Article

Same brain, different look? – The impact of scanner, sequence and preprocessing on diffusion imaging outcome parameters

Ronja Thieleking[1,*], Rui Zhang[1], Maria Paerisch[1], Kerstin Wirkner[2,3], Alfred Anwander[4], Frauke Beyer[1], Arno Villringer[1,5,6] and A. Veronica Witte[1,5,*]

[1]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany
[2]Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, 04107 Leipzig, Germany
[3]Leipzig Research Center for Civilization Diseases (LIFE), University of Leipzig, 04103 Leipzig, Germany
[4]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany
[5]Day Clinic of Cognitive Neurology, University of Leipzig, 04103 Leipzig, Germany
[6]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10117 Berlin, Germany
* Correspondence: thieleking@cbs.mpg.de (R.T.); witte@cbs.mpg.de (A.V.W.)

Abstract: In clinical diagnostics and longitudinal studies, the reproducibility of MRI assessments is of high importance in order to detect pathological changes, but developments in MRI hard- and software often outrun extended periods of data acquisition and analysis. This could potentially introduce artefactual changes or mask pathological alterations. However, if and how changes of MRI hardware, scanning protocols or preprocessing software affect complex neuroimaging outcomes from e.g. diffusion weighted imaging (DWI) remains largely understudied. We therefore compared DWI outcomes and artefact severity of 121 healthy participants (age range 19-54 years) who underwent two matched DWI protocols (Siemens product and Center for Magnetic Resonance Research sequence) at two sites (Siemens 3T Magnetom Verio and Skyra[fit]). After different preprocessing steps, fractional anisotropy (FA) and mean diffusivity (MD) maps,obtained by tensor fitting, were processed with tract-based spatial statistics (TBSS). Inter-scanner and inter-sequence variability of skeletonised FA values reached up to 5% and differed largely in magnitude and direction across the brain. Skeletonised MD values differed up to 14% between scanners. We here demonstrate that DTI outcome measures strongly depend on imaging site and software, and that these biases vary between brain regions. These regionally inhomogeneous biases may exceed and considerably confound physiological effects such as ageing, highlighting the need to harmonise data acquisition and analysis. Future studies thus need to implement novel strategies to augment neuroimaging data reliability and replicability.

Keywords: Diffusion Magnetic Resonance Imaging, White Matter, Fractional anisotropy, Multi-centre, Reproducibility, Imaging artefacts, Ageing

Introduction

Diffusion-weighted imaging (DWI) is a widely established, powerful and non-invasive *in-vivo* magnetic resonance imaging (MRI) technique used in human clinical and non-clinical applications (Leemans, 2010; Johansen-Berg & Behrens, 2014). DWI measures water diffusion in biological tissue which is hindered and restricted, for example, by fibre bundles, cell membranes and other cell structures in the brain. This renders DWI a valuable tool to acquire *in-vivo* information of brain properties at a microscopic scale (Basser, Mattiello, & Le Bihan, 1994). For example, diffusion tensor imaging (DTI) uses the diffusion of water molecules to determine the static anatomy of the brain (not influenced by brain function), yielding different tensor-based measures such as fractional anisotropy (FA), which is the degree of directionality of water diffusion within brain tissue, and mean diffusivity (MD), which describes the molecular diffusion rate of water within brain tissue. Thereof, axonal fibre tract coherence and structural connectivity as well as microstructural properties of the white and grey matter can be estimated (Basser, Mattiello, & Le Bihan, 1994; Basser & Jones, 2002; Assaf & Pasternak, 2008). DWI/DTI is both noise-sensitive and prone to imaging artefacts due to e.g. eddy currents, susceptibility-induced distortions, Nyquist ghosting or physiologically related factors (e.g. cardiac pulsation and subject motion) (Tournier et al., 2011). Therefore, technicians and scientists put continuously high effort into developing improvements for hardware and software (Tournier et al., 2011). Since the introduction of DTI in the mid 1990s (Basser, Mattiello, & Le Bihan, 1994), MRI techniques developed towards higher magnetic fields, stronger gradients and more sensitive detectors; thereby the signal-to-noise ratio and spatial resolution of MR images in general and of DTI in particular improved. This has led to a better understanding of structural connectivity (Horsefield & Jones, 2002; Jellison et al., 2004) and to the discovery of changes in microstructure due to e.g. ageing and neurodegenerative diseases (Goveas et al., 2015; de Groot et al., 2015 & 2016; Branzoli et al., 2016) as well as experience-dependent plasticity (Scholz et al., 2009; Blumenfeld-Katzir et al., 2011; Zatorre et al., 2012; Sampaio-Baptista & Johansen-Berg, 2017). With increasing availability of research-oriented MRI assessments on a larger scale in the last decades, such as in (multi-centre) longitudinal clinical trials and epidemiological cohorts reaching hundreds and even thousands of measurements (e.g. Human Connectome Project (van Essen et al., 2013), UK Biobank (Miller et al., 2016), German National Cohort MRI Study (Schlett et al., 2016)), however, developments in MRI hard- and software start to outrun the extended periods of data acquisition and analysis of these studies. Thus, DTI studies often experience changes during data acquisition like improvements of scanning protocols, the development of new sequences or minor to major hardware changes such as scanner upgrades. Nevertheless, it has been more and more acknowledged that not only obvious MRI artefacts per se but also subtle changes in sequence parameters and scanner hardware can systematically affect outcome measures (Hasan et al., 2001; Alexander & Barker, 2005; Ni et al., 2006; Giannelli et al., 2010; Schilling et al., 2017). Therefore, ensuring the comparability of DTI-derived outcome measures across-scanners - but also within-scanner and across-sequences - is of uttermost importance.

Previous attempts to estimate the inter-site reproducibility of DTI-derived measures reported divergent results, e.g. with regard to how large a potential inter-site difference would be (Prohl et al., 2019; Schwartz et al., 2019; Tax et al., 2019; Palacios et al., 2017; Fortin et al., 2017; Pohl et al., 2016; Mirzaalian et al., 2016; Belli et al., 2016; Kuhn et al., 2016; Buchanan et al., 2014; Zhan et al., 2014; Malyarenko et al., 2013; Fox et al., 2012; Teipel et al., 2011; Vollmar et al., 2010; Pfefferbaum et al., 2003). Outcomes from a common approach to assess human brain white matter microstructure (i.e., tract-based spatial statistics (TBSS; Smith et al., 2006) of FA maps) comprise for example coefficients of variation (CoV) ranging from 1.0% (Vollmar et al., 2010) to 4.1% (Palacios et al., 2017) to 14.4% (Teipel et al., 2011) for inter-site reproducibility. However, certain methodological shortcomings limit the validity of these studies. For example, if only a phantom was scanned on several imaging sites (Belli et al., 2016; Malyarenko et al., 2013; Teipel et al., 2011), it is hard

