



# Multi-center reproducibility of structural, diffusion tensor, and resting state functional magnetic resonance imaging measures

S. Deprez<sup>1,2</sup> · Michiel B. de Ruiter<sup>3</sup> · S. Bogaert<sup>4</sup> · R. Peeters<sup>1,2</sup> · J. Belderbos<sup>5</sup> · D. De Ruyscher<sup>6,7</sup> · S. Schagen<sup>3</sup> · S. Sunaert<sup>1,2</sup> · P. Pullens<sup>8</sup> · E. Achten<sup>4</sup>

Received: 1 November 2017 / Accepted: 19 March 2018  
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

## Abstract

**Purpose** The aim of this study is to assess multi-center reproducibility and longitudinal consistency of MRI imaging measurements, as part of a phase III longitudinal multi-center study comparing the neurotoxic effect following prophylactic cranial irradiation with hippocampal avoidance (HA-PCI), in comparison with conventional PCI in patients with small-cell lung cancer.

**Methods** Harmonized MRI acquisition protocols from six participating sites and two different vendors were compared using both physical and human phantoms. We assessed variability across sites and time points by evaluating various phantoms and data including hippocampal volume, diffusion metrics, and resting-state fMRI, from two healthy volunteers.

**Results** We report average coefficients of variation (CV) below 5% for intrascanner, intravendor, and intervender reproducibility for both structural and diffusion imaging metrics, except for diffusion metrics obtained from tractography with average CVs ranging up to 7.8%. Additionally, resting-state fMRI showed stable temporal SNR and reliable generation of subjects DMN across vendors and time points.

**Conclusion** These findings indicate that the presented multi-site MRI acquisition protocol can be used in a longitudinal study design and that pooling of the acquired data as part of the phase III longitudinal HA-PCI project is possible with careful monitoring of the results of the half-yearly QA assessment to follow-up on potential scanner-related longitudinal changes in image quality.

**Keywords** Multi-center · Reproducibility · Structural MRI · Diffusion tensor imaging · Resting state fMRI

## Introduction

Longitudinal multicenter MRI studies are becoming increasingly important to study structural and functional changes in the brain following pathophysiological conditions or

therapeutic treatment. This multi-center approach is particularly important when only limited numbers of patients can be recruited at a single research site and permits to collect data from larger groups of participants in a reasonable time frame [1]. However, combining images acquired at different time

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00234-018-2017-1>) contains supplementary material, which is available to authorized users.

✉ Michiel B. de Ruiter  
m.d.ruiter@nki.nl

<sup>1</sup> Department of Radiology, University Hospitals Leuven, Leuven, Belgium

<sup>2</sup> Department of Imaging and Pathology, KU Leuven, Leuven, Belgium

<sup>3</sup> Division of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

<sup>4</sup> Department of Radiology, Ghent University Hospital, Ghent, Belgium

<sup>5</sup> Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands

<sup>6</sup> Department of Radiation Oncology (Maastricht clinic), GROW School, Maastricht University Medical Center, Maastricht, The Netherlands

<sup>7</sup> Radiation Oncology, KU Leuven, Leuven, Belgium

<sup>8</sup> Department of Radiology, University Hospital Antwerp & Antwerp University, Edegem, Belgium

points and different scanners require comparable and stable MR imaging measurements over time and across sites. This is important as potential differences due to scanner-dependent variability may confound true changes in brain structure and function.

The aim of this study is to assess (1) multi-center reproducibility and (2) longitudinal consistency of MRI imaging measurements as part of a phase III longitudinal multi-center study (registered at [clinicaltrials.gov](https://clinicaltrials.gov) no. NCT01780675). This study will compare the radiotherapy-induced neurocognitive effect and the structural and functional changes in the brain following prophylactic cranial irradiation with hippocampal avoidance (HA-PCI) in comparison with conventional PCI in patients with small-cell lung cancer (SCLC). In this phase III study, 3D-T1w structural imaging will be used to assess whether HA-PCI is associated with prevention of hippocampal atrophy as compared to whole brain PCI. Additionally, diffusion tensor imaging (DTI) and resting state functional MRI (rsfMRI) will be acquired to assess differences in white matter microstructure and functional brain connectivity, following HA-PCI and PCI. Finally, the study aims (3) to propose a framework for the evaluation of the image quality of multi-modal neuroimaging in a longitudinal study design.

Longitudinal multi-center studies have to consider different sources of variability. Longitudinal studies are prone to across-session variability induced by MRI scanner instabilities (scanner drift over time, e.g., [2], by scanner maintenance and software or hardware upgrades of the system [3], which we will refer to as intrascanner reproducibility. Additionally, when including scanners of multiple centers, each technical issue needs to be considered, such as variability related to differences in field strength, MRI acquisition protocols, scanner hardware and software versions, room temperature, and differences in scanner manufacturers. Therefore, combining images acquired at different centers requires comparable MR imaging measurements and an across-sites validation of techniques assessing interscanner reproducibility.

Currently, only a limited number of studies have investigated the stability of multi-center multimodal MRI measurements. Earlier studies have focused on inter- or intrascanner neuroimaging measurements of one imaging modality only (rsfMRI: [4]; T1-w: [5–8], and DTI: [3, 9–11] or were restricted to scanners from a single vendor [11–14]. However, studies examining the variability of longitudinal multimodal measurements (T1-w, DTI, and rsfMRI data) including scanners from different vendors, and combining both human and phantom imaging parameters, are lacking. Huang et al. presented a framework to quantify the reproducibility of multimodal neuroimaging studies (structural, BOLD, and DTI) collected across identical scanners based on a traveling “human” phantom. Belli and colleagues [15] used an isotropic diffusion phantom to assess multi-center reproducibility of diffusion weighted imaging on scanners of different vendors. Zhu

et al. [11] combined both scans of an isotropic diffusion phantom and human volunteers to assess inter-site variations of DTI scans acquired on scanners of the same vendor, while Teipel et al. [10] combined both DTI scans from an anisotropic diffusion phantom and healthy volunteers to assess inter-site variations of scans from different vendors. Similarly, Jovicich et al [4] combined both phantom and human scans to study rsfMRI variability including scanners from different vendors.

In this study, we wanted to go beyond this, by measuring imaging reproducibility in anatomical, diffusion, and fMRI scans across different time points and scanners from different vendors, both in phantoms as well as in healthy volunteers.

First, we assessed structural imaging variability across different scanners and time points by (1) evaluating geometrical deformations on the American College of Radiology (ACR) MRI phantom and (2) calculating both Freesurfer volume estimates and manually delineated hippocampal volumes from two volunteers. Second, we assessed variability in diffusion metrics across scanners and time points from (1) both isotropic and anisotropic diffusion phantoms and (2) two healthy volunteers in both manually and automatically defined ROI's. Finally, we evaluated across-scanner and across-session reliability of rsfMRI by assessing temporal signal-to-noise ratio (tSNR) of phantom scans and the generation of the default-mode network (DMN) in two healthy volunteers.

## Methods

### Subjects and phantoms

Two healthy volunteers who provided written informed consent (female, age 25 and 30 years, both right handed) were scanned on 6 different 3.0 Tesla MRI scanners from two major vendors (Siemens and Philips) using the head coil with the highest number of channels available (Table 1). Subject 1 was scanned a second time 1 month later on a representative scanner from each vendor (sites 1 and 3) to investigate intrascanner reproducibility. To further assess intrascanner stability of the hippocampal volume measurements, subject 2 was scanned five additional times on the same scanner with the same protocol in site 1, with repositioning and slice re-angulation in between each scan.

Three physical phantoms were used to assess variability and reproducibility between scanners for the different sequences (Supp. Fig. 1). The ACR MRI accreditation phantom was used to assess structural imaging variability (MRI Accreditation Program Requirements; The American College of Radiology, Reston, VA. Available at: <http://www.acr.org>). In order to fit the size of the head coil, the large and small MRI ACR phantoms were used in sites 3 to 6 and sites 1 and 2, respectively.

**Table 1** Description of hardware, software, and head coils of the participating sites

Scanner	B0	Vendor	Scanner type	Software	RF coil type	Acceleration algorithm
1	3.0 T	Siemens	Trio	VB17	32ch	GRAPPA
2	3.0 T	Siemens	Trio	VB17	32ch	GRAPPA
3	3.0 T	Philips	Achieva	5.1.2.0	32ch	SENSE
4	3.0 T	Philips	Achieva	5.1.2.0	16ch	SENSE
5	3.0 T	Philips	Achieva	3.2.1.0	8ch	SENSE
6	3.0 T	Philips	Achieva	3.2.1.0	8ch	SENSE

An anisotropic diffusion tensor imaging (DTI) phantom (HQ imaging, Heidelberg, Germany [www.hq-imaging.de](http://www.hq-imaging.de)) was used to evaluate differences in diffusion measures. The phantom consists of a polyamide fiber ring with constant anisotropy at each radial position, embedded in a homogeneous medium [16].

Additionally, an isotropic diffusion phantom, consisting of a 500-ml bottle of dodecane  $\text{CH}_3(\text{CH}_2)_{10}\text{CH}_3$  ReagentPlus® ≥99% (Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), encased in a plastic container and fixed with polyurethane foam as a thermal insulator [17], was used to assess stability of the employed diffusion gradients in each scanner, to verify if correction for gradient variability was required, and to evaluate differences in temporal SNR of the functional MRI scans [17, 18].

Both the ACR and anisotropic DTI phantom traveled from site to site, and were scanned four times as part of a half-yearly quality assurance (QA) process at each scanner in the study. The first scan at each site was performed by the same experienced radiographer (SB). Local radiographers who were hands-on trained during this first session performed all subsequent QA scans at the three following time points. In the second year, we added the isotropic diffusion phantom to the half-yearly QA procedure. Each site received its own isotropic phantom (all dodecane was obtained from the same production batch), and two time points were included in this study.

All image acquisitions were conducted after the phantoms were acclimatized to the scanner room temperature for at least 6 h. For every time point, the phantoms were positioned in the isocenter of the magnetic field following strict written guidelines.

## MR imaging acquisition

We will briefly discuss the MR protocols used in this study. MRI acquisition parameters can be found in Table 2.

**Human protocol** High-resolution 3D-T1 weighted images were acquired following the ADNI protocol developed for multi-center intervender acquisitions with parameters optimized for contrast between white and gray matter and cerebrospinal fluid [19]. Diffusion-weighted imaging was performed with a spin-echo echo-planar-imaging (SE-EPI)

sequence using 60 different diffusion directions and a b-value of  $1300 \text{ s/mm}^2$ . The diffusion directions were kept the same between vendors. Site 4, however, did not include the DTI scan as the sequence with 60 diffusion directions could not be set-up. Additionally, rsfMRI data were acquired using whole brain T2\*-weighted gradient-echo EPI, sensitive to blood oxygenation level-dependent (BOLD) contrast, with the instruction to keep the eyes closed and relax, but not to fall asleep. B0 field maps were acquired to correct the other images for possible geometrical distortions.

**Phantom protocol** The same high-resolution 3D-T1, diffusion-weighted imaging, and rsfMRI scans as described above were acquired for the ACR, DTI, and dodecane phantom, respectively. Additionally, following the recommendations of the Routine Assurance Pipeline for Imaging of Diffusion (RAPID) (De Santis, Evans, & Jones, 2013), the dodecane phantom was scanned using a standard DTI sequence with 3.0 mm isovolumetric voxels, 12 directions, and 6 b-values (0, 350, 700, 1050, 1400, and  $1750 \text{ s/mm}^2$ ) in all sites except site 4.

