Appendix A.4. Intraclass Correlations After Residualizing on Potential Confounds

One possible source of low intraclass correlations could be systematic inaccuracies with certain kinds of participants. For example, one of the two methods could tend to underestimate volumes when given brains that have experienced severe atrophy, which would decrease the agreement of the methods. To assess this, correlations between the amygdala volumes and the "simple" confounds within the UKB were calculated (Alfaro-Almagro et al., 2021).

Several of the correlations appeared to be non-zero (Figure A2). However, regressing these confounds from the estimated amygdala volumes did not improve the intraclass correlations (Table A2).

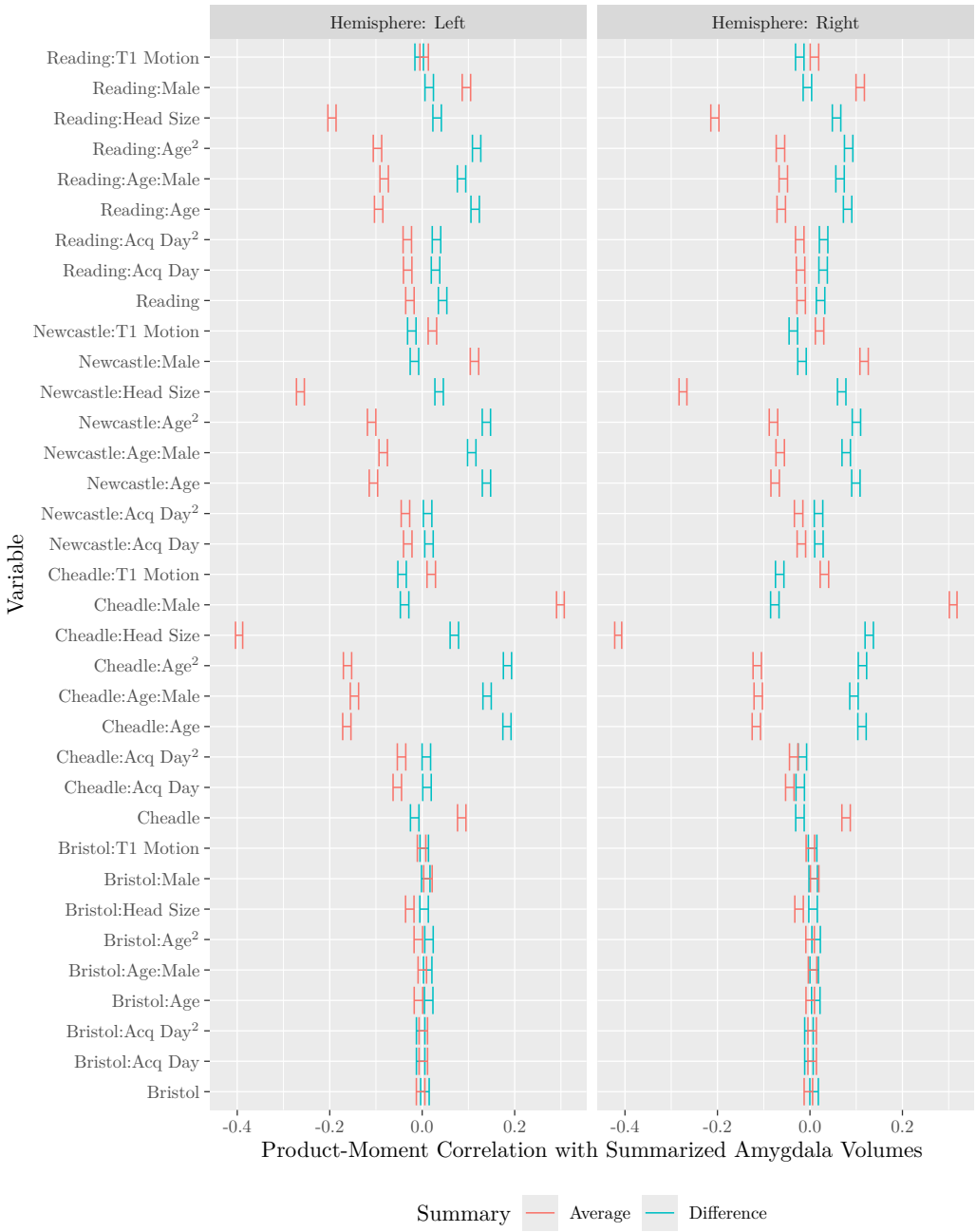

Figure A2: Correlation Between Potential Confounders and Amygdala Volume Measurements. Summary describes how the the volumes were combined across methods. Summaries were either an average or a difference (FSL-FreeSurfer). Note that the variable "Head Size" corresponds to a scaling factor, and that larger values imply smaller brains.

Structure	Hemisphere	ICC(C,1)	ICC(A,1)
Accumbens	Left	$_{0.42}0.43_{0.44}$	$_{0.01}0.30_{0.51}$
Accumbens	Right	$_{0.44}0.45_{0.45}$	$_{-0.05}0.27_{0.51}$
Amygdala	Left	$_{0.07}0.08_{0.09}$	$_{-0.01}0.04_{0.10}$
Amygdala	Right	$_{0.04}0.05_{0.06}$	$_{-0.01}0.01_{0.04}$
Caudate	Left	$_{0.81}0.81_{0.81}$	$_{0.68}0.78_{0.84}$
Caudate	Right	$_{0.82}0.82_{0.82}$	$_{0.52}0.76_{0.86}$
Hippocampus	Left	$_{0.57}0.58_{0.59}$	$_{0.36}0.53_{0.64}$
Hippocampus	Right	$_{0.58}0.59_{0.59}$	$_{0.23}0.50_{0.66}$
Pallidum	Left	$_{0.55}0.56_{0.57}$	$_{-0.09}0.30_{0.59}$
Pallidum	Right	$_{0.54}0.54_{0.55}$	$_{-0.01}0.38_{0.61}$