to interpret how differences in magnetic field homogeneity or gradient fields would affect the results for a human brain or body. In addition, in studies analysing human brain DW images, subject number was either low (n=1 (Kuhn et al., 2016), n=1 (Palacios et al., 2017), n=1 (Prohl et al., 2019), n=1 (Schwartz et al., 2019), n=2 (Fox et al., 2012), n=3 (Pohl et al., 2016), n=5 (Jovicich et al., 2014)) or different individuals underwent DWI at the imaging sites that were compared (Zavaliangos-Petropulu et al., 2019; Fortin et al., 2017; Mirzaalian et al., 2016; Pohl et al., 2016). In the latter case, results do not reflect scanner intrinsic variances but are biased by differences in the individual microstructure. Also, time passed between scans was up to one year (Pohl et al., 2016) or even up to 22 months (Tax et al., 2019) which makes differences in brain microstructure difficult to attribute solely to differences between MR scanners (and not to e.g. physiological changes). Other limitations are the choice of a rather narrow age range (67-84 y.o. (Zavaliangos-Petropulu et al., 2019); 8-19 y.o. (Fortin et al., 2017); 50-58 y.o. (Buchanan et al., 2014)) as well as the inclusion of a neurologically non-healthy subject (relapsing-remitting multiple sclerosis (Schwartz et al., 2019); mild cognitive impairment and Alzheimer's Disease (Zavaliangos-Petropulu et al., 2019)) thus results cannot be generalised. Collecting DW images with a variety of manufacturers (Palacios et al., 2017; Belli et al., 2016; Mirzaalian et al., 2016; Pohl et al., 2016; Prohl et al., 2019), sequence parameters (Mirzaalian et al., 2016; Pohl et al., 2016) and field strengths (Belli et al., 2016) introduces even more uncertainties. Due to technical advances, a significant amount of studies is affected by scanner upgrades. Nevertheless, to our best knowledge, there has only been one investigation of DTI metrics on scanner platform effects (pre- and post-upgrade), and only for 3T General Electric MRI scanners with a limited sample size (Zhan et al., 2014). To summarize, these methodological considerations highlight the need for more comprehensive inter-site comparability studies.

Besides acquisition-related differences in DTI-derived outcome measures that possibly originate from physically inherent differences between scanners and protocols, these differences may in addition fortify due to imaging artefacts. While several a posteriori correction methods have been implemented in commonly used preprocessing software to mitigate such imaging artefacts (Smith et al., 2004; Woolrich et al., 2009; Andersson & Sotiropoulos, 2016), one of the most ubiquitous artefacts, the Gibbs ringing (GR), received less attention. Only in recent years, attempts addressing the removal of this artefact have been published (Perrone et al., 2015; Kellner et al., 2016; Veraart et al., 2016a; Zhang et al., 2019; Zhao et al. 2020; Muckley et al., 2021). GR appears due to a k-space truncation along finite image sampling and presents as signal oscillations at sharp intensity transitions leading to physically implausible signals (PIS) and erroneous FA values (e.g. FA > 1), thus potentially wrong interpretations of the underlying microstructure. Physically implausible FA values show as lines of bright voxels at tissue boundaries in the FA images. Even though this adverse impact of GR on DTI-derived metrics has been recognised for almost three decades (Constable & Henkelmann, 1991; Pan, 2001; Gelb & Archibald, 2002; Perrone et al., 2015; Veraart et al., 2016a; Muckley et al., 2021), until recently, no state-of-the-art preprocessing pipeline such as FSL (Smith et al., 2004; Woolrich et al., 2009) had a GR artefact removal tool included yet. Whether such preprocessing would indeed lessen acquisition-related differences in DTI-derived outcome measures has not been evaluated yet. Nevertheless, throughout the evaluation of a novel GR artefact removal tool - the "Kellner Method" (Kellner et al., 2016) - it has already been integrated in the MRtrix3.0 (Tournier et al., 2019) preprocessing pipeline.

We aimed to systematically determine the effects of different scanner versions and preprocessing pipelines on DTI-derived outcomes using a large sample size. Specifically, we compared a Siemens 3T Magnetom Verio with its upgraded version Siemens 3T Magnetom Skyra and evaluated the "Kellner Method" (a Gibbs ringing artefact removal tool; Kellner et al., 2016) in comparison to the standard low-pass window filtering technique available on the Siemens scanner, noise reduction (MRtrix3.0; Tournier et al., 2019; Veraart et al., 2016b) and a pipeline without correction. We chose a within-subject design, short time gap between scans, high number of subjects and matched scanning protocols at both

imaging sites – scanning protocols are publicly available at https://osf.io/vnuqp/. The main research questions of the current study comprise:

1. What is the reproducibility of DTI-derived measures across-scanners (with differing upgrade versions) using high-resolution diffusion-weighted MRI on two 3T high-field scanner systems?

2. What is the intra-site but across-DWI-sequences comparison of DTI outcome measures from two sequences with matched protocols?

3. What is the impact of different preprocessing tools on measurement reproducibility (image denoising, GR artefact reduction, default low-pass window filtering)?

4. What are the conclusions to be drawn from the above mentioned results in relation to physiological effects (such as ageing) on white matter FA?

Methodology

Participants

121 healthy participants (60 female, age range 19 to 54 years, 29.9 $\pm$ 8.2 y.o.) were invited to undergo two head MRI acquisitions lasting about 75 minutes each. Exclusion criteria were MRI contraindications such as implanted medical device, metal fragment in the body, or claustrophobia as well as pregnancy, neurological or psychiatric conditions and centrally effective medication. Participants were scanned at two different imaging sites. Five participants did not re-turn for the second appointment (rescanning), resulting in a total of 116 participants for analyses. The interval between individual scanning sessions ranged from 2 to 139 days, and all scans were acquired within a 5-months-period. The study was approved by the Research Ethics Committee of the University of Leipzig and was conducted in accordance with the Declaration of Helsinki. All subjects gave written informed consent and received reimbursement for participation.

MR image acquisition

DWI scans were performed on two common 3T Siemens MRI scanners, namely Magnetom Verio (Syngo MR B17) and Magnetom Skyra^{fit} (Syngo MR E11) (Siemens Healthineers GmbH, Erlangen, Germany). These two scanner versions are often linked through an upgrade from Verio to Skyra^{fit}. The upgrade would include the replacement of all hard-and software parts except for the main magnet and the gradient coil. All 121 participants were at first scanned on Verio at the Day Clinic of Cognitive Neurology at the University of Leipzig and then on Skyra at the Max Planck Insti-tute for Human Cognitive and Brain Sciences, Leipzig. A counterbalanced order of scanners could unfortunately not be realised for organisational reasons. To assure a reproducible image acquisition, the brains of all participants were carefully positioned in the centre of the gradient system with a standardised head positioning procedure in order to minimise distortions and b-value variations caused by gradients non-linearities.

on both scanners, we used a 32-channel head coil and two double spin-echo encoding sequences which lasted 16 minutes 8 seconds. Throughout this work, we will refer to the two different protocols with "MPIL" (Siemens product sequence) and "CMRR" (developed by Moeller et al., 2010, at the**C**enter for **M**agnetic **R**esonance **R**esearch, University of Minnesota) ([TR]/[TE]: 13800/100 ms, 72 slices, 60 diffusion directions (b=1000 s/mm^2), 7 non-diffusion-weighted volumes (b=0 s/mm^2), EPI-factor: 128 (resolution 128 × 128), FoV: 220 × 220 × 123 mm^3, voxel size: (1.7mm)^3, Phase Partial Fourier: 6/8 (MPIL) and 7/8 (CMRR)). Parallel imaging was performed in both protocols with a generalised auto-calibrating partially parallel acquisition (GRAPPA), reconstruction algorithm and an acceleration factor of 2. The

CMRR protocol was identical at both scanners, only the MPIL protocol had to be slightly adjusted at the 3T Magnetom Skyra to [TR]: 14,400 ms due to the duty cycle limitations of the Skyra system relative to Verio. Nevertheless, this difference should not affect diffusion imaging as the white and grey matter spin systems ought to be relaxed to equilibrium after any [TR] > 10s and therefore, the MPIL protocol can also be viewed as identical at both scanners. The CMRR protocol was run with Siemens product low-pass window filtering option of the raw data to reduce high frequency imaging artefacts such as GR whereas the MPIL protocol was run without it and DW images were in addition retrospectively reconstructed on the scanner console with the Siemens product low-pass window filter. Thereby, we were able to assess the quality of this low-pass windowfilter.