## Image processing and statistical analysis

Prior to analysis, all MRI images were visually checked for signal homogeneity and artifacts including Nyquist ghosting, motion, and eddy current artifacts.

### Structural MRI

To study the structural imaging variability across scanners and time points, we analyzed the high-resolution 3D-T1 weighted images of the ACR phantom and healthy volunteers.

**ACR phantom** All 3D-T1 datasets were axially reconstructed and analyzed according to the ACR guidelines (MRI Accreditation Program Requirements; The American College of Radiology, Reston, VA. Available at: <http://www.acr.org>). Parameter estimates for geometric accuracy, slice thickness accuracy, slice position accuracy, image intensity uniformity, percent signal ghosting, and low-contrast detectability were compared to the synthetic values of the phantom.

**Table 2** Sequence parameters of the HA-PCI protocol

Sequence	Purpose	Vendor	Orientation	TR/TE (ms)	FA (°)	Voxel size (mm <sup>3</sup> )	# slices	BW (Hz/pixel)	PE acceleration	Accel. factor PE
MPRAGE	Human, ACR	Siemens	Sagittal	2500/2.84	9	1.1 × 1.1 × 1.2	160	240	GRAPPA	2
RS fMRI	Human, Dodec	Philips	Sagittal	2500/shortest	9	1.1 × 1.1 × 1.2	170	240	SENSE	2
		Siemens	Axial	2000/27	90	3.0 × 3.0 × 3.0	38	2170	GRAPPA	2
DTI	Human, DTI	Philips	Axial	2000/27	80	3.0 × 3.0 × 3.0	37	2018	SENSE	2
		Siemens	Sagittal	7600/81	90	2.5 × 2.5 × 2.5	66	2264	GRAPPA	2
DTI 6 b-values	Dodec	Philips	Sagittal	7600/80	90	2.5 × 2.5 × 2.5	58	2226	SENSE	2.5
		Siemens	Axial	4000/93	90	3.0 × 3.0 × 3.0	18	2790	GRAPPA	2
		Philips	Axial	4000/95	90	3.0 × 3.0 × 3.0	18	2413	SENSE	2
Coil normalization	Coil combination	EPI factor#	Total acquisition matrix	Diffusion scheme	Extra	TA				
Prescan normalize ON	Adaptive combine		232 × 256		TI 900 ms	05:19				
CLEAR	–		220 × 180			05:34				
Prescan normalize OFF	Sum of squares	72	72 × 72		Dyn scans 200	06:48				
CLEAR	–	41	80 × 76			06:51				
Prescan normalize OFF	Adaptive combine	80	80 × 96		60 dir/b-values 0, 1300 s/mm <sup>2</sup>	08:06				
CLEAR	–	41	80 × 96		Monopolar	10:31				
Prescan normalize OFF	Adaptive combine	128	128 × 126		12 dir/b-values 0, 50, 700, \1050, 1400, 1750 s/mm <sup>2</sup>	04:18				
CLEAR	–	63	128 × 126		Monopolar	04:12				

#In case of Siemens, the final EPI factor after reconstruction is given, while Philips reports the EPI factor before SENSE reconstruction. After reconstruction, the EPI factors are the same

**Hippocampal volume assessment** Hippocampal volume estimates were obtained using both a manual and an automatic delineation protocol. A radiographer (SB)—trained and supervised by an experienced neuroradiologist—traced every slice containing hippocampal tissue on a high-resolution monitor, using a sagittal ray-tracing protocol [20] and dedicated software (syngo.MR General v.VB10A, Siemens AG, Erlangen-Germany) that allowed interpolation. The delineated hippocampal tissue included the internal white matter between the cornu ammonis regions and excluded the fimbria. The surface of the ROI encircling the hippocampus was multiplied by the slice thickness to obtain hippocampal volume in cubic centimeters. Additionally, images were automatically processed with the FreeSurfer v5.3 (Martinos Center for Biomedical Imaging, Harvard-MIT, Boston-USA) longitudinal stream to obtain hippocampal volumes and were visually checked for accuracy [21].

To assess intrascanner, intravendor, and intervender variability, coefficients of variation (CV) (=standard deviation/mean  $\times 100\%$ ) were calculated [22] for both manual and automatic delineations. Intrascanner variability was based on the repeated scans for subject 1 in sites 1 and 3 and the five times repeated scan of subject 2 in site 1.

Intravendor CVs were calculated based on the acquired data for each vendor separately. To take possible sampling bias into account, an additional CV was calculated for vendor B, including an equal number of sites as those for vendor A. As calculations for vendor A were based on data from two sites with the same head coil configuration, the additional calculation for vendor B was based on data from two out of the four participating sites with the same headcoil configuration (sites 5 and 6). Finally, intervender variability was calculated based on the mean hippocampal volumes of all six included sites.

## Diffusion imaging

To assess the variability of diffusion imaging across scanners and time points, we evaluated diffusion imaging metrics obtained from the isotropic and anisotropic diffusion phantoms, as well as from the two healthy volunteers.

**Anisotropic diffusion phantom** DTI images obtained at the four QA time points in all sites were pre-processed using ExploreDTI [23]. FA and MD maps were generated using a non-linear- least-square tensor estimation procedure. Subsequently, mean FA values for each phantom scan were extracted manually and in automatically obtained ROIs (Supp. Fig. 1d).

Eight octahedral ROIs were manually delineated with 3 voxels diameter ( $7 \times 8 = 56$  voxels/phantom) positioned on the FA-map of the fiber ring. Additionally, to automatically obtain a mask of the fiber ring, Otsu segmentation [24] was

applied on an image composed of the stdev of DWI images (without b0), thresholded above  $0.4 \times \text{maximum over std(DWI)}$  and smoothed with a Gaussian kernel of  $5 \times 5$  voxels and  $\sigma = 0.6$ . The obtained masks of the fiber ring were then eroded to remove voxels affected by partial voluming, yielding an average of 70 voxels (range 39 to 113) per phantom.

Next, intrascanner, intravendor, and intervender coefficients of variation were calculated based on FA values of all voxels contained in both masks.

**Isotropic diffusion phantom** To assess diffusion gradient performance of the different scanners on a periodic basis (at regular intervals), we used the “Routine Assurance Pipeline for Imaging of Diffusion (RAPID)” proposed by [25]. Isotropic diffusion phantom scans were first processed in MRtrix3 (J-D Tournier, Brain Research Institute, Melbourne, Australia, <https://github.com/MRtrix3/mrtrix3>) to generate ADC maps. ADC was extracted at each time point from a circular ROI with a diameter of 8 voxels in the middle slice. To account for the effect of ambient temperature on the diffusion properties, the temperature in the scanner room was recorded and used to correct the obtained DWI values for temperature fluctuations. Holz et al. [26] provide a calibration curve obtained with Pulsed Gradient NMR Spin Echo measurements at well-controlled temperatures for Dodecane diffusion coefficients:

$$D(T) = \exp(5.3193 - 1648.3/T)$$

with  $D$  in ( $10^{-9} \text{ m}^2/\text{s}$ ) and  $T$  in Kelvin. The ADC measurements in this study were compared against this calibration curve (Supp. Fig. 2).

The RAPID pipeline (<http://www.mathworks.com/matlabcentral/fileexchange/36463>) was used to calculate deviations in x, y, and z diffusion gradient performance. The RAPID output, which is a gradient correction factor  $g' = [g'x \ g'y \ g'z]$  in the x, y, and z directions was then applied to the b-matrix:  $B' = g' \times B$ . The tensor was re-estimated with the corrected gradient table and ADC was again extracted from the same ROI. Corrected and uncorrected ADC values were compared using a paired  $t$  test in Matlab.

**Human data** To evaluate reproducibility of regional fractional anisotropy (FA) and mean diffusivity (MD), DTI scans were analyzed using two different approaches: automated tract-based spatial statistics (TBSS) and semi-automated tractography.

**TBSS** DTI scans from both subjects were analyzed using FMRIB’s diffusion toolbox (FDT) implemented in FSL5.0 (FMRIB Analysis Group, Oxford, UK, <http://www.fmrib.ox.ac.uk/analysis/research/tbss>). First, eddy current-induced morphing was corrected by affine registration of the diffusion-weighted images to the average b0 image. Then, a



diffusion tensor model was fit to the data to generate FA and MD maps. Subsequently, every FA map was aligned to every other one to identify the most representative one, and use this as the target image. This target image was then affine-aligned into MNI152 standard space, and every FA map was transformed into  $1 \times 1 \times 1$  mm MNI152 space by combining the nonlinear transform to the target FA image with the affine transform from that target to MNI152 space. This procedure was performed separately for both subjects. The mean FA map was “skeletonized” and FA values  $> 0.2$  were retained. To create skeletonized FA maps for each subject, individual FA maps were projected onto the mean FA skeleton. Finally, MD maps were warped into standard space and skeletonized with the same parameters that were used for the FA maps. Eight tracts were selected for extraction of FA and MD values: forceps major; forceps minor; corticospinal tract (CST, l/r); temporal component of the superior longitudinal fasciculus (SLF temp, l/r); and hippocampal part of the cingulum (cing hippo, l/r). ROIs were carefully selected to ensure that only white matter was retained. ROIs were derived from the Johns Hopkins University probabilistic white-matter tractography atlas provided within FSL thresholded at 20%.

**Semi-automated tractography** DTI preprocessing was done in ExploreDTI 4.8.4 consisting of motion and distortion correction with reorientation of the b-matrix. After generation of FA and MD maps, whole brain tractography was performed using constrained spherical deconvolution (CSD). The same eight fiber tracts (forceps major; forceps minor; corticospinal tract (CST, l/r); temporal component of the superior longitudinal fasciculus (SLF temp, l/r); hippocampal part of the cingulum (cing hippo, l/r)) were reconstructed according to Wakana et al. [27]. Manual tractography was performed for data from one site per subject by two independent operators (SD and MR). Then, automated atlas-based tractography (ExploreDTI) was performed for each subject to reconstruct the tracts for the data of the remaining sites [28] using the manually defined regions-of-interest and the representative single subject FA map as a template. All subject FA maps were registered to the subject template map using a non-affine transformation followed by tensor reorientation. All tracts were visually inspected for accuracy and spurious tracts were removed by defining additional “not” ROIs. Mean FA and MD were calculated for all tract segments between the delineated ROIs.

To assess intrascanner, intravendor, and intervender variability of the obtained DTI metrics, coefficients of variation were calculated in the same way as described for hippocampal volume measurements.

Additionally, inter-observer reliability was estimated by calculating the intraclass correlation (ICC) within SPSS (IBM SPSS Statistics for Macintosh, Version 23.0) using a two-way mixed model, absolute agreement, and single measures.

## Resting-state fMRI

**Isotropic diffusion phantom** The EPI-BOLD scans of the dodecane phantom were used to estimate the temporal stability of the measured time course of the rsfMRI sequence. Temporal SNR (tSNR) measures were generated using the quality control report available from the FBIRN QA tools ([http://www.nitrc.org/projects/bxh\\_xcede\\_tools](http://www.nitrc.org/projects/bxh_xcede_tools); fmriqa\_generate.pl). This QA report includes tSNR values for a ROI in the middle slice. The tSNR is calculated by dividing the average signal over time by the standard deviation of the residuals after temporally detrending the images [2].

