Test-retest reliability and sample size estimates after MRI scanner relocation

Tracy R. Melzer, Ross J. Keenan, Stephen Kingston-Smith, Simon A. Felton, Sarah K. Green, Kerry J. Henderson, Nickolas J. Palmer, Reza Shoorangiz, Mustafa M. Almuqbel, Daniel J. Myall

Objective: Many factors can contribute to the reliability and robustness of MRI-derived metrics. In this study, we assessed the reliability and reproducibility of three MRI modalities after an MRI scanner was relocated to a new hospital facility.

Methods: Twenty healthy volunteers (12 females, mean age (standard deviation) = 41 (11) years, age range [25–66]) completed three MRI sessions. The first session (S1) was one week prior to the 3T GE HDxt scanner relocation. The second (S2) occurred nine weeks after S1 and at the new location; a third session (S3) was acquired 4 weeks after S2. At each session, we acquired structural T1-weighted, pseudo-continuous arterial spin labelled, and diffusion tensor imaging sequences. We used longitudinal processing streams to create 12 summary MRI metrics, including total gray matter (GM), cortical GM, subcortical GM, white matter (WM), and lateral ventricle volume; mean cortical thickness; total surface area; average gray matter perfusion, and average diffusion tensor metrics along principal white matter pathways. We compared mean MRI values and variance at the old scanner location to multiple sessions at the new location using Bayesian multi-level regression models. K-fold cross validation allowed identification of important predictors. Whole-brain analyses were used to investigate any regional differences. Furthermore, we calculated within-subject coefficient of variation (wsCV), intraclass correlation coefficient (ICC), and dice similarity index (SI) of cortical segmentations across scanner relocation and within-site. Additionally, we estimated sample sizes required to robustly detect a 4% difference between two groups across MRI metrics.

Results: All global MRI metrics exhibited little mean difference and small variability (bar cortical grey matter perfusion) both across scanner relocation and within-site repeat. T1- and DTI-derived tissue metrics showed < 0.3% mean difference and <1.2% variance across scanner location and <0.4% mean difference and <0.8% variance within the new location, with between-scan intraclass correlation coefficient (ICC) > 0.80 and within-subject coefficient of variation (wsCV) < 1.4%. Mean cortical grey matter perfusion had the highest between-session variability (6.7% [0.3, 16.7]), estimate [95% uncertainty interval]), and hence the smallest ICC (0.71 [0.44,0.92]) and largest wsCV (13.4% [5.4, 18.1]). No global metric exhibited evidence of a meaningful mean difference between scanner locations. However, surface area showed evidence of a mean difference within-site repeat (between S2 and S3). Whole-brain analyses revealed no significant areas of difference between scanner relocation or within-site. For all metrics, we found no support for a systematic difference in variance across relocation sites compared to within-site test-retest reliability. Necessary sample sizes to detect a 4% difference between two independent groups varied from a maximum of n = 362 per group (cortical grey matter perfusion),
1. Introduction

The importance and interest in magnetic resonance imaging (MRI) techniques to glean information about the state of the brain in health and disease cannot be understated. However, in order to provide useful information for researchers and clinicians, it is imperative that MRI-derived metrics are accurate and reproducible. Encouragingly, measures of cortical thickness and gray matter volume derived from T1-weighted MRI, resting arterial spin labeling (ASL) perfusion imaging, and diffusion tensor imaging (DTI) metrics generally report good reproducibility and reliability (Boekel et al., 2017; Dickerson et al., 2008; Hodkinson et al., 2013; Iscan et al., 2015; Madan and Kensinger, 2017; Madhyastha et al., 2014; Shahim et al., 2017; Saali et al., 2016; Vollmar et al., 2010).

Specifically, previous studies of gray matter volume, cortical thickness, and surface area have demonstrated good to excellent reliability and reproducibility (high intraclass correlation coefficients (ICC) > 0.8) and small mean differences between sessions (Dickerson et al., 2008; Eggert et al., 2012; Iscan et al., 2015; Jovicich et al., 2009, 2013; Madan and Kensinger, 2017; Wonderlick et al., 2009). This consistently high reliability is reported across a number of scenarios, including different pulse sequences (Jovicich et al., 2009, 2013), acquisition parameters (Wonderlick et al., 2009), and scanner software upgrades (Jovicich et al., 2009). Different vendors, magnetic field strengths, and software version can result in a difference in volume estimates (Chen Zhiye and Lin, 2017; Lee et al., 2019), but the variability appears comparable to within-subject measures (Jovicich et al., 2009).

Similarly, resting ASL studies (both pulsed and continuous ASL) have shown good reliability over time and consistency across centers (Gevers et al., 2011; Jiang et al., 2011; Petersen et al., 2010; Sousa et al., 2014). Pseudo-continuous ASL (PCASL) sequences display good reliability (Gevers et al., 2011; Saali et al., 2016), agreement with \(^{15}\)O-water PET (Xu et al., 2010), and comparable perfusion measurements across manufacturers, although small differences in sequence parameters can have large effects on reproducibility (Mutsaerts et al., 2014).

Standard diffusion tensor imaging (DTI) also exhibits relative stability, over both hours and weeks (Boekel et al., 2017; Shahim et al., 2017). As expected, variability of DTI metrics is smaller within scanner than across different scanners (Magnotta et al., 2012; Palacios et al., 2017; Vollmar et al., 2010), and different processing pipelines have a large influence on DTI metrics (Madhyastha et al., 2014; Vollmar et al., 2010). However, consistent processing approaches, such as Tract-based spatial statistics (TBSS) (Smith et al., 2006), have shown good test-retest reliability across standard DTI metrics along the white matter skeleton (ICCs up to 0.8 (Madhyastha et al., 2014) and intra-site coefficient of variation < 0.73% (Palacios et al., 2017)).