For anatomical reference, a high-resolution 3D structural image was acquired for all participants at each scanning site. Therefore, we used a magnetisation-prepared 180 degrees radio-frequency pulses and rapid gradient-echo (MPRAGE) sequence with the following parameters: [TR]/[TI]/[TE]: 2300/900/2.98 ms, 176 slices, flip angle: 9° , FoV: 256 × 240 × 176 mm^3, voxel size: $(1mm)^3$, GRAPPA-factor 2. Scanning protocols are available at https://osf.io/vnuqp/.

Image Processing

Raw image data was exported as DICOMs and transformed to NIfTI format (Li et al., 2016). During this step b-values and b-vectors were extracted. Further, we applied a denoising tool from MRtrix3.0-rc1 (Tournier et al., 2019) in order to reduce signal fluctuations originating in thermal noise. This preprocessing step is supposed not to target Gibbs ringing artefacts. It is recommended to denoise the images before approaching the removal of the Gibbs ringing artefact as the denoising tool (MRrtix3.0 (“ DWI Denoising — MRtrix 3.0 Documentation”); Veraart et al., 2016b) detects and removes noise characteristics which would be altered by any additional preprocessing step. To evaluate different preprocessing techniques for diffusion-weighted images, we compared four different preprocessing pipelines as shown in figure 1. Main focus was to compare the Siemens low-pass window filtering and the “Kellner Method” (Kellner et al., 2016) which address the removal of the Gibbs ringing artefact. The other two comprised neither noise nor Gibbs ringing correction nor only noise correction.

Further preprocessing steps included the segmentation and removal of non-brain tissue with bet (**B**rain **E**xtraction **T**ool) embedded in FSL (Smith, 2002). With the FSL software eddy (Andersson et al., 2016), we implemented the replacement of slices showing signal drop-outs due to subject head motion. Next, we applied motion correction, and rigid body registration to each participant’s own skull-stripped and AC-PC-reoriented T1-weighted image in one step together with the interpolation of the target isotropic resolution of 1mm with a tool developed in-house called Lipsia (Lohmann et al., 2001). In the final preprocessing step, the diffusion tensor was modelled and metrics like FA and MD were estimated at each voxel using Lipsia again. To account for motion-attributed changes in the DTI parameters, we estimated frame-to-frame head motion by calculating the frame-wise displacement (FD, in mm) across volumes (Power et al., 2012) using the 6-parameter motion output generated from eddy (Andersson & Sotiropoulos 2016). This mean FD was used as a covariate to correct for head motion in statistical analysis (Beyer et al., 2017).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

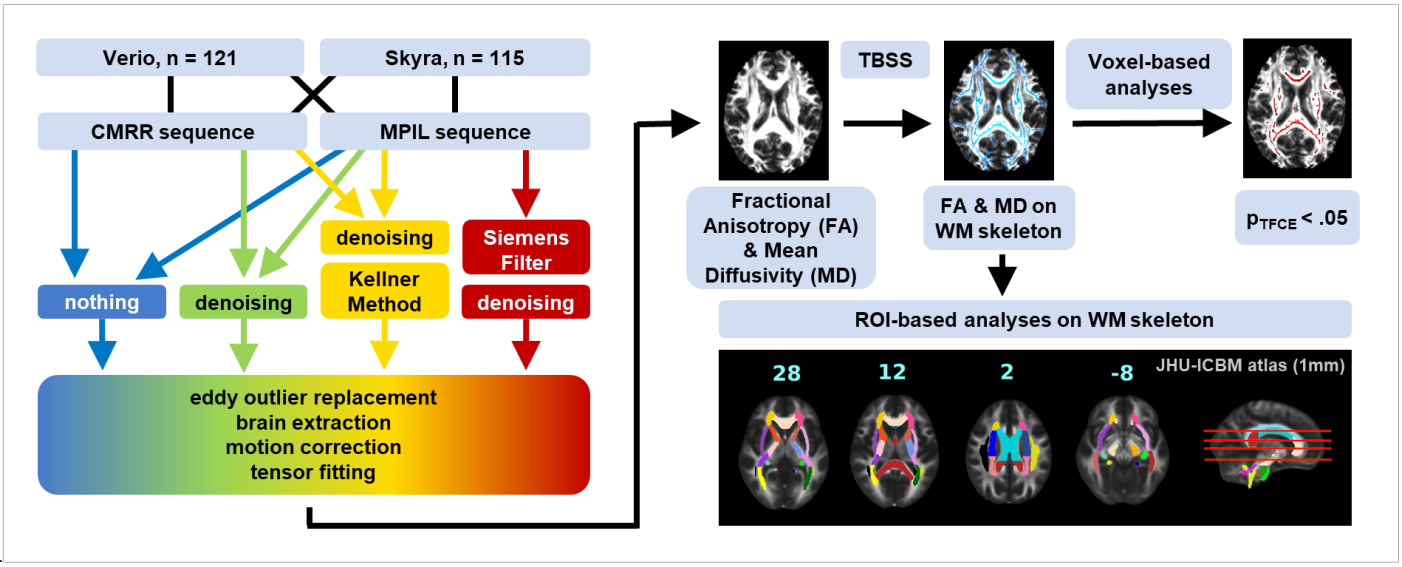


Figure 1: Analysis outline. DW images were collected at 3T Siemens Magnetom Verio and Skyra^{fit} with CMRR (Moeller et al., 2010) and MPIL (Siemens product sequence) sequence respectively, resulting in four datasets. Each dataset from the MPIL sequence was processed with four different pipelines: no filtering (blue), denoising (green), denoising + unringing by "Kellner Method" (yellow) (Kellner et al., 2016) and Siemenslow-pass window filtering + denoising (red); datasets from the CMRR sequence were processed with three different pipelines (Siemens low-pass window filtering was not applied). Standard preprocessing steps followed these pre-filtering steps. By tensor fitting obtained FA and MD maps were skeletonised with tract-based spatial statistics (TBSS; Smith et al., 2006) and fed into voxel-wise as well as ROI-based analyses on white matter (WM) skeleton.

Quality Assessment

Initial visual Quality Assessment (QA) was conducted to ensure data fidelity. No individuals had to be excluded due to structural abnormalities. We conducted further quality checks after every main preprocessing step. To be exact, we assessed the overall quality of the diffusion data by inspecting the signal-to-noise-ratio maps of the b0 images and the contrast-to-noise-ratio maps of the b1000 images as well as the residuals (difference between the observation and Gaussian process predictions) with FSL's eddy for imaging artefacts. Before statistical analyses, we reviewed the registration of all FA maps to the common template (FMRIB58_FA in MNI space from FSL) (Smith, 2002) to confirm the precission of this step which is crucial for our region of interest approach.