**Human data** Resting state fMRI identifies the temporal synchronicity between time courses of several regions of interest as a measure of connectivity. We explored qualitatively the generation of the default mode network using independent component analysis (ICA) across sites and time points. RsfMRI images were preprocessed in SPM8 (Wellcome Department of Imaging Neuroscience, London, UK, 2009) including slice-time correction, realignment to correct for small head movements and unwarping, coregistration with structural volumes, normalization to standard MNI space and spatial smoothing with a  $6\text{-mm}^3$  full-width-at-half-maximum (FWHM) isotropic Gaussian kernel. The time-series were further optimized by removal of linear trend from the signal and by temporal band-pass filtering (cut-off frequency value =  $0.01\text{ Hz} < f < 0.1\text{ Hz}$ ;  $\text{TR} = 2\text{ s}$ ) with the rsfMRI data analysis toolkit v1.8 (REST) [29] in Matlab r2013b. No brain tissue mask was selected during calculation. A group ICA [30] (based on subject 1 and 2) was performed with the Group ICA/VA of MRI toolbox v2.0a (GIFT) (<http://mialab.mrn.org/software/gift>) in Matlab r2013b to identify spatially distinct, temporally coherent components from resting state data per site. Twenty independent components were estimated using the Infomax algorithm. The component representing the Default Mode Network (DMN) was extracted for each subject and used to produce a spatial map per individual per site for visual assessment.

## Results

### Structural MRI

#### ACR phantom

All participating MRI scanners passed the ACR criteria at the four included time points, except for two deviations: The MRI scanner in sites 5 and 6 had a small deviation for the geometric accuracy test at time point 1; however all subsequent time points were within norms. The percent image uniformity (PIU) never passed the 82% criterium for scanners operating

at 3.0 Tesla in sites 3, 5, and 6 and passed only in one out of four time points in sites 1 and 2. PIU results in site 4 met the criterion. Parameter estimates for geometric accuracy, spatial resolution, slice thickness accuracy, slice position accuracy, image intensity uniformity, percent signal ghosting, and low-contrast detectability are detailed in Table 3.

### Hippocampal volume assessment

**Intrascanner variability** Hippocampal volume estimates obtained from both manual and automated processes showed good test-retest reliability over time. Intrascanner CV obtained from the repeated scans of subject one with 1-month interval ranged from 0.18 to 2.62%. Freesurfer results were most consistent with three out of four CVs < 1.0% as compared to one out of four after manual delineation. The intrascanner consistency for Freesurfer hippocampal volume estimations was confirmed with a CV of 0.72% for the right and 1.16% for the left hippocampus for the five times repeated scan of subject 2 (Table 4, Fig. 1).

**Intravendor variability** Manual delineation resulted in both the lowest and the highest intravendor coefficients of variation ranging from 0 to 5.03% with an average of 1.13% for vendor A and 2.35% for vendor B, whereas automatic segmentation ranged more consistently (0.03 to 2.10%) with an average of 1.39% for vendor A and 0.98% for vendor B (Table 4).

**Intervendor variability** When comparing hippocampal volumes across scanners, we noticed that hippocampal volumes resulting from the automated process (but not the manual process) were systematically higher (mean difference = 9.5%, (7.88 to 11.72%)) when images were acquired with vendor B relative to vendor A (Fig. 2). This systematic difference in automated calculated hippocampal volumes between vendors resulted in on average better intervender CV for the manual process (2.55%) then for the automated process (4.39%) (Table 4).

Automatically traced hippocampal volumes were systematically higher than manually delineated volumes (Fig. 3).

### Diffusion imaging

#### Anisotropic diffusion phantom

Boxplots presenting the distribution of FA values in the defined ROIs are shown in Fig. 4. There was a high similarity between values obtained with the manual and automatic procedure. Intervendor CVs calculated from the manual and automated procedure were 3.83 and 3.45% respectively (Table 5). Intrascanner CVs ranged from 2.61 to 4.96%, while intravendor CVs range from 2.95 to 3.96% (Table 5).

#### Isotropic diffusion phantom

For each isotropic diffusion measurement, gradient correction factors were calculated using the RAPID pipeline. In Fig. 5a, correction factors are shown for each site and time point. The data from site 5, time point 2, were incorrectly acquired with 6 diffusion directions instead of the 12 directions that were prescribed. For site 6, time point 2, the correction factors could not be calculated due to the fact that the last two diffusion volumes were not acquired.

Gradient correction factors were above 0.97 for all sites and time points except for site 2, time point 3, indicating good scanner calibration. At sites 1 and 2 (Vendor A), the x-gradient was always 1 (meaning optimal calibration) while the y and z gradient were < 1. For sites 3, 5, and 6, gradient calibration varied between the x, y, and z gradients. The gradient correction factors were used to calculate calibrated ADC values. Tensor and ADC maps were then re-calculated. We did not find significant differences between corrected and uncorrected ADC data (paired *t* test) for any site or time point.

ADC values measured in a circular ROI in the middle slice are shown in Fig. 5b for both corrected and uncorrected datasets. Mean ADC ranged from 750 to  $871 \cdot 10^{-6} \text{ mm}^2/\text{s}$ , median ADC from 742 to  $871 \cdot 10^{-6} \text{ mm}^2/\text{s}$ .

#### Human data

**Intrascanner variability** The repeated scans of subject one with 1-month interval showed good test-retest reliability for diffusion metrics over time in most white matter tracts: for both TBSS and semi-automated tractography, average intrascanner CV ranged between 1.2 and 2.9% (Table 6). Mean FA values for the different tracts aggregated across scanners are shown in Table 7.

TBSS results showed average intrascanner CV across ROIs of 1.2 (MD) and 2.3% (FA) for vendor A and 1.1 (MD) and 2.5% (FA) for vendor B. For six out of eight individual tracts, CVs ranged from 0.1 to 2.1%. For the left and right cinguli, however, variability was higher and CVs ranged from 0.9 to 10%.

More variability was observed when using the semi-automated tractography process. While average CVs across tracts and raters were 2.2% (MD) and 2.9% (FA) for vendor A and 1.9% (MD) and 2.9% (FA) for vendor B, individual tracts showed more variability. For 6 out of 8 tracts, CVs ranged from 0.0 to 6.1% for vendor A and from 0.0 to 9.4% for vendor B with one exceptionally large CV of 13.7% for FA of the right SLF of subject 1, for one operator. Also here, a similar pattern was shown for the left and right cingulum with larger CVs ranging from 0.2 to 9.0%.

**Intravendor variability** When evaluating variability of DTI measures obtained on different scanners from the same



**Table 3** ACR phantom parameter estimates for the different timepoints and scanners

Center		Small ACR phantom (Siemens)												Large ACR phantom (Philips)											
		1	2	3	4	1	2	3	4	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4
Geometric accuracy (mm)		1	2	3	4	1	2	3	4	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4
Geometric accuracy (mm)	Localizer	100 +/- 2	+1.3	-0.4	-0.6	-0.1	0.3	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.4
	Slice 1 <sup>1</sup>	100 +/- 2	0.4	0.7	0.5	0.4	0.9	0.4	0.5	0.4	0.9	0.4	0.5	0.4	0.9	0.4	0.9	0.4	0.5	0.4	0.9	0.4	0.9	0.4	0.8
	Slice 1 <sup>2</sup>	100 +/- 2	1.1	-0.1	-0.1	0.1	1.2	-0.1	-0.1	0.1	1.2	-0.1	-0.1	0.1	1.2	-0.1	-0.1	0.1	0.7	0.1	0.7	-0.1	0.1	1.1	1.1
	Slice 3 <sup>1</sup>	100 +/- 2	0.4	0.4	0.4	0.7	0.1	0.7	0.7	0.1	0.1	0.7	0.7	0.1	0.1	0.1	0.1	0.7	0.1	0.1	0.3	0.1	0.1	1.1	1.7
Slice thickness accuracy (mm)	Slice 3 <sup>2</sup>	100 +/- 2	1.5	-0.1	0.3	0.3	0.2	-0.1	0.3	0.1	0.2	0.1	0.3	0.5	1.1	0.1	0.1	0.7	0.1	0.1	0.3	0.1	0.1	1.1	1.2
	Slice 3 <sup>3</sup>	100 +/- 2	1.4	0	0	0.3	0.2	0	0.3	0.5	0.2	0.3	0.3	0.5	1.1	0.1	0.1	0.7	0.1	0.1	0.3	0.1	0.1	1.1	1.2
	Slice 3 <sup>4</sup>	100 +/- 2	1.6	0	-0.2	0.5	0	-0.2	0.5	0.2	-0.2	-0.2	0.5	0.2	-0.2	-0.2	-0.2	-0.2	0.5	0.2	-0.2	-0.2	-0.2	0.9	0.9
	Slice 1	5 +/- 0.7	+0.5	+0.1	+0.1	+0.1	+0.5	+0.1	+0.1	+0.5	+0.5	+0.1	+0.1	+0.5	+0.5	+0.1	+0.5	+0.1	+0.1	+0.5	+0.1	+0.1	+0.5	+0.1	+0.5
Percent image uniformity (%)	Slice 1 <sup>5</sup>	<5.0	0	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1	-0.3	+0.1
	Slice 4	>82%	76.4 <sup>a</sup>	91.1	89.2	86.4	80.0 <sup>a</sup>	83.9	89.2	86.4	80.0 <sup>a</sup>	83.9	89.2	86.4	80.0 <sup>a</sup>	83.9	89.2	86.4	80.0 <sup>a</sup>	83.9	89.2	86.4	80.0 <sup>a</sup>	83.9	90.7
Percent signal ghosting (%)	Slice 5	≤2.5%	0.1	0.3	0.4	0.2	0.2	0.3	0.4	0.2	0.2	0.3	0.4	0.2	0.2	0.3	0.4	0.2	0.2	0.3	0.4	0.2	0.3	0.3	0.3
	Low-contrast object detectability	≥9	19	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
Center		3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2
Geometric accuracy (mm)	Localizer	148 +/- 2	-1.8	-1.0	-1.4	-1.3	-1.3	-1.3	-1.4	-1.3	-1.3	-1.3	-1.4	-1.3	-1.3	-1.3	-1.3	-1.3	-1.4	-1.3	-1.3	-1.3	-1.3	-1.3	-1.3
	Slice 1 <sup>1</sup>	190 +/- 2	-0.6	-0.3	-0.6	-0.3	-0.3	-0.1	-0.6	-0.3	-0.1	-0.1	-0.6	-0.3	-0.1	-0.1	-0.1	-0.1	-0.6	-0.3	-0.1	-0.1	-0.1	-0.1	-0.1
	Slice 1 <sup>2</sup>	190 +/- 2	+0.2	+0.5	+0.2	+0.5	-0.1	+0.8	+0.2	+0.5	-0.1	+0.2	+0.2	+0.5	-0.1	+0.2	+0.5	-0.1	+0.2	+0.5	-0.1	+0.2	+0.5	-0.1	-0.1
	Slice 5 <sup>1</sup>	190 +/- 2	-0.3	-0.9	-0.1	-0.6	-0.9	0	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3
Slice thickness accuracy (mm)	Slice 5 <sup>2</sup>	190 +/- 2	+0.2	+0.5	+0.2	+0.5	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5	+0.2	+0.5
	Slice 5 <sup>3</sup>	190 +/- 2	+0.3	-0.4	-0.1	-0.1	-0.1	+0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3	-0.1	-0.6	-0.3
	Slice 5 <sup>4</sup>	190 +/- 2	+0.1	-0.3	+0.3	+0.6	+0.6	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7
	Slice 1	5 +/- 0.7	+0.6	+0.5	+0.6	+0.6	+0.6	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7	+0.7
Percent image uniformity (%)	Slice 1 <sup>5</sup>	<5.0	+0.8	0	0	0	0	+0.2	+0.3	0	0	+0.2	+0.3	0	0	+0.2	+0.3	0	+0.2	+0.3	0	+0.2	+0.3	0	+0.2
	Slice 11 <sup>5</sup>	<5.0	+0.5	-0.6	-0.1	-0.1	-0.1	0	-0.1	+0.1	-0.1	-0.1	+0.1	-0.1	-0.1	-0.1	+0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Percent signal ghosting (%)	Slice 7	>82%	80.5 <sup>a</sup>	80.8 <sup>a</sup>	81.7 <sup>a</sup>	80.2	82.9	83.1	83.4	82.1	76.3 <sup>a</sup>	79.2 <sup>a</sup>	74 <sup>a</sup>	79.7	78.5 <sup>a</sup>	77.8 <sup>a</sup>	76.7 <sup>a</sup>	77.5	76.7 <sup>a</sup>	77.8 <sup>a</sup>	76.7 <sup>a</sup>	77.5	76.7 <sup>a</sup>	77.5	77.5
	Low-contrast object detectability	≥37	39	39	39	39	39	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37