While previous literature suggests that structural MRI, PCASL, and DTI metrics are reliable and reproducible across varied circumstances, to the best of our knowledge, there is no published information about the effect of scanner relocation on these sequences. Recently, the MRI scanner we utilize in a number of ongoing research studies was relocated from a private radiology clinic to a new hospital. This involved completely ramping down the scanner magnet, removal from one shielded MRI suite, transport across the city to the new facility, installation in the new MRI environment, re-establishing the superconducting magnet, and a physical re-shim by the vendor. While no hardware or software was changed during the move, we were unsure of the influence the move and new scanning environment would have on an established neuroimaging protocol. Thus, in this study we assessed the reliability and reproducibility of three MRI modalities (T1-weighted, PCASL, and DTI) across scanner relocation. Specifically, we compared both global and regional MRI values and variability at the old scanner location to multiple sessions at the new location. Furthermore, using the variability observed in these global MRI metrics, we estimated sample sizes needed to robustly detect a 4% difference between two groups across the various MRI metrics.

2. Material and methods

Twenty healthy volunteers (12 females, mean age (standard deviation) = 41 (11) years, age range [25–66]) provided written informed consent. Participants completed three MRI sessions. The first session (S1) occurred one week prior to scanner relocation; the second (S2) occurred nine weeks after S1 and at the new location, while the third session (S3) occurred 4 weeks after S2. This design allowed us to investigate variability associated with moving the scanner (S1–S2), as well as standard between-session test-retest reliability (S2–S3).

2.1. Magnetic resonance imaging acquisition

Imaging was conducted on a 3T General Electric HDxt scanner (GE Healthcare, Waukesha, USA) with an eight-channel head coil. The following images were acquired at each session (S1, S2, and S3), per previous work (Melzer et al., 2011, 2012, 2013, 2015).

2.1.1. Conventional structural

Images were acquired with a volumetric T1-weighted 3D inversion-prepared spoiled gradient recalled echo (SPGR) acquisition (TE/TR = 2.8/6.6 ms, TI = 400 ms, flip angle = 15 deg, acquisition matrix = 256 × 256 × 170, FOV = 250 mm, slice thickness = 1 mm, voxel size = 0.98 × 0.98 × 1.0 mm\(^3\)).

2.1.2. Pseudo-continuous arterial spin labelling

A stack of spiral, fast spin echo acquired images were prepared with pseudo-continuous arterial spin labelling and background suppression to measure whole brain perfusion quantitatively (Dai et al., 2008): TR = 6 s, echo spacing = 9.2 ms, post-labeling delay = 1.525 s, labelling duration = 1.5 s, eight interleaved spiral arms with 512 samples at 62.5 kHz bandwidth and 30 phase encoded 5 mm thick slices, NEX = 5. Participants were asked to close their eyes during the PCASL acquisition.

2.1.3. Diffusion tensor imaging

A 2D diffusion-weighted, spin echo, echo planar imaging sequence was used to measure white matter microstructural integrity, with diffusion weighting in 28 uniformly distributed directions (b = 1000 s/mm\(^2\)) and 4 acquisitions without diffusion weighting (b = 0 s/mm\(^2\)): TE/TR = 86.4/13000 ms, flip angle = 90 deg, acquisition matrix = 128 × 128 × 48, FOV = 240 mm, slice thickness = 3 mm, reconstructed voxel size = 1.88 × 1.88 × 3.0 mm\(^3\), NEX = 1, ungated.

to total gray matter volume (n = 114), average fractional anisotropy (n = 23), total gray matter volume normalized by intracranial volume (n = 19), and axial diffusivity (n = 3 per group).

Conclusion: Cortical gray matter perfusion was the most variable metric investigated (necessitating large sample sizes to identify group differences), with other metrics showing substantially less variability. Scanner relocation appeared to have a negligible effect on variability of the global MRI metrics tested. This manuscript reports within-site test-retest variability as a tool for calculating sample size in future investigations. Our results suggest that when all other parameters are held constant (e.g., sequence parameters and MRI processing), the effect of scanner relocation is indistinguishable from within-site variability, but may need to be considered depending on the question being investigated.
2.2. MRI processing

2.2.1. Structural MRI processing

To extract estimates of cortical thickness, surface area, and volume, T1-weighted images were processed using the longitudinal stream in FreeSurfer (v6.0.0; http://surfer.nmr.mgh.harvard.edu/) (Reuter et al., 2012). An unbiased within-subject template space and image (Reuter and Fischl, 2011) were created using robust, inverse consistent registration (Reuter et al., 2010). Further processing included skull stripping, Talairach transforms, atlas registration as well as spherical surface maps and parcellations, which were initialized with common information from the within-subject template. We extracted global metrics (mean cortical thickness, total surface area, total gray matter (GM) volume, cortical and subcortical GM volume, total white matter (WM) volume, and lateral ventricle volume) in each participant at each timepoint; regional thickness and surface area were extracted from the Desikan-Killiany FreeSurfer parcellations at each timepoint (Desikan et al., 2006). To facilitate whole-brain investigation of regional test-retest reliability, we created difference maps normalized to S2 ((S2−S1)/S2)—between site difference; and (S2−S3)/S2—within site difference—which were warped to fsaverage space and smoothed with a circularly symmetric Gaussian kernel across the surface with a full width at half maximum of 10 mm. Hippocampal volume was calculated using the longitudinal hippocampal subfield segmentation stream (Iglesias et al., 2016).

In addition to the surface-based FreeSurfer processing, we also calculated global T1 and PCASL-derived metrics using a volumetric, voxel-based approach using CAT12 (r934, http://www.neuro.uni-jena.de/cat/), a toolbox of SPM12 (v7912, http://www.fil.ion.ucl.ac.uk/spm/); see Supplementary Methods.

2.2.2. PCASL processing

At each timepoint, quantified cerebral blood perfusion images were co-registered to the structural image using FreeSurfer’s bregister (with default values). These registration parameters were combined with the structural longitudinal warping parameters to register the perfusion image to the within-subject template. To account for any remaining misalignment, perfusion images were smoothed (using FSL, with sigma = 2). Global, regional cortical (from Desikan-Killiany parcellation after projection onto the cortical surface), and subcortical perfusion values were extracted for each subject at each timepoint. We also created normalized difference maps ((S2−S1)/S2 and (S2−S3)/S2) that were warped to fsaverage space.

