Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream

Emily P Hedges1, Mihail Dimitrov2, Uzma Zahid1, Barbara Brito Vega1, Shuqing Si1, Hannah Dickson2, Philip McGuire1, Steven Williams3, Gareth J Barker3, Matthew J Kempton1

**Affiliations**

1 Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, United Kingdom SE5 8AF

2 Department of Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, United Kingdom SE5 8AF

3 Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, United Kingdom SE5 8AF

Corresponding author: Emily P. Hedges; Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, 16 De Crespigny Park, SE5 8AF; emily.p.hedges@kcl.ac.uk

# Abstract

**Background:** Large-scale longitudinal and multi-centre studies are used to explore neuroimaging markers of normal aging, and neurodegenerative and mental health disorders. Longitudinal changes in brain structure are typically small, therefore the reliability of automated techniques is crucial. Determining the effects of different factors on reliability allows investigators to control those adversely affecting reliability, calculate statistical power, or even avoid particular brain measures with low reliability. This study examined the impact of several image acquisition and processing factors and documented the test-retest reliability of structural MRI measurements.

**Methods:** In Phase I, 20 healthy adults (11 females; aged 20-30 years) were scanned on two occasions three weeks apart on the same scanner using the ADNI-3 protocol. On each occasion, individuals were scanned twice (repetition), after re-entering the scanner (reposition) and after tilting their head forward. At one year follow-up, nine returning individuals and 11 new volunteers were recruited for Phase II (11 females; aged 22-31 years). Scans were acquired on two different scanners using the ADNI-2 and ADNI-3 protocols. Structural images were processed using FreeSurfer (v5.3.0, 6.0.0 and 7.1.0) to provide subcortical and cortical volume, cortical surface area and thickness measurements. Intra-class correlation coefficients (ICC) were calculated to estimate test-retest reliability. We examined the effect of repetition, reposition, head tilt, time between scans, MRI sequence and scanner on reliability of structural brain measurements. Mean percentage differences were also calculated in supplementary analyses.

**Results:** Using the FreeSurfer v7.1.0 longitudinal pipeline,we observed high reliability for subcortical and cortical volumes, and cortical surface areas at repetition, reposition, three weeks and one year (mean ICCs>0.97). Cortical thickness reliability was lower (mean ICCs>0.82). Head tilt had the greatest adverse impact on ICC estimates, for example reducing mean right cortical thickness to ICC=0.74. In contrast, changes in ADNI sequence or MRI scanner had a minimal effect. We observed an increase in reliability for updated FreeSurfer versions, with the longitudinal pipeline consistently having a higher reliability than the cross-sectional pipeline.

**Discussion:** Longitudinal studies should monitor or control head tilt to maximise reliability. We provided the ICC estimates and mean percentage differences for all FreeSurfer brain regions, which may inform power analyses for clinical studies and have implications for the design of future longitudinal studies.

*Key words:* Structural MRI, Longitudinal, Reproducibility, Morphology

# Introduction

Longitudinal and multi-centre studies have become increasingly popular to explore neuroimaging markers of normal aging, and neurodegenerative and mental health disorders (Misra et al., 2009, Scahill et al., 2003, van Erp et al., 2016). Due to the large number of participants in such studies, quantifying brain structures by time-consuming manual tracing methods becomes unfeasible (Morey et al., 2009). Instead, automated brain segmentation techniques are used to process magnetic resonance imaging (MRI) scans (Jovicich et al., 2013, Morey et al., 2010). FreeSurfer is an open-source automated segmentation software package (Martinos Center for Biomedical Imaging, Harvard-MIT, Boston; <https://surfer.nmr.mgh.harvard.edu/>) that has been widely used in brain volume studies, such as UK Biobank (Elliott et al., 2018), ENIGMA Consortium (Thompson et al., 2014) and Lifebrain (Walhovd et al., 2018). FreeSurfer is appealing to longitudinal and multi-centre studies as it offers a fully automated longitudinal processing stream for structural MRI images (Reuter et al., 2012) that extracts reliable brain morphometry estimates (Jovicich et al., 2013, McCarthy et al., 2015). In addition, by using the same template system, FreeSurfer provides consistent results as it is not subject to inter-rater variability, unlike manual tracing (Lehmann et al., 2010, Morey et al., 2009). Nevertheless, combining data in longitudinal studies may contribute additional sources of variance, such as changes in MRI sequence, that could limit the reliability of brain measurements. Therefore, it is important for clinical neuropsychiatric research to determine factors that may affect the reliability of FreeSurfer-derived measurements so we can increase the sensitivity of longitudinal MRI to real disease-related changes (Jovicich et al., 2006).

Previous reliability studies of structural MRI have demonstrated that changes in image acquisition variables, such as scan session, MRI sequence (Jovicich et al., 2009) and head motion during scanning (Alexander‐Bloch et al., 2016, Reuter et al., 2015) may affect brain measurement reliability. Moreover, changes in FreeSurfer segmentation methods, such as inter-version variation (Chepkoech et al., 2016, Whelan et al., 2016) and processing stream (Jovicich et al., 2013), may also affect the reproducibility of brain measurements. To assess the reliability of repeated measures, intra-class correlation coefficients (ICC) (McGraw and Wong, 1996) are a favourable metric that have been widely used in research from genetics (Gibert et al., 1998) to clinical rating scales (Nuechterlein et al., 2008), as well as in neuroimaging studies (Du et al., 2020, Morey et al., 2010, Sederevičius et al., 2021, Takao et al., 2021). However, reliability studies to date are significantly limited: studies have focused on older versions of FreeSurfer software and commonly report on only a few regional volume or thickness estimates. Yet, it is imperative to quantify and report the reliability of all segmented brain regions, as well as examining the effects of a range of image acquisition and processing factors on reliability. This will allow future studies to address factors that may reduce the reliability of MRI-derived measurements and so increase sensitivity of longitudinal MRI to real changes of interest. In addition, ICC reliability values may be used to inform power analyses to ensure clinical studies have sufficient statistical power to detect small group differences (Liem et al., 2015).