<sup>a</sup> Measurement does not satisfy ACR criteria<sup>1</sup> Horizontal dimension<sup>2</sup> Vertical dimension<sup>3</sup> Diagonal dimension, from upper right to lower left<sup>4</sup> Diagonal dimension, from upper left to lower right<sup>5</sup> Plus sign if the slice is displaced superiorly, minus sign if the slice is displaced inferiorly



**Table 4** Intrascanner, intravendor, and intervender coefficients of variation (%), delineated both manually and automatically

											
		Manual		Automatic		Manual		Automatic		Manual	Automatic
	Vendor	L HC	R HC	L HC	R HC	L HC	R HC	L HC	R HC	Mean CV	Mean CV
Intrascanner 1 (2 scans)	Siemens	2.50	1.12	0.18	2.62	-	-	-	-	1.81	1.40
Intrascanner 3 (2 scans)	Philips	0.69	1.72	0.90	0.71	-	-	-	-	1.20	0.81
Intrascanner 1 (5 scans)	Siemens	-	-	-	-	-	-	1.16	0.72	-	0.94
Intravendor (Inter scanner 1 & 2)	Siemens	0.00	0.00	1.15	1.32	0.73	3.79	2.10	0.99	1.13	1.39
Intravendor (Inter scanner 5 & 6)	Philips	2.06	1.55	0.68	0.79	5.03	0.99	0.03	1.04	2.41	0.64
Intravendor (Inter scanner 3 - 6)	Philips	1.76	1.47	1.04	0.88	3.73	2.44	0.96	1.02	2.35	0.98
Intervendor (Scanner 1 - 6)	Siemens/Philips	1.68	1.82	4.37	5.45	4.16	2.54	3.79	3.96	2.55	4.39

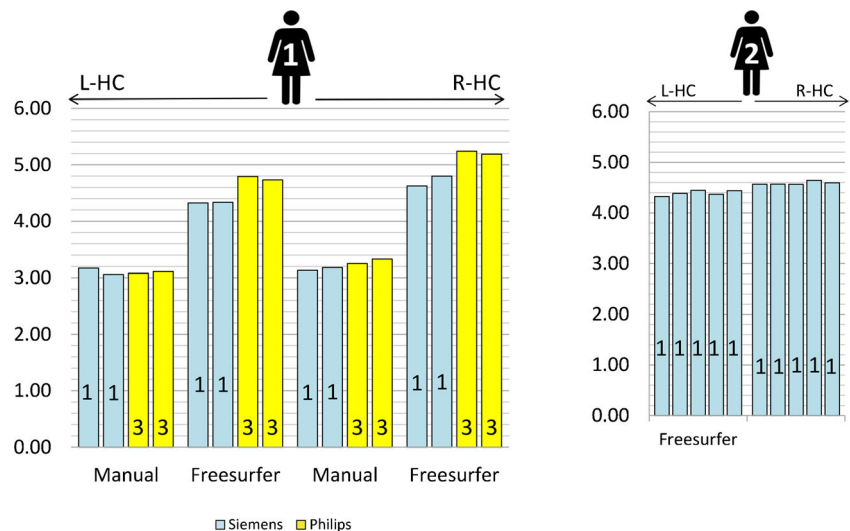
vendor, we observed good reliability in most white matter tracts. For TBSS and semi-automated tractography, average intravendor CV ranged from 2.3 to 6.6% for vendor A and from 1.2 to 3.5% (equal number of sites) for vendor B.

TBSS results showed average intravendor CV across ROIs of 2.8% (MD) and 6.6% (FA) for vendor A and slightly lower CVs for vendor B, 1.2% (MD) and 1.7% (FA) when including equal number of sites as vendor A, and 1.8% (MD) and 2.5% (FA) when including all vendor B sites. For six out of eight individual tracts, CVs ranged from 0.0 to 6.5%. For the left and right cinguli, however, also here variability was higher and CVs ranged from 0.3 to 13.2% with one exceptionally high CV for FA in the right cingulum of subject 1 of 29.8% for vendor A.

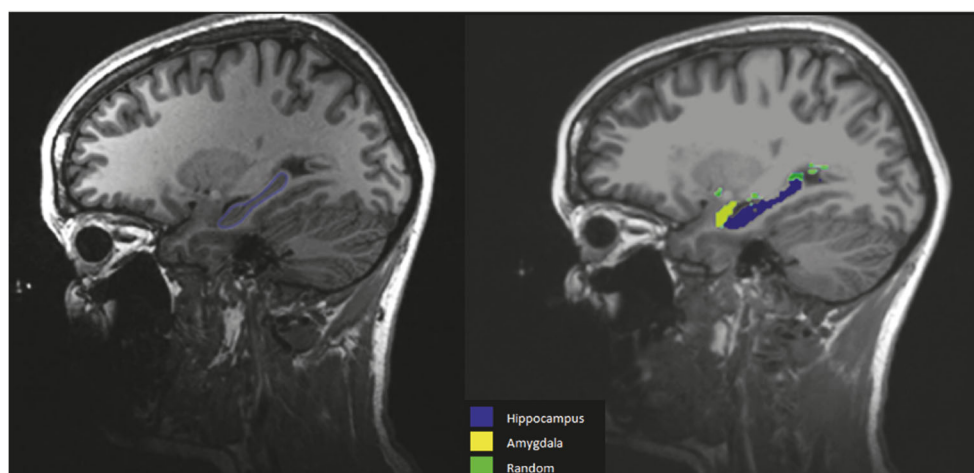
Semi-automated tractography showed average CVs across tracts, raters, and subjects of 3.2% (MD) and 4.4% (FA) for vendor A and 2.3% (MD) and 3.1% (FA) for vendor B (equal number of sites). Individual tracts, however, showed more variability. For six out of eight individual tracts, CVs ranged from 0.0 to 8.7%. For the left and right cinguli, CVs showed more variability and ranged from 1.3 to 15.7% with one exceptionally high CV of 23.3% (vendor B, subject 2, operator MR, FA left cingulum).

**Intervendor variability** Average interscanner CV ranged from 4.2 to 6.0% (TBSS) and from 4.2 to 7.8% (semi-automated tractography) (Table 6). Average FA values for the different tracts and scanners are shown in Table 7.

**Fig. 1** Left panel: left and right hippocampal volumes ( $\text{cm}^3$ ) of the repeated scans of subject 1 with 1 month interval acquired in centers 1 and 3, delineated both manually and automatically. Right panel: left and right hippocampal volumes ( $\text{cm}^3$ ) of five scans of subject 2 acquired on the same day in center 1



**Fig. 2** Manual versus automated (Freesurfer) delineation surface of the hippocampus



TBSS results showed average interscanner CV across ROIs and subjects of 5.0 (MD) and 5.5% (FA). CVs of individual tracts ranged from 1.0 to 7.9% for six out of eight tracts. For the cinguli, however, CVs were higher ranging from 2.1 to 10.5% with one exceptionally high CV for the FA of the right cingulum of subject 1 (15.6%). Higher variability was observed when using the semi-automated tractography process. While average CVs across tracts and raters were 5.1% (MD) and 6.7% (FA), individual tracts showed more variability. For 6 out of 8 individual tracts, CVs, ranged from 2.2 to 9.3 with one CV higher than 10% (right SLF of subject 1 for one of the operators). For the left and right cingulum, reproducibility was less satisfactory with CVs ranging from 2.9 to 17.2%.

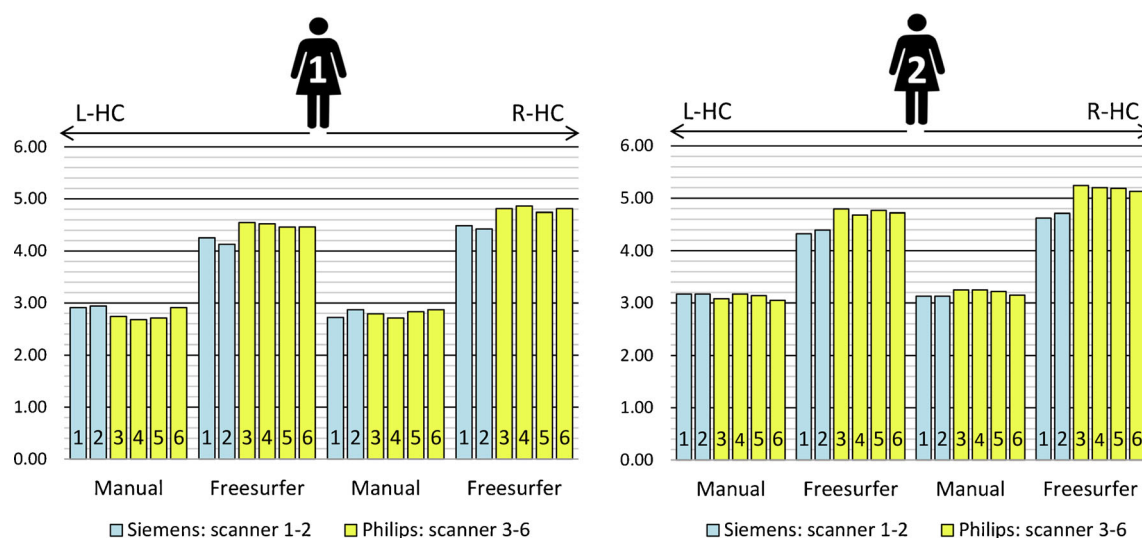
Interobserver reliability was high for forceps major, forceps minor, left and right SLF, and left cingulum with ICC ranging from 0.85 to 0.93 (all  $p$ s < .001). For the right cingulum, ICC was 0.74 ( $p$  < .005). Interobserver reliability was low for both

right (ICC: 0.52 ( $p$  < .05)) and left CST (ICC was 0.32 ( $p$  < .1)).

