Estimating the effect of a scanner upgrade on measures of gray matter structure for longitudinal designs

Evelyn Medawar, MSc, [1] Ronja Thieleking, MSc, [1] Iryna Manuilova, BSc, [1]
Maria Paerisch, MSc, [1] Arno Villringer, Prof., [1][2][3]
A. Veronica Witte, PhD, [1][2] Frauke Beyer, PhD, [1][2]

1 Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig
2 CRC 1052 "Obesity Mechanisms", Subproject A1, University of Leipzig
3 Day Clinic for Cognitive Neurology, University Clinic Leipzig

Contents

1 Abstract

Longitudinal imaging studies are crucial for advancing the understanding of brain development over the lifespan. Thus, more and more studies acquire imaging data at multiple time points or with long follow-up intervals. In these studies changes to magnetic resonance imaging (MRI) scanners often become inevitable which may decrease the reliability of the MRI assessments and introduce biases.
We therefore investigated the difference between MRI scanners with subsequent versions (3 Tesla Siemens Verio vs. Skyra) on the cortical and subcortical measures of grey matter in 116 healthy, young adults using the well-established longitudinal FreeSurfer stream for T1-weighted brain images. We found excellent between-scanner reliability for cortical and subcortical measures of grey matter structure (intra-class correlation coefficient > 0.8). Yet, paired t-tests revealed statistically significant differences in at least 75% of the regions, with percent differences up to 5%, depending on the outcome measure. Offline correction for gradient distortions only slightly reduced these biases. Further, quality measures reflecting gray-white matter contrast systematically differed between scanners.
We conclude that scanner upgrades during a longitudinal study introduce bias in measures of cortical and subcortical grey matter structure. Therefore, before upgrading a MRI scanner during an ongoing study, researchers should prepare to implement an appropriate correction method for these effects.

2 Introduction

Many longitudinal neuroimaging studies of aging and development investigate changes in local grey matter volume (GMV) over time to identify biomarkers relevant to health and disease. Notably, in the past decade many large-scale studies have implemented longitudinal designs in the general population (with at least two timepoints: (Bycroft et al. 2018; Ikram et al. 2015), second timepoint currently being acquired: (Loeffler et al. 2015; Bamberg et al. 2015)).

Such longitudinal imaging studies assess within-subject differences and thereby benefit from reduction of error variance and confounding. Yet, scanner changes often become inevitable with long follow-up intervals (4-6 years) in these studies, entailing issues of reliability because of changes in signal-to-noise ratio or image intensity (Preboske et al. 2006; Takao et al. 2010; Ewers et al. 2006; Chen et al. 2014). This is especially problematic in the case of two-visit longitudinal imaging studies where measurement occasion may be collinear with scanner upgrade, making it difficult to draw unbiased conclusions on within-subject change. In contrast, scanner upgrades will affect cross-sectional designs less as scanner version can be modelled like a site effect (Fortin et al. 2018).

Before the follow-up of the LIFE-Adult Study, a two-visit longitudinal imaging study with a long inter-visit interval (5-7 years), we had to decide on the upgrade of the study scanner from Verio to Skyra (Loeffler et al. 2015). At the time (end of 2017), most studies on the effects of scanner upgrades had investigated small samples ($n<15$) or voxel-based morphometry estimates of grey matter (GM) structure, with varying estimates of reliability and bias (Jovicich et al. 2009; Shuter et al. 2008; Takao, Hayashi, and Ohtomo 2013). Thus, the impact of a scanner upgrade on region- and vertex-wise measures of cortical GM (thickness, area and volume) as well as subcortical GM volume still lacked quantification. Also, these studies did not take into account gradient distortion correction which has been shown to partly account for variation between scanners (Jovicich et al. 2006; Cannon et al. 2014).

Here, we therefore investigated the difference between scanners with subsequent versions (3 Tesla Verio vs. Skyra) on the cortical and subcortical measures of GM in a large sample of healthy, young adults. Differences between the systems included the changes introduced by software and hardware upgrades (update to syngo MR E11 software, a new Tim 4G body coil and installation of DirectRF) and side-specific variations in the scanner hardware. Using the validated longitudinal FreeSurfer stream, we expected the reliability of whole-brain and regional GM measures to be similar to previous studies investigating between-site reliability (Reuter et al. 2012; Keshavan et al. 2016; Jovicich et al. 2013). Based on previous upgrade studies, we hypothesized a systematic bias with varying effect sizes and direction in cortical and subcortical regions (Jovicich et al. 2009; Han et al. 2006). Finally, we expected gradient distortion correction to improve reliability and reduce bias.

3 Methods

3.1 Sample

121 healthy participants (median age = 28 years, range = 19 - 54 years; NA females) were scanned on two different 3 Tesla MRI scanners (Magnetom Verio syngo MR B17 (Siemens Healthcare, Erlangen, Germany), Magnetom Skyra fit syngo MR E11 (Siemens Healthcare, Erlangen, Germany). Scanners are referred to as Verio and Skyra throughout the manuscript. Due to a pending version upgrade of the Verio scanner, all participants were first scanned at the Verio and then at the Skyra scanner. The median time between sessions was 1.92 months (range: 0.12 - 4.56 months).
5 participants did only participate in the first scanning session at the Verio and were therefore excluded in the following analysis. The study was approved by the local ethics committee at the University of Leipzig and all participants gave written informed consent according to the Declaration of Helsinki.

3.2 Differences between scanners and Image Acquisition

The upgrade from Verio to Skyra includes an extensive retrofit of hardware and software components (e.g. new body coil, new RF transmit and receive signal transmission system and change to syngo MR E11 software). (https://cdn0.scrvt.com/39b415fb07de4d9656c7b516d8e2d907/1800000004906630/4b03e3ba438b/siemens-healthineers_mri_magnetom-skyra_fit_brochure_2018-02_1800000004906630.pdf] Further, we compare two different scanners which therefore also differ in the main B0-field and other hardware components, which would be untouched by a true upgrade.
On both scanners, anatomical T1-weighted imaging was performed with a magnetization-prepared rapid gradient-echo (MPRAGE) sequence (TR=2300 ms,TE=2.98 ms, TI=900 ms, parallel imaging: GRAPPA with factor 2 and adaptive coil combination, flip angle: 9°, imaging matrix 256 x 240 x 176 and voxel size= 1 mm^3, with prescan normalize option) according to the ADNI protocol (Jack et al. 2011). On both scanners, a 32 channel head coil was used. On the Skyra scanner, both online 3D gradient distortion-corrected images (D) and images not corrected for distortions (ND) were available. The Verio scanner delivered the images without gradient-distortion correction (ND).

3.3 Image Processing

3.3.1 FreeSurfer analysis

To extract reliable volume and thickness estimates, we processed the T1-weighted images with the longitudinal stream in FreeSurfer (Reuter et al. 2012). Within this pipeline, an unbiased within-subject template space is created using robust and inverse consistent registration (Reuter, Rosas, and Fischl 2010; Reuter and Fischl 2011). The longitudinal stream increases the reliability of cortical and subcortical GM estimates compared to the cross-sectional stream and is thus appropriate for longitudinal studies (Jovicich et al. 2013). We used FreeSurfer version 6.0.0p1 with the default parameters `recon-all -all -parallel -no-isrunning -openmp 8`, which include non-parametric non-uniform intensity normalization with the MINC tool `nu_correct`. We ran the `recon-all` longitudinal stream with Verio ND and Skyra ND images as input, and repeated the analysis with Verio ND and Skyra D images.

3.3.2 Gradient distortion correction

Gradient distortion correction has been shown to contribute to measurement error in repeated sessions of anatomical brain imaging (Takao et al. 2010). Accordingly, correcting for distortion correction can improve the reproducibility of intensity data significantly (Jovicich et al. 2006). For the Verio scanner, the vendor provided no online distortion correction while the Skyra system offered online 3D-distortion correction. To assess the effect of this processing step on reliability and bias, we applied an identical tool for offline gradient distortion correction on the ND sequences from both scanners.
Gradient unwarping calculates the geometric displacement based on the spherical expansion of the magnetic gradient fields and applies it to the image (Glasser et al. 2013; Jovicich et al. 2006). We used the gradunwarp implementation [(https://github.com/Washington-University/gradunwarp)] v1.1.0 in Python 2.7. We visually

compared the original and the `gradunwarp` result files to determine the appropriate number of sampling points and interpolation order. Based on this, we chose 200 sampling points and 4th order interpolation (`--fovmin -0.2 --fovmax 0.2 --numpoints 200 --interp_order 4`) because this yielded most similar intensity distributions. After unwarping, we repeated FreeSurfer's cross-sectional and longitudinal stream for these images. Then, we assessed the reliability and bias in cortical and subcortical ROI measures between the `gradunwarp` distortion corrected Skyra ND and Verio ND images.