Region of Interest Approach

In order to assess differences in mean FA and mean MD values introduced by imaging site, sequence and composition of the preprocessing pipeline not only on the whole brain level but also for different regions, we extracted mean FA and MD values from fibre tracts that were defined in line with (Vollmar et al., 2010) - namely the splenium of the corpus callosum (SCC) (selected manually), left superior longitudinal fascicle (LSLF) and left uncinate fascicle (LUF) (ROIs highlighted in Supplementary Figure S1). Before masking the mean FA and MD skeleton to obtain mean values for the above listed ROIs, we non-linearly warped the ROIs from the JHU-ICBM atlas (1mm) to the FMRIB58_FA MNI space (both from FSL).

Statistical Analysis

To test whether differences in scanners (Verio vs. Skyra), sequences (MPIL vs. CMRR) or preprocessing tools (unfiltered vs. denoised vs. Siemens low-pass window filtering vs. unringing by Kellner method) affect DTI-derived out-

 doi:10.20944/preprints202107.0134.v2

come measures, we analysed whole-brain voxel-wise and ROI-based mean FA and MD values within the white matter skeleton (see Figure 1).

Through tract based spatial statistics (TBSS; Smith et al., 2006), we obtained FA and MD maps of the white matter skeleton for each subject in each condition. Briefly, all FA maps were co-registered using affine and non-linear transformations to standard space and the individual local maximal FA values were projected onto the standard FA skeleton to match individual's anatomy. The threshold for these standardised white matter fibre tract maps was set at 0.2. In order to obtain the MD skeleton maps for each subject, we applied the non-linear warps and skeleton projection from the FA processing to the MD data. The FA and MD skeleton maps were lastly fed into voxel-wise analysis of FA and MD for statistical comparison using the randomise tool by FSL version 5.0.1. We used 1000/2000 permutations and threshold-free cluster enhancement as test statistic. With this tool, we conducted voxel-based paired two-sample t-tests on white matter skeletons of each subject to detect locations which differed significantly (p-value (FWE) < 0.05). Voxel-wise analysis was conducted on a whole brain level and the FD estimates across volumes were included as a covariate of no interest. In addition, we extracted and compared the average skeletonised FA and MD values in the three different ROIs (Figure S1) to compare broader regional variations.

On the whole brain level, we further compared mean FA and MD values of the WM skeleton with Bayesian linear modelling and Bayesian paired two-sample t-tests with the "BayesFactor" package included in R. Bayesian statistics in a nutshell: Bayes Factor $BF = \frac{likelihood\ of\ data\ given\ H1}{likelihood\ of\ data\ given\ H0}$. Conventionally, the alternative hypothesis H1 ("there are one or more effects") is more likely if $BF > 3$ and the null hypothesis H0 ("data is random noise") is accepted if $BF < \frac{1}{3}$. A Bayes Factor between $\frac{1}{3}$ and 3 suggests that given the data are not informative about which hypotheses should be accepted. Bayesian statistics were additionally applied on FA and MD values of the WM skeleton in selected ROIs (see "Region of Interest Approach" above).

Inter-scanner variability

For the analysis of the scanner comparison (Verio vs. Skyra), we focused on the DW images recorded with the CMRR sequence from 115 subjects in order to guarantee high statistical power and on the "state-of-the-art" preprocessing pipeline including denoising. Data from the CMRR sequence of one of the 116 subjects scanned at both imaging sites were corrupted and the subject had to be excluded. We applied both, the whole brain and ROI-based FA and MD approach using TBSS. We further calculated the differences of the mean FA and MD values in percentages on a per subject basis: $\frac{meanFAvalue(Skyra) - meanFAvalue(Verio)}{meanFAvalue(Verio)} * 100$ or $\frac{meanMDvalue(Skyra) - meanMDvalue(Verio)}{meanMDvalue(Verio)} * 100$.

Inter-sequence variability

Regarding the sequence comparison (MPIL vs. CMRR), we excluded datasets recorded which were acquired with different reconstruction parameters, leaving us with 51 subjects for the MPIL sequence and their matched scans from the CMRR sequence (Verio n=23, Skyra n=28). Deviations from the measurement protocol comprised the missing retrospective reconstruction with the Siemens product window filtering (Verio, MPIL sequence, n=93) and the application of data interpolation during reconstruction (Skyra, MPIL sequence, n=88). Voxel-wise statistical comparison of the sequences was then conducted with TBSS' randomise tool on the FA and MD skeleton maps as described above.

Gibbs ringing (GR) artefact

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

We visually assessed the reduction of GR artefacts by the different preprocessing pipelines (MPIL sequence, Verio n=23, Skyra n=28). In order to quantify the GR artefact, we extracted the amount of voxels with implausible fractional anisotropy values (FA > 1) which are introduced by GR. Those voxels were clearly affected by GR and the amount of those voxels provided a conservative estimation to analyse if this number differs between scanners and preprocessing approaches using Bayesian statistics.

Motion effects

To ensure comparability of studies conducted at different scanners, we also looked into possible differences in subject head motion quantified as frame-wise displacement (FD, in mm) between scanners (with CMRR sequence, n=115). Thereto, we fed mean FD values into Bayesian linear modelling. We further investigated if motion effects can be attenuated by certain preprocessing approaches, using Bayesian linear modelling and post-hoc paired two-sample t-tests.

Age effect

We investigated age as a biological phenotype of interest and evaluated size differences of the negative effect of age on voxel-wise and whole brain mean FA and MD (CMRR sequence, n=115). To this end, we performed additional analysis of data from the LIFE Adult Study (n=1255; Loeffler et al., 2015; Zhang et al., 2018). Based on this cross-sectional data, the negative age effect could be estimated to a decrease in mean FA of the WM skeleton of 0.14% per year. In order to simulate the case of data collection at different imaging sites, we compared not only the age effect on the dataset from Verio with the one from Skyra but also with a dataset consisting of randomly chosen DW images from Verio and Skyra (1:1 ratio, n=50 from each scanner).

Harmonisation attempt

In line with Pohl et al. (2016), we calculated the ratio between the mean FA value of the whole brain's WM skeleton from Skyra and Verio in order to possibly harmonise FA values across scanners. The ratio would serve as a correction factor (cf) to harmonise data before statistical analyses ($meanFA(Verio) = cf * meanFA(Skyra)$). Adding to the whole brain analysis, we also looked into the ratios for the selected ROIs.

Coefficient of Variance

Previous studies on DTI test–retest replicability commonly reported the coefficient of variation (CoV) as statistical measure (Pfefferbaum et al., 2003; Vollmar et al., 2010; Teipel et al., 2011; Mirzaalian et al., 2016; Prohl et al., 2019). The CoV, defined as the ratio of the measurements standard deviation σ divided by the mean μ and multiplied by 100 ($CoV = \frac{\sigma}{\mu} * 100$), served as an estimate of data dispersion expressed as relative percentage independent of the absolute measurement values. For the assessment of the inter-scanner variability, we calculated the CoV of WM skeleton mean difference FA and MD values ($|Verio - Skyra|$ in %, single subject, voxel-based) after preprocessing with denoising on a whole brain level and in selected ROIs. Regarding the GR artefact and motion effects, we report respectively the CoV of the amount of voxels with FA>1 and of the mean FD values.