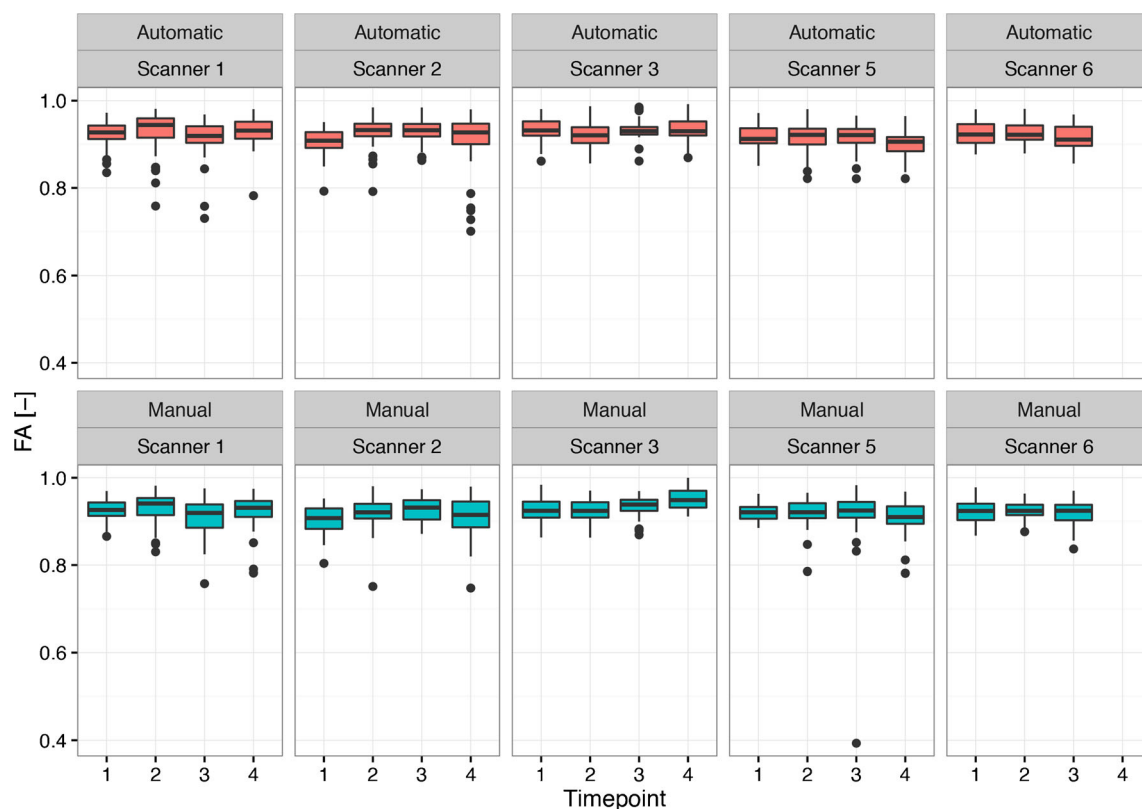
## Resting-state fMRI

### Temporal SNR measured on the isotropic diffusion phantom

tSNR was calculated for each scan of the isotropic diffusion phantom, except for site 4 and site 5 time point 2, where the data could not be analyzed due to incorrect acquisition. Temporal SNR ranged between 171.2 and 254.4 (arbitrary units). These values are in the same range as reported earlier in FBIRN phantoms [2]. tSNR values in vendor A were notably smaller than in vendor B, while longitudinal variability appeared to be more stable in vendor A, where the difference in tSNR between the two time points was  $-1.2\%$ , (site 1) and  $3.2\%$  (site 2) versus  $10.5\%$  (site 3) and  $7.7\%$  (site 6) for vendor B (Table 8).



**Fig. 3** Left and right hippocampal volumes ( $\text{cm}^3$ ) of subjects 1 and 2 acquired in centers 1–6 at time point 1, delineated both manually and automatically



**Fig. 4** Boxplots presenting the distribution of FA values in the defined ROIs in the anisotropic diffusion phantom with the automatic and manual procedure (upper and lower panel, respectively)

### Human data

The DMN could be reliably extracted across sites and time points using a 20 component ICA. Spatial maps of the coherent resting-state fluctuations are shown in Supp. Fig. 3.

### Discussion

While multicenter MRI studies are becoming increasingly important to collect data from larger groups of patients in a

reasonable time frame, potential differences due to scanner-dependent variability might confound true changes in brain structure and function.

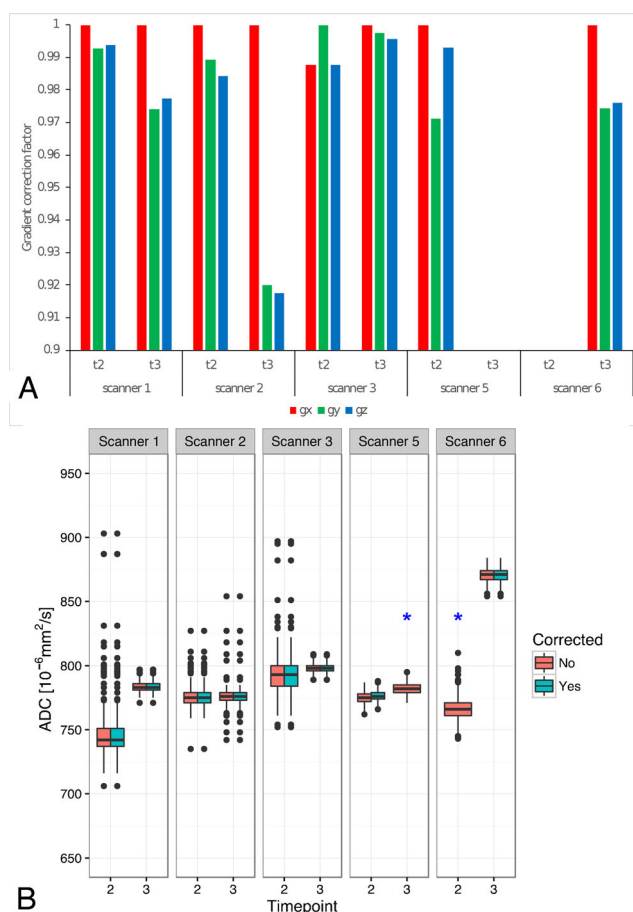
We investigated multi-center variability and longitudinal consistency of multimodal (T1-w structural, DTI and rsfMRI) imaging measurements by assessing variability across scanners from six participating sites and two different vendors using aligned MRI acquisition protocols. Data from three different phantoms and two healthy volunteers scanned at several time points were included in this study. We report average coefficients of variation (CV) below 5% for intrascanner, intravendor, and intervender reproducibility for both structural and diffusion imaging metrics, except for diffusion metrics obtained from tractography with average CVs ranging up to 7.8%. Additionally, resting-state fMRI showed stable temporal SNR and reliable generation of subjects' DMN across vendors and time points.

**Table 5** Coefficients of variation (CV) of manually and automatically delineated ROI of the anisotropic diffusion phantom

	Manual CV FA (%)	Automatic CV FA (%)
Intrascanner 1	3.83	3.78
Intrascanner 2	3.90	4.12
Intrascanner 3	2.85	2.79
Intrascanner 5	4.96	2.96
Intrascanner 6	2.61	2.79
Intravendor Siemens (center 1 and 2)	3.88	3.96
Intravendor Philips (center 3–6)	3.78	2.95
Intervendor Siemens and Philips (center 1–6)	3.83	3.45

### Structural MRI

We studied structural imaging variability across scanners and time points by assessing (1) geometrical deformations using ACR phantom scans and (2) hippocampal volume estimates from two healthy volunteers (the hippocampi are the area of interest for our ongoing HA-PCI study,



**Fig. 5** **a** Gradient correction factors for x, y, and z gradients as calculated by the RAPID method in all sites and time points. For site 5 time point 3 and site 6 time point 2, the correction factors could not be calculated due to incorrect DWI acquisition. **b** Boxplots of corrected and uncorrected ADC in a circular ROI of the middle slice in the isotropic phantom. Boxes show median and first quartiles of the data, while the whiskers show the third quartile. Dots indicate outliers. Mean ADC ranged from 751 to 871 across sites, and median ADC ranged from  $742$  to  $871 \times 10^{-6} \text{ mm}^2/\text{s}$

NCT01780675) using both manual delineation procedures and automatic freesurfer morphometry.

To be consistent with the clinical scan sequences, we acquired a sagittal 3D-T1 scan of the ACR phantom instead of the ACRs prescribed axial T1 and T2. All participating sites had scanners operating conforming to the ACR recommended acceptance criteria except for image uniformity. Several factors could influence image uniformity such as signal ghosting [31], phantom positioning, and non-linearity of the gradient system. It is unlikely, however, that one of the above-listed factors was the source here. Signal ghosting—which can be caused by vibrations of the scanner table and phantom—was low in all participating sites and sites 1 and 2 with the highest signal ghost percentages (due to docked tables) seem to suffer the least from inhomogeneous signal intensity. This might be explained by the fact that the phantom used for sites 1 and 2 was smaller, which might have led to lower RF penetration problems. Additionally, the phantom was carefully positioned

in each site by the same radiographer and checked for centrality in the head coil. Non-linearity of the gradient systems on the other hand, should have been solved after recalibration during service. We believe that lower image uniformity in several sites could be related to the use of correction methods and head coils designed to produce fairly uniform signal near the middle of the coil when loaded with an inhomogeneous human head rather than a homogeneous phantom [32].

We achieved consistent hippocampal volume measures across time points and vendors for both manual and automatic procedures. Automatically traced hippocampal volumes, however, were systematically higher than manually delineated volumes. This overestimation is well-described in the literature [33]. To longitudinally compare hippocampal volumes before and after radiotherapy in patients, the automatic procedure seems to be a good alternative for the time-consuming manual delineation. With mean intrascanner CVs around 1%, one would expect the technique to be sensitive enough to distinguish subtle differences in hippocampal volume. These results are in line with previously published findings on Freesurfer [5, 34]. In the case of manual delineation, the intervendor variability is often lower than the intravendor and even intrascanner variability. This suggests that manual delineation is relatively insensitive to differences caused by different hardware configurations and that a manual delineation protocol allows grouping of results of these two vendors. Hippocampal volumes obtained with Freesurfer on the other hand are systematically higher for vendor B than for vendor A, resulting in higher intervendor variabilities; a similar systematic between-vendor bias for Freesurfer segmentations was described by Jovicich et al. [35]. This should be taken into account in future cross-sectional comparisons, e.g., by including vendor type as a stratification factor in RCTs.

## Diffusion weighted MRI

Variability in diffusion metrics across scanners and time points were assessed in manually and automatically defined ROIs from an isotropic and anisotropic diffusion phantom and two human subjects.