3.3.3 Outcomes

We selected cortical thickness (CT), area (CA) and volume (CV) estimates for regions of interests defined by the Desikan-Killiany (DK) cortical parcellation (64 ROI for both hemispheres) as outcomes. Subcortical volumes were extracted from FreeSurfer's subcortical segmentation ("aseg.mgz", 18 bilateral ROI). We analyzed all ROI per hemisphere. Subcortical volumes were not adjusted for head size because during the longitudinal stream, both images are normalized to the same head size.

3.3.4 Quality Assessment

We visually checked the cross-sectional as well as the longitudinal runs for errors in white matter segmentation and misplaced pials (Klapwijk et al. 2019). There were 17 cases where the pial surface expanded into non-brain tissue. These were corrected by either editing the brainmask in the longitudinal template or by correcting the cross-sectional runs. After correction, we re-ran the longitudinal template creation step and the longitudinal timepoints. No issues regarding white matter segmentation were noticed.
To quantify potential differences in image quality between scanners, we compared different quality control measures provided by `mriqc` (version 0.15.0) (Esteban et al. 2017). We used signal-to-noise ratio (SNR) to assess overall signal qualiy and compared contrast-to-noise ratio (CNR) to quantify the difference between grey and white matter intensity distributions. Furthermore, we used coefficient of joint variation (CJV) which also reflects grey-to-white matter contrasts and entropy focus criterion (EFC) to describe the amount of ghosting and blurring induced by head motion (Esteban et al. 2017). We performed `mriqc` on the Verio ND, Skyra ND and Skyra D images.

3.4 Analysis

In the main analysis, we compared Verio ND and Skyra ND as they correspond to the same stage of image reconstruction and are therefore most comparable. Additionally, we investigated the effects of offline `gradunwarp` distortion correction on Verio ND and Skyra ND images. In a supplementary analysis, we compared Verio ND and Skyra D outcomes as these are the default reconstructions available at the respective scanners (see the diagram of the study flow in Figure **??**).

All statistical analysis were performed in R version 3.6.1 (R Core Team 2017). The package `fsbrain v.0.3.0` was used to plot vertex- and ROI-wise results (Schaefer and Ecker 2020).

3.4.1 Reliability and percent difference of cortical and subcortical GM measures

To assess the reliability of the grey matter (GM) estimates, we calculated the intra-class correlation coefficient (ICC), an established measure of agreement between raters. The ICC is calculated as the proportion of overall variance that is explained by between-subject variance, and thereby gives an estimate of the variance introduced by systematic differences and error between raters (Liljequist, Elfving, and Skavberg Roaldsen 2019).

$$\rho_{3A} = \frac{\sigma_{r\check{s}}}{\sigma_{r\check{s}} + \Theta_{c\check{s}} + \sigma_{\nu\check{s}}}$$

Here, $\sigma_{r\check{s}}$ is the population variance, $\Theta_{c\check{s}}$ is the variance of fixed biases and $\sigma_{\nu\check{s}}$ is the error variance. We used the two-way mixed effect ICC model for single measures with absolute agreement (Shrout and Fleiss 1979), implemented in the package `psy` to calculate ICC for each cortical DK and subcortical ROI and reported the estimate and 95% confidence interval, derived by bootstrapping. According to (Cicchetti 1994), we considered an ICC below .4 to be poor, between .40 and .59 to be fair; .60 and .74 to be good and between .75 and 1.00 to be excellent.

ICC depends on the between-subject variance (i.e. when between-subject variance is low, ICC decreases even if rater bias remains similar) and does not provide an estimation of bias and difference between measurements. Therefore, we used Bland-Altman plots with 95% limits of agreement to visually compare the agreement between the two scanners (Bland and Altman 1986).
To quantify the relative difference of GM measures between scanners, we calculated percent difference (PD) (also termed variability error (Jovicich et al. 2013; Iscan et al. 2015)). We calculated the mean of the PD for each ROI j across n participants according to

$$PD_j = \frac{2}{n} \sum_{i=1}^{n} \frac{V_{ij} - S_{ij}}{V_{ij} + S_{ij}}$$

where V_{ij} is the GM measure of a ROI measured on the Verio, S_{ij} is the GM measure of a ROI measured on the Skyra.
Finally, we performed paired t-tests to inform about the direction and statistical significance of potential systematic differences between scanners. Here, we used Benjamini-Hochberg correction to adjust p-values per cortical GM measure and deemed differences to be significant at p_{adj} <0.05 (Benjamini and Hochberg 1995). We reported T-value, uncorrected and corrected p-values.
We assessed the improvement by comparing the ICC and PD measures of CT and subcortical volume between the gradient distortion correction analysis (gradunwarp Skyra ND, gradunwarp Verio N), the original analysis (Skyra ND, Verio ND) and the secondary analysis (Skyra D, Verio ND).

3.4.2 Vertex-wise estimation of reliability and percent difference

For whole-brain visualization, we performed vertex-wise calculations on the fsaverage template following (Liem et al. 2015) in Matlab version 9.7 (2019b). We calculated ICC and PD for cortical thickness, area and volume to visualize reliability and difference between scanners on a vertex-wise level.

3.4.3 Quality metrics

For the quality metrics from `mriqc`, we used linear mixed models (LMM) to assess differences between scanners (Verio, Skyra) and acquisitions (D, ND) using `lmerTest`. Significance was defined based on model comparisons (using Chi-square test with R's `anova`) between LMM including either scanner or acquisition as a fixed effect and null models only including the random effects of subject. Significance was defined as $p < 0.05$. We also tested whether CNR was associated with regional CT, independent of scanner, using a LMM with both factors. We reported β estimates, raw and Benjamini-Hochberg adjusted p-values.

3.4.4 Data and Code availability

Region of interest data and code used for this publication are available on github (https://github.com/fBeyer89/life_upgrade). Under certain conditions, the authors may also provide access to the MRI data.

4 Results

4.1 Differences in cortical GM measures between scanners

Figures 1, 2 and 3 summarize the results for CT, CA and CV, respectively.

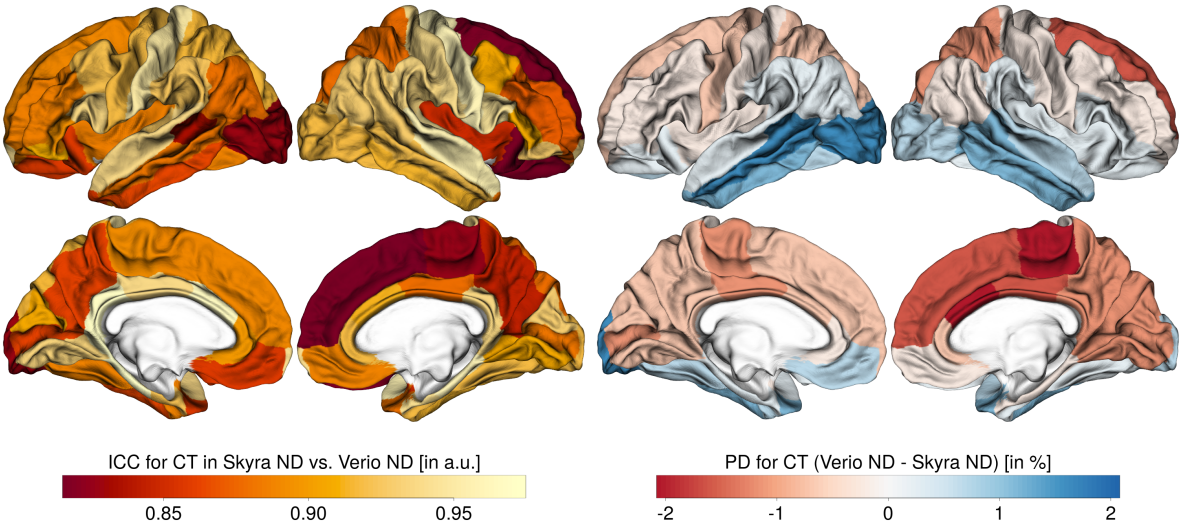

Figure 1: left panel: CT ICC, right panel: CT PD (for each panel, left column shows lateral and medial view of right hemisphere, right column shows lateral and medial view of left hemisphere), negative values:Skyra>Verio, positive values: Verio>Skyra