The purpose of the present study was to examine the effects of image acquisition variables (i.e., scan session, subject positioning and head tilt, MRI sequence and scanner) and FreeSurfer processing variables (i.e., software version and processing stream) on brain morphometry estimates. For the Precision in Neuroimaging study, we adopted a prospective, longitudinal design where healthy participants were scanned at baseline, three weeks and one year. For MRI acquisition, we chose to use the third generation of the Alzheimer’s Disease Neuroimaging Iniative (ADNI) protocol (<http://adni.loni.usc.edu/>) as the ADNI-3 sequence was specifically developed for longitudinal multi-centre studies (Jack Jr et al., 2008). However, for comparison, we also collected MRI data using the second generation of the ADNI protocol as the ADNI-2 sequence has already been used in the ADNI consortium and several international studies of mental health disorders [e.g., EU-GEI High Risk Study (Modinos et al., 2020) and PSYSCAN (Tognin et al., 2020)]. Processing of structural MRI data was independently carried out for the cross-sectional and longitudinal streams of three FreeSurfer versions (v5.3.0, v6.0.0, v7.1.0). We have conducted a comprehensive assessment of reliability for all FreeSurfer-segmented cortical and subcortical volumes, cortical surface area, and cortical thickness estimates. Results were derived from ICC estimates of individual brain regions that are freely available in a supplement to this paper. Furthermore, the defaced neuroimaging data for a subset of participants who provided consent has been made publicly available for researchers to download ([https://sites.google.com/view/pinstudy](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fsites.google.com%2Fview%2Fpinstudy&data=04%7C01%7Cemily.p.hedges%40kcl.ac.uk%7Cf5482be0c6724bc4175208d9898fec72%7C8370cf1416f34c16b83c724071654356%7C0%7C0%7C637692073495293136%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=2fl%2Bj%2F6iZrkcDKvZOccbWmdxgT%2Fy%2FMTNX5ixX%2B%2FiiF4%3D&reserved=0)).

# Methods

## 2.1 Participants

The Precision in Neuroimaging (PIN) study consisted of two phases. Participant recruitment for Phase 1 began in December 2017 and Phase II in March 2019. At Phase I, twenty participants were recruited from King’s College London and from the general public. At Phase II, nine Phase I participants were successfully re-recruited. Phase I participants who did not wish to take part in Phase II were replaced with new participants.

For Phase I and Phase II, participants were screened to determine study eligibility which consisted of the following criteria: (1) aged 18-31 years old; (2) mentally and physically healthy; (3) no contraindication to MRI (e.g., no pacemaker or metal in the body), and (4) registered with a general practitioner in the UK. Ethical approval was obtained from King’s College London Research Ethics Committee and participants gave written informed consent to take part in the study. In addition, a subset of participants provided consent for their demographic and defaced neuroimaging data to be made publicly available to other researchers at the end of the study.

## 2.2 Experimental design

At Phase I and II, participants attended King’s College London for two visits approximately three weeks apart for an MRI scan.

At Phase I, all scans were acquired on Scanner 1 (details below). Participants first received two identical MRI scans (A1 & A2). Participants were then asked to get out and immediately re-enter the scanner (A3), which intended to replicate longitudinal study designs where each scan might be associated with persons’ positioning within the scanner. For the last scan, participants were instructed to tilt their head down by approximately ten degrees (A4); *see* Supplementary S1 eFigure 1for graph showing mean head tilt (pitch) movement of 7.61o. This movement in pitch, and the resulting change in vertical translation of the head, has been reported as the most common head movement of adults during an MRI scan (Cusack et al., 2017). Localizer scans were administered prior to the A1 and A3 scans. This procedure was repeated at the second visit (B1 – B4).

For Phase II visits, participants received three MRI scans on both Scanner 1 (C1 – C3) and Scanner 2 (D1 – D3). Scanners 1 and 2 were different scanners but the same model (details below). The order of scanner used first was manually balanced for the participants. This procedure aimed to examine inter-scanner variability. The third scan of each visit (C3 & D3) was acquired using a different MRI sequence compared to all other scans to explore inter-sequence variability (*see* Figure 1 for data acquisition process).

##

*Figure 1: Illustration of the structural MRI data acquisition process in Phase I and II of the PIN study. In Phase II, the order of visits was balanced for participants so the D1 – D3 scans (on Scanner 2) were acquired first for ten participants. C3 and D3 scans were acquired using a different MRI sequence.*

## 2.3 MRI data acquisition

Structural MRI data were acquired using 3T MR750 Discovery GE (GE Healthcare, Chicago, USA) scanners: Scanner 1 (Software Version DV25.0\_R02) and Scanner 2 (Software Version DV25.1\_R02). MRI scanners did not undergo software updates during the course of the study and the GE standard distortion correction, “Gradwarp”, was enabled on the scanners.

The primary acquisition used was the T1 ADNI-3 sequence; an accelerated sagittal IR-FSPGR (Repetition Time (TR) = 7.36ms, Echo Time (TE) = 3.05ms, Inversion Time (TI) = 400ms, Flip Angle (FA) = 11o, Voxel Size (VS) = 1x1x1mm, Matrix Size = 256 x 256, Field of View (FOV) = 256, 196 slices). The ADNI-3 sequence was chosen as it has been specifically developed for longitudinal multi-centre studies where similar image qualities, such as reliability, contrast-to-noise, and spatial resolution, must be achieved across sites and over time (Jack Jr et al., 2008).

At Phase II, C3 and D3 scans were acquired using the ADNI-2 non-accelerated sagittal T1-weighted sequence (TR = 7.34ms, TE = 3.04ms, TI = 400ms, FA = 11o, VS = 1x1x1.2mm, Matrix Size = 256 x 256, FOV = 260, 196 slices). In addition to the ADNI consortium, the ADNI-2 sequence has previously been used in several international studies of mental health disorders [e.g., EU-GEI (Modinos et al., 2020) and PSYSCAN (Tognin et al., 2020)].

## 2.4 MRI data processing

To examine the effect of software version, FreeSurfer v5.3.0, v6.0.0 and v7.1.0 (Martinos Center for Biomedical Imaging) were used for structural MRI data pre-processing, segmentation and parcellation. The FreeSurfer automated stream consists of several processing steps, such as skull stripping, Talairach transformation, subcortical structure labelling, surface extraction, spherical registration and cortical parcellation. For each subject, we used the FreeSurfer cross-sectional and longitudinal streams. For cross-sectional analysis, the scans at every time point are processed independently for each subject. For longitudinal analysis, FreeSurfer uses results from the cross-sectional analysis of all time points to create an unbiased within-subject template; both of which are then used to initialise the final longitudinal processing stage (Reuter et al., 2010, Reuter et al., 2012). This three-stage procedure reduces variability and increases morphometry reproducibility compared to the cross-sectional analysis (Reuter et al., 2012). Scans were processed for the longitudinal stream in four groups: (1) Phase I (A1 – B4); (2) Phase I and II for nine participants (A1 – C2); (3) Phase II C1 and C3 data; and (4) Phase II ADNI-3 scans (C1, C2, D1 & D2).