Mean ADC values obtained from the isotropic diffusion phantom varied between  $750.2$  and  $860.6 \times 10^{-6} \text{ mm}^2/\text{s}$  across sites and time points. That is a difference of 15% between the lowest and highest ADC value. Belli et al [15] report on a multi-center DWI doped water phantom study, where the largest variation from the standard was 9.3%, and 80% of the ADCs measured were within 5% of the reference value when corrected to a temperature of  $20^\circ \text{C}$ . The ADC versus temperature reference standard was obtained on a NMR spectrometer. Another study reports ADC variations up to 3% near isocenter in an ice-water phantom [36]. The higher variation in mean ADC values in the present study could be related to incorrect temperature measurements or phantom positioning.

**Table 6** Coefficient of variation for FA and MD values resulting from TBSS and tractography analysis

		TBSS				Tractography							
		Subject 1		Subject 2		Subject 1		Subject 1		Subject 2		Subject 2	
						Operator SD		Operator MR		Operator SD		Operator MR	
		FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)
Test/retest Siemens	Forceps major	0.7	0.1			1.4	0.7	0.7	5.8				
	Forceps minor	1.3	0.8			3.9	1.7	4.9	1.5				
	L CST	0.1	1.7			1.6	1.0	0.5	0.3				
	R CST	0.4	0.6			1.5	0.3	2.5	0.0				
	L tSLF	0.8	2.1			1.7	0.5	3.0	0.5				
	R tSLF	1.6	0.5			6.1	0.2	4.7	0.3				
	L cing	6.6	2.5			2.3	9.0	2.8	6.0				
	R cing	6.7	1.1			8.4	4.6	0.4	2.9				
	Mean	2.3	1.2			3.4	2.2	2.5	2.2				
Test/retest Philips	Forceps major	0.6	1.6			1.1	1.7	3.4	1.9				
	Forceps minor	1.5	1.4			1.2	0.5	0.4	0.6				
	L CST	0.2	0.4			1.2	1.2	0.0	0.5				
	R CST	0.7	0.7			4.1	2.1	0.4	0.2				
	L tSLF	0.9	1.5			0.9	2.6	0.5	2.5				
	R tSLF	2.1	1.0			13.7	0.7	9.4	1.0				
	L cing	10.0	0.9			4.1	3.6	4.6	0.5				
	R cing	4.4	1.5			1.7	2.4	0.2	7.6				
	Mean	2.5	1.1			3.5	1.9	2.4	1.9				
Intra Siemens	Forceps major	5.8	0.0	1.6	0.3	3.6	3.4	2.7	1.0	3.7	0.6	3.2	0.7
	Forceps minor	2.7	2.7	4.4	2.8	0.4	3.7	2.2	3.4	2.0	4.4	1.3	4.6
	L CST	2.5	3.3	0.9	4.4	0.0	3.4	1.4	4.0	0.3	3.5	2.5	2.8
	R CST	6.5	0.1	5.5	1.5	3.3	1.3	3.0	1.0	4.2	1.9	6.1	0.2
	L tSLF	2.6	5.6	2.8	4.0	1.1	3.7	1.0	4.1	3.4	4.2	7.5	8.7
	R tSLF	2.6	0.6	2.2	0.7	6.5	1.3	5.5	1.5	1.5	0.7	8.1	0.4
	L cing	0.3	1.	7.2	2.5	7.7	3.3	3.9	1.3	13.6	10.0	10.4	2.9
	R cing	29.8	9.0	13.2	2.5	5.1	4.6	3.8	6.5	13.2	2.6	9.8	5.0
	Mean	6.6	2.8	4.7	2.3	3.5	3.1	3.0	2.8	5.3	3.5	6.1	3.2
Intra Philips	Forceps major	0.5	1.2	0.1	1.1	6.9	2.7	3.4	3.7	0.9	6.2	4.7	5.4
	Forceps minor	0.4	1.3	2.2	3.0	1.4	0.3	0.3	0.7	3.9	2.1	8.3	1.4
	L CST	0.9	1.0	0.1	0.6	4.1	4.0	0.6	1.3	1.2	0.0	2.0	0.1
	R CST	0.3	1.3	1.5	1.0	1.7	1.3	1.7	1.1	5.1	0.2	0.7	0.2
	L tSLF	0.9	0.2	2.0	0.1	5.7	1.5	6.2	1.8	2.6	0.2	1.6	0.4
	R tSLF	2.5	0.3	3.4	0.9	0.1	0.6	2.0	1.3	0.5	0.2	0.5	0.4
	L cing	3.2	0.9	3.7	2.2	1.2	4.2	6.2	6.6	0.1	5.3	3.4	6.8
	R cing	5.1	3.5	1.7	0.3	4.5	5.2	7.0	1.	2.9	3.6	6.9	3.4
	Mean	1.7	1.2	1.8	1.2	3.2	2.5	3.4	2.2	2.1	2.2	3.5	2.3
Intra Philips 3 measurements	Forceps major	1.9	2.6	1.6	1.1	4.9	2.0	2.6	2.7	0.6	4.7	3.5	5.0
	Forceps minor	0.7	0.9	1.6	2.2	2.3	6.8	0.3	0.9	3.0	1.5	5.9	3.9
	L CST	0.8	2.6	0.5	1.8	3.6	3.0	1.1	1.6	3.8	2.8	2.1	0.3
	R CST	0.6	1.7	1.8	1.8	2.4	1.7	1.4	0.8	3.7	0.2	0.5	0.1
	L tSLF	1.7	0.7	1.6	1.2	4.1	1.7	4.4	1.7	3.9	0.4	1.6	0.6
	R tSLF	2.3	0.4	2.8	1.3	5.4	1.1	4.5	1.7	0.8	0.4	0.9	0.3
	L cing	4.8	2.1	7.7	3.3	2.0	3.3	4.9	5.0	14.4	8.9	23.3	8.7
	R cing	7.1	3.8	1.6	0.3	5.0	7.6	9.7	15.7	3.4	6.5	7.2	3.2



**Table 6** (continued)

		TBSS		Tractography									
		Subject 1		Subject 2		Subject 1		Subject 1		Subject 2		Subject 2	
						Operator SD		Operator MR		Operator SD		Operator MR	
		FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)	FA (%)	MD (%)
Intervendor 5 measurements	Mean	2.5	1.8	2.4	1.6	3.7	3.4	3.6	3.8	4.2	3.2	5.6	2.8
	Forceps major	3.7	2.3	1.9	3.5	4.6	2.2	2.9	3.8	3.0	3.4	3.5	3.7
	Forceps minor	4.5	5.3	4.5	6.2	6.6	4.9	7.3	4.7	7.8	6.4	6.9	6.1
	L CST	2.5	2.6	1.0	3.8	4.6	4.0	4.1	3.8	4.5	6.4	3.5	3.4
	R CST	4.6	2.9	5.1	4.2	5.3	2.8	6.3	2.6	5.1	3.3	6.9	4.0
	L tSLF	4.6	7.9	3.8	6.5	6.1	5.3	6.8	5.5	6.3	5.4	5.0	5.9
	R tSLF	3.6	5.6	5.5	7.4	11.1	4.7	8.5	4.6	5.6	6.4	9.3	7.4
	L cing	9.0	2.1	10.5	10.4	5.2	2.9	6.2	3.6	12.3	6.5	17.2	6.9
	R cing	15.6	5.2	8.2	4.5	7.0	6.5	10.4	11.1%	5.8	7.3	10.3	6.3
	Mean	6.0	4.2	5.1	5.8	6.3	4.2	6.6	5.0	6.3	5.6	7.8	5.5

We found that the temperatures reported by some sites in our study are implausible, either by incorrect temperature measurements and/or the phantom not being in thermal equilibrium with the scanner room (i.e., the phantom was likely not

**Table 7** Mean FA values resulting from TBSS and tractography analysis aggregated across scanners

FA	TBSS				Tractography			
	Subject 1		Subject 2		Subject 1		Subject 2	
	Mean	std	Mean	std	Mean	std	Mean	std
Forceps major	0.50	0.02	0.50	0.01	0.54	0.02	0.62	0.02
Forceps minor	0.52	0.02	0.53	0.02	0.53	0.04	0.54	0.04
L CST	0.56	0.01	0.61	0.01	0.55	0.02	0.56	0.02
R CST	0.54	0.02	0.60	0.03	0.55	0.03	0.56	0.03
L tSLF	0.48	0.02	0.53	0.02	0.49	0.03	0.53	0.03
R tSLF	0.50	0.02	0.52	0.03	0.44	0.04	0.49	0.04
L cing	0.40	0.04	0.40	0.04	0.27	0.02	0.33	0.05
R cing	0.39	0.06	0.46	0.04	0.30	0.03	0.34	0.03
Mean	0.49	0.03	0.52	0.02	0.46	0.03	0.50	0.03
MD	Mean	std	Mean	std	Mean	std	Mean	std
Forceps major	0.74	0.02	0.72	0.03	0.95	0.03	0.82	0.03
Forceps minor	0.71	0.04	0.70	0.04	0.76	0.04	0.74	0.05
L CST	0.66	0.02	0.63	0.02	0.69	0.03	0.66	0.03
R CST	0.67	0.02	0.64	0.03	0.69	0.02	0.68	0.02
L tSLF	0.67	0.05	0.64	0.04	0.67	0.04	0.64	0.04
R tSLF	0.66	0.04	0.64	0.05	0.67	0.03	0.65	0.04
L cing	0.71	0.01	0.69	0.07	0.82	0.03	0.83	0.06
R cing	0.68	0.04	0.65	0.03	0.80	0.07	0.80	0.05
Mean	0.69	0.03	0.66	0.04	0.76	0.03	0.73	0.04

*CST* corticospinal tract, *tSLF* temporal component of the superior longitudinal fasciculus, *cing* hippocampal part of the cingulum

stored in the scanner room as requested). A second source of variation in ADC across sites could be the positioning of the phantom in the scanner. An offset from isocenter can result in a variation of ADC values by as much as 20% [37]. Correction for gradient mismatch, however, showed only very small differences between the x, y, and z gradient and did not result in significant changes in ADC values. This is in line with another study using an n-Undecane isotropic diffusion phantom reporting very small differences between gradients [38]. Based on these findings, we suggest that it will not be necessary to schedule additional isotropic diffusion phantom scans along with each patient scan to calculate gradient correction factors for use in future analysis. Instead, a regular half-yearly QA scan and after scanner maintenance should be sufficient to follow-up on gradient stability in each participating scanner.

Anisotropic phantom scans showed intervender coefficients of variation of FA and MD values below 4%. These values are lower than CVs reported for a similar anisotropic phantom by Teipel et al. [10]. In that study, however, both 1.5 and 3 T scanners were included and scan sequence parameters showed more variability between sites.

There was a high similarity between values obtained with the manual and automatic procedure. This suggests that the automatic segmentation is a good alternative for the manually positioned ROIs. Further, the intravender and intra-site variability was in some cases higher than the intervender

**Table 8** Summary of tSNR values for the different sites

tSNR (–)	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
t2	188.1	171.2	254.4	217.3		219.6
t3	190.3	176.6	227.6		192.8	202.6

t2 timepoint 2, t3 timepoint 3

variability. This implies only a modest effect of differences in hardware across sites and vendors relative to intra-site variability across time points, suggesting that the selected DTI protocol is suitable for multi-center longitudinal studies.