Overall, the ICC or scan-rescan reliability was excellent (CT: mean=0.91, min=0.82, max=0.97; CA: mean=0.98, min=0.91, max=1; CV: mean=0.98, min=0.93, max=0.99).
The PD was around 2-3% for all measures (CT: mean=0.2, min=-1.79, max=2.06; CA: mean=0.82, min=-2.99, max=3.42; CV: mean=1.51, min=-2.8, max=3.3) with a significant bias. The most pronounced bias was found for CT, with lower CT in Verio compared to Skyra in medial frontal and central regions, and higher values in Verio compared to Skyra in lateral occipital and inferior temporal regions. For CA and CV, the bias pattern was more related to gyrification, with higher CA/CV for Skyra compared to Verio in sulci, and the reverse pattern in gyri (see Figure @ref(fig:inflated_pd_maps)). Overall, the CT bias followed an medial-frontal to lateral-occipital pattern, and CA and CV differed between gyral-sulcal areas.

Bland-Altman plots confirmed the bias of Verio versus Skyra measurements. Exemplary plots of superior frontal and lateral occipital regions are shown in Figure 5.

For the superior frontal region, CT, CA and CV are larger for Skyra compared to Verio with 95% of CT differences were between 0.04 and -0.05 mm, 95% of CA differences were between 103.12 and -79.01 mm^2, and 95% of CV differences were between 694.23 and -453.16 mm^3. The inverse pattern was present in the lateral occipital region with 95% of CT differences were between -0.03 and -0.12 mm, 95% of CA differences were between 55.32 and -45.9 mm^2, and 95% of CV differences were between 86.04 and -600.19 mm^3. Accordingly, paired t-tests indicated systematic differences between scanners for the majority of regions of interest (FDR-corrected, CT: 67.2% of all 64 bilateral ROI, CA: 92.2%, CV: 90.6%).
For detailed results per cortical region see Tables 1.1, 2.1 and 3.1 in the Supplementary Material.

4.2 Differences in subcortical measures between scanners

As shown in Table 1, subcortical regions, similar to cortical areas, showed excellent reliability for all regions of interest (mean=0.95, min=0.81, max=0.99). The PD was around 2-3% (mean=2.76%, min=1.34%, max=9.56%), with an outlying PD of 9.5% for left Accumbens. Higher values were measured on Skyra compared to Verio for all regions, and these were significant for most regions (FDR-corrected, 85.7% of all 14 bilateral ROI).

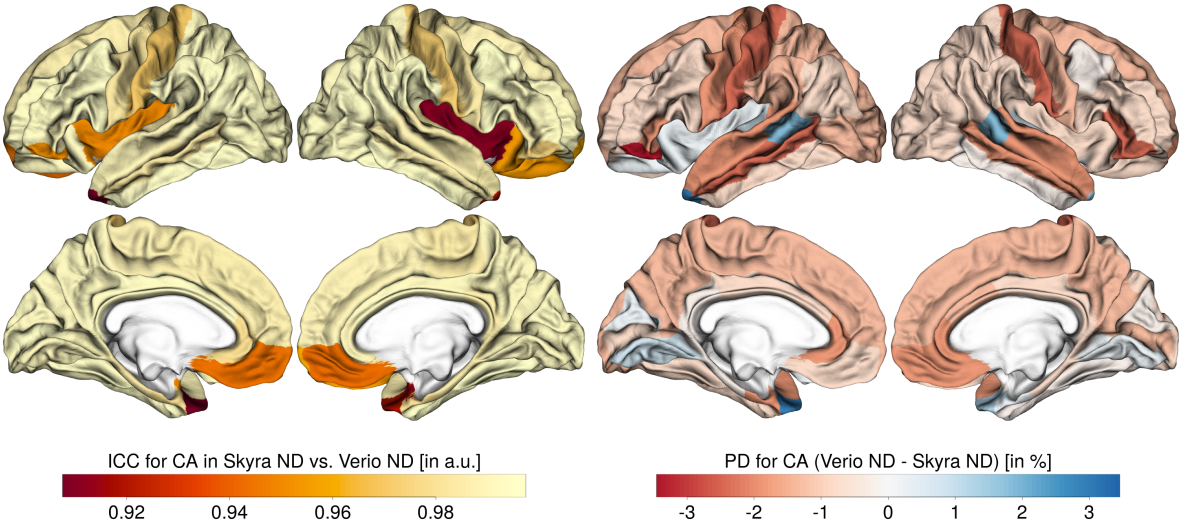

Figure 2: left panel: CA ICC, right panel: CA PD (for each panel, left column shows lateral and medial view of right hemisphere, right column shows lateral and medial view of left hemisphere), negative values:Skyra>Verio, positive values: Verio>Skyra

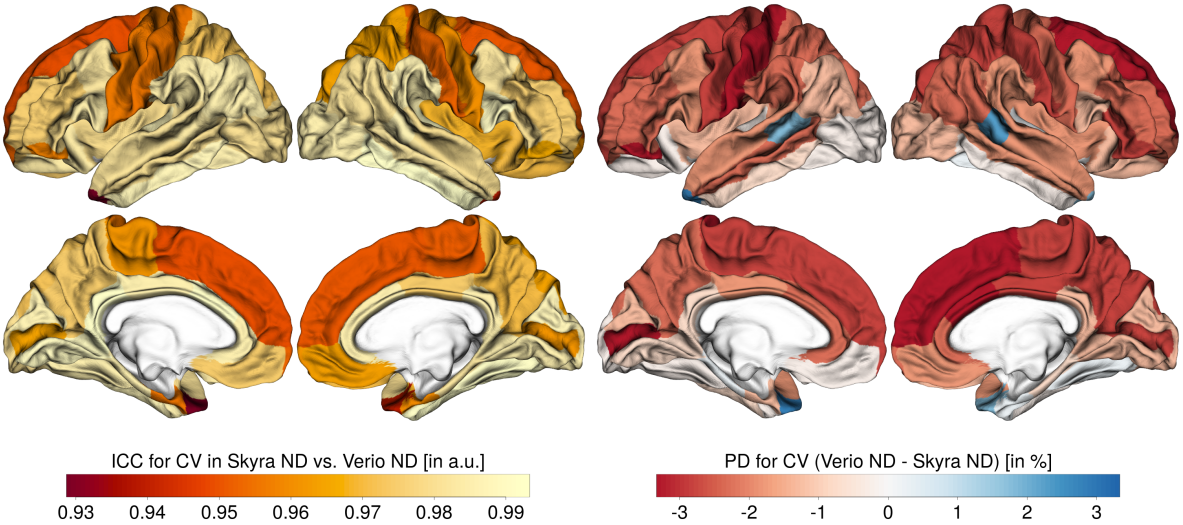

Figure 3: left panel: CV ICC, right panel: CV PD (for each panel, left column shows lateral and medial view of right hemisphere, right column shows lateral and medial view of left hemisphere), negative values:Skyra>Verio, positive values: Verio>Skyra