Results

Inter-scanner variability

We observed a significant difference between 3T Verio and Skyra scanners after preprocessing with denoising in whole brain white matter skeleton mean FA values (Figure 2 and 4; CMRR sequence; Bayesian linear modelling,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

n=115: BF[mean FA value ~ scanner] = 33.9) with a CoV of about 7.1%; this means that the alternative hypothesis H1 is 33 times more likely than the null hypothesis. On the whole brain level, Skyra delivered slightly higher mean FA values (see Table 1). However, scanner differences in mean FA value were not consistent across white matter tracts (Figure 2, Table 1). Central fibre tracts with high FA values such as the splenium of the corpus callosum (SCC) delivered significant differences between the mean FA values of the two scanners (Bayesian linear modelling, n=115: BF[mean FA value ~ scanner] = $1.1 * 10^{32}$, CoV 30.7%) with Verio showing much higher values. More lateral such as the left uncinate fascicle (LUF) did not show significant differences between scanners (BF = 1.1) but a CoV of about 31.5%. However, in fibre tracts with mean FA values of the same scale ($FA \approx 0.5$) but of a longer range such as the left superior longitudinal fascicle (LSLF), scanner differences were significant (BF = $3.3 * 10^{12}$) with Skyra showing higher values and a CoV around 11.1%. Differences of the mean FA values in percentages $\frac{meanFAvalue(Skyra) - meanFAvalue(Verio)}{meanFAvalue(Verio)} * 100$: whole brain: ~1%, SCC: ~ -5.2%, LUF: ~1%, LSLF: ~1.1%. In addition to the analysis of scanner differences in the FA skeleton for the different brain regions, we also tested for scanner differences in MD values (Figure 3 and 5). With TBSS, we found a clear whole brain difference with Skyra showing higher MD values than Verio (Figure 3; CMRR sequence; preprocessed with denoising). This clear direction of the scanner difference is supported by the whole brain mean MD value comparison with Bayesian linear modelling (Figure 5, Table 2; n=115: BF[mean MD value ~ scanner] = $1.4*10^{8}$) with a CoV of about 11.1%. Central fibre tracts such as the splenium of the corpus callosum (SCC) and lateral fibre tracts such as the left uncinate fascicle (LUF) exhibit the same pattern (Skyra showing higher MD values than Verio) with differences in magnitude (SCC: BF[mean MD value ~ scanner] = $1.3*10^{50}$, CoV = 27.4%; LUF: BF[mean MD value ~ scanner] = $1.4*10^{4}$, CoV = 28.3%). Only in longer fibre tracts comprising many different and crossing fibre orientations, scanner differences are not evident (BF[mean MD value ~ scanner] = 0.16, CoV = 13.9%). Differences of the mean MD values in percentages $\frac{meanMDvalue(Skyra) - meanMDvalue(Verio)}{meanMDvalue(Verio)} * 100$: whole brain ~2%, SCC: ~14%, LUF: ~3%, LSLF: ~ -0.2%.

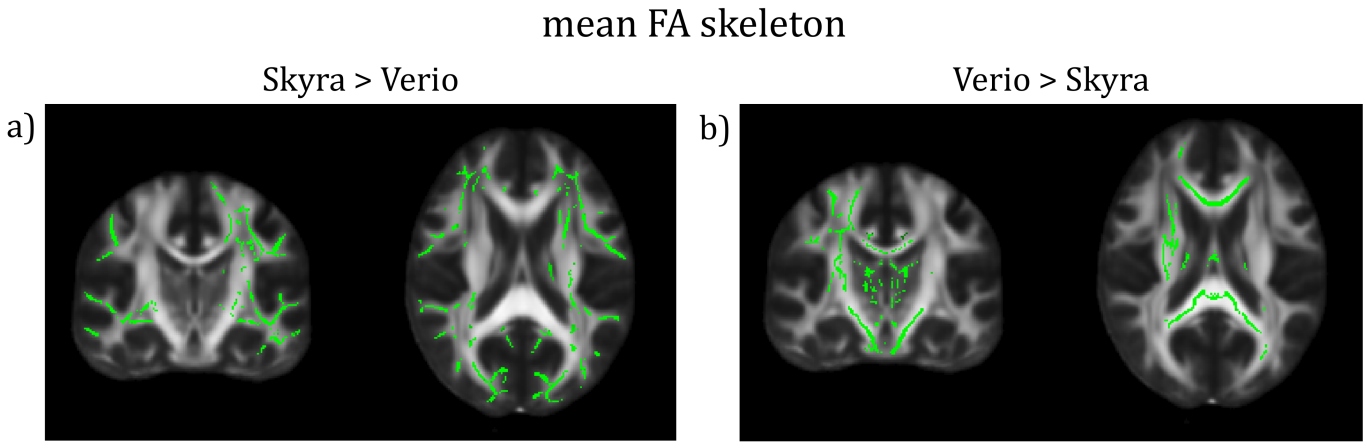


Figure 2: TBSS on the FA skeleton of scanner differences (CMRR sequence) after preprocessing with denoising (TFCE corrected, highlighted tracts: $p < .05$, [y z] = [-18 19]). a) More superficial WM tracts show higher values in Skyra than in Verio. b) Rather deep WM tracts show higher values in Verio than in Skyra.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

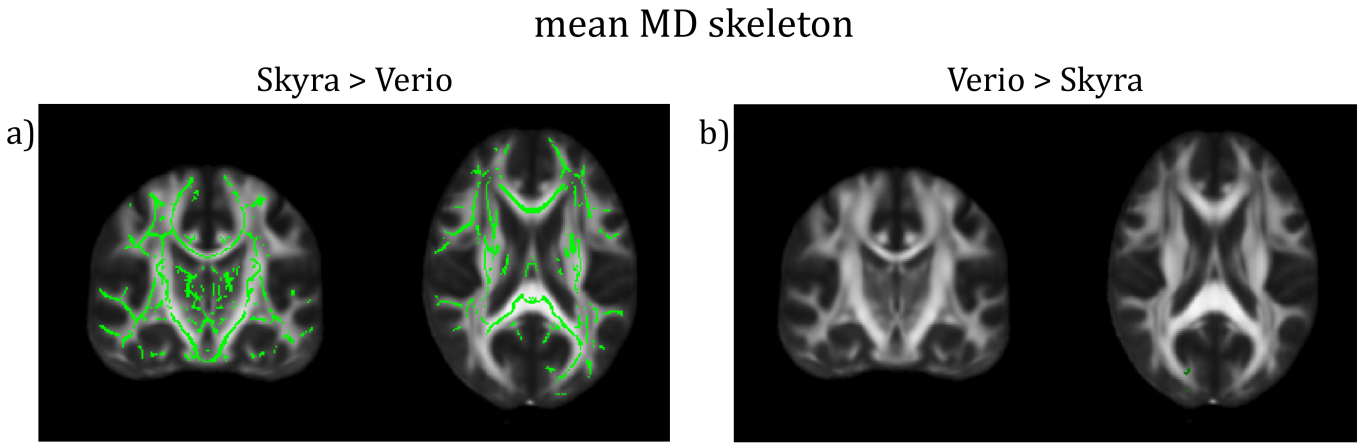


Figure 3: TBSS on the MD skeleton of scanner differences (CMRR sequence) after preprocessing with denoising (TFCE corrected, highlighted tracts: $p < .05$, [y z] = [-18 19]). a) The whole brain WM skeleton (except for a small part in the right occipital lobe) shows higher values in Skyra than in Verio. b) Only a small white matter region in the right occipital lobe show higher values in Verio than in Skyra.