2.2.3. DTI processing

Preprocessing was performed similarly to previous longitudinal DTI studies using FSL (v5.0.9) (Engvig et al., 2012; Madhyastha et al., 2014; Melzer et al., 2015). At each timepoint, this included motion- and eddy current distortion-correction; rotation of the b matrix accordingly; motion quantification via root mean square deviation between each pair of realigned diffusion images and averaging over all pairs to create a single, ‘relative’ motion metric; smoothing (using fslmaths with a 1 voxel box kernel and -median flag), which has been shown to increase reliability (Madhyastha et al., 2014); brain extraction; and fitting a diffusion tensor to produce fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD, the principal diffusion eigenvalue), and radial diffusivity (RD, the mean of the second and third eigenvalues) images. For each participant, we used mrtrixis robust template (Reuter et al., 2012) (part of FSL) to create an unbiased, within-subject FA median template as well as robustly registering FA/MD/AD/RD images from S1, S2, and S3 to this within-subject template. The within-subject FA template was then entered into a tract-based spatial statistics (TBSS; Smith et al., 2006) analysis to create the group-wise FA skeleton (thinned at FA > 0.25), representing the centers of all tracts common to all participants. Midpoint-space registered FA images (S1, S2, and S3) were projected onto the skeleton to create FA skeletons containing S1, S2, and S3 values. Using the transforms derived from the FA procedure, we created separate MD, AD, and RD skeletons at each timepoint. Average DTI metrics (FA/MD/AD/RD) along the skeleton, and regionally within 17 white matter tracts (defined by the JHU ICBM-DTI-81 White-Matter Labels atlas (Mori et al., 2008)), were extracted for each participant at each timepoint. Lastly, we created normalized difference maps along the skeleton ((S2−S1)/S2 and (S2−S3)/S2) for each DTI metric.

2.3. Statistical analysis

2.3.1. Global MRI metric analysis

We assessed each of 12 global metrics independently for reliability; total GM, cortical GM, subcortical GM, total WM, and lateral ventricle volume; average cortical thickness and total surface area; average cortical perfusion; and average FA, MD, AD, and RD along the skeleton. Bayesian models were used for analysis and were fit using the ‘brms’ (v2.9.0) package (Buerkner, 2017; Carpenter et al., 2017) in R (v3.6.2). In each model, four chains with 10000 iterations each were used to generate posterior samples. In the baseline model an intercept, varying intercept by subject, session (referenced to S2, the first scan at the new location), and between-session variance by subject were included. To determine if allowing between-session variance to differ between the scanner relocation and within-site repeat was useful at explaining variance in the data, we performed model comparison using k-fold cross validation to estimate the expected log pointwise predictive density (ELPD) (Vehtari et al., 2017). The initial model fit (which included a constant variance across all sessions) was compared to a second model in which variance was allowed to vary between the scanner locations. A higher ELPD indicated the model provided a better fit to the data. The standard error in the difference of the ELPD between models gave a measure of the uncertainty. When the difference in ELPD was increased by at least twice the standard error of the estimated difference, we took this as reasonable evidence that the better-performing model should be preferred. ELPD values are reported relative to the initial model (where variance across scanner location was held constant). In order to account for the multiple models and comparisons used to investigate the 12 global metrics, we included a student t prior to shrink estimates in all models, thereby reducing the potential multiple comparisons issue (degrees of freedom = 3, mean = 0, standard deviation = 0.5% of the average of the specific MRI metric at S2).

2.3.2. Regional analysis

We investigated regional test-retest reliability in two ways. First, we plotted regional thickness, surface area, perfusion, and all diffusion metrics to visualize variability across scanner relocation and within-site repeat. Secondly, we used one sample t-tests of difference images to investigate any significant differences between (S2−S1) or within-site (S2−S3) repeat. For all metrics, we investigated four contrasts (S1−S2, S1−S2, S3−S2, and S3−S2). For thickness, surface area, and perfusion, we used FreeSurfer’s permutation-based cluster-wise correction for multiple comparisons (5000 permutations, cluster forming threshold of p < 0.0001, corrected p < 0.00025 (Greve and Fischl, 2018). All voxel-wise DTI comparisons were performed using a permutation-based inference tool for nonparametric thresholding (FSL’s “randomise” (Winkler et al., 2014)). For each contrast, the null distribution was generated over 5000 permutations and the alpha level set at p < 0.00025, corrected for multiple comparisons (family-wise error correction using threshold-free cluster-enhancement (Smith and Nichols, 2009)). Our corrected threshold (p < 0.00025) was our original cluster-corrected threshold (p < 0.01) Bonferroni-corrected for the number of independent vertex/voxel-wise comparisons performed ((3 metrics (CTh, SA, CBF) × 2 hemispheres × 4 contrasts) + 4 DTI metrics × 4 contrasts) = 40).

2.3.3. Additional test-retest metrics

To further investigate reliability, we also calculated within-subject coefficient of variation (wsCV) (Bland and Altman, 1996), the
intraclass correlation coefficient (ICC) (McGraw and Wong, 1996), and the dice similarity index (SI) (Dice, 1945). SI was used to quantify the similarity between cortical segmentations across imaging sessions. For ICC, we used a two-way mixed effects model with absolute agreement (participant was modelled as a random effect and scanner as a fixed effect), using the R packages brms and sjstats (v0.16.0, https://doi.org/10.5281/zenodo.1284472).

2.4. Sample size estimates

Based on the observed within-site variance, we performed frequentist sample size calculations with 80% power at alpha = 0.05 to detect a 4% difference in all global MRI metrics; sample sizes were also calculated for global volumes normalized by ICV. Sample size was calculated for a difference between two independent groups, assuming equal group sizes, allowing 5% type I error, 2-tailed significant test p < 0.05 (Mak et al., 2017).

2.5. Data availability

Analysis code and data are available at https://github.com/nzbri/scanner-test-retest.

3. Results

One of 20 participants (12 females, mean age (standard deviation) = 41 (11) years, age range [25–66]) did not complete S3, leaving a total of 59 imaging sessions. No structural or PCASL images were excluded, however three DTI datasets (from two subjects) were excluded due to excessive motion (defined as mean relative motion greater than three times the standard deviation of the group).