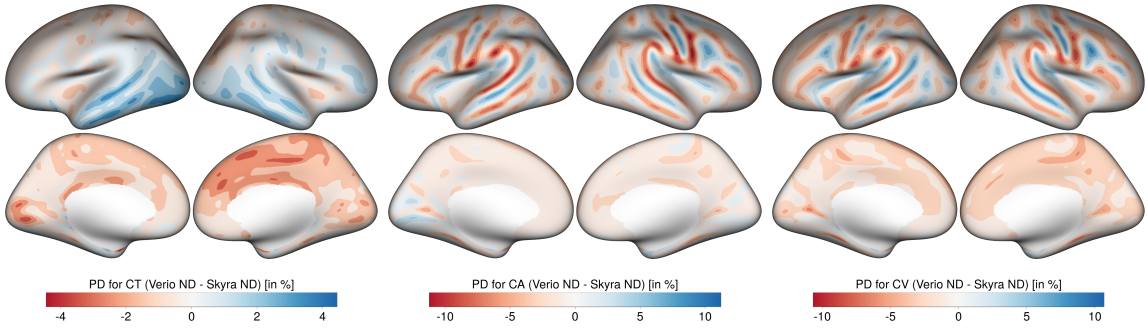

Figure 4: PD for volume, area and thickness shown on inflated fsaverage subject

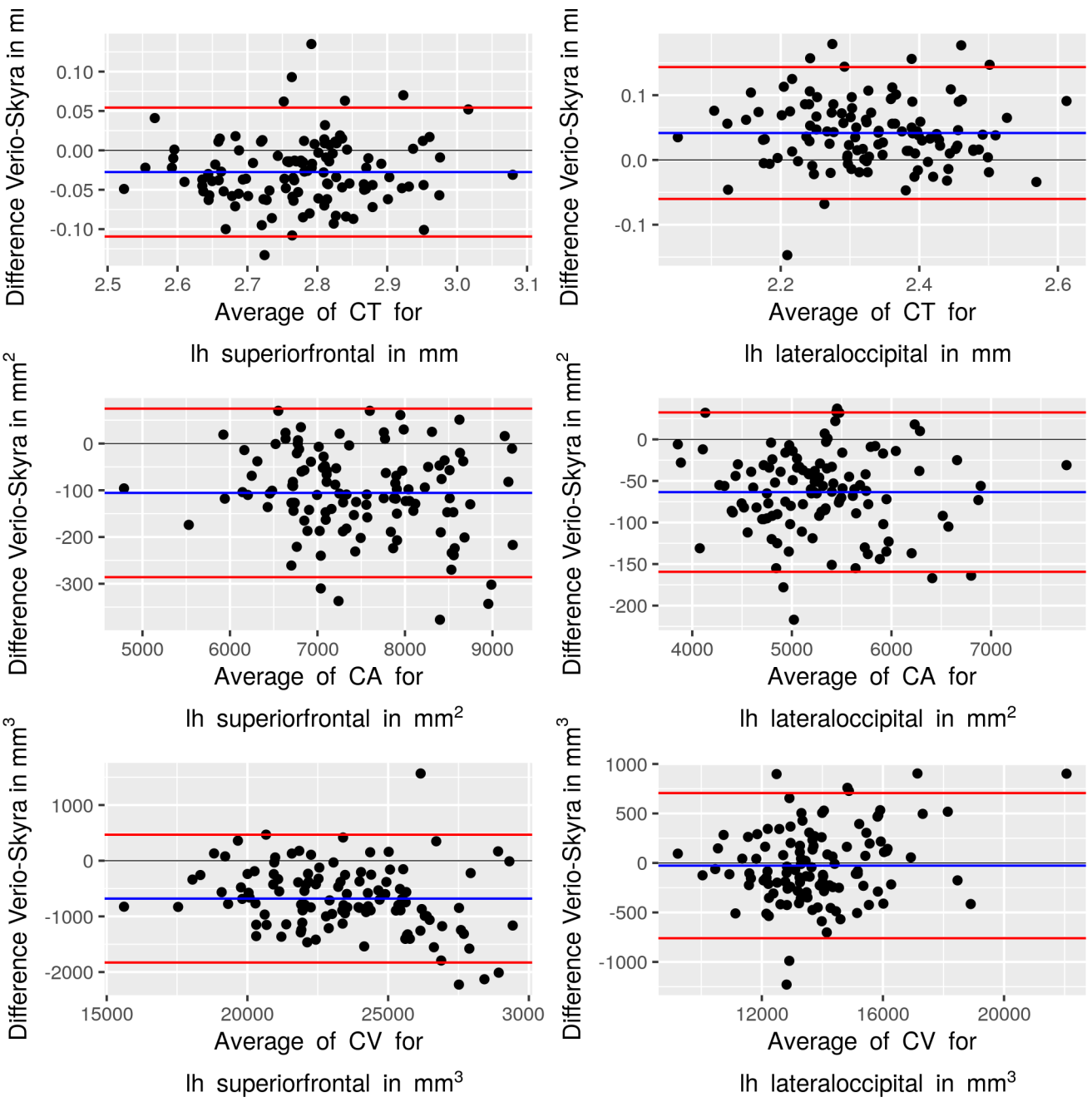

Figure 5: Bland Altman plot showing differences of Verio Skyra against means for superior frontal (left column) and lateral occipital cortex (right column) of the left hemisphere. Top row shows cortical thickness, middle row shows cortical area and bottom row shows cortical volume. Limits of agreement at 95% of standard deviation

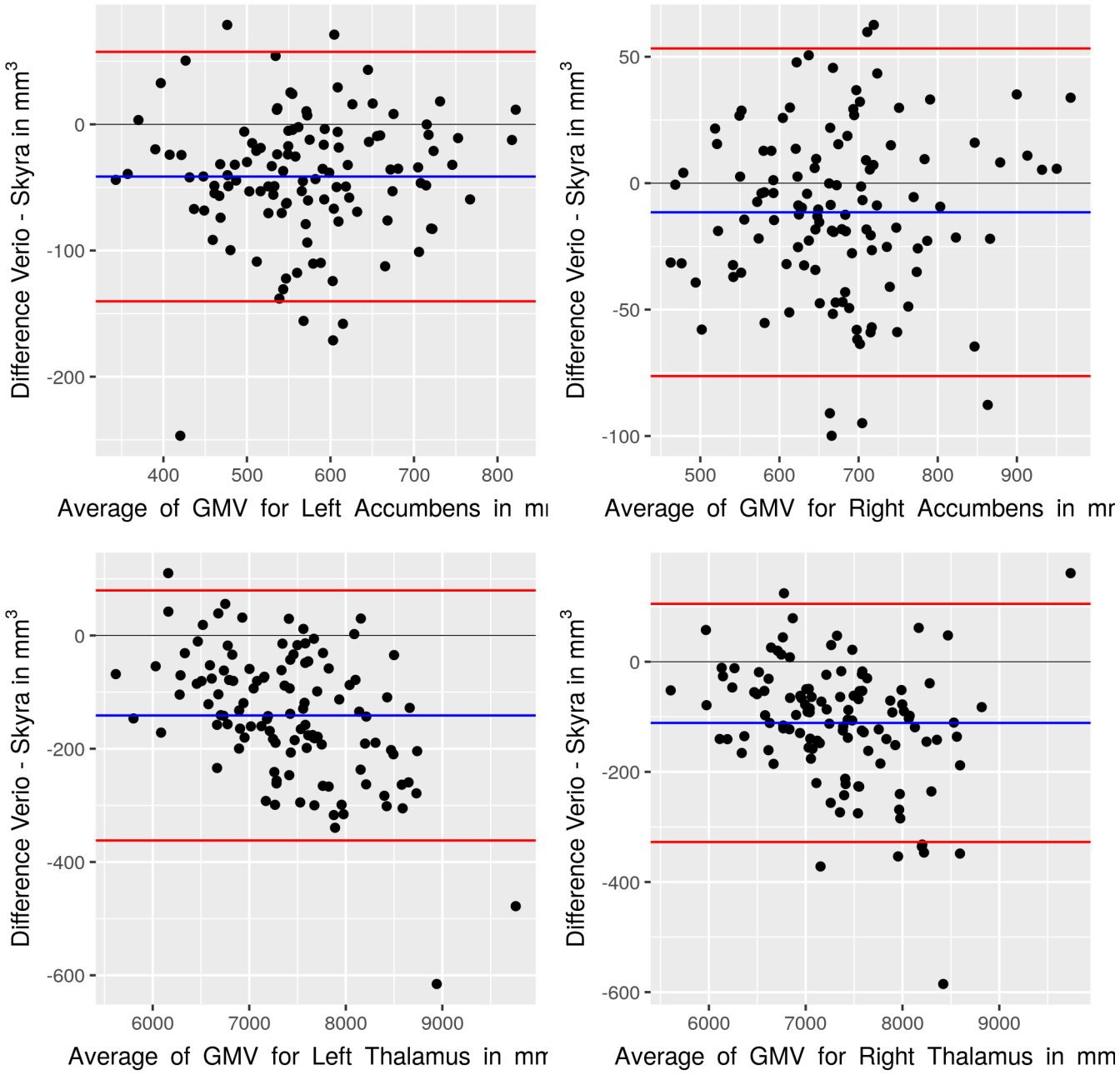

Figure 6: Bland-Altman plot showing differences of Verio-Skyra against means for Accumbens (top row) and Thalamus (bottom row). Limits of agreement at 95% of standard deviation