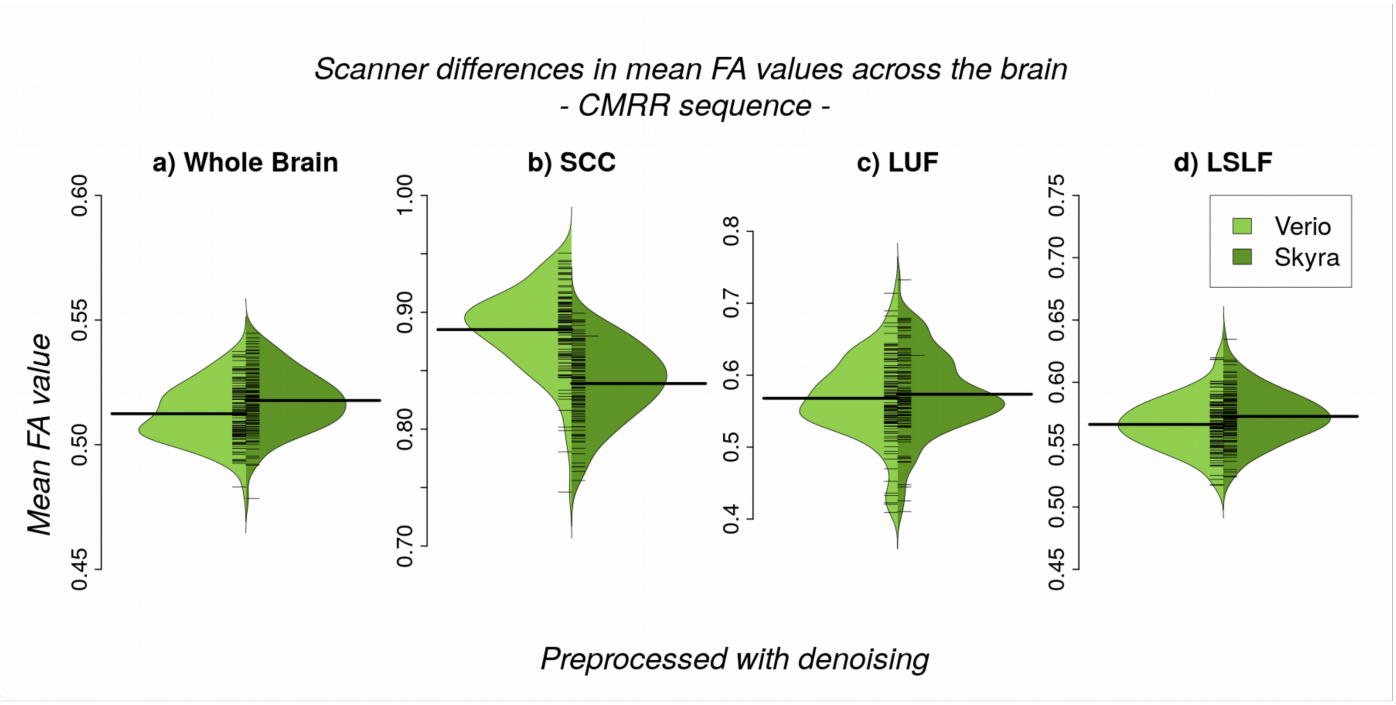
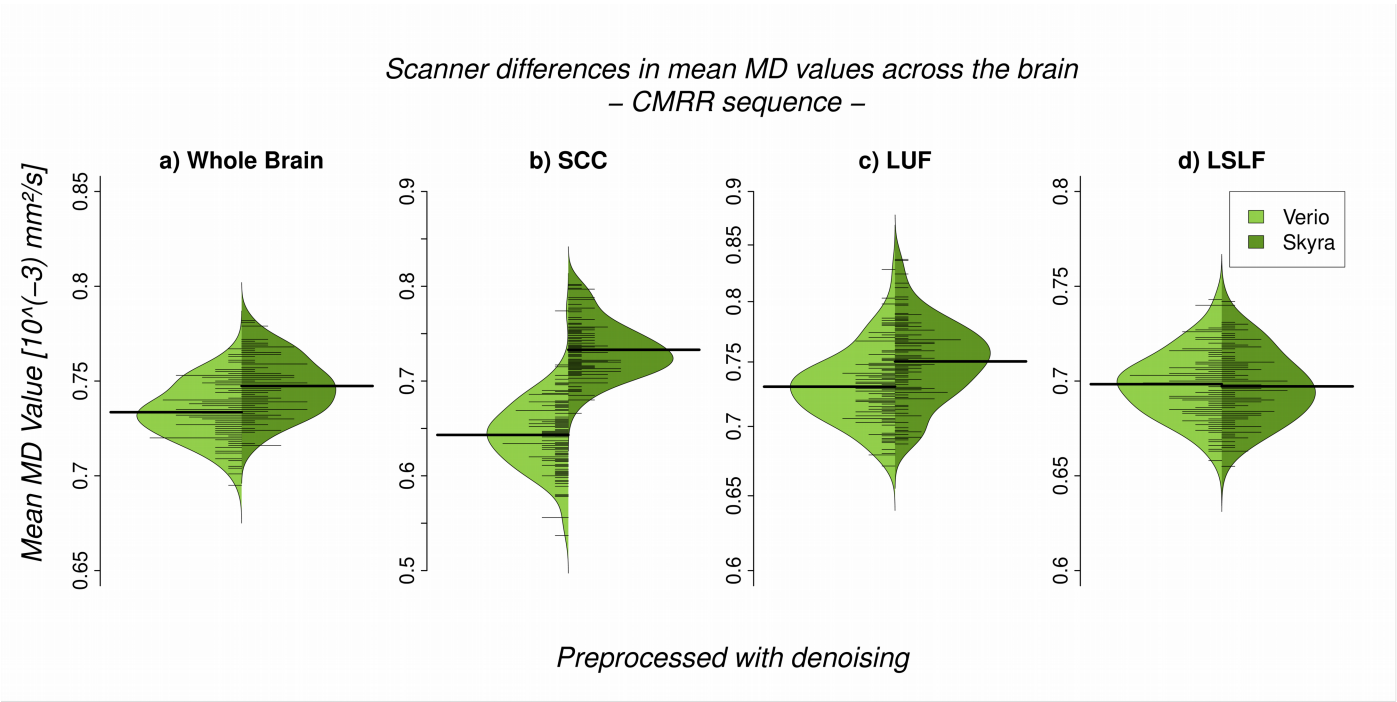


Figure 4: Scanner differences in the mean FA value of the white matter (WM) skeleton (CMRR sequence, after denoising) for a) whole brain and region-of-interest analyses (b), c) and d)). a) On the whole brain level, Skyra delivers higher FA values than Verio (~1%, BF>>3). b) The splenium of the corpus callosum (SCC), c) left uncinate fascicle (LUF) and d) the left superior longitudinal fascicle (LSLF) highlight the differences in direction and magnitude of the scanner differences across ROIs: differences between scanners in percentages $\frac{meanFAvalue(Skyra) - meanFAvalue(Verio)}{meanFAvalue(Verio)} * 100$: SCC: ~ -5.2% (BF>>3), LUF: ~1% (BF=1.1), LSLF: ~1.1% (BF>>3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2