We observed no evidence of a difference in mean MRI metrics between scanner locations, however there was evidence that SA varied within-site (from S2 to S3). Fig. 1 and Table 1 present global MRI metrics across sessions. Across all MRI metrics, there was no indication that allowing between-session variance to be different between scanner locations improved the model fit (Table 2). That is, we found no evidence of a systematic difference in variance pre- or post-relocation. Furthermore, ICC was high and within-subject coefficient of variation was low across all metrics (Table 1).

3.1. T1-weighted results

3.1.1. Global results

All T1-derived metrics demonstrated excellent reproducibility and reliability across scanner relocation and within-site repeat (Table 1, Fig. 1). Cortical segmentation was very reliable; dice similarity indices (SI) were high ([S3–S2]: left hemisphere [95% uncertainty interval] = 0.937 [0.933, 0.942]; right hemisphere = 0.934 [0.929, 0.940]. [S1–S2]: left hemisphere = 0.936 [0.928, 0.944]; right hemisphere = 0.933 [0.923, 0.943]), with no evidence of a difference in SI between cortical segments at S1 or S3 relative to S2. While we estimated a 0.0% change [-0.3, 0.3] in SA between scanner relocation, we did observe a 0.3% [0.1, 0.6] increase in SA across the within-site repeat (S3>S2).

3.1.2. Regional results

Regional CTh and SA from Desikan-Killiany cortical parcellations across sessions are presented in Supplementary Figs. 1 and 2, respectively; subcortical volumes are presented in Supplementary Fig. 3. Vertex-wise, we identified no significant differences in CTh or SA between or within-site repeat.

![Fig. 1. Relative MRI metrics between sessions.](image_url)

MRI metrics for all participants are displayed relative to session 2 (S2, the first session at the new location). All individual trajectories (n = 20) are displayed in gray, while the mean for each session is overlaid in color. Each panel displays one of the 12 MRI metrics investigated. CBF demonstrates the largest variability, while the other 11 metrics demonstrate robust consistency across relocation and within-site test-retest reliability. S1 = session 1 (scanner at old location), S2 = session 2 (first scan at scanner’s new location), S3 = session 3 (second scan at scanner’s new location), Cortical Vol = total cortical gray matter volume, SubCort Vol = total subcortical gray matter volume, GM Vol = total gray matter volume, WM Vol = total white matter volume, Vent Vol = total lateral ventricular volume, CTh = average cortical thickness, SA = total surface area, CBF = average cortical gray matter cerebral blood flow (perfusion), FA = average fractional anisotropy along principal white matter tracts, MD = average mean diffusivity along principal white matter tracts, AD = average axial diffusivity along principal white matter tracts, RD = average radial diffusivity along principal white matter tracts.
Table 1: Reliability and reproducibility across global MRI metrics.

| Metric                        | Population Estimate | Between-scene variability ICC | Between-scene variability wsCV |
|-------------------------------|----------------------|-------------------------------|--------------------------------|
| Total GM mean (cm³)           | 672 [641, 703]       | –                             | –                              |
| PreMove                       | –0.1% [-0.6, 0.5]    | 1.2% [0.1, 2.2]               | 0.93 [0.85, 0.99]              |
| Repeat                        | +0.2% [0.3, 0.6]     | 0.5% [0.0, 1.2]               | 0.97 [0.93, 0.99]              |
| Cortical GM mean (cm³)        | 495 [471, 520]       | –                             | –                              |
| PreMove                       | –0.0% [-0.6, 0.6]    | 0.8% [0.0, 1.9]               | 0.93 [0.86, 0.99]              |
| Repeat                        | –0.4% [-0.2, 1.0]    | 0.5% [0.0, 1.6]               | 0.95 [0.89, 0.99]              |
| Subcortical GM mean (cm³)     | 60.1 [58.2, 61.9]    | –                             | –                              |
| PreMove                       | –0.3% [-0.7, 0.1]    | 0.4% [0.0, 1.0]               | 0.99 [0.98, 1.0]               |
| Repeat                        | +0.1% [-0.3, 0.5]    | 0.4% [0.0, 1.0]               | 0.99 [0.98, 1.0]               |
| WM mean (cm³)                 | 463 [439, 487]       | –                             | –                              |
| PreMove                       | –0.2% [-0.9, 0.4]    | 1.0% [0.1, 1.8]               | 0.98 [0.95, 1.0]               |
| Repeat                        | +0.1% [-0.3, 0.5]    | 0.5% [0.0, 1.0]               | 0.99 [0.98, 1.0]               |
| Lateral ventricle mean (cm³)  | 16.7 [12.9, 20.5]    | –                             | –                              |
| PreMove                       | +0.5% [-0.9, 2.0]    | 2.4% [0.1, 5.6]               | 1.0 [0.99, 1.0]                |
| Repeat                        | –0.0% [-1.3, 1.3]    | 1.7% [0.1, 4.0]               | 1.0 [0.99, 1.0]                |
| Cortical thickness mean (mm)  | 2.61 [2.57, 2.64]    | –                             | –                              |
| PreMove                       | –0.1% [-0.6, 0.5]    | 0.6% [0.0, 1.6]               | 0.89 [0.77, 0.98]              |
| Repeat                        | +0.1% [-0.4, 0.6]    | 0.5% [0.0, 1.3]               | 0.91 [0.80, 0.98]              |
| Surface area (cm²)            | 848 [807, 889]       | –                             | –                              |
| PreMove                       | +0.0% [-0.3, 0.3]    | 0.6% [0.1, 1.0]               | 0.98 [0.95, 0.99]              |
| Repeat                        | +0.3% [0.1, 0.6]     | 0.2% [0.0, 0.6]               | 0.99 [0.98, 1.0]               |
| CBF mean (ml/100 g/min)       | 51.6 [47.2, 56.0]    | –                             | –                              |
| PreMove                       | +0.3% [-1.0, 2.2]    | 6.9% [0.3, 16.7]              | 0.71 [0.44, 0.92]              |
| Repeat                        | –0.1% [-1.7, 1.3]    | 4.3% [0.2, 12.2]              | 0.81 [0.63, 0.95]              |