Table 1: Reliability and percent differences for subcortical volumes

ROI	hemi	ICC	lower ICC	upper ICC	PD	T	p	adj.p
Thalamus	Left	0.97	0.97	0.98	1.77	-11.38	0.00	**0**
Thalamus	Right	0.98	0.98	0.98	1.51	-10.06	0.00	**0**
Caudate	Left	0.99	0.99	0.99	1.43	-9.34	0.00	**0**
Caudate	Right	0.98	0.98	0.98	2.00	-14.23	0.00	**0**
Putamen	Left	0.98	0.97	0.98	1.78	-8.69	0.00	**0**
Putamen	Right	0.99	0.98	0.99	1.34	-6.64	0.00	**0**
Pallidum	Left	0.96	0.96	0.97	2.51	-3.48	0.00	**0**
Pallidum	Right	0.95	0.94	0.96	2.79	-3.34	0.00	**0**
Hippocampus	Left	0.96	0.95	0.97	2.20	-13.54	0.00	**0**
Hippocampus	Right	0.97	0.97	0.98	1.71	-8.03	0.00	**0**
Amygdala	Left	0.93	0.91	0.94	3.19	-0.56	0.57	0.57
Amygdala	Right	0.94	0.93	0.95	2.90	-1.55	0.12	0.13
Accumbens	Left	0.81	0.77	0.83	9.56	-10.25	0.00	**0**
Accumbens	Right	0.94	0.93	0.96	3.96	-4.85	0.00	**0**

Bland-Altman plots for subcortical regions confirmed the systematic bias and further indicated that differences in variability between subjects influenced ICC estimates. For example, there was high between-subject variability in the Thalamus, so that, despite large differences between measurements, ICC was high. Similarly, differences between scanners were less pronounced in the Accumbens, yet, due to lower between-subject variability the ICC of this region was lower (see Figure 6, and supplementary Figure 5). For the Accumbens, 95% of differences between scanners were between -109.21 and 48.74 mm^3, for Thalamus 95% of differences were between -324.2 and 105.46 mm^3.

4.3 QA measures

First, we compared SNR, CNR, CJV and EFC, three quality measures from mriqc between Verio ND and Skyra ND acquisitions. We aimed to determine whether differences in basic signal properties might underlie the observed differences in measures of GM structure.

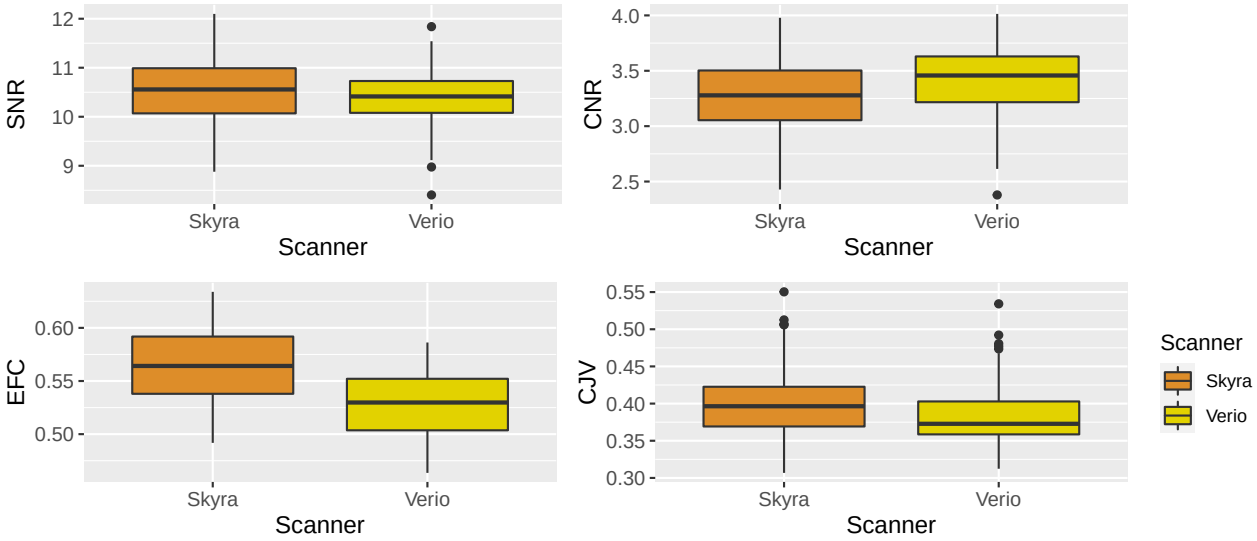

Figure 7: Quality metrics (CNR (left panel), EFC (middle panel) and CJV (right panel)) compared between Skyra ND (orange) and Verio ND (yellow) acquisitions, showing overall higher data quality on the Verio scanner

We found that there was no significant difference in SNR between Verio and Skyra (β=-0.16, p = 0). Yet, Verio ND T1-weighted images had higher CNR (β=0.14, p < 0.001), lower EFC (β= -0.04, p < 0.001) and lower CJV (β=-0.02, p < 0.001) compared to Skyra ND images, also see Figure 7. This indicates higher contrast between WM and GM and less blurring on the Verio scanner. Similar to (Shuter et al. 2008), we investigated whether increased CNR would predict differences in CT. Here, we found that higher CNR across both scanners was associated with higher CT for most regions (see Figure **??**, left panel). Moreover, scanner predicted CT independent of CNR in the same regions as shown above (see Figure **??**, right panel, and Table 4.1 in the Supplementary Material).

Figure 8 shows the association of CNR and CT for two exemplary regions with contrary scanner effects (superior frontal and lateral occipital).

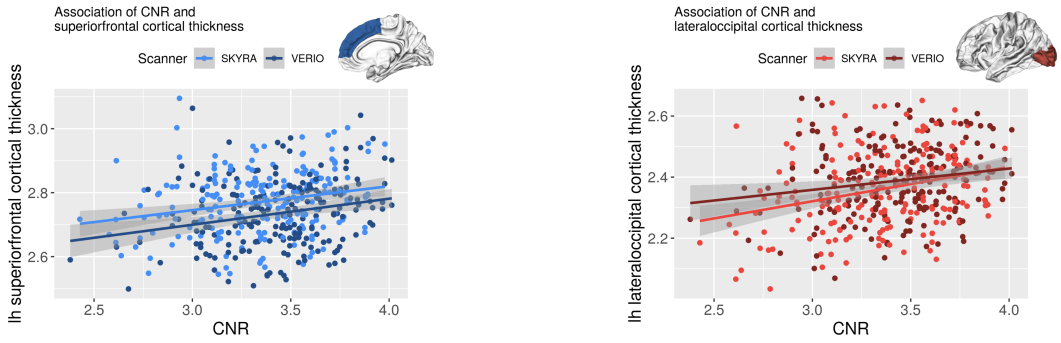

Figure 8: Association of CNR and cortical thickness in left superior frontal and lateral occipital cortex

4.4 Effect of offline gradient distortion correction

We examined whether the differences in cortical and subcortical GM measures arise from the difference in gradient distortion between the two scanners. We corrected both ND files using vendor-provided information on gradient distortions with `gradunwarp`.