331

332**Figure 5:** Scanner differences in the mean MD value of the white matter (WM) skeleton (CMRR sequence, after denois-
333ing) for a) whole brain and region-of-interest analyses (b), c) and d)). a) On the whole brain level, Skyra delivers
334higher MD values than Verio (~2%, BF >>3). b) The splenium of the corpus callosum (SCC) and c) the left uncinate
335fascicle (LUF) show as well higher MD values for Skyra (SCC: ~14% (BF>>3), LUF: ~3% (BF>>3)). d) Only in the left
336superior longitudinal fascicle (LSLF) scanner differences are not pronounced: ~ -0.2% (BF=0.16). Percentages reflect re-
337lative differences between scanners: $\frac{meanMDvalue(Skyra) - meanMDvalue(Verio)}{meanMDvalue(Verio)} * 100$

338**Table 1:** Mean FA values of the white matter skeleton at the whole-brain level and in different ROIs (CMRR sequence,
339after denoising). Bayesian linear modelling shows significant scanner differences for ROIs in the centre (SCC) and
340long white matter tracts (LSLF) but not for ROIs with curved lateral tracts (LUF).

WM Skeleton (after denoising)	Scanner	Mean FA value	SD	CoV | Verio - Skyra | (%)	Linear Model Bayes Factor (Mean FA value ~ Scanner, n=115)
whole brain	Verio	0.5124	0.0112	7.1	33.9
	Skyra	0.5177	0.0121		
SCC	Verio	0.7796	0.0168	30.7	$1.1 * 10^{32}$
	Skyra	0.7631	0.0172		
LUF	Verio	0.5679	0.0566	31.5	1.1
	Skyra	0.5735	0.0571		
LSLF	Verio	0.5663	0.0204	11.1	$3.3 * 10^{12}$
	Skyra	0.5727	0.0208		

341**Table 2:** Mean MD values of the white matter skeleton at the whole-brain level and in different ROIs (CMRR se-
342quence, after denoising). Bayesian linear modelling shows significant scanner differences on the whole brain level and
343for central and lateral ROIs (SCC and LUF) but scanner differences in the LSLF are not evident.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

WM Skeleton (after denoising)	Scanner	Mean MD value [10^{-3}mm/s^2]	SD [10^{-3}mm/s^2]	CoV | Verio - Skyra | (%)	Linear Model Bayes Factor (Mean MD value ~ Scanner, n=115)
whole brain	Verio	0.7336	0.0149	11.1	$1.4 * 10^8$
	Skyra	0.7474	0.0155		
SCC	Verio	0.6431	0.0370	27.4	$1.3 * 10^{50}$
	Skyra	0.7330	0.0289		
LUF	Verio	0.7310	0.0285	28.3	$1.4 * 10^4$
	Skyra	0.7511	0.0321		
LSLF	Verio	0.6983	0.0173	13.9	0.16
	Skyra	0.6972	0.0186		

Inter-sequence variability

Investigating the effect of different imaging sequences run at the same scanner (with harmonised protocol parameters, MPIL vs. CMRR, Verio n=23, Skyra n=28), TBSS (TFCE and motion corrected, $p<.05$) detected that regional mean FA values differed significantly in several WM tracts dependent on scanner. Sequence differences were more pronounced in FA maps from Verio: CMRR showed higher FA values in central brain-areas, mainly in the CC, whereas MPIL showed higher FA values in cortical tracts in the left hemisphere (Figure 6a)+b)). Data from Skyra though showed no WM tracts in which CMRR delivered higher FA values than MPIL but MPIL indicated higher FA values in both hemispheres, cortically and sub-cortically (Figure 6c)+d)). Regarding the comparison of the sequences based on MD maps, patterns were less pronounced: the CMRR sequence appeared to deliver higher MD values in both hemispheres cortically and sub-cortically (Figure 7 a)+c)) whereas the MPIL sequence shows higher MD values cortically and rather frontally and in the right hemisphere at Verio (Figure 7 b)) but more occipitally and in the right hemisphere at Skyra (Figure 7d)).

 doi:10.20944/preprints202107.0134.v2

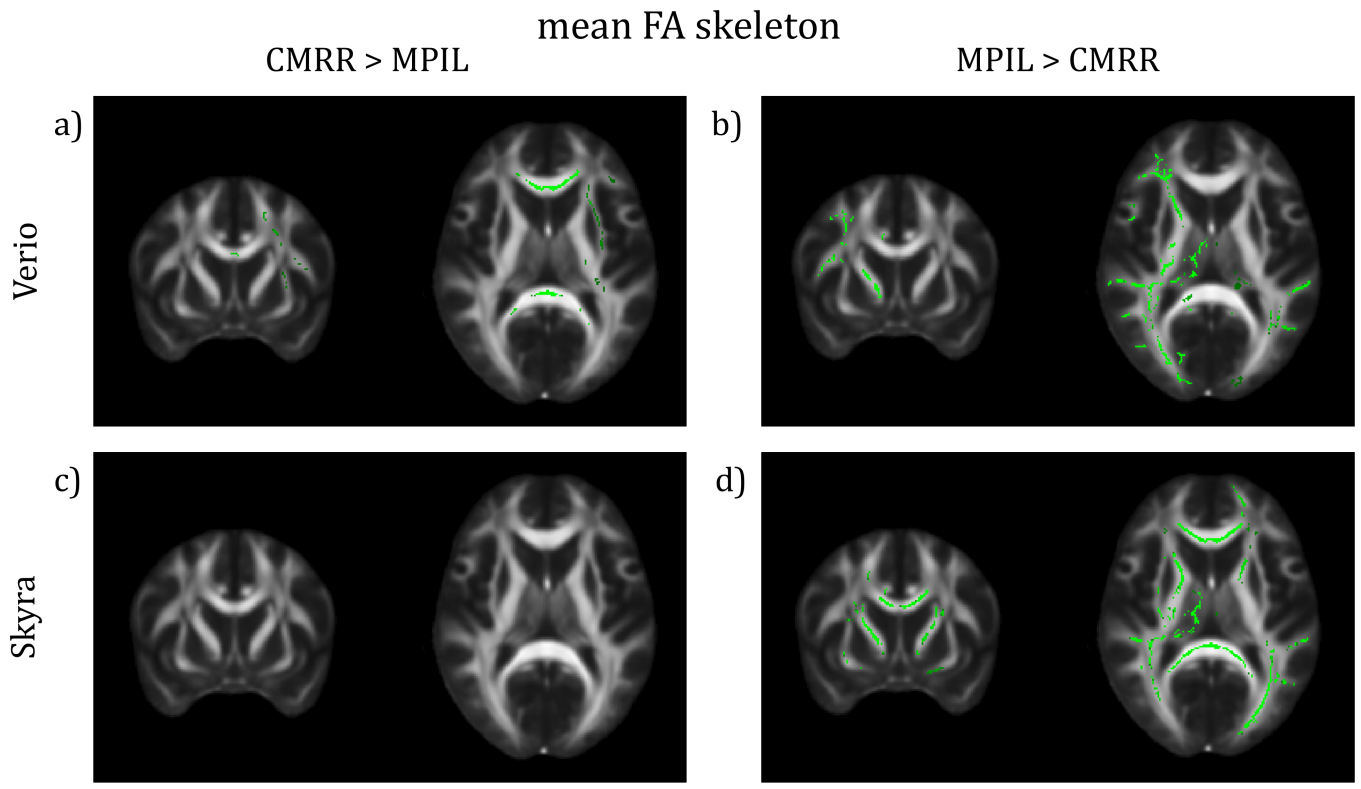


Figure 6: TBSS of differences on the FA skeleton between sequences after preprocessing with denoising (TFCE corrected, highlighted white matter areas: $p < .05$, [y z] = [9 10]). a)+b) Verio: CMRR shows higher FA values in central brain areas, mainly in the CC, whereas MPIL shows higher FA values in lateral areas in the right hemisphere. c)+d) Skyra: MPIL delivers higher FA values in both hemispheres.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