Table 1 (continued)

| Metric                        | Population Estimate | Between-scene variability ICC | Between-scene variability wsCV |
|-------------------------------|----------------------|-------------------------------|--------------------------------|
| FA mean (× 10^-5)            | 42.7 [41.8, 43.7]    | –                             | –                              |
| PreMove                       | +0.2% [-0.2, 0.6]    | 0.3% [0.0, 1.1]               | 0.98 [0.94, 1.0]               |
| Repeat                        | –0.0%               | 0.6% [0.0, 1.5]               | 0.95 [0.91, 0.99]              |
| MD mean (mm²/s × 10^-5)       | 81.3 [80.1, 82.0]    | –                             | –                              |
| PreMove                       | –0.1% [-0.6, 0.3]    | 0.5% [0.0, 1.4]               | 0.80 [0.63, 0.95]              |
| Repeat                        | –0.2% [-0.7, 0.3]    | 0.6% [0.0, 1.5]               | 0.79 [0.59, 1.06]              |
| AD mean (mm²/s × 10^-5)       | 121 [120, 122]       | –                             | –                              |
| PreMove                       | –0.1% [-0.4, 0.3]    | 0.5% [0.0, 0.9]               | 0.87 [0.75, 0.97]              |
| Repeat                        | –0.2% [-0.5, 0.2]    | 0.4% [0.0, 0.9]               | 0.85 [0.71, 0.97]              |
| RD mean (mm²/s × 10^-5)       | 61.5 [60.6, 62.4]    | –                             | –                              |
| PreMove                       | –0.2% [-0.8, 0.4]    | 0.7% [0.0, 1.9]               | 0.84 [0.69, 0.96]              |
| Repeat                        | –0.3% [-0.9, 0.4]    | 0.8% [0.0, 2.1]               | 0.84 [0.69, 1.06]              |

Reliability and reproducibility across global MRI metrics. Session S2, the first session at the new scanner location, was the reference. Thus, the population estimate is the estimate for each MRI metric at the new location. The ‘PreMove’ population estimate quantifies the mean difference at the old scanner location (S1), as a percentage of the reference population estimate (the new scanner location, S2); this quantifies the mean difference due to scanner relocation. The ‘Repeat’ population estimate quantifies the mean difference between a repeat scan at the new location (S3), relative to the first session at the new location (S2). PreMove between-session variability quantifies the standard error between S1 and S2, as a percentage of the population estimate; repeat variability quantifies S2 and S3. 95% uncertainty intervals are reported for population estimate, between-session variability, and ICC; 95% confidence intervals are reported for wsCV. Note, Total GM, cortical GM, subcortical GM, WM, ventricular volume, and SA were normalized to intracranial volume for calculation of ICC. AD = axial diffusivity, FA = fractional anisotropy, GM = gray matter, ICC = Intraclass correlation coefficient, MD = mean diffusivity, RD = radial diffusivity, WM = white matter, wsCV = within-subject coefficient of variation.

3.2. PCASL results

3.2.1. Global results

Average cortical gray matter perfusion had the highest magnitude between-session variability, both between sites and within-site (Table 1, Fig. 1). Consequently, there was no evidence of a significant systematic difference in mean signal nor any difference in variance between sites or for the repeat.

3.2.2. Regional results

Regional cortical and subcortical mean perfusion values across sessions are presented in Supplementary Figs. 4 and 5, respectively. We identified no areas of significantly different perfusion across the cortex between scanner relocation or within-site repeat.
3.3.1. Global results

DTI metrics were highly reproducible and reliable. Fig. 1 also shows relative skeletal FA, MD, AD, and RD across the three scanning sessions. We estimated mean differences between sites and variability between scanner relocation and within-site repeat to be <0.3% for all metrics (Table 1).

3.3.2. Regional results

Regional DTI metrics from white matter tracts are presented in Supplementary Figs. 6–9. Corrected voxel-wise analyses along the white matter skeleton showed no differences in DTI metrics between relocation or within-site repeat.

3.4. Sample size calculations

Estimated sample sizes required to detect a 4% difference between two independent groups are summarized in Fig. 2.

Necessary sample size to detect a 4% difference between two independent groups given the observed variance in different global MRI metrics, at 80% power, and a 2-sided 0.05 level of significance. AD = average skeletal axial diffusion, RD = average skeletal radial diffusion, MD = average skeletal mean diffusivity, CTh = average cortical thickness, SA = total surface area, ICC = indicates that the metric has been divided by intracranial volume, GM Vol = total gray matter volume, FA = average fractional anisotropy, Cortical Vol = total cortical gray matter volume, WM Vol = total white matter volume, Subcort Vol = total subcortical volume, CBF = average cortical gray matter perfusion. Note: Lateral ventricular volume has been excluded from this plot as the necessary sample size (2686) was too large to appropriately display on the same axes.

3.5. Volume-based processing results

Test-retest reliability of the four global metrics processed using volumetric, CAT12-derived processing are presented in Supplementary Fig. 10 and Supplementary Table 1. As with the surface-based processing, we found no evidence of a systematic difference in variance pre- or post-relocation (Supplementary Table 2). Of the four metrics derived using CAT12 processing, only WM volume exhibited evidence for a small difference in mean volume (−0.5% [−0.9, −0.04]) between scanner locations (see Supplementary Results).

4. Discussion

An MRI scanner we use for ongoing longitudinal research was relocated from a private radiology clinic to a new hospital. In this study, we showed that structural, perfusion, and diffusion MRI metrics showed good to excellent reproducibility and reliability both globally and regionally. We observed no evidence for a systematic difference in mean signal (a bar a small change within-site for total SA) nor any difference in variance across scanner relocation or within-site repeat. Both between-site and within-site test-retest metrics for volume, thickness, and surface area—including mean difference, ICC, wsCV, and SI—were equivalent or higher to those reported in the within-site literature (Eggert et al., 2012; Iscan et al., 2015; Jovicich et al., 2013; Madan and Kensing, 2017), which suggests that these structural MRI metrics are not necessarily compromised by relocation of the MRI scanner.

While we were not necessarily powered to investigate regional variability, we did so in a conservative manner. Neither vertex-based nor voxel-based analysis revealed any region-specific variation between site or within-site. Our global results suggest there was no global bias as a consequence of the scanner relocation; these regional results, supported by Supplementary Figs. 1–9, suggest that there were also no additional region-specific differences in test-retest reliability.