Putamen	Left	$_{0.68}0.69_{0.69}$	$_{0.38}0.62_{0.75}$
Putamen	Right	$_{0.73}0.74_{0.74}$	$_{0.41}0.66_{0.79}$
Thalamus	Left	$_{0.64}0.65_{0.65}$	$_{-0.08}0.27_{0.60}$
Thalamus	Right	$_{0.64}0.65_{0.65}$	$_{-0.09}0.28_{0.60}$

Table A2: Intraclass Correlation After Deconfounding. Prior to calculating consistency, volumes were deconfounded by a version of the "simple" parameter set described by Alfaro-Almagro et al. (2021). For the list of variables, see Figure A2. Subscripts indicate 95% confidence intervals.

Appendix A.4.1. Adjusting for ICV

When reporting differences in volume between groups, it is common to adjust for either head size or cerebral volume (Barnes et al., 2010; Mathalon et al., 1993; Voevodskaya et al., 2014). Adjusting by intracranial volume as estimated by FreeSurfer did not improve the intraclass correlations (Table A3).

Structure	Hemisphere	ICC(C,1)	ICC(A,1)
Accumbens	Left	$_{0.88}0.88_{0.88}$	$_{0.40}0.81_{0.91}$
Accumbens	Right	$_{0.90}0.90_{0.90}$	$_{0.23}0.81_{0.92}$
Amygdala	Left	$_{0.27}0.28_{0.29}$	$_{-0.03}0.17_{0.34}$
Amygdala	Right	$_{0.20}0.21_{0.22}$	$_{-0.05}0.07_{0.20}$
Caudate	Left	$_{0.79}0.80_{0.80}$	$_{0.67}0.77_{0.83}$
Caudate	Right	$_{0.81}0.81_{0.81}$	$_{0.52}0.75_{0.85}$
Hippocampus	Left	$_{0.63}0.63_{0.64}$	$_{0.42}0.58_{0.69}$
Hippocampus	Right	$_{0.63}0.64_{0.64}$	$_{0.28}0.55_{0.71}$
Pallidum	Left	$_{0.63}0.63_{0.64}$	$_{-0.09}0.36_{0.66}$
Pallidum	Right	$_{0.60}0.61_{0.61}$	$_{0.02}0.45_{0.68}$
Putamen	Left	$_{0.72}0.72_{0.72}$	$_{0.44}0.66_{0.78}$
Putamen	Right	$_{0.76}0.76_{0.76}$	$_{0.46}0.70_{0.81}$
Thalamus	Left	$_{0.75}0.75_{0.76}$	$_{-0.09}0.39_{0.72}$
Thalamus	Right	$_{0.75}0.76_{0.76}$	$_{-0.09}0.40_{0.73}$