For the human subjects, intrascanner variability (longitudinal test) for DTI across ROIs/tracts was below 4% for both TBSS and semi-automatic tractography, indicating that both techniques are generally suitable for longitudinal study designs. Semi-automatic tractography results were somewhat less reproducible than TBSS results. The interscanner variability (multi-center test) generally showed higher levels of variability than the longitudinal test. Within the same vendor (intravendor variability), TBSS results were below 7 and 3% across ROIs for vendors A and B, respectively. This implies that interscanner variability is somewhat lower for vendor B than vendor A using TBSS, also when scanners with different configurations are taken into account for vendor B. Similar results to ours were observed by Grech-Sollars et al. [39], where the intra- and interscanner reproducibility in white matter had CVs ranging between 1 and 7.4%. With respect to the tractography results across tracts, intravendor variability remained below 7% for vendor A and 6% for vendor B.

When considering results from all five sites that contributed DTI data (intervendor variability), variability for TBSS data across ROIs was up to 6% with variability for separate tracts (except the cinguli) remaining below 8%. For tractography, across-tract variability was below 8% whereas for separate tracts except the cinguli variability remained below 10% with one exception. The fact that we observed higher variability for the cinguli could be related to the smaller size of the tract. Larger tracts such as the cortico-spinal tract and the corpus callosum contain more voxels, which can result in lower variability.

Although overall both TBSS and tractography yielded good results, TBSS generally showed somewhat lower variability than tractography, which is in agreement with a previous report including both ROI-based and tractography data [14]. Likewise, we observed that ROIs and tracts with higher FA values generally correspond to lower variability across sessions and scanners [10, 40]. This indicates that fiber tracts with one main direction of fibers have better reproducibility than fiber tracts with crossing fibers. This could also explain why we observe higher reproducibility for the anisotropic phantom in comparison with the human subjects.

Our findings indicate that, when focusing on changes over time within subjects that are scanned on the same scanner, both TBSS and tractography can be used as reliable tools to study effects of PCI on white matter integrity in various white matter tracts/ROIs. Also, Takao et al. reported relatively stable FA and MD measures with repeated scans obtained on the same scanner using TBSS

[13]. As expected, between-scanner differences were more pronounced resulting in higher variability, particularly for the smaller tracts such as the cingulum when using tractography.

## Resting-state functional MRI

Temporal signal-to-noise ratio (tSNR) is an important factor influencing fMRI reliability in resting-state studies and gives a good measure of functional image quality. We evaluated phantom tSNR, providing a well-accepted measure of temporal stability having direct relevance to rsfMRI image analysis [41]. Furthermore, Jovicich et al [4] showed a positive correlation between brain and phantom tSNR measures, suggesting that in vivo, inter-site tSNR differences are primarily driven by MRI system parameters. In contrast to structural and diffusion imaging measures, metrics from in vivo resting state fMRI data, such as connectivity strength or spatial extent of specific brain networks, strongly depend on the cognitive and physiological status of the subject during image acquisition and are therefore not ideal for testing reliability of differences in imaging hardware across centers. Huang et al [12] showed that the measurement error introduced by the different scanners is small relative to the noise introduced by the different subjects. Therefore, in the present study, we focused on phantom tSNR measurements and qualitative visual assessment of in-vivo rsfMRI DMN extraction using a data-driven method ICA. ICA-based methods have been reported to show a high level of consistency in detecting the default mode network during rest [42] and to show similar results as seed-based methods in a group of healthy subjects [43]. Additionally, ICA appears to yield more reliable DMN measurements relative to seed-based analysis [44].

Our measurement showed that tSNR values in a phantom for the different scanners across time points were high, allowing reliable rsfMRI analysis. Across sites median tSNR was around 200, which is in line with previous studies [2, 4]. However, tSNR values in vendor A were smaller than in vendor B, which might be related to differences in image reconstruction and preprocessing, which is different between vendors, leading to possible differences in measured SNR values. Within vendor B, scanners with the highest number of receive channels (32-channels head-coil) had higher tSNR than scanners with 8 and 16 channel coils, which is as expected. In contrast, longitudinal variability appeared to be more stable in vendor A than vendor B. Factors such as differences in phantom positioning and phantom temperature instabilities could have played a role.

Despite the slight differences in tSNR and variability across vendors, we could reliably extract the DMN consistently across sites and time points using a 20-component ICA.

## Possible impact of scanner maintenance and proposed framework for QA monitoring

Scanner maintenance can have a major impact on image quality. During preventive maintenance, the system will be recalibrated, restoring possible drifted settings. The set of parameters that are typically checked and corrected can have an influence on the image geometry, temporal stability, and overall image quality. In the lifetime of an MRI scanner, several system parameters possibly need to be recalibrated yearly. The irradiated RF power of the RF amplifiers and the transmit body coil can change over time, resulting in incorrect flip angles being generated, leading to signal intensity changes in the images. The main field homogeneity can degrade over time due to small particles (possibly containing iron) ending up in the scanner hardware. During the period of a longitudinal study, broken parts of the scanner might need replacement, which can also influence scanner performance and/or image quality.

Systematic differences between scans acquired at different time points, resulting from scanner maintenance interventions, could be misinterpreted as real brain changes. It is, therefore, crucial for longitudinal multi-center studies to have a good QA monitoring system in place to assess image quality reproducibility on a regular basis.

We implemented a Quality Assurance framework where the traveling ACR phantom and both the isotropic and traveling anisotropic diffusion phantoms are scanned in 6-month intervals by a local trained radiographer. Based on the acquired phantom scans, we will follow-up reproducibility of measures obtained from the structural, diffusion, and rsfMRI scans twice a year. The ACR phantom is used to assess structural imaging variability, the isotropic diffusion phantom to assess the stability of the diffusion gradients and to evaluate differences in temporal SNR of the rsfMRI scans, and the anisotropic diffusion phantom to evaluate differences in DTI measures. The obtained measures in combination with an accurate registration of all scanner maintenance events of the included scanners will allow for a systematic follow-up of scanner maintenance related variations in image quality. This will be critical for subsequent statistical analysis and interpretation of longitudinal multi-center neuroimaging study results.

## Methodological considerations

A major strength of this study is the combined use of both traveling human and physical phantoms, which were scanned at the different sites and time points. The use of traveling phantom scans guarantees minimal variability related to physical or anatomical differences of the phantom and, therefore, implies better reflection of intrinsic scanner variations. This was an advantage when harmonizing the scan protocols between the different sites. This approach permitted to compare

scans with minimal inter-individual differences of the subjects/phantoms resulting in low interscanner variability of the different protocols.

We recognize, however, that there are some limitations in this study. For budget reasons, we could only include two traveling human phantoms in this study who were scanned at different time points in all sites. To have a better estimate of the variance of a single scanner, future studies should possibly include larger sample sizes scanned at more time points. For rsfMRI, this is even more important, as intra-individual differences between different time points (both physically and mentally) can be high, resulting in high variability in resting-state activity even when scanned on the same scanner. To take this into account in the present study, we assessed temporal SNR based on physical phantom scans acquired at several time points.

Additionally, the anisotropic diffusion phantom used in this study had a high intrinsic FA value, allowing higher reproducibility. White matter bundles in the human brain, however, can have both high (e.g., corpus callosum) and lower FA values. Ideally, another anisotropic diffusion phantom should be added to also assess reproducibility of fibers with low FA values. Furthermore, the QA procedure for assessing structural imaging variability could be strengthened by adding a phantom with an irregular shape of known volume inside.

We recognize that our scan protocols are not fully harmonized between vendors. Slight differences in acquisition parameters will always remain because of the differences in hard- and software between vendors. Furthermore, because the scan protocol is aimed at patients undergoing irradiation therapy, we are limited to using the MRI scanners available in the hospitals where the patients are treated; we, therefore, cannot fully control the hard- and software available to this research project, and this introduces more variability as when we would be able to use dedicated resources.

Finally, we would like to note that the results of this study cannot be generalized to acquisition protocols other than those used and described in this manuscript.

## Conclusion

With the implemented quality assurance program, the results of this study do show to what extent variability can be attributed to procedural inaccuracies and the use of different scanners. The results also show that variability between sites and between scanners is of the same order as intra-site and intrascanner variability. Although some systematic differences were found, which are unavoidable due to differences in scanner hardware and MRI sequence implementations, this will have less influence in finding changes over time. Therefore, the conclusion is that, within the limits of variability that we provide in this paper, the longitudinal MRI

follow-up of patients after PCI, with or with hippocampal avoidance, can be safely conducted in a multi-center, multivendor set-up, and that results acquired from pooling of the data will not be confounded if the necessary precautions for quality are taken into account.

**Acknowledgements** We are grateful to Dr. Bram Stieltjes (Universitätsspital Basel, CH) for his advice and help with the DTI anisotropic phantom.

## Compliance with ethical standards

**Funding** This study was funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (Grant Number IWT 130262), the Vlaamse Liga Tegen Kanker (VLK) and the Dutch Cancer Society.