Mean perfusion from the PCASL acquisition had the highest between-session variability, and hence the smallest ICC and largest within-subject coefficient of variation. This variability may be associated with the nature of the PCASL acquisition; the measurement is a result of a difference image, which is susceptible to motion and low signal to noise ratio (Alsop et al., 2015; Dai et al., 2008). Additionally, this high variability likely reflects the more functional nature of perfusion imaging. As opposed to structural metrics, such as T1 or diffusion MRI, perfusion imaging is capable of capturing differences associated with short term brain state. This is both an advantage and a challenge. For example, scanning with eyes open versus eyes shut produces large perfusion differences in the visual cortex. Here, we asked participants to close their eyes; however, we did not monitor compliance. Circadian rhythm has also been shown to associate with cerebral perfusion (Hodkinson et al., 2014). We therefore attempted to scan participants at the same time of day over the three sessions, but there were situations where this was not practical. Both chronic and acute caffeine intake can change perfusion values by up to

| Table 2 | Model comparisons. |
|---------|--------------------|
| Model Component | Baseline | Relative to baseline model |
| Variance | Common | Separate |
| Total GM | −213 ± 7.8 | −3.0 ± 2.0 |
| Cortical GM | −618 ± 7.0 | 5.7 ± 5.4 |
| Subcortical GM | −450 ± 4.3 | 0.8 ± 1.7 |
| WM | −176 ± 6.2 | −2.9 ± 2.6 |
| Ventricle | 74.6 ± 8.4 | −2.6 ± 6.7 |
| CTh | 122 ± 5.5 | −2.9 ± 1.2 |
| SA | −55 ± 7.9 | 4.4 ± 5.2 |
| CBF | −199 ± 7.1 | −0.5 ± 2.4 |
| FA | −36 ± 8.0 | −5.5 ± 3.7 |
| MD | −78 ± 7.5 | −0.5 ± 3.9 |
| AD | −347 ± 5.8 | −1.3 ± 1.4 |
| RD | −351 ± 5.4 | −1.8 ± 1.9 |

ELPD ± standard error. A positive difference indicates a better fit than the baseline model and a negative difference, a worse fit (generally indicating overfitting). When this difference is greater than twice the standard error, there is reasonable evidence that the better-performing model should be preferred. In all cases there was no strong evidence that a model that allowed a predictor for each session or the between-session variance to be different improved the model fit, and in many cases there was evidence that the simpler model was preferred. ‘Common’ modelled the variance between scanner relocation and repeat with a single parameter; ‘Separate’ allowed the variance between scanner relocation and repeat to vary.

GM = Gray matter, WM = White matter, Ventricle = lateral ventricular volume, CTh = Cortical thickness, SA = Surface area, CBF = Cerebral blood flow, AD = Axial diffusivity, RD = Radial diffusivity, ELPD = Estimated log pointwise predictive density.
24% (Clement et al., 2018). Acquisition at different times of day, with varying levels of caffeine intake, in addition to any of the other 58 documented perfusion modifiers—like levels of drowsiness (Poudel et al., 2012), changes in mood, alcohol use, and temporal proximity to physical exercise—may help explain the surprisingly high PCASL variability (Clement et al., 2018). This greater variability suggests that larger effect sizes or larger sample sizes are required in order to identify effects of interest, and that subtle effects may be extremely difficult to identify robustly with PCASL.

Despite the higher variability of PCASL relative to T1- and DTI-derived metrics, the technique still exhibits much better reliability than both task-fMRI and fMRI-based functional connectivity measures. A recent study that performed a meta-analysis and an empirical test-retest reliability analysis of 11 commonly used task-fMRI experiments in the Human Connectome Project and the Dunedin Longitudinal Study reported poor reliability, with ICCs = 0.067–0.485 (Elliott et al., 2019). Similarly, a recent systematic review and meta-analysis of functional connectivity reported that on average, individual connections exhibited a “poor” ICC of 0.29 (Noble et al., 2019). In the current study, within-site cortical GM perfusion ICC was 0.81 (0.63, 0.95), which still falls within the “excellent” range (>0.75). Similar to previous work directly comparing PCASL to task-fMRI and resting state fMRI (Holiga et al., 2018), our results suggest that PCASL should not be abandoned and provides much higher test-retest reliability than either task-fMRI or resting state functional connectivity.

DTI with TBSS processing also exhibited excellent reliability in terms of both bias and variance across relocation and test-retest within the new location. Substantial variability in DTI metrics has been reported across different scanners, acquisition protocols, and processing methodologies (Magnotta et al., 2012; Palacios et al., 2017; Shahim et al., 2017; Vollmar et al., 2010; Wang et al., 2012). However, metrics appear highly reproducible when test-retest reliability is examined on the same scanner, with the same acquisition, and same processing protocol, with wsCV generally below 7% (Shahim et al., 2017; Vollmar et al., 2010; Zhou et al., 2018), and down to 0.73% (FA) and 0.16% (MD) (Palacios et al., 2017), and ICCs >0.9 (Boekel et al., 2017) when global values from TBSS on a single scanner were used (Jovicich et al., 2013). With our consistent acquisition and processing, all investigated DTI metrics showed equivalent reliability both across and within relocation sites, suggesting that the scanner relocation had a negligible effect on average DTI values both regionally and globally along the white matter skeleton.

Calculated sample sizes required to robustly identify a 4% difference between two groups varied widely, depending on MRI metric (Fig. 2). For example, in order to show a 4% difference in cortical gray matter perfusion, groups of 362 participants per arm would be needed; for the same mean difference, only 23 participants per group would be needed for average FA and only 9 per group for cortical thickness. These dramatically different sample size requirements further highlight the real-world consequences of metrics exhibiting higher variance.

Regional brain volumes are frequently normalized by some measure of intracranial volume to reduce variability (Hansen et al., 2015; Whitwell et al., 2001). Our sample size calculations demonstrate and reinforce this important methodological procedure; while 114 participants per group would be needed if total gray matter volume were used, normalizing by ICV provides much more power, requiring only 19 per group.