Table A3: Consistency of Volumes when Adjusting by Intracranial Volume. Adjustments were done by residualizing with respect to ICV. Subscripts indicate 95% confidence intervals.

Appendix A.5. Simulated Experiments

To simulate hypothetical data, the models described in the previous sections were used. Sample sizes were set between 10 to 100 in steps of 10. At each combination of parameters, experiments were repeated 1,000,000 times. Simulated experiments with the UKB data were performed analogously to those with hypothetical data (UKB samples were taken with replacement).

In all simulations with artificial data, several parameters described in Section Appendix A.1 would not influence results after setting an intraclass and interclass correlation and so were set to arbitrary values: $\nu = 0$, $\sigma_\delta = 1$, and $\sigma_c = 0$. The remaining parameter σ_λ was set to a value that was estimated from the full UKB with a linear mixed-effects model that was estimated by restricted maximum likelihood as implemented by `lme4` (Bates et al., 2015): 0.019.

Across repetitions, the rates of each effect were estimated by analytic Bayesian methods (binomial likelihood with Beta prior whose shape parameters were set to 1/2). For the effect of "Different Significance", the posterior was based on the number of simulations in which one correlation exhibited a p-value less than 0.05 and the other was above 0.05 (successes among relevant simulations) and the number of simulations in which at least one p-value was below 0.05 (total relevant simulations). The effect of "Different Direction" was calculated similarly, but used simulations in which both correlations had

opposite magnitudes out of those in which both were significant. In the main text, ranges of uncertainty refer to 95% equal-tailed intervals and proportions refer to posterior medians.

The 95% equal-tailed interval of the posteriors for the simulations are shown in Figure A3.