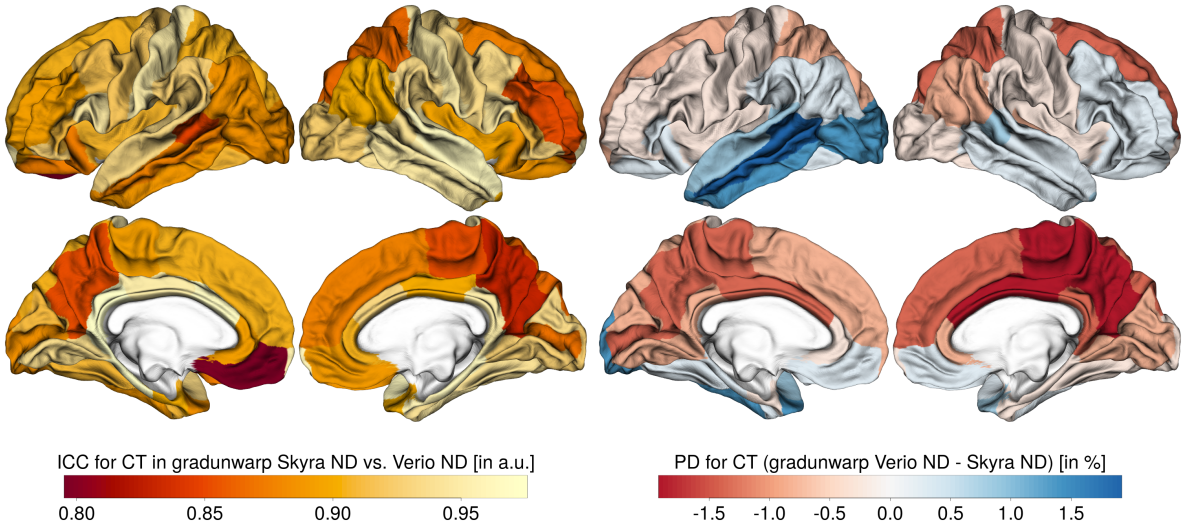

Figure 9: Comparison of CT results from gradunwarp-corrected data. Left panel: CT ICC, right panel: CT PD (for each panel, left column shows lateral and medial view of right hemisphere, right column shows lateral and medial view of left hemisphere), negative values:Skyra>Verio, positive values: Verio>Skyra

Figure 9 shows the results for CT derived from the `gradunwarp` distortion corrected data (also see Table 5.1 in the Supplementary Material). The ICC was excellent throughout all ROI (mean= 0.91, min=0.8,

Table 2: Reliability and percent difference for subcortical volumes from gradient non-linearity corrected data (T<0 reflects Skyra >Verio , T>0 reflects Verio >Skyra)

ROI	hemi	ICC	lower ICC	upper ICC	PD	T	p	adj.p
Thalamus	Left	0.97	0.97	0.98	1.93	-13.33	0.00	**0**
Thalamus	Right	0.98	0.97	0.98	1.64	-10.98	0.00	**0**
Caudate	Left	0.99	0.99	0.99	1.46	-8.34	0.00	**0**
Caudate	Right	0.99	0.99	0.99	1.78	-13.43	0.00	**0**
Putamen	Left	0.98	0.98	0.99	1.63	-7.12	0.00	**0**
Putamen	Right	0.99	0.98	0.99	1.39	-8.57	0.00	**0**
Pallidum	Left	0.97	0.96	0.97	2.49	-2.92	0.00	**0**
Pallidum	Right	0.94	0.92	0.95	2.87	-2.29	0.02	**0.02**
Hippocampus	Left	0.95	0.95	0.96	2.29	-13.95	0.00	**0**
Hippocampus	Right	0.96	0.96	0.97	1.91	-10.35	0.00	**0**
Amygdala	Left	0.91	0.91	0.93	3.58	-2.78	0.01	**0.01**
Amygdala	Right	0.94	0.92	0.95	2.99	-2.93	0.00	**0**
Accumbens	Left	0.81	0.78	0.84	9.13	-8.72	0.00	**0**
Accumbens	Right	0.94	0.93	0.95	4.11	-3.94	0.00	**0**

max=0.98), and as expected, it was higher for the gradient distortion corrected data compared to the previous analysis of Verio ND vs Skyra ND (mean ICC `gradunwarp` Skyra D vs Verio D: 0.914, mean ICC Skyra ND vs Verio ND: 0.906, paired t-test: T = -3.39, p 0.001).
Gradient distortion correction reduced PD to 1-2% (mean=0.25%, min=-1.89%, max=1.79), which was significant compared to ND data (mean PD `gradunwarp` Skyra D vs Verio D: 0.63, mean PD Skyra ND vs Verio ND: 0.72, paired t-test: T = 2.53, p = 0.014). Yet,the inferior-superior pattern of biases remained similar, and there were still significant differences after `gradunwarp` for the majority of regions of interest (FDR-corrected, 62.5% of 64 bilateral cortical ROI). In addition, we saw that the systematic bias was largest when comparing the default scanner output images Skyra D (online gradient distortion correction) and Verio ND (mean ICC `gradunwarp` Skyra D vs Verio D: 0.91, mean ICC Skyra ND vs Verio ND: 0.91, mean ICC Skyra D vs Verio ND: 0.89, ANOVA F-test: F = 3.7, p = 0.026).

Table 2 shows the results for subcortical volumes derived from the gradient distortion corrected data.
The ICC is excellent in all regions, similar to the cortical analysis (mean=0.95, min=0.81, max=0.99). For subcortical volumes, gradient distortion correction did not lead to a further improvement in ICC (mean ICC `gradunwarp` Skyra D vs Verio D = 0.95, mean ICC Skyra ND vs Verio ND = 0.95, paired t-test: T = 1, p= 0.34).
The PD was around 2-3% (mean=2.8%, min=1.39%, max=9.13%) and did not differ from the original analysis (mean PD `gradunwarp` Skyra ND vs Verio ND = 2.8%, mean PD Skyra ND vs Verio ND = 2.76%, paired t-test: T= -0.74, p= 0.47). There were significant differences after `gradunwarp` for all regions of interest (FDR-corrected, 100% of 14 bilateral subcortical regions).

5 Discussion

Summary

In this paper, we aimed to investigate the reliability and bias in GM structure induced by a scanner upgrade in a longitudinal study. We compared outcomes of FreeSurfer's longitudinal pipeline between two different MRI scanners with subsequent versions. We found between-scanner reliability measured with ICC to be excellent. Yet, Bland-Altman plots and paired t-tests revealed statistically significant differences, i.e. biases, in cortical thickness and subcortical GM volumes, as well as for cortical area and volume in a large number of regions. Offline correction for gradient distortions based on vendor-provided gradient information reduced this bias significantly, yet it was not fully removed. T1-imaging based quality measures differed systematically between scanners, also when adjusting for gradient distortions. We conclude that scanner upgrades during a longitudinal study introduce bias in measures of cortical and subcortical grey matter structure and make it difficult to detect true effects when these are subtle like in the case of healthy aging, e.g. ~ 1% annual hippocampal volume loss in older healthy adults (Fraser, Shaw, and Cherbuin 2015). Therefore, before upgrading a MRI system during an ongoing longitudinal study, researchers should prepare to implement an appropriate correction method, such as deriving scaling factors from repeated measures before/after the upgrade or statistical adjustment methods.

Comparison to previous reliability studies

The results of our study are in line with previous findings which have indicated systematic effects of scanner upgrade on GM imaging outcomes (Lee et al. 2019; Han et al. 2006; Jovicich et al. 2009; Brunton et al. 2015).

In recent studies, scanner upgrades induced a significant bias in cortical and subcortical GM measures, while in one study, no systematic bias was reported for hippocampus volume (Potvin et al. 2019; Plitman et al. 2020; Brown et al. 2020). Similar to our findings, ICC values for cortical measures were good to excellent. While the size of biases was comparable to our results (around 1-6% for cortical PD for CV and CT in (Potvin et al. 2019)), the location of the biased regions was different. In (Potvin et al. 2019; Plitman et al. 2020) GM estimates in prefrontal and temporal regions increased with the upgrade, which might be driven by upgrade-related increases in SNR in these regions, which typically show relatively poor within-subject reliability (Liem et al. 2015; Iscan et al. 2015). In our study, we found a medial-frontal to lateral-occipital gradient, with medial-frontal CT as well as subcortical volumes biased towards higher CT and GM volume in Skyra compared to Verio, while lateral-occipital CT was higher in Verio. For CA and CV we saw a gyrification dependent pattern, with higher CA and CV in sulci for Verio compared to Skyra, and higher CA and CV in gyri for Skyra compared to Verio. These pattern also shaped the ICC estimates. Image quality differed between the scanners, yet in contrast to previous studies, we found higher image quality on the (older) Verio scanner (Potvin et al. 2019; Plitman et al. 2020). *Some words why we may see this.* Still, while we found higher CNR to be associated with higher CT, the CT bias pattern was independent of differences in global measures of image quality (Shuter et al. 2008; Potvin et al. 2019).