A systematic difference between relocation means of 0.1% in total GM volume (from the current study) will most likely not change interpretations of group-level cross-sectional studies where large effect sizes are expected (e.g., in advanced Parkinson’s disease (Melzer et al., 2012)). Even in situations where effects of interest are relatively small—for example, tracking aging in a group of 60-year-olds, with an expected total GM volume loss in healthy aging of ~0.45% per year (Schippling et al., 2017)—this systematic bias may not significantly influence interpretations. Our results suggest that systematic bias as a consequence of scanner relocation appears negligible. However, this type of scanner change could (and should) be included in statistical modelling when possible (Lee et al., 2019). We did not investigate the reliability of more advanced diffusion acquisitions and models, and are therefore unable to comment on the interesting question of reliability of high angular resolution diffusion imaging (HARDI) sequences (Zhou et al., 2018).

The surface-based (Freesurfer) and volume-based (CAT12, Supplementary Material) processing streams showed very similar results. For example, the population estimate, mean difference, and between-session variability were virtually identical for both estimates of total GM volume and average GM perfusion. However, we observed a small effect of scanner relocation for CAT12-processed WM volume (~0.5% [−0.9, −0.04]), but the difference for Frestsuer-derived WM volume between sites (0.2% [−0.9, 0.4]) did not meet our statistical threshold. Frestsuer-derived metrics required smaller sample sizes per group than the equivalent CAT12-derived metrics (Supplementary Results). Frestsuer-derived total GM volume divided by ICV required n = 19 per group to robustly identify a 4% difference between groups, while the CAT12-derived equivalent required n = 45 per group, a 2.4-fold increase. This suggests that the longitudinal Freesurfer pipeline allows for smaller sample sizes (Supplementary Discussion).

There may be instances in the future when a scanner needs to be relocated. Here we provide an indication of what scanning changes can be expected. With appropriate vendor support and re-commissioning of the scanning environment (re-shim), we have shown that MRI change due to relocation appears equivalent to within-site test-retest reliability. Beyond relocation, we provide estimates of test-retest reliability (both mean and variance), which allows others to estimate variance of specific metrics, facilitating power analyses. This manuscript also re-inforces the robustness of T1 and DTI metrics, but highlights the high variability of cortical perfusion. This is important to consider when comparing the relative importance of structural versus PCASL-based investigations and the influence these have on interpretation of results. Despite this higher variability for the PCASL acquisition, it still compares favorably to task-fMRI and functional connectivity in terms of reliability.

In this work, we investigated the effect of MRI scanner relocation on the reliability of global and regional MRI metrics derived from structural, perfusion, and diffusion acquisitions. Cortical gray matter perfusion was the most variable metric investigated, with all other metrics showing substantially less variability. We observed no evidence for a systematic difference in mean signal or variance between-site or within-site repeat (bar a small SA mean difference within-site). These results suggest that when all other parameters are held constant (e.g., sequence parameters and MRI processing), the effect of scanner relocation is small, but may need to be considered depending on the question at hand.

Data/code availability statement

Analysis code and data are available at https://github.com/nzbri/scanner-test-retest.

Ethics

Twenty healthy volunteers provided written informed consent.

Statistical analysis

TR Melzer and DJ Myall.

Funding

Research reported in this publication was supported by Pacific Radiology Christchurch. This work was performed while TRM was supported by a Sir Charles Hercus Early Career Development Fellowship from the Health Research Council of New Zealand (17/039).
Role of funder

The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declaration of competing interest

COI

CRediT authorship contribution statement

Tracy R. Melzer: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. Ross J. Keenan: Resources, Writing - review & editing. Gareth J. Leeper: Investigation, Writing - review & editing. Stephen Kingston-Smith: Investigation, Writing - review & editing. Simon A. Felton: Investigation, Writing - review & editing. Sarah K. Green: Investigation, Writing - review & editing. Kerry J. Henderson: Investigation, Writing - review & editing. Nicholas J. Palmer: Investigation, Writing - review & editing. Reza Shoorangiz: Methodology, Writing - review and editing. Mustafa M. Almuqbel: Project administration, Methodology, Investigation, Writing - review & editing. Daniel J. Myall: Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing.

Acknowledgements

Authors would like to thank Pacific Radiology Christchurch for research support and volunteers for their time. This work was performed while TRM was supported by a Sir Charles Hercus Early Career Development Fellowship from the Health Research Council of New Zealand (17/039).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116608.

References

Alois, D.C., Detre, J.A., Golay, X., Günther, M., Hendriks, J., Hernandez-Garcia, L., Lu, H., MacIntosh, B.J., Parkes, L.M., Smits, M., van Osch, M.J.P., Wang, D.J.J., Wong, E.C., Zaharchuk, G., 2015. Recommended implementation of arterial spin-labeled perfusion MRI for clinical applications: a consensus of the ISMBM perfusion study group and the European consortium for ASL in dementia. Magn. Reson. Med. 73, 102–116. https://doi.org/10.1002/mrm.25607.

Bland, J.M., Altman, D.G., 1996. Statistics Notes: Measurement error proportional to the mean. BMJ 313, 106. https://doi.org/10.1136/bmj.313.7049.106.

Boekel, W., Forstmann, B.U., Keuken, M.C., 2017. A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control. Psychophysiology 54, 24–23. https://doi.org/10.1111/tpyp.12769.

Buergner, P., 2017. brms: An R package for Bayesian multilevel models using Stan. J. Stat. Software 80, 1–28. https://doi.org/10.18637/jss.v080.i01.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Software 76, 1–37. https://doi.org/10.18637/jss.v076.i01.

Chen Zhiye, L.M., Lin, Ma, 2017. Surface-based morphometry of human brain: intra-administration, Methodology, Investigation, Writing - review. Buerkner, P., 2017. brms: an R package for Bayesian multilevel models using Stan. J. Stat. Software 76, 1

Declaration of competing interest

COI

manuscript for publication.

Appendix A. Supplementary data

(17/039).

opment Fellowship from the Health Research Council of New Zealand

 interviewed the healthy subjects

Larsson, E.-M., Achten, E., 2018. Variability of physiological brain perfusion in a systematic review of modi

Lee, H., Nakamura, K., Narayanan, S., Brown, R.A., Arnold, D.L., Alzheimer’s Disease Neuroimaging, 1. 2019. Estimating and accounting for the effect of MRI scanner changes on longitudinal whole-brain volume change measurements. Neuroimage 184, 555–565. https://doi.org/10.1016/j.neuroimage.2018.09.057.