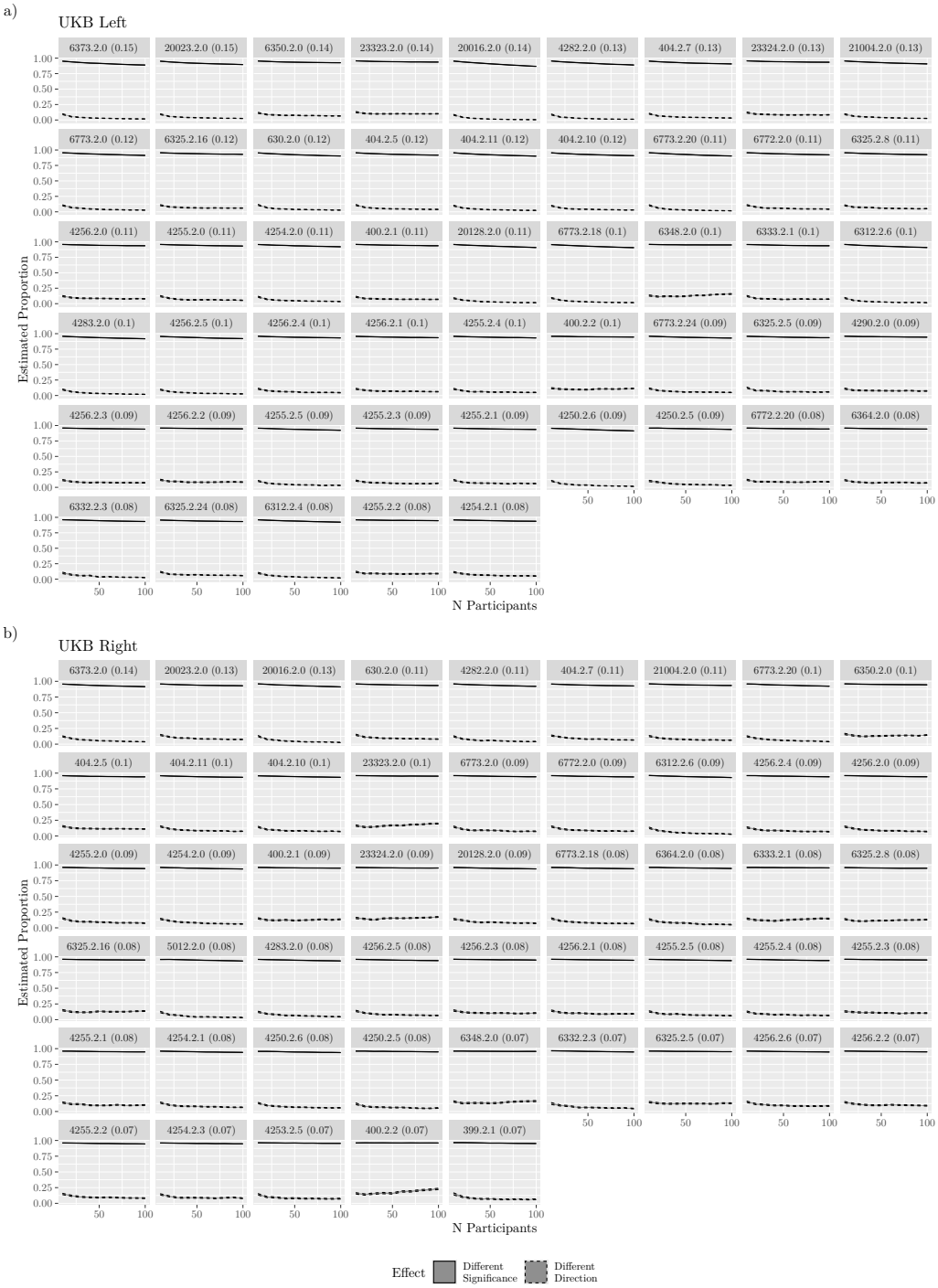

Figure A3: Effects of Low Measurement Consistency on the UKB in Left and Right Hemispheres. Ribbons span 95% equal-tailed interval estimated from simulated experiments. See also Figure 2, where the medians are represented with color.

Appendix A.6. Differing Significance Minimally Affected by Increasing Sample Sizes When ICC is Low

As described in the main text, when intraclass correlations were low, increasing the sample size affected the rates of differing significance only minimally Figure 2. This lack of influence can be understood by inspecting the distribution of simulated correlations Figure A4. When the intraclass correlation is low Figure A4, the significance of one correlation is nearly uninformative about the significance of the other (that is, the distributions of the two correlations are nearly circular). Moreover, the power to detect small correlations with even 100 participants is low, and so the power for two tests is very low. But when the intraclass correlation is higher Figure A4, the two correlations cluster, which improves the power of a second test conditioning on one test being significant.