The data was inspected for hard failures (i.e., where pre-processing is aborted) and screened for the presence of soft failures by manual visual inspection following the protocol of Iscan and colleagues (2015). No hard failures were seen, and no manual edits were made. Structural brain measurements were not impacted by normal variation in physiological variables, including the time of day of scans, blood pressure and hydration levels (Zahid et al., in press).

For the reliability analysis, we analysed volume, surface area and thickness morphometry measurements generated from FreeSurfer’s subcortical segmentation (ASEG) and cortical parcellation (APARC). We excluded some subcortical structures that were highly skewed, such as WM-hypointensities and 5th ventricle (cavum septum pellucidum). Further, we did not include the estimated total intracranial volume as this measure remains fixed for a given subject during the longitudinal processing stream (*see* Supplementary Materials S2 and S3 for brain regions included). In the Results Section 3.2.1, we determined that the head tilt condition was associated with reduced cortical thickness. To investigate this further, we used local values of the contrast-to-noise ratio and the number of surface holes (which indicate topological defects in the segmentation of the cortex); both of which are reported by FreeSurfer.

## 2.5 Statistical analyses

Reliability analysis of morphometric measurements was conducted in RStudio (RStudio Team, 2021) using the *irr* (Gamer et al., 2019) and *ggseg* (Mowinckel and Vidal-Piñeiro, 2020) packages. We calculated two-way mixed model intra-class correlation coefficients with absolute agreement of single measurements (McGraw and Wong, 1996):

ICC = $\frac{MSr-MSe}{MSr+\left(k-1\right)MSe+\frac{k}{n}(MSc-MSe) }$

where **MSr** is the mean square for rows (i.e., between-scan measurements), **MSe** is the mean square for error, **MSc** is the mean square for columns (i.e., within-scan measurements) and **k** is the number of measurements. Identical brain volumes produce an ICC value of 1.0. For volumetric studies, an ICC value greater than 0.9 indicates excellent reliability (Schnack et al., 2004).

Supplementary analyses were also carried out to calculate the mean absolute percentage differences (MPD) using the following equation:

MPD = $\frac{\sum\_{i=1}^{n}|\frac{(x\_{j}- y\_{j})}{x\_{j}}|}{n}×100$

where **x** is the reference scan, i.e., A1, **b** is the scan of interest, i.e., A2, and **j** represents the participant. Low MPD values indicate low variability and therefore, values closest to 0 are considered to be the most reproducible (McGuire et al., 2017). A random sample of ICCs and MPDs were validated in IBM SPSS Statistics 26 (IBM) to verify accuracy.

To compare the reliability of FreeSurfer processing, ICC and MPD values were calculated for data generated from both the cross-sectional and longitudinal streams of each FreeSurfer version. For each morphometric measure, seven reliability estimates were calculated. Structural T1-weighted ADNI-3 A1 data was compared to: (1) A2 (referred to as repetition), (2) A3 (reposition), (3) B1 (three-weeks), (4) C1 (one-year). In addition, A3 data was compared to A4 data (head tilt), C1 to C3 data (sequence change) and C1 to D1 data (scanner change) (*refer* back to Figure 1). To reduce the number of analyses, we focused on comparisons that may have implications for longitudinal studies, therefore B2–B4, C2 and D2-D3 scans were not compared in the current paper.

Post-hoc statistical power calculations were carried out to determine the sample size required to detect longitudinal changes of predetermined magnitudes in regional thickness measurements, based on recent work by Laguna and colleagues (2020). Specifically, we calculated the minimum number of participants needed to detect 0.5%, 1.0% and 2.0% group-level changes from baseline to (a) three-weeks and (b) after head tilt. These percentage changes are in line with those reported in examples of neurological and psychiatric disorders (Olabi et al., 2011; Frings et al., 2014). To note, the selected values relate to 0.5%, 1.0% and 2.0% changes in FreeSurfer measurement and may not directly correspond to the same biological change. For the power calculations, we obtained the effect size relating to a paired *t*-test on the difference between two dependent means (Cohen, 1988). We used the *in-vivo* MRI data for (a) A1 and B1 scans for three weeks and (b) A3 and A4 scans for head tilt to estimate the population means (μx and μy), population variances (σx2 and σy2) and the correlation between the observations (ρxy). To calculate changes of a predetermined magnitude, we replaced | μx - μy | with | μx \* *pc*| (for more details *see* Laguna et al., 2020) so that:

*d*z = $\frac{| μ\_{x} \* pc | }{\sqrt{σ\_{x}^{2}+ σ\_{y}^{2} – 2ρ\_{xy}σ\_{x}^{2}σ\_{y}^{2}}}$

where ***pc*** = 0.005, 0.01, 0.02.In G\*Power (Faul et al., 2007), we entered the computed effect size, *d*z, statistical significance level α = 0.05 (two-tailed) and statistical power of at least 1 − β = 0.8.

# 3. Results

## 3.1 Participants

Twenty healthy participants (11 female and 9 male, aged = 20-30, mean age = 24.0±2.9) took part in Phase I of the study. Mean interval between participants’ first and second visit of Phase I was 20.3±16.1 days. Nine Phase I participants and 11 new participants completed Phase II (11 female and 9 male, aged = 22-31, mean age = 24.7±2.5). For the nine returning participants, the mean number of days from the start of Phase I to Phase II was 474.7±43.2. For Phase II, mean interval between 20 participants’ first and second visit was 9.1±4.0 days. Demographic and structural MRI data for a subset of 24 participants is publicly available to download at <https://sites.google.com/view/pinstudy>.

## 3.2 Morphometric measure reliability

### 3.2.1 Impact of image acquisition factors

Table 1 shows mean ICC values of subcortical and cortical volumes, cortical thickness, and cortical surface area for the seven planned comparisons using FreeSurfer v7.1.0 longitudinal stream (*see* Supplementary Materials S2 for ICC values of individual brain regions). There were high values of reliability for all subcortical and cortical volumes and cortical surface area comparisons, with values ranging from 0.964 – 0.990 (95% confidence intervals (CI): 0.804 – 0.997). Cortical thickness appears to be the least reliable morphometric measure with an ICC range of 0.736 – 0.926 (95% CI: 0.352 – 0.970). As a result, using data presented in Supplementary Materials S2, we calculated the mean ICC (across the seven planned comparisons) for thickness estimates of individual regions. Five regions with highest cortical thickness reliability were the right hemisphere caudal anterior cingulate, precentral gyrus, isthmus cingulate, and right and left hemisphere parahippocampus (ICCs of 0.937 – 0.969). Those with the lowest reliability estimates were the right hemisphere supramarginal gyrus, superior parietal lobe, temporal pole, and right and left hemisphere inferior parietal lobe (ICCs of 0.714 – 0.754).