Similarly, while gradient distortion correction reduced the bias and increased reliability for cortical measures, the overall medial-frontal to lateral-occipital bias in CT and the gyrification-dependent pattern in CA and CV remained similar. Thus, while gradient distortions impact the reliability of GM estimates, they do not fully explain the bias between Skyra and Verio scanner. Instead, we speculate that differences in scaling or signal intensities, e.g. between white and gray matter, have led to the observed differences (Clarkson et al. 2009). This would be compatible with the fronto-medial to occipital-lateral bias pattern (medial and subcortical regions biased toward higher CT and GM volume in Skyra compared to Verio) and the bias following the gyrification in CA. Upon visual inspections of the longitudinal runs (i.e. when both had been registered to a common template), we noticed a subtle expansion of the brain in Skyra compared to Verio for some exemplary subjects. Taken together, we believe the systematic biases between Verio and Skyra stem from both scaling and image intensity differences, and were strongly related to scanner hardware.

While our results certainly overestimate the effects of a real upgrade as discussed above, they still support previous studies on the biasing effects of a scanner upgrade and urge for the use of an adequate correction method if an upgrade becomes necessary during a longitudinal study. One possibility is to measure the same subjects shortly before and after the upgrade and to derive scaling factors like in (Keshavan et al. 2016). Another possibility, which does not require additional data acquisition, is longitudinal ComBat correction, which takes into account biased mean and scaling due to systematic scanner differences (Beer et al. 2020) or

the use of a deep-learning-based harmonisation framework (Dinsdale, Jenkinson, and Namburete 2021).

Limitations

The main limitation of our study is that we did not assess the impact of a true upgrade (i.e. repeated measurements on the same scanner), instead we performed a site-comparison in which the MRI scanners at the two sides were as similar as possible. Another limitation is that we did not randomize the order of participants across scanners and that we could not assess test-retest reliability as we only performed on scan on each system. Yet, previous studies indicated that PD of cortical and subcortical GM are comparable to our results (i.e. PD for subcortical volumes around 2-4 % on Skyra and Verio scanners (Jovicich et al. 2013; Yan et al. 2020)). ICC is a common yet somewhat flawed measure of reliability. ICC does not reflect differences in inter-individual variability, as underlined by Bland-Altman plots of subcortical volumes, and was high in this study even though substantial bias was present.

Strengths

Our study includes around 10 times more participants than previous reliability studies (Potvin et al. 2019; Plitman et al. 2020). This gave us the power to detect small-to-medium systematic differences. For example, Cohen's d of superior-frontal CT difference was 0.33, which would lead to a minimum number of 73 subjects needed to detect this effect with 80% power at $p = 0.05$. In population neuroimaging studies such as LIFE-Adult, we are interested in small effects, which is why it is relevant to assess systematic bias in an adequately powered sample. Another strength of our study is that we applied region-and brain-wide analyses, adjusted for gradient distortions and calculated complementary measures of reliability. Additionally, we present quantitative quality control measures derived from `mriqc`, a state-of-the-art quality control software.

Conclusions

Taken together, in this study, we investigated the impact of a scanner upgrade on longitudinal cortical and subcortical GM measures. We found high reliability but strong regional biases in most regions of interest. While we possibly overestimated the effects of a real upgrade, this study urges for careful monitoring of scanner upgrades and adjustment of biases in longitudinal imaging studies. This may be achieved by deriving scaling factors immediately before/after the upgrade or by using longitudinal batch correction.

References

Bamberg, Fabian, Hans-Ulrich Kauczor, Sabine Weckbach, Christopher L. Schlett, Michael Forsting, Susanne C. Ladd, Karin Halina Greiser, et al. 2015. "Whole-Body Mr Imaging in the German National Cohort: Rationale, Design, and Technical Background." *Radiology* 277 (1): 206–20. https://doi.org/10.1148/radiol.2015142272.

Beer, Joanne C., Nicholas J. Tustison, Philip A. Cook, Christos Davatzikos, Yvette I. Sheline, Russell T. Shinohara, and Kristin A. Linn. 2020. "Longitudinal Combat: A Method for Harmonizing Longitudinal Multi-Scanner Imaging Data." *NeuroImage* 220: 117129. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117129.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal Article. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bland, J Martin, and DouglasG Altman. 1986. "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement." *The Lancet* 327 (8476): 307–10.

Brown, Emma M., Meghan E. Pierce, Dustin C. Clark, Bruce R. Fischl, Juan E. Iglesias, William P. Milberg, Regina E. McGlinchey, and David H. Salat. 2020. "Test-Retest Reliability of Freesurfer Automated Hippocampal Subfield Segmentation Within and Across Scanners." *NeuroImage* 210: 116563. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.116563.

Brunton, Simon, Cerisse Gunasinghe, Nigel Jones, Matthew J. Kempton, Eric Westman, and Andrew Simmons. 2015. "A Voxel-based Morphometry Comparison of the 3.0 T Adni-1 and Adni-2 Volumetric Mri Protocols." Journal Article. *International Journal of Geriatric Psychiatry* 30 (5): 531–38.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The Uk Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9.

Cannon, Tyrone D., Frank Sun, Sarah Jacobson McEwen, Xenophon Papademetris, George He, Theo G. M. van Erp, Aron Jacobson, et al. 2014. "Reliability of Neuroanatomical Measurements in a Multisite Longitudinal Study of Youth at Risk for Psychosis." Journal Article. *Human Brain Mapping* 35 (5): 2424–34. https://doi.org/10.1002/hbm.22338.

Chen, Jiayu, Jingyu Liu, Vince D. Calhoun, Alejandro Arias-Vasquez, Marcel P. Zwiers, Cota Navin Gupta, Barbara Franke, and Jessica A. Turner. 2014. "Exploration of Scanning Effects in Multi-Site Structural Mri Studies." Journal Article. *Journal of Neuroscience Methods* 230: 37–50. https://doi.org/https://doi.org/10.1016/j.jneumeth.2014.04.023.

Cicchetti, Domenic V. 1994. "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment* 6 (4): 284.

Clarkson, Matthew J., Sébastien Ourselin, Casper Nielsen, Kelvin K. Leung, Josephine Barnes, Jennifer L. Whitwell, Jeffrey L. Gunter, et al. 2009. "Comparison of Phantom and Registration Scaling Corrections Using the Adni Cohort." *NeuroImage* 47 (4): 1506–13. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.05.045.

Dinsdale, Nicola K., Mark Jenkinson, and Ana I. L. Namburete. 2021. "Deep Learning-Based Unlearning of Dataset Bias for Mri Harmonisation and Confound Removal." *NeuroImage* 228: 117689. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117689.

Esteban, Oscar, Daniel Birman, Marie Schaer, Oluwasanmi O Koyejo, Russell A Poldrack, and Krzysztof J Gorgolewski. 2017. "MRIQC: Advancing the Automatic Prediction of Image Quality in Mri from Unseen Sites." *PloS One* 12 (9): e0184661.

Ewers, M., S. J. Teipel, O. Dietrich, S. O. Schönberg, F. Jessen, R. Heun, P. Scheltens, L. Van de Pol, N. R. Freymann, and H. J. Moeller. 2006. "Multicenter Assessment of Reliability of Cranial Mri." Journal Article. *Neurobiology of Aging* 27 (8): 1051–9.

Fortin, Jean-Philippe, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, et al. 2018. "Harmonization of Cortical Thickness Measurements Across Scanners and Sites." Journal Article. *NeuroImage* 167: 104–20. https://doi.org/https://doi.org/10.1016/j.neuroimage.2017.11.024.

Fraser, Mark A., Marnie E. Shaw, and Nicolas Cherbuin. 2015. "A Systematic Review and Meta-Analysis of Longitudinal Hippocampal Atrophy in Healthy Human Ageing." *NeuroImage* 112: 364–74. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.03.035.

Glasser, Matthew F, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, et al. 2013. "The Minimal Preprocessing Pipelines for the Human Connectome Project." *Neuroimage* 80: 105–24.