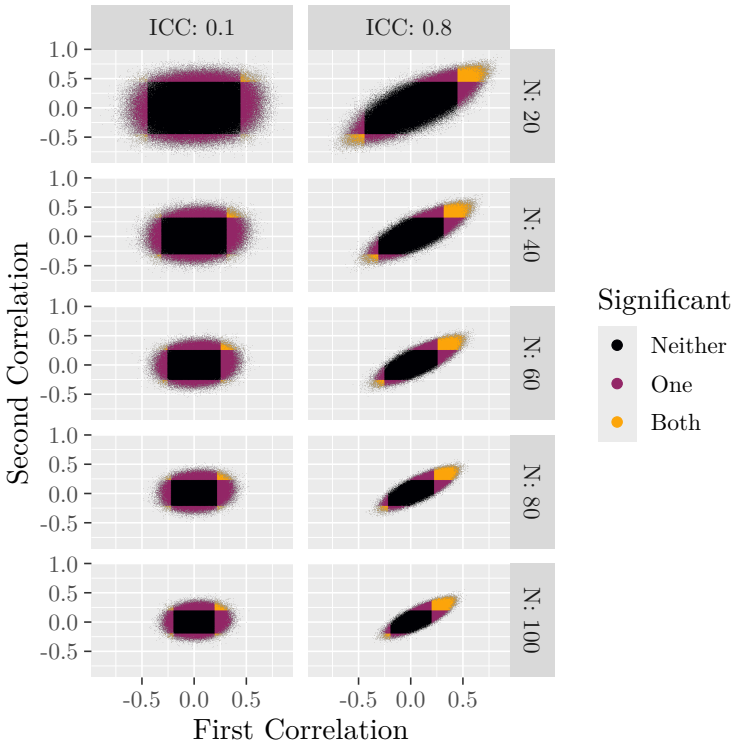

Figure A4: Correlations from Simulations with Artificial Data. Points correspond to simulations and are colored based on statistical significance. The figure shows only a subset of the simulated sample sizes (rows), only the smallest and largest intraclass correlations (columns), and only simulations in which the true effect size (correlation) was 0.1.

Appendix A.7. Measurement Error and Reduced Correlation Magnitude

Using the model from Appendix A.1, we can build a measurement noise model (e.g., Frost and Thompson, 2000). Call the final target for which we hope to find a relationship with the volume y. The relationship between that value and the volume was given by an ordinary linear regression with error δ.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 \lambda_i + \delta_i \\
\delta_i &\sim N(0, \sigma_\delta^2)
\end{aligned}
$$

But since we do not know λ, the volumes estimated by the tools are used instead, changing the regression coefficient as follows

$$
y_i = \beta_0 + \tilde{\beta}_1 x_i + \delta_i
$$

Dilution occurs because the coefficient $\tilde{\beta}_1$ estimated in this model tends to be closer to zero, decreased by a factor related to the intraclass correlation (Frost and Thompson, 2000).

$$
\tilde{\beta}_1 = \beta_1 \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2}
$$

Correspondingly, the desired correlation, $\rho = cor(y, \lambda)$, will also be biased.

$$
\begin{aligned}
\beta_1 &= \rho \frac{\sigma_\delta}{\sigma_\lambda} \\
\tilde{\beta}_1 &= \tilde{\rho} \frac{\sigma_\delta}{sd(x)} \\
&= \tilde{\rho} \frac{\sigma_\delta}{\sqrt{\sigma_\epsilon^2 + \sigma_\lambda^2}} \\
&\implies \\
\rho \frac{\sigma_\delta}{\sigma_\lambda} &= \tilde{\rho} \frac{\sigma_\delta}{\sqrt{\sigma_\epsilon^2 + \sigma_\lambda^2}} \frac{\sigma_\lambda^2 + \sigma_\epsilon^2}{\sigma_\lambda^2} \\
&\implies \\
\rho &= \tilde{\rho} \frac{\sqrt{\sigma_\epsilon^2 + \sigma_\lambda^2}}{\sigma_\lambda}
\end{aligned}
$$

Appendix A.8. Correlations Between Volumes of the Amygdala and Cognitive Variables

As described in the main text, a set of 50 "cognitive" variables from the UKB was selected for each hemisphere. Selection was based on the rank correlation between the variable and the average (across methods) amygdalar volume. The correlations between those variables and the original volume estimates are shown in Figure A5. For both hemispheres, the magnitude of the correlations with the volumes as reported by FreeSurfer tended to be higher than those reported by FSL.