LH, left hemisphere; RH, right hemisphere

1generated from subcortical segmentation (aseg.stats) of FreeSurfer v7.1.0 longitudinal stream

2generated from cortical parcellation (aparc.stats) of FreeSurfer v7.1.0 longitudinal stream

3calculated for each comparison as the mean ICC value of subcortical volume and cortical volume, thickness, and surface area measurements

4n=9 participants

*Table 1: Mean ICC values of subcortical and cortical volumes, cortical thickness, and cortical surface area morphometric measurements from FreeSurfer v7.1.0 longitudinal stream*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   |   | RepetitionA1 vs A2  | RepositionA1 vs A3 | Head tiltA3 vs A4 | Three weeksA1 vs B1 | One year4A1 vs C1 | SequenceC1 vs C3 | ScannerC1 vs D1 |
| Subcortical volume1 | 0.984 | 0.980 | 0.976 | 0.976 | 0.972 | 0.970 | 0.975 |
| Cortical volume2 | LH | 0.987 | 0.986 | 0.965 | 0.979 | 0.972 | 0.979 | 0.980 |
| RH | 0.988 | 0.983 | 0.964 | 0.979 | 0.974 | 0.982 | 0.982 |
| Cortical thickness2 | LH | 0.924 | 0.926 | 0.769 | 0.892 | 0.864 | 0.862 | 0.873 |
| RH | 0.908 | 0.889 | 0.736 | 0.858 | 0.821 | 0.874 | 0.859 |
| Cortical surface area2 | LH | 0.990 | 0.987 | 0.976 | 0.987 | 0.986 | 0.986 | 0.983 |
| RH | 0.988 | 0.984 | 0.978 | 0.985 | 0.988 | 0.984 | 0.986 |
| Comparison mean value3 | 0.967 | 0.962 | 0.909 | 0.951 | 0.940 | 0.948 | 0.948 |

From Table 1, asking participants to tilt their head in the scanner (mean pitch movement of 7.61o) resulted in lower reliability of morphometric measurements (overall mean ICC = 0.909) compared to follow-up scans at three weeks (ICC = 0.951) or one year (ICC = 0.940). The reliability of individual brain regions is shown for ‘head tilt’ (Figure 2) and ‘three weeks’ (Figure 3).



*Figure 2: ICC values of subcortical and cortical volumes, cortical thickness, and cortical surface area morphometric measurements for ‘head tilt’ comparison from FreeSurfer 7.1.0 longitudinal stream. This change in pitch has been reported as the most common type of head movement seen in adult groups during an MRI scan (Cusack et al., 2017). For cortical thickness measurements (Image C), 16 regions had an ICC<0.7 (minimum value = 0.402) (see Supplementary Materials S2).*



*Figure 3: ICC values of subcortical and cortical volumes, cortical thickness, and cortical surface area morphometric measurements for ‘three-week’ comparison from FreeSurfer 7.1.0 longitudinal stream. For cortical thickness measurements (Image C), two regions had an ICC<0.7 (minimum value = 0.598) (see Supplementary Materials S2).*

To understand the low reliability of cortical thickness measurements after the head tilt, seen in Figure 2C, we first plotted the thickness estimates from A3 and A4 scans for a cortical region which was particularly impacted by head movement. Figure 4 indicates that head tilt movement is associated with an approximate 3 – 4% decrease in cortical thickness measurement. Using a paired samples t-test, we further observed a significant global reduction in mean cortical thickness following the head tilt (t[19] = -6.17, *p*<.001). Second, we tested whether the cortical thinning was associated with the difference in contrast-to-noise ratio or the number of surface holes after the head tilt among five cortical regions with the lowest reliability (listed in Section 3.2.1 above). The contrast-to-noise ratio was significantly reduced in the head tilt condition for four of the cortical regions (all *p* <.01). Results indicated that lower contrast-to-noise ratio after the head tilt was associated with a reduction in cortical thickness for all five regions (Pearson’s *r* of 0.46 to 0.62, all *p* <.05). In contrast, no significant correlations were reported for the difference in cortical thickness and surface hole measures before and after the head tilt (all *p* >.05).

*Figure 4: Left: scatterplot of right superior parietal* ***thickness*** *(mm) showing the effect of head tilt (A4 scan), which is associated with relatively low reliability (ICC=0.40) and an apparent reduction in cortical thickness. Right: scatterplot for* ***volume*** *(mm3)**comparison of the same region is shown where there is higher reliability (ICC=0.96).*

Post-hoc statistical power analyses were conducted to determine the minimum number of participants needed to detect 0.5%, 1.0% and 2.0% longitudinal changes in cortical thickness from baseline to (a) three-week scan compared to (b) scan after the head tilt. For a wide range of reliability, the five cortical regions showing the highest and lowest reliability in Section 3.2.1 were used. Figure 5 shows that a larger number of participants are typically required to detect these changes as a result of a head tilt after the localizer scan.

### 3.2.2 Impact of image processing factors

*Figure 5: Statistical power of ‘three-week’ and ‘head tilt’ comparisons across ten regional thicknesses of the highest and lowest reliability. Graph displays the sample size needed to detect longitudinal changes in thickness at p<0.05 significance level and 1-*β *statistical power = 0.80 (two-tailed). For example, for a study to be sufficiently powered to detect a 1% longitudinal change in the left parahippocampal thickness, 36 participants would be required using data from the three-week scan. However, 50 participants would be needed if there was a change in head tilt.*

Table 2 shows mean ICC values of subcortical and cortical volumes, cortical thickness, and cortical surface area for the seven planned comparisons using FreeSurfer v7.1.0 *cross-sectional* stream. Mean ICCs ranged from 0.706 – 0.950 (*see* Supplementary Materials S2 for ICC values of individual brain regions).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   |   | RepetitionA1 vs A2  | RepositionA1 vs A3 | Head tiltA3 vs A4 | Three weeksA1 vs B1 | One year4A1 vs C1 | SequenceC1 vs C3 | ScannerC1 vs D1 |
| Subcortical volume1 | 0.950 | 0.945 | 0.943 | 0.942 | 0.921 | 0.922 | 0.943 |
| Cortical volume2 | LH | 0.916 | 0.898 | 0.897 | 0.922 | 0.897 | 0.886 | 0.920 |
| RH | 0.926 | 0.908 | 0.884 | 0.915 | 0.843 | 0.842 | 0.887 |
| Cortical thickness2 | LH | 0.819 | 0.831 | 0.723 | 0.787 | 0.714 | 0.718 | 0.774 |
| RH | 0.839 | 0.821 | 0.706 | 0.774 | 0.739 | 0.729 | 0.780 |
| Cortical surface area2 | LH | 0.923 | 0.904 | 0.920 | 0.942 | 0.938 | 0.918 | 0.916 |
| RH | 0.935 | 0.912 | 0.906 | 0.929 | 0.846 | 0.872 | 0.887 |
| Comparison mean value3 | 0.901 | 0.888 | 0.854 | 0.887 | 0.843 | 0.841 | 0.872 |