Lee, H., Nakamura, K., Narayanan, S., Brown, R.A., Arnold, D.L., Alzheimer’s Disease Neuroimaging, 1. 2019. Estimating and accounting for the effect of MRI scanner changes on longitudinal whole-brain volume change measurements. Neuroimage 184, 555–565. https://doi.org/10.1016/j.neuroimage.2018.09.057.

Madan, C.R., Kinsinger, E.A., 2017. Test-retest reliability of brain morphology estimates. Brain Inf 4, 107–121. https://doi.org/10.4108/j4inf.2016.0060-4.

Madhyastha, T., Mérillat, S., Hirsiger, S., Bezzola, L., Liem, F., Grabowski, T., Jancke, L., 2012. Longitudinal reliability of object-based spatial statistics in diffusion tensor imaging. Hum. Brain Mapp. 35, 4544–4555. https://doi.org/10.1002/hbm.23493.
Magnotta, V.A., Matsui, J.T., Liu, D., Johnson, H.J., Long, J.D., Bolster Jr., B.D.,
McCullough, R.A., Lim, K., Mori, S., Helmer, K.G., Turner, J.A., Reading, S.,
Van Horn, L., Aylward, E., Flashman, L.A., Bonett, G., Paulsen, J.S., 2012. Multicenter reliability of diffusion tensor imaging. Brain 2, 345–355. https://doi.org/10.1098/
brain.2012.0112.

Mak, E., Su, L., Williams, G.B., Firbank, M.J., Lawson, R.A., Yarnall, A.J., Duncan, G.W.,
Mollenhauer, B., Owen, A.M., Khoo, T.K., Brooks, D.J., Rowe, J.B., Barker, R.A.,
Burn, D.J., O’Brien, J.T., 2017. Longitudinal whole-brain atrophy and ventricular enlargement in non-demented Parkinson’s disease. Neurobiol. Aging 55, 78–90. https://doi.org/10.1016/j.neurobiolaging.2017.03.012.

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

Melzer, T.R., Myall, D.J., MacAskill, M.R., Pitcher, T.L., Livingston, L., Watts, R.,
Keenan, R.J., Dalrymple-Alford, J.C., Anderson, T.J., 2015. Tracking Parkinson’s disease over one year with multimodal magnetic resonance imaging in a group of older patients with moderate disease. PloS One 10, e0143923. https://doi.org/
10.1371/journal.pone.0143923.

Melzer, T.R., Watts, R., MacAskill, M.R., Pearson, J.F., Rueger, S., Pitcher, T.L.,
Livingston, L., Graham, C., Keenan, R., Shankaranarayanan, A., Allop, D.C.,
Dalrymple-Alford, J.C., Anderson, T.J., 2011. Arterial spin labelling reveals an abnormal cerebral perfusion pattern in Parkinson’s disease. Brain 134, 845–855. https://doi.org/10.1093/brain/awq257.

Melzer, T.R., Watts, R., MacAskill, M.R., Pitcher, T.L., Livingston, L., Keenan, R.,
Dalrymple-Alford, J.C., Anderson, T.J., 2013. White matter microstructure deteriorates across cognitive stages in Parkinson disease. Neurology 80, 1841–1849. https://doi.org/10.1212/NEU.0b013e3182929f62.

Melzer, T.R., Watts, R., MacAskill, M.R., Pitcher, T.L., Livingston, L., Keenan, R.,
Dalrymple-Alford, J.C., Anderson, T.J., 2012. Grey matter atrophy in cognitively impaired Parkinson’s disease. J. Neurol. Neurosurg. Psychiatry 83, 183–194. https://doi.org/10.1136/jnnp.2011.300828.

Mori, S., Ohki, K., Jiang, H., Jiang, L., Li, X., Akhter, K., Hua, K., Faria, A.V.,
Mahmood, A., Woods, R., Toga, A.W., Pike, G.B., Neto, P.R., Evans, A., Zhang, J.,
Huang, H., Miller, M.L., van Zijl, P., Mazziotta, J., 2008. Stereotoxic white matter atlas based on diffusion tensor imaging in an ICBM template. Neuroimage 40, 570–582. https://doi.org/10.1016/j.neuroimage.2007.12.035.

Mutsaerts, H.J.M.M., Steketee, R.M.E., Heijtjes, D.F.R., Kuijer, J.P.A., van Osch, M.J.P.,
Majoie, C.B.L.M., Smits, M., Nederveen, A.J., 2014. Inter-vendor reproducibility of site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. Neuroimage 51, 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046.

Wang, J.Y., Abdi, H., Bakhadirov, K., Díaz-Arrastia, R., Devous, M.D., 2012. A comprehensive reliability assessment of quantitative diffusion tensor tractography. Neuroimage 60, 1127–1138. https://doi.org/10.1016/j.neuroimage.2011.12.062.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060.

Wonderlick, J.S., Ziegler, D.A., Hossein-Zarimkhorshidi, P., Locascio, J.J., Bakkour, A., van der Kourwe, A., Triantafyllou, C., Corkin, S., Dickerson, R.C., 2009. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. Neuroimage 44, 1324–1333. https://doi.org/10.1016/j.neuroimage.2008.10.037.

Xu, G., Rowley, H.A., Wu, G., Allop, D.C., Shankaranarayanan, A., Dowling, M.,
Christian, B.T., Oakes, T.R., Johnson, S.C., 2010. Reliability and precision of pseudo-continuous arterial spin labeling perfusion MRI on 3.0 T and comparison with 15O-water PET in elderly subjects at risk for Alzheimer’s disease. NMR Biomed. 23, 286–293. https://doi.org/10.1002/nbm.1462.

Zhou, X., Sakaie, K.E., Debbins, J.P., Narayanan, S., Fox, R.J., Lowe, M.J., 2018. Scan-
rescan repeatability and cross-scanner comparability of DTI metrics in healthy subjects in the SPRINT-MS multicenter trial. Magn. Reson. Imaging 53, 105–111. https://doi.org/10.1016/j.mri.2018.07.011.