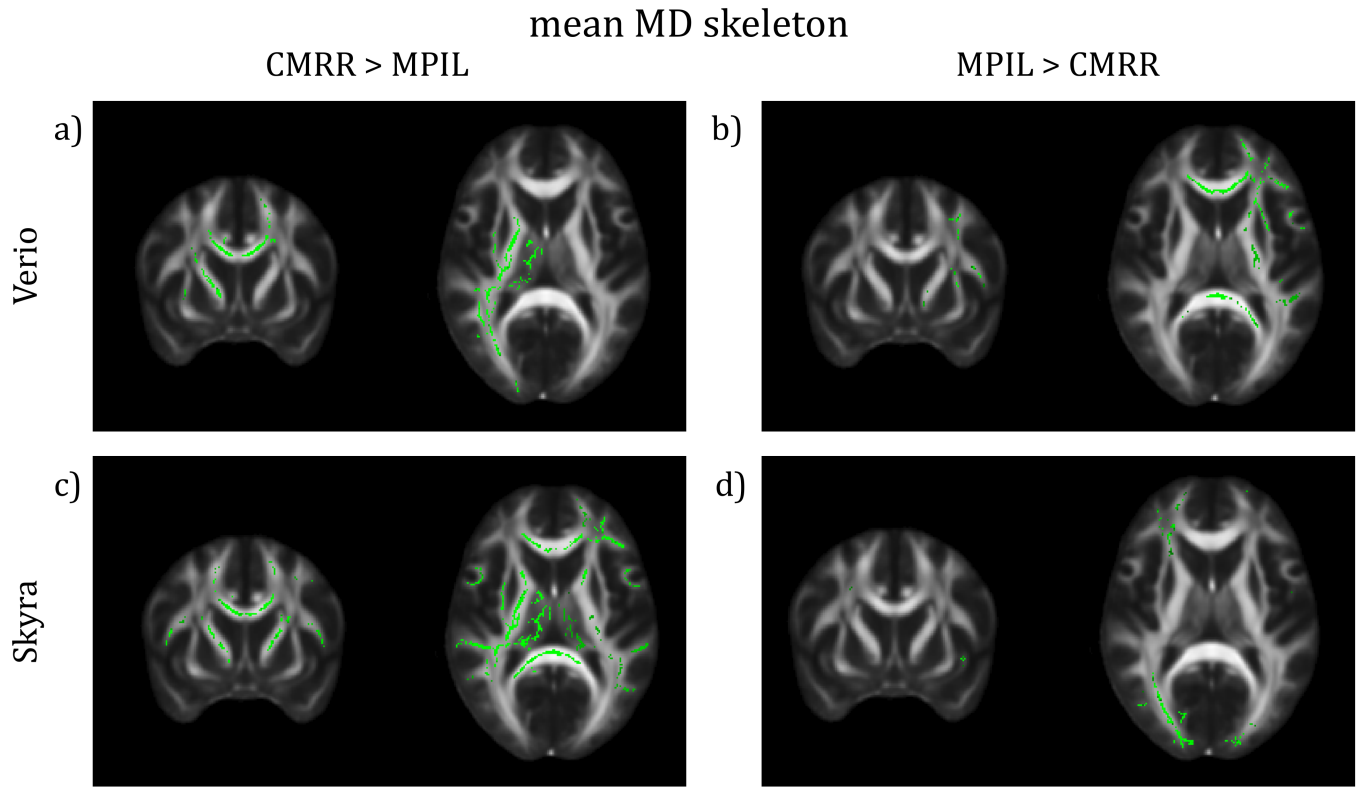
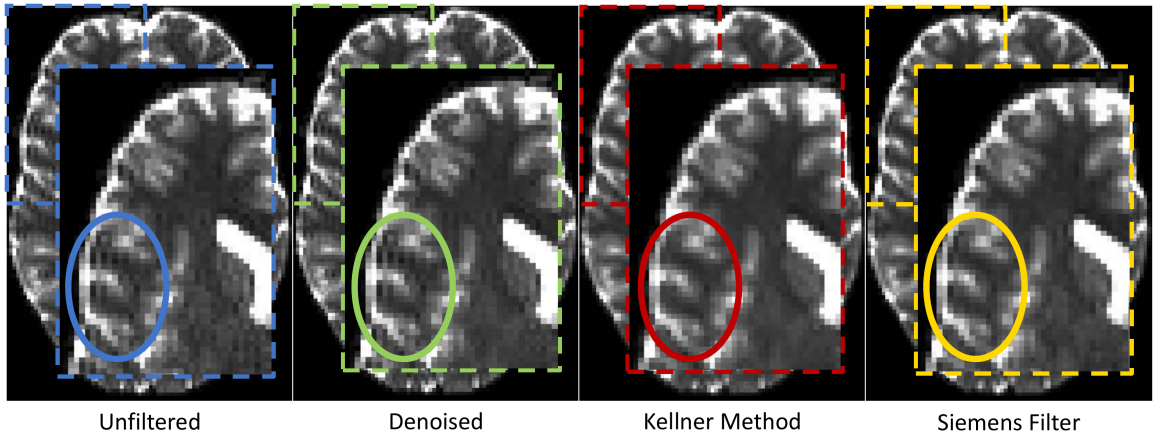


Figure 7: TBSS of differences on the MD skeleton between sequences after preprocessing with denoising (TFCE corrected, highlighted tracts: $p < .05$, [y z] = [9 10]). a)+c) The CMRR sequence seems to deliver higher MD values in both hemispheres in central and lateral brain regions whereas the MPIL sequence shows higher MD values b) in central and frontal regions and in the left hemisphere at Verio but d) more occipitally and in the right hemisphere at Skyra.

GR artefact in DW images

The qualitative visual data control of the MPIL sequence (Skyra) revealed that the GR artefact appeared very strong in the unfiltered b0 (T2-weighted) images. After preprocessing, different levels of GR reduction were visually detected (Figure 8). Specifically, while denoising did not reduce GR artefacts, the unringing tool by Kellner et al. (2016) seemed toclearly reduce the GR artefact. The low-pass window filtering by Siemens introduced a global blurring but the oscillations starting from the bright cortico-spinal-fluid (CSF) surrounding the brain were still visible.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

Figure 8: Appearance of Gibbs ringing artefact in b0 (T2-weighted) images after applying different preprocessing tools. Note that only the unringing tool (red) from Kellner et al. (2016) considerably reduced GR as is most evident in the circled part of the b0 image.

In line with visual assessment, quantitative analyses indicated that different preprocessing pipelines (Skyra, MPIL sequence) differ significantly in their efficiency of reducing