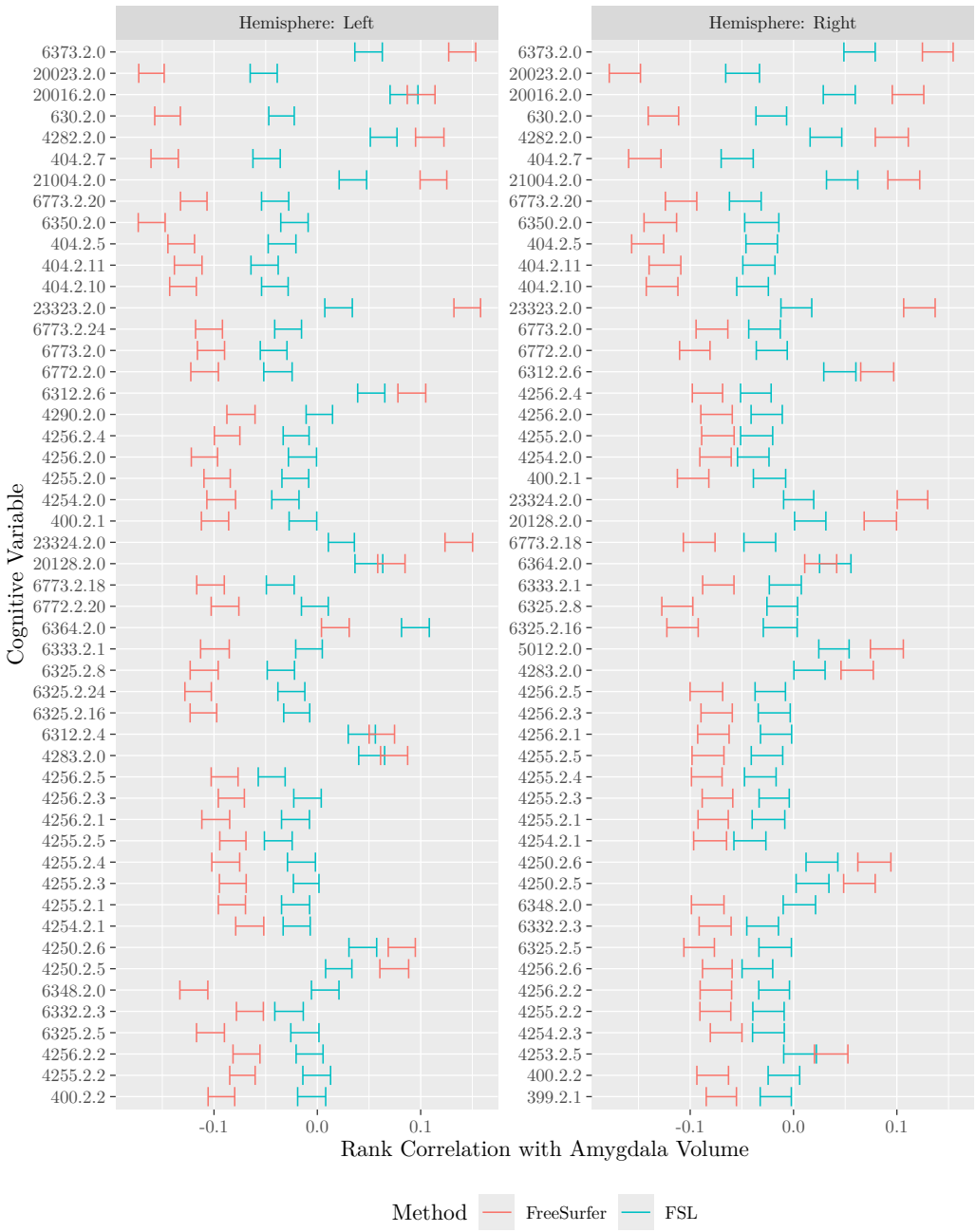

Figure A5: Correlations Between Cognitive Factors and Estimated Volumes of the Left and Right Amygdala. Variables are ordered by decreasing rank correlation using average of left hemisphere volume estimates. Note that variables were selected based on the magnitude of their correlation with amygdalar volumes, which differed across hemispheres, and so the variables in left and right panels differ. Error bars span 95% confidence intervals (bootstrapped with 1000 samples).