LH, left hemisphere; RH, right hemisphere

1generated from subcortical segmentation (aseg.stats) of FreeSurfer v7.1.0 cross-sectional stream

2generated from cortical parcellation (aparc.stats) of FreeSurfer v7.1.0 cross-sectional stream

3calculated for each comparison as the mean ICC value of subcortical volume and cortical volume, thickness, and surface area measurements

4n=9 participants

*Table 2: Mean ICC values of subcortical and cortical volumes, cortical thickness, and cortical surface area morphometric measurements from FreeSurfer v7.1.0 cross-sectional stream*

From Table 3, we can begin to compare the reliability of the different FreeSurfer processing streams using data from the ‘three weeks’ comparison. FreeSurfer v7.1.0, v6.0.0 and v5.3.0 longitudinal streams displayed high reliability across morphometry measurements (overall FreeSurfer mean ICCs of 0.930 – 0.951) and demonstrated greater reliability than the cross-sectional streams (overall FreeSurfer mean ICCs of 0.829 – 0.887).

|  |  |  |  |
| --- | --- | --- | --- |
|  |   | FreeSurferlongitudinal stream | FreeSurfercross-sectional stream |
|   |   | 7.1.0 | 6.0.0 | 5.3.0 | 7.1.0 | 6.0.0 | 5.3.0 |
| Subcortical volume1 | 0.976 | 0.976 | 0.973 | 0.942 | 0.947 | 0.905 |
| Cortical volume2 | LH | 0.979 | 0.978 | 0.960 | 0.922 | 0.885 | 0.875 |
| RH | 0.979 | 0.971 | 0.964 | 0.915 | 0.857 | 0.812 |
| Cortical thickness2 | LH | 0.892 | 0.902 | 0.845 | 0.787 | 0.770 | 0.752 |
| RH | 0.858 | 0.859 | 0.802 | 0.774 | 0.759 | 0.722 |
| Cortical surface area2 | LH | 0.987 | 0.985 | 0.983 | 0.942 | 0.900 | 0.896 |
| RH | 0.985 | 0.982 | 0.985 | 0.929 | 0.868 | 0.841 |
| FreeSurfer mean value3 | 0.951 | 0.950 | 0.930 | 0.887 | 0.855 | 0.829 |

LH, left hemisphere; RH, right hemisphere

1generated from subcortical segmentation (aseg.stats)

2generated from cortical parcellation (aparc.stats)

3calculated for each FreeSurfer processing stream as the mean ICC value of subcortical volume and cortical volume, thickness, and surface area measurements

*Table 3: Mean ICC values of subcortical and cortical volume, cortical thickness, and cortical surface area morphometric measurements from three-weeks (A1 vs B1) comparison*

## 3.3 Supplementary analyses using mean absolute percentage difference

Focusing on results from the FreeSurfer v7.1.0 longitudinal stream, the mean absolute percentage difference (MPD) values for cortical volume, surface area and thickness measurements supported the ICC results, where head tilt had the greatest adverse effect on reproducibility. Generally, the MPD values were higher for subcortical volume measurements (1.948% to 2.852%), which were particularly impacted by poor inferior lateral ventricle reproducibility, and were more affected by the inter-scan interval, and change in MRI sequence or scanner (Supplementary S1 eTable 1). From the MPD results, we observed that brain measurements processed by FreeSurfer v7.1.0 longitudinal stream typically had higher reproducibility (mean MPD of 1.978%) than the cross-sectional stream and earlier FreeSurfer versions (mean MPDs of 2.046% to 5.038%) (Supplementary S1 eTables 2 and 3). Results for individual brain regions have been provided in Supplementary Materials S3.

# 4. Discussion

We reported a minimal decrease in reliability of longitudinal MRI measures from within-session to three weeks and one year. Generally, head tilt had a greater adverse impact on reliability of brain measurements compared to participant repositioning and changes in MRI sequence or scanner. Notably, the study demonstrated the reliability advantages of the FreeSurfer longitudinal stream compared to the cross-sectional stream. These findings were supported by supplementary MPD analyses. To our knowledge, this is the first study to provide ICC reliability and MPD values for all FreeSurfer-segmented regions, which may inform power calculations and study design for future longitudinal MRI studies.

We first demonstrated how the test-retest reliability of MRI-derived measurements are impacted by image acquisition factors, namely scan session (repetition), inter-scan interval (three weeks and one year), participant positioning, head tilt, MRI sequence and scanner. Focusing on results obtained from the FreeSurfer v7.1.0 longitudinal stream, repetition had the greatest reliability but the ICCs at three weeks and one year still indicated high reliability of structural brain measurements. These results are in line with early findings by Jovicich and colleagues (2009) who reported comparable within- and between-session reliability of subcortical volume estimates from FreeSurfer v4.0. In our study, the small decrease in reproducibility of brain measurements at one year may reflect physiological changes, such as normal aging (Scahill et al., 2003). In general, the reliability of brain measurements, and particularly cortical thickness, seemed to be most affected by participants’ head tilt. Furthermore, post-hoc statistical power calculations showed that, on average, an increased number of participants would be needed to detect longitudinal thickness changes following a head tilt compared to three-week scan. We have demonstrated evidence that reduced cortical thickness in the head tilt condition may arise due to reduced contrast-to-noise ratio in the MRI data. Although other head rotations could also have had an impact, we chose to examine the change in pitch (i.e., chin moving towards the neck) as this is a common motion over time during long scans. The head tilt (of seven degrees) occurred after the localizer image, but before the start of the scan, which may occur in a clinical study if the structural scan is acquired later in the protocol and the participant’s head has since moved. Previous research has only reported on the effects of head motion during a scan, which can also lead to systematic biases in brain estimation (Alexander‐Bloch et al., 2016, Reuter et al., 2015, Savalia et al., 2017). As well as reducing head motion during the scan (Reuter et al., 2015), we recommend that studies monitor or control for head tilt after the localizer scan. Furthermore, we reported that comparisons of different MRI sequences and scanners produced high reliability of brain measurements. Similarly, Jovicich and colleagues (2009) have shown comparable reliability results of two imaging sequences (MPRAGE and multi-echo FLASH), indicating that FreeSurfer segmentation of brain volumes is robust across similar image contrast properties. In a more recent study of 24 participants scanned across three models of Siemens MRI scanners, Sederevičius and colleagues (2021) reported high test-retest reliability of FreeSurfer v7.1.0 automated subcortical segmentation (ASEG). However, ASEG showed increased inter-scanner variability of repeated measures compared to Sequence Adaptive Multimodal Segmentation approach, newly released as part of FreeSurfer v7.1 software. As the ICC values for both the MRI sequence and scanner comparisons are similar to those at three weeks (using the same sequence and scanner), this suggests that these sequences and GE MR750 scanner could be combined in a longitudinal MRI study.

The present study examined the test-retest reliability of subcortical and cortical volume, cortical surface area, and cortical thickness estimates. For FreeSurfer v7.1.0 longitudinal processing, we found that volume and surface area measurements appeared to be more reliable measures, whereas cortical thickness estimates had lower reliability. Specifically, our findings indicated high reproducibility of cortical thickness measurements for scans obtained within the same session (repetition and subject repositioning), but reliability estimates did vary by brain region and were more affected by changes in acquisition and processing factors, such as head tilt, one year scan and MRI scanner or sequence change. This finding is similar to previous studies (Han et al., 2006, McGuire et al., 2017). Authors reported that cortical thickness measurements showed excellent ICC values but were impacted by imaging sequence and field strength (Han et al., 2006) and were lower for certain brain structures, including the entorhinal, insula and medial orbitofrontal (McGuire et al., 2017). For particular brain regions with low ICC reliability, it may be beneficial to use other software or even manual tracing measures to increase precision (Du et al., 2020). Overall, we determined that FreeSurfer-derived measurements are highly reproducible, particularly when acquisition and processing factors are controlled for. Therefore, it is important to consider these factors during the design of a longitudinal MRI study (Han et al., 2006).

Looking at the effect of image processing factors on reliability, our study highlighted the advantages of using the FreeSurfer v7.1.0 longitudinal stream. In line with previous findings (Han et al., 2006, Jovicich et al., 2013), our results showed that the longitudinal stream offers an improvement of reliability for both within- and between-session brain measurements compared to the cross-sectional stream. Furthermore, our results demonstrated increasing reliability of software versions over time (from v5.3.0 to v7.1.0), although the longitudinal streams of v7.1.0 and v6.0.0 showed similar reliability. Earlier reliability studies have also reported an effect of FreeSurfer software version (from v4.1.0 to v5.1.0) on cortical thickness estimates (Chepkoech et al., 2016, Gronenschild et al., 2012), suggesting that reliability is affected by FreeSurfer processing conditions. Due to version differences, we recommend that consistent FreeSurfer versions should be used for MRI data processing in ongoing longitudinal studies.

The present study has several strengths and limitations. First, we have conducted and shared a comprehensive reliability assessment of structural brain measurements in a group of healthy participants. Further, we have made the raw, defaced MRI data for a subset of participants publicly available to other researchers ([https://sites.google.com/view/pinstudy](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fsites.google.com%2Fview%2Fpinstudy&data=04%7C01%7Cemily.p.hedges%40kcl.ac.uk%7Cf5482be0c6724bc4175208d9898fec72%7C8370cf1416f34c16b83c724071654356%7C0%7C0%7C637692073495293136%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=2fl%2Bj%2F6iZrkcDKvZOccbWmdxgT%2Fy%2FMTNX5ixX%2B%2FiiF4%3D&reserved=0)). A third strength is that we selected the widely used ADNI-3 sequence for MRI acquisition which has been specifically developed for multi-site studies (Jack Jr et al., 2008) and is likely to be used in future longitudinal MRI studies. A limitation is that our sample consisted of 20 healthy individuals, which limits the interpretation and generalisability of our results (i.e., to samples of older adults and individuals with neurodegenerative disorders). However, this study recruited young adults as our clinical research focuses on mental health disorders, where the typical age of onset is in late adolescence and early adulthood (Kessler et al., 2008) and where we expect to see much smaller structural brain changes over time (Olabi et al., 2011). In addition, our sample size of healthy participants is similar to, if not greater than, previous reliability studies (Du et al., 2020, Jovicich et al., 2009, McGuire et al., 2017, Morey et al., 2010, Reuter et al., 2015). Furthermore, longitudinal neuroimaging studies require smaller sample sizes to detect small effect size differences in brain structure compared to cross-sectional studies (Steen et al., 2007). For example, for brain regions with ICCs close to 1.0, a longitudinal study requires approximately ten participants to be powered to detect small to large effect sizes (Morey et al., 2010). Another limitation is that, although we obtained data from two MRI scanners, they were both GE Discovery MR750 scanners. Therefore, our results cannot necessarily be generalised to MRI data combined from different scanner manufacturers in multi-centre studies.

As accurate and reliable automated segmentation is essential for longitudinal studies (Sederevičius et al., 2021), it is crucial to examine and report the reproducibility of quantitative brain morphometry results (Jovicich et al., 2006). Overall, the present study demonstrated high test-retest reliability of FreeSurfer-derived measurements and showed superiority of FreeSurfer v7.1.0 and v6.0.0 longitudinal streams. However, reliability of measurements still varied depending upon other image acquisition factors. Cortical thickness reliability estimates were lower for certain brain regions and so researchers may want to consider using alternative software, manual tracing methods, or even excluding these regions in their analysis. As head tilt seemed to have the greatest impact on reliability of structural brain measurements, we suggest that future longitudinal studies monitor and/or control head tilt after the localizer scan.

## Data availability statement

Data is available from [https://sites.google.com/view/pinstudy](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fsites.google.com%2Fview%2Fpinstudy&data=04%7C01%7Cemily.p.hedges%40kcl.ac.uk%7Cf5482be0c6724bc4175208d9898fec72%7C8370cf1416f34c16b83c724071654356%7C0%7C0%7C637692073495293136%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=2fl%2Bj%2F6iZrkcDKvZOccbWmdxgT%2Fy%2FMTNX5ixX%2B%2FiiF4%3D&reserved=0) for a subset of participants who provided consent for their demographic and defaced neuroimaging data to be made publicly available.

## CRediT authorship contribution statement

**Emily P Hedges:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Mihail Dimitrov:** Methodology, Software, Investigation, Data curation, Writing – review & editing. **Uzma Zahid:** Methodology, Investigation, Data curation, Writing – review & editing. **Barbara Brito Vega:** Investigation, Writing – review & editing. **Shuqing Si:** Validation, Formal analysis, Writing – review & editing, Visualization. **Hannah Dickson:** Writing – review & editing. **Philip McGuire:** Writing – review & editing. **Steve Williams:** Writing – review & editing, Funding acquisition. **Gareth J Barker:** Methodology, Writing – review & editing. **Matthew J Kempton:** Conceptualization, Methodology, Resources, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

## Declaration of competing interest

The authors declare no conflict of interest.

## Acknowledgements

M Kempton was supported a Medical Research Council Career Development Fellowship (grant MR/J008915/1). This study was partly funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

# References

ALEXANDER‐BLOCH, A., CLASEN, L., STOCKMAN, M., RONAN, L., LALONDE, F., GIEDD, J. & RAZNAHAN, A. 2016. Subtle in‐scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human brain mapping,* 37**,** 2385-2397.

CHEPKOECH, J. L., WALHOVD, K. B., GRYDELAND, H., FJELL, A. M. & INITIATIVE, A. S. D. N. 2016. Effects of change in FreeSurfer version on classification accuracy of patients with Alzheimer's disease and mild cognitive impairment. *Human brain mapping,* 37**,** 1831-1841.

COHEN, J. 1988. Chapter 2. The t Test for Means. In Statistical Power Analysis for the Behavioral Sciences. Academic Press, CY, pp. 19-74.

CUSACK, R., LINKE, A. C., ZUBIAURRE-ELORZA, L., DUFFY, H., HERZMANN, C., STOJANOSKI, B., HAN, V. K., LEE, D. S. C. & WILD, C. 2017. Differences in the spatial and temporal patterns of head motion during MRI of adults and infants. *bioRxiv*, 114447.

DU, J., LIANG, P., HE, H., TONG, Q., GONG, T., QIAN, T., SUN, Y., ZHONG, J. & LI, K. 2020. Reproducibility of volume and asymmetry measurements of hippocampus, amygdala, and entorhinal cortex on traveling volunteers: a multisite MP2RAGE prospective study. *Acta Radiologica***,** 0284185120963919.

ELLIOTT, L. T., SHARP, K., ALFARO-ALMAGRO, F., SHI, S., MILLER, K. L., DOUAUD, G., MARCHINI, J. & SMITH, S. M. 2018. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature,* 562**,** 210-216.

FAUL, F., ERDFELDER, E., LANG, A.-G. & BUCHNER, A. 2007. G\* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39, 175-191.

FRINGS, L., YEW, B., FLANAGAN, E., LAM, B. Y., HÜLL, M., HUPPERTZ, H.-J., HODGES, J. R. & HORNBERGER, M. 2014. Longitudinal grey and white matter changes in frontotemporal dementia and Alzheimer’s disease. *PloS one,* 9, e90814.

GAMER, M., LEMON, J., FELLOWS, I. & SINGH, P. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement.* R package version 0.84.1. https://CRAN.R-project.org/package=irr

GIBERT, P., MORETEAU, B., MORETEAU, J.-C. & DAVID, J. R. 1998. Genetic variability of quantitative traits in Drosophila melanogaster (fruit fly) natural populations: analysis of wild-living flies and of several laboratory generations. *Heredity,* 80**,** 326-335.

GRONENSCHILD, E. H., HABETS, P., JACOBS, H. I., MENGELERS, R., ROZENDAAL, N., VAN OS, J. & MARCELIS, M. 2012. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one,* 7**,** e38234.

HAN, X., JOVICICH, J., SALAT, D., VAN DER KOUWE, A., QUINN, B., CZANNER, S., BUSA, E., PACHECO, J., ALBERT, M. & KILLIANY, R. 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage,* 32**,** 180-194.

ISCAN, Z., JIN, T. B., KENDRICK, A., SZEGLIN, B., LU, H., TRIVEDI, M., FAVA, M., MCGRATH, P. J., WEISSMAN, M. & KURIAN, B. T. 2015. Test–retest reliability of freesurfer measurements within and between sites: Effects of visual approval process. *Human brain mapping,* 36**,** 3472-3485.

JACK JR, C. R., BERNSTEIN, M. A., FOX, N. C., THOMPSON, P., ALEXANDER, G., HARVEY, D., BOROWSKI, B., BRITSON, P. J., L. WHITWELL, J. & WARD, C. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine,* 27**,** 685-691.

JOVICICH, J., CZANNER, S., GREVE, D., HALEY, E., VAN DER KOUWE, A., GOLLUB, R., KENNEDY, D., SCHMITT, F., BROWN, G. & MACFALL, J. 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage,* 30**,** 436-443.

JOVICICH, J., CZANNER, S., HAN, X., SALAT, D., VAN DER KOUWE, A., QUINN, B., PACHECO, J., ALBERT, M., KILLIANY, R. & BLACKER, D. 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage,* 46**,** 177-192.

JOVICICH, J., MARIZZONI, M., SALA-LLONCH, R., BOSCH, B., BARTRÉS-FAZ, D., ARNOLD, J., BENNINGHOFF, J., WILTFANG, J., ROCCATAGLIATA, L. & NOBILI, F. 2013. Brain morphometry reproducibility in multi-center 3 T MRI studies: a comparison of cross-sectional and longitudinal segmentations. *Neuroimage,* 83**,** 472-484.

KESSLER, R. C., AGUILAR-GAXIOLA, S., ALONSO, J., ANGERMEYER, M. C., ANTHONY, J. C., BERGLUND, P. A., CHATTERJI, S., DE GIROLAMO, G., DE GRAAF, R. & DEMYTTENAERE, K. 2008. Lifetime prevalence and age of onset distributions of mental disorders in the World Mental Health Survey Initiative. *The WHO world mental health surveys: Global perspectives on the epidemiology of mental disorders.* Cambridge University Press.

LAGUNA, P. A. L., COMBES, A. J., STREFFER, J., EINSTEIN, S., TIMMERS, M., WILLIAMS, S. C. & DELL'ACQUA, F. 2020. Reproducibility, reliability and variability of FA and MD in the older healthy population: A test-retest multiparametric analysis. *NeuroImage: Clinical*, 26.

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.

LIEM, F., MÉRILLAT, S., BEZZOLA, L., HIRSIGER, S., PHILIPP, M., MADHYASTHA, T. & JÄNCKE, L. 2015. Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *Neuroimage,* 108**,** 95-109.

MCCARTHY, C. S., RAMPRASHAD, A., THOMPSON, C., BOTTI, J.-A., COMAN, I. L. & KATES, W. R. 2015. A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in neuroscience,* 9**,** 379.

MCGRAW, K. O. & WONG, S. P. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods,* 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.

MISRA, C., FAN, Y. & DAVATZIKOS, C. 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *Neuroimage,* 44**,** 1415-1422.

MODINOS, G., KEMPTON, M. J., TOGNIN, S., CALEM, M., PORFFY, L., ANTONIADES, M., MASON, A., AZIS, M., ALLEN, P. & NELSON, B. 2020. Association of adverse outcomes with emotion processing and its neural substrate in individuals at clinical high risk for psychosis. *JAMA psychiatry,* 77**,** 190-200.

MOREY, R. A., PETTY, C. M., XU, Y., HAYES, J. P., WAGNER II, 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.

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.

MOWINCKEL, A. M. & VIDAL-PIÑEIRO, D. 2020. Visualization of brain statistics with R Packages ggseg and ggseg3d. *Advances in Methods and Practices in Psychological Science,* 3, 466-483.

NUECHTERLEIN, K. H., GREEN, M. F., KERN, R. S., BAADE, L. E., BARCH, D. M., COHEN, J. D., ESSOCK, S., FENTON, W. S., FRESE III, P. D., FREDERICK J & GOLD, J. M. 2008. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. *American Journal of Psychiatry,* 165**,** 203-213.

OLABI, B., ELLISON-WRIGHT, I., MCINTOSH, A. M., WOOD, S. J., BULLMORE, E. & LAWRIE, S. M. 2011. Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biological psychiatry*, 70, 88-96.

REUTER, M., ROSAS, H. D. & FISCHL, B. 2010. Highly accurate inverse consistent registration: a robust approach. *Neuroimage,* 53**,** 1181-1196.

REUTER, M., SCHMANSKY, N. J., ROSAS, H. D. & FISCHL, B. 2012. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage,* 61**,** 1402-1418.

REUTER, M., TISDALL, M. D., QURESHI, A., BUCKNER, R. L., VAN DER KOUWE, A. J. & FISCHL, B. 2015. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage,* 107**,** 107-115.

RSTUDIO TEAM. 2021. *RStudio: Integrated Development Environment for R.* [Online]. RStudio, PBC, Boston, MA. Available: http://www.rstudio.com/ [Accessed].

SAVALIA, N. K., AGRES, P. F., CHAN, M. Y., FECZKO, E. J., KENNEDY, K. M. & WIG, G. S. 2017. Motion‐related artifacts in structural brain images revealed with independent estimates of in‐scanner head motion. *Human brain mapping,* 38**,** 472-492.

SCAHILL, R. I., FROST, C., JENKINS, R., WHITWELL, J. L., ROSSOR, M. N. & FOX, N. C. 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Archives of neurology,* 60**,** 989-994.

SCHNACK, H. G., VAN HAREN, N. E., HULSHOFF POL, H. E., PICCHIONI, M., WEISBROD, M., SAUER, H., CANNON, T., HUTTUNEN, M., MURRAY, R. & KAHN, R. S. 2004. Reliability of brain volumes from multicenter MRI acquisition: a calibration study. *Human brain mapping,* 22**,** 312-320.

SEDEREVIČIUS, D., VIDAL-PIÑEIRO, D., SØRENSEN, Ø., VAN LEEMPUT, K., IGLESIAS, J. E., DALCA, A. V., GREVE, D. N., FISCHL, B., BJØRNERUD, A. & WALHOVD, K. B. 2021. Reliability and sensitivity of two whole-brain segmentation approaches included in FreeSurfer–ASEG and SAMSEG. *NeuroImage,* 237**,** 118113.

STEEN, R., HAMER, R. & LIEBERMAN, J. 2007. Measuring brain volume by MR imaging: impact of measurement precision and natural variation on sample size requirements. *American journal of neuroradiology,* 28**,** 1119-1125.

TAKAO, H., AMEMIYA, S., ABE, O. & INITIATIVE, A. S. D. N. 2021. Reliability of Changes in Brain Volume Determined by Longitudinal Voxel‐Based Morphometry. *Journal of Magnetic Resonance Imaging,* 54**,** 672.

THOMPSON, P. M., STEIN, J. L., MEDLAND, S. E., HIBAR, D. P., VASQUEZ, A. A., RENTERIA, M. E., TORO, R., JAHANSHAD, N., SCHUMANN, G. & FRANKE, B. 2014. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior,* 8**,** 153-182.

TOGNIN, S., VAN HELL, H. H., MERRITT, K., WINTER-VAN ROSSUM, I., BOSSONG, M. G., KEMPTON, M. J., MODINOS, G., FUSAR-POLI, P., MECHELLI, A. & DAZZAN, P. 2020. Towards precision medicine in psychosis: benefits and challenges of multimodal multicenter studies—PSYSCAN: translating neuroimaging findings from research into clinical practice. *Schizophrenia bulletin,* 46**,** 432-441.

VAN ERP, T. G., HIBAR, D. P., RASMUSSEN, J. M., GLAHN, D. C., PEARLSON, G. D., ANDREASSEN, O. A., AGARTZ, I., WESTLYE, L. T., HAUKVIK, U. K. & DALE, A. M. 2016. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular psychiatry,* 21**,** 547-553.

WALHOVD, K. B., FJELL, A. M., WESTERHAUSEN, R., NYBERG, L., EBMEIER, K. P., LINDENBERGER, U., BARTRÉS-FAZ, D., BAARÉ, W. F., SIEBNER, H. R. & HENSON, R. 2018. Healthy minds 0–100 years: Optimising the use of European brain imaging cohorts (“Lifebrain”). *European Psychiatry,* 50**,** 47-56.

WHELAN, C. D., HIBAR, D. P., VAN VELZEN, L. S., ZANNAS, A. S., CARRILLO-ROA, T., MCMAHON, K., PRASAD, G., KELLY, S., FASKOWITZ, J. & DEZUBIRACAY, G. 2016. Heritability and reliability of automatically segmented human hippocampal formation subregions. *Neuroimage,* 128**,** 125-137.

ZAHID, U., HEDGES, E. P., DIMITROV, M., MURRAY, R. M., BARKER, G. J. & KEMPTON, M. J. 2021. Impact of physiological factors on longitudinal structural MRI measures of the brain, *in press.*