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

## References

1. Van Horn JD, Toga AW (2009) Multisite neuroimaging trials. *Curr Opin Neurol* 22:370–378. <https://doi.org/10.1097/WCO.0b013e32832d92de>
2. Friedman L, Glover GH, Fbirm C (2006) Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *NeuroImage* 33:471–481. <https://doi.org/10.1016/j.neuroimage.2006.07.012>
3. Fox RJ, Sakaie K, Lee JC, Debbins JP, Liu Y, Arnold DL, Melhem ER, Smith CH, Philips MD, Lowe M, Fisher E (2012) A validation study of multicenter diffusion tensor imaging: reliability of fractional anisotropy and diffusivity values. *AJNR Am J Neuroradiol* 33:695–700. <https://doi.org/10.3174/ajnr.A2844>
4. Jovicich J, Minati L, Marizzoni M, Marchitelli R, Sala-Llonch R, Bartrés-Faz D, Arnold J, Benninghoff J, Fiedler U, Roccatagliata L, Picco A, Nobili F, Blin O, Bombois S, Lopes R, Bordet R, Sein J, Ranjeva JP, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargalló N, Ferretti A, Caulo M, Aiello M, Cavaliere C, Soricelli A, Parnetti L, Tarducci R, Floridi P, Tsolaki M, Constantinidis M, Drevelegas A, Rossini PM, Marra C, Schönknecht P, Hensch T, Hoffmann KT, Kuijper JP, Visser PJ, Barkhof F, Frisoni GB, PharmaCog Consortium (2016) Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fMRI study. *NeuroImage* 124:442–454. <https://doi.org/10.1016/j.neuroimage.2015.07.010>
5. Jovicich J, Marizzoni M, Sala-Llonch R, Bosch B, Bartrés-Faz D, Arnold J, Benninghoff J, Wiltfang J, Roccatagliata L, Nobili F, Hensch T, Tränkle A, Schönknecht P, Leroy M, Lopes R, Bordet R, Chanoine V, Ranjeva JP, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargalló N, Blin O, Frisoni GB, PharmaCog Consortium (2013) Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. *NeuroImage* 83:472–484. <https://doi.org/10.1016/j.neuroimage.2013.05.007>
6. Chalavi S, Simmons A, Dijkstra H, Barker GJ, Reinders AATS (2012) Quantitative and qualitative assessment of structural magnetic resonance imaging data in a two-center study. *BMC Med Imaging* 12:27. <https://doi.org/10.1186/1471-2342-12-27>
7. Weiskopf N, Suckling J, Williams G, Correia MM, Inkster B, Tait R, Ooi C, Bullmore ET, Lutti A (2013) Quantitative multi-parameter mapping of R1, PD(\*), MT, and R2(\*) at 3T: a multi-center validation. *Front Neurosci* 7:95. <https://doi.org/10.3389/fnins.2013.00095>
8. Keshavan A, Paul F, Beyer MK, Zhu AH, Papinutto N, Shinohara RT, Stern W, Amann M, Bakshi R, Bischof A, Carriero A, Comabella M, Crane JC, D'Alfonso S, Demaerel P, Dubois B, Filippi M, Fleischer V, Fontaine B, Gaetano L, Goris A, Graetz C, Gröger A, Groppa S, Hafler DA, Harbo HF, Hemmer B, Jordan K, Kappos L, Kirkish G, Llufrü S, Magon S, Martinelli-Boneschi F, McCauley J, Montalban X, Mühlau M, Pelletier D, Pattany PM, Pericak-Vance M, Cournu-Rebeix I, Rocca MA, Rovira A, Schlaeger R, Saiz A, Sprenger T, Stecco A, Uitdehaag BMJ, Villoslada P, Wattjes MP, Weiner H, Wuerfel J, Zimmer C, Zipp F, International Multiple Sclerosis Genetics Consortium. Electronic address: AIVINSON@PARTNERS.ORG, Hauser SL, Oksenberg JR, Henry RG (2016) Power estimation for non-standardized multisite studies. *NeuroImage* 134:281–294. <https://doi.org/10.1016/j.neuroimage.2016.03.051>
9. Jovicich J, Marizzoni M, Bosch B, Bartrés-Faz D, Arnold J, Benninghoff J, Wiltfang J, Roccatagliata L, Picco A, Nobili F, Blin O, Bombois S, Lopes R, Bordet R, Chanoine V, Ranjeva JP, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargalló N, Ferretti A, Caulo M, Aiello M, Ragucci M, Soricelli A, Salvadori N, Tarducci R, Floridi P, Tsolaki M, Constantinidis M, Drevelegas A, Rossini PM, Marra C, Otto J, Reiss-Zimmermann M, Hoffmann KT, Galluzzi S, Frisoni GB, PharmaCog Consortium (2014) Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. *NeuroImage* 101:390–403. <https://doi.org/10.1016/j.neuroimage.2014.06.075>
10. Teipel SJ, Reuter S, Stieltjes B, Acosta-Cabronero J, Ernemann U, Fellgiebel A, Filippi M, Frisoni G, Hentschel F, Jessen F, Klöppel S, Meindl T, Pouwels PJW, Hauenstein KH, Hampel H (2011) Multicenter stability of diffusion tensor imaging measures: a European clinical and physical phantom study. *Psychiatry Res Neuroimaging* 194:363–371. <https://doi.org/10.1016/j.psychres.2011.05.012>
11. Zhu T, Hu R, Qiu X, Taylor M, Tso Y, Yiannoutsos C, Navia B, Mori S, Ekholm S, Schifitto G, Zhong J (2011) Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study. *NeuroImage* 56:1398–1411. <https://doi.org/10.1016/j.neuroimage.2011.02.010>
12. Huang L, Wang X, Baliki MN, Wang L, Apkarian AV, Parrish TB (2012) Reproducibility of structural, resting-state BOLD and DTI data between identical scanners. *PLoS One* 7:e47684. <https://doi.org/10.1371/journal.pone.0047684>
13. Takao H, Hayashi N, Ohtomo K (2011) Effect of scanner in asymmetry studies using diffusion tensor imaging. *NeuroImage* 54:1053–1062. <https://doi.org/10.1016/j.neuroimage.2010.09.023>
14. Vollmar C, O'Muircheartaigh J, Barker GJ et al (2010) Identical, but not the same: intra-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>
15. Belli G, Busoni S, Ciccarone A, Coniglio A, Esposito M, Giannelli M, Mazzoni LN, Nocetti L, Sghedoni R, Tarducci R, Zatelli G, Anoja RA, Belmonte G, Bertolino N, Betti M, Biagini C,



- Ciarmatori A, Cretti F, Fabbri E, Fedeli L, Filice S, Fulcheri CPL, Gasperi C, Mangili PA, Mazzocchi S, Meliàdò G, Morzenti S, Noferini L, Oberhofer N, Orsingher L, Paruccini N, Princigalli G, Quattrocchi M, Rinaldi A, Scelfo D, Freixas GV, Tenori L, Zucca I, Luchinat C, Gori C, Gobbi G, for the Italian Association of Physics in Medicine (AIFM) Working Group on MR Intercomparison (2016) Quality assurance multicenter comparison of different MR scanners for quantitative diffusion-weighted imaging. *J Magn Reson Imaging* 43:213–219. <https://doi.org/10.1002/jmri.24956>
16. Laun FB, Huff S, Stieltjes B (2009) On the effects of dephasing due to local gradients in diffusion tensor imaging experiments: relevance for diffusion tensor imaging fiber phantoms. *Magn Reson Imaging* 27:541–548. <https://doi.org/10.1016/j.mri.2008.08.011>
17. Jones DK (2011) Diffusion MRI: theory, methods, and applications. Oxford University Press, Oxford
18. Dowell NG, Tofts PS (2008) Simple reliable and precise quantitative quality assurance of in-vivo brain ADC. *Proc. Int. Soc. Magn. Reson. Med.* 16th Annu. Meet. 3152
19. Jack CR Jr, Bernstein MA, Fox NC et al (2008) The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J Magn Reson Imaging* 27:685–691. <https://doi.org/10.1002/jmri.21049>
20. Achten E, Deblaere K, De Wagter C et al (1998) Intra- and inter-observer variability of MRI-based volume measurements of the hippocampus and amygdala using the manual ray-tracing method. *Neuroradiology* 40:558–566
21. Reuter M, Schmansky NJ, Rosas HD, Fischl B (2012) Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage* 61:1402–1418. <https://doi.org/10.1016/j.neuroimage.2012.02.084>
22. Jones R, Payne B (1997) Clinical investigation and statistics in laboratory medicine. ACB Venture Publications, London
23. Leemans A (2009) ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. 17th Annu Meet Intl Soc Mag Reson Med 3537
24. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9:62–66
25. De Santis S, Evans CJ, Jones DK (2013) RAPID: a routine assurance pipeline for imaging of diffusion. *Magn Reson Med* 70:490–496. <https://doi.org/10.1002/mrm.24465>
26. Holz M, Heil SR, Sacco A (2000) Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements. *Phys Chem Chem Phys* 2:4740–4742. <https://doi.org/10.1039/B005319H>
27. Wakana S, Caprihan A, Panzenboeck MM, Fallon JH, Perry M, Gollub RL, Hua K, Zhang J, Jiang H, Dubey P, Blitz A, van Zijl P, Mori S (2007) Reproducibility of quantitative tractography methods applied to cerebral white matter. *NeuroImage* 36:630–644. <https://doi.org/10.1016/j.neuroimage.2007.02.049>
28. Lebel C, Walker L, Leemans A, Phillips L, Beaulieu C (2008) Microstructural maturation of the human brain from childhood to adulthood. *NeuroImage* 40:1044–1055. <https://doi.org/10.1016/j.neuroimage.2007.12.053>
29. Song XW, Dong ZY, Long XY, Li SF, Zuo XN, Zhu CZ, He Y, Yan CG, Zang YF (2011) REST: a toolkit for resting-state functional magnetic resonance imaging data processing. *PLoS One* 6:e25031. <https://doi.org/10.1371/journal.pone.0025031>
30. Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2001) A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp* 14:140–151
31. Chen CC, Wan YL, Wai YY, Liu HL (2004) Quality assurance of clinical MRI scanners using ACR MRI phantom: preliminary results. *J Digit Imaging* 17:279–284. <https://doi.org/10.1007/s10278-004-1023-5>
32. Tropp J (2004) Image brightening in samples of high dielectric constant. *J Magn Reson* 167:12–24. <https://doi.org/10.1016/j.jmr.2003.11.003>
33. Wenger E, Mårtensson J, Noack H, Bodammer NC, Kühn S, Schaefer S, Heinze HJ, Düzel E, Bäckman L, Lindenberger U, Lövdén M (2014) Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. *Hum Brain Mapp* 35:4236–4248. <https://doi.org/10.1002/hbm.22473>
34. Maclaren J, Han Z, Vos SB, Fischbein N, Bammer R (2014) Reliability of brain volume measurements: a test-retest dataset. *Sci Data* 1:140037. <https://doi.org/10.1038/sdata.2014.37>
35. Jovicich J, Czanner S, Han X, Salat D, van der Kouwe A, Quinn B, Pacheco J, Albert M, Killiany R, Blacker D, Maguire P, Rosas D, Makris N, Gollub R, Dale A, Dickerson BC, Fischl B (2009) MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage* 46:177–192. <https://doi.org/10.1016/j.neuroimage.2009.02.010>
36. Malyarenko D, Galban CJ, Londy FJ et al (2013) Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. *J Magn Reson Imaging* 37:1238–1246. <https://doi.org/10.1002/jmri.23825>
37. Malyarenko DI, Newitt D, JW L et al (2016) Demonstration of nonlinearity bias in the measurement of the apparent diffusion coefficient in multicenter trials. *Magn Reson Med* 75:1312–1323. <https://doi.org/10.1002/mrm.25754>
38. Wu YC, Alexander AL (2007) A method for calibrating diffusion gradients in diffusion tensor imaging. *J Comput Assist Tomogr* 31:984–993. <https://doi.org/10.1097/rct.0b013e31805152fa>
39. Grech-Sollars M, Hales PW, Miyazaki K, Raschke F, Rodriguez D, Wilson M, Gill SK, Banks T, Saunders DE, Clayden JD, Gwilliam MN, Barrick TR, Morgan PS, Davies NP, Rossiter J, Auer DP, Grundy R, Leach MO, Howe FA, Peet AC, Clark CA (2015) Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. *NMR Biomed* 28:468–485. <https://doi.org/10.1002/nbm.3269>
40. Marengo S, Rawlings R, Rohde GK, Barnett AS, Honea RA, Pierpaoli C, Weinberger DR (2006) Regional distribution of measurement error in diffusion tensor imaging. *Psychiatry Res* 147:69–78. <https://doi.org/10.1016/j.psychres.2006.01.008>
41. Parrish TB, Gitelman DR, LaBar KS, Mesulam MM (2000) Impact of signal-to-noise on functional MRI. *Magn Reson Med* 44:925–932
42. Damoiseaux JS, Rombouts SA, Barkhof F et al (2006) Consistent resting-state networks across healthy subjects. *Proc Natl Acad Sci U S A* 103:13848–13853. <https://doi.org/10.1073/pnas.0601417103>
43. Rosazza C, Minati L, Ghielmetti F, Mandelli ML, Bruzzone MG (2012) Functional connectivity during resting-state functional MR imaging: study of the correspondence between independent component analysis and region-of-interest-based methods. *AJNR Am J Neuroradiol* 33:180–187. <https://doi.org/10.3174/ajnr.A2733>
44. Jovicich J, Czanner S, Greve D et al (2006) Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *NeuroImage* 30:436–443