Han, Xiao, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, et al. 2006. "Reliability of Mri-Derived Measurements of Human Cerebral Cortical Thickness: The Effects of Field Strength, Scanner Upgrade and Manufacturer." Journal Article. *NeuroImage* 32 (1): 180–94. https://doi.org/https://doi.org/10.1016/j.neuroimage.2006.02.051.

Ikram, M Arfan, Aad van der Lugt, Wiro J Niessen, Peter J Koudstaal, Gabriel P Krestin, Albert Hofman, Daniel Bos, and Meike W Vernooij. 2015. "The Rotterdam Scan Study: Design Update 2016 and Main Findings." *European Journal of Epidemiology* 30 (12): 1299–1315.

Iscan, Zafer, Tony B. Jin, Alexandria Kendrick, Bryan Szeglin, Hanzhang Lu, Madhukar Trivedi, Maurizio Fava, Patrick J. McGrath, Myrna Weissman, and Benji T. Kurian. 2015. "Test–Retest Reliability of Freesurfer Measurements Within and Between Sites: Effects of Visual Approval Process." Journal Article. *Human Brain Mapping* 36 (9): 3472–85.

Jack, Clifford R., Marilyn S. Albert, David S. Knopman, Guy M. McKhann, Reisa A. Sperling, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. 2011. "Introduction to the Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease." Journal Article. *Alzheimer's & Dementia* 7 (3): 257–62. https://doi.org/https://doi.org/10.1016/j.jalz.2011.03.004.

Jovicich, Jorge, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre van der Kouwe, Randy Gollub, David Kennedy, et al. 2006. "Reliability in Multi-Site Structural Mri Studies: Effects of Gradient Non-Linearity Correction on Phantom and Human Data." Journal Article. *NeuroImage* 30 (2): 436–43. https://doi.org/https://doi.org/10.1016/j.neuroimage.2005.09.046.

Jovicich, Jorge, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, et al. 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." Journal Article. *NeuroImage* 46 (1): 177–92. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.02.010.

Jovicich, Jorge, Moira Marizzoni, Roser Sala-Llonch, Beatriz Bosch, David Bartrés-Faz, Jennifer Arnold, Jens Benninghoff, et al. 2013. "Brain Morphometry Reproducibility in Multi-Center 3T Mri Studies: A Comparison of Cross-Sectional and Longitudinal Segmentations." Journal Article. *NeuroImage* 83: 472–84. https://doi.org/https://doi.org/10.1016/j.neuroimage.2013.05.007.

Keshavan, Anisha, Friedemann Paul, Mona K. Beyer, Alyssa H. Zhu, Nico Papinutto, Russell T. Shinohara, William Stern, et al. 2016. "Power Estimation for Non-Standardized Multisite Studies." Journal Article. *NeuroImage* 134: 281–94. https://doi.org/https://doi.org/10.1016/j.neuroimage.2016.03.051.

Klapwijk, Eduard T., Ferdi Van De Kamp, Mara Van Der Meulen, Sabine Peters, and Lara M. Wierenga. 2019. "Qoala-T: A Supervised-Learning Tool for Quality Control of Freesurfer Segmented Mri Data." Journal Article. *NeuroImage* 189: 116–29.

Lee, Hyunwoo, Kunio Nakamura, Sridar Narayanan, Robert A. Brown, Douglas L. Arnold, and Initiative Alzheimer's Disease Neuroimaging. 2019. "Estimating and Accounting for the Effect of Mri Scanner Changes on Longitudinal Whole-Brain Volume Change Measurements." Journal Article. *NeuroImage* 184: 555–65.

Liem, Franziskus, Susan Mérillat, Ladina Bezzola, Sarah Hirsiger, Michel Philipp, Tara Madhyastha, and Lutz Jäncke. 2015. "Reliability and Statistical Power Analysis of Cortical and Subcortical Freesurfer Metrics in a Large Sample of Healthy Elderly." Journal Article. *Neuroimage* 108: 95–109.

Liljequist, David, Britt Elfving, and Kirsti Skavberg Roaldsen. 2019. "Intraclass Correlation – a Discussion and Demonstration of Basic Features." Journal Article. *PLOS ONE* 14 (7): e0219854. https://doi.org/10.1371/journal.pone.0219854.

Loeffler, M., C. Engel, P. Ahnert, D. Alfermann, K. Arelin, R. Baber, F. Beutner, et al. 2015. "The Life-Adult-Study: Objectives and Design of a Population-Based Cohort Study with 10,000 Deeply Phenotyped Adults in Germany." Journal Article. *BMC Public Health* 15 (1): 691. https://doi.org/10.1186/s12889-015-1983-z.

Plitman, Eric, Aurelie Bussy, Vanessa Valiquette, Alyssa Salaciak, Raihaan Patel, Marie-Lise Béland, Stephanie Tullo, et al. 2020. "The Impact of the Siemens Trio to Prisma Upgrade and Volumetric Navigators on Mri Indices: A Reliability Study with Implications for Longitudinal Study Designs." *bioRxiv*.

Potvin, Olivier, April Khademi, Isabelle Chouinard, Farnaz Farokhian, Louis Dieumegarde, Ilana Leppert, Rick Hoge, et al. 2019. "Measurement Variability Following Mri System Upgrade." *Frontiers in Neurology* 10: 726. https://doi.org/10.3389/fneur.2019.00726.

Preboske, Gregory M., Jeff L. Gunter, Chadwick P. Ward, and Clifford R. Jack. 2006. "Common Mri Acquisition Non-Idealities Significantly Impact the Output of the Boundary Shift Integral Method of Measuring Brain Atrophy on Serial Mri." *NeuroImage* 30 (4): 1196–1202. https://doi.org/https://doi.org/10.1016/j.neuroimage.2005.10.049.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reuter, Martin, and Bruce Fischl. 2011. "Avoiding Asymmetry-Induced Bias in Longitudinal Image Processing." *NeuroImage* 57 (1): 19–21. https://doi.org/https://doi.org/10.1016/j.neuroimage.2011.02.076.

Reuter, Martin, H. Diana Rosas, and Bruce Fischl. 2010. "Highly Accurate Inverse Consistent Registration: A Robust Approach." *NeuroImage* 53 (4): 1181–96. https://doi.org/https://doi.org/10.1016/j.neuroimage.2010.07.020.

Reuter, Martin, Nicholas J. Schmansky, H. Diana Rosas, and Bruce Fischl. 2012. "Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis." Journal Article. *NeuroImage* 61 (4): 1402–18. https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.02.084.

Schaefer, Tim, and Christine Ecker. 2020. "Fsbrain: An R Package for the Visualization of Structural Neuroimaging Data." https://doi.org/10.1101/2020.09.18.302935.

Shrout, Patrick E, and Joseph L Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86 (2): 420.

Shuter, Borys, Ing Berne Yeh, Steven Graham, Chris Au, and Shih-Chang Wang. 2008. "Reproducibility of Brain Tissue Volumes in Longitudinal Studies: Effects of Changes in Signal-to-Noise Ratio and Scanner Software." *NeuroImage* 41 (2): 371–79. https://doi.org/https://doi.org/10.1016/j.neuroimage.2008.02.003.

Takao, Hidemasa, Osamu Abe, Naoto Hayashi, Hiroyuki Kabasawa, and Kuni Ohtomo. 2010. "Effects of Gradient Non-Linearity Correction and Intensity Non-Uniformity Correction in Longitudinal Studies Using Structural Image Evaluation Using Normalization of Atrophy (Siena)." *Journal of Magnetic Resonance Imaging* 32 (2): 489–92. https://doi.org/10.1002/jmri.22237.

Takao, Hidemasa, Naoto Hayashi, and Kuni Ohtomo. 2013. "Effects of the Use of Multiple Scanners and of Scanner Upgrade in Longitudinal Voxel-Based Morphometry Studies." *Journal of Magnetic Resonance Imaging* 38 (5): 1283–91. https://doi.org/10.1002/jmri.24038.

Yan, Shuang, Tianyi Qian, Bénédicte Maréchal, Tobias Kober, Xianchang Zhang, Jinxia Zhu, Jing Lei, Mingli Li, and Zhengyu Jin. 2020. "Test-Retest Variability of Brain Morphometry Analysis: An Investigation of Sequence and Coil Effects." *Annals of Translational Medicine* 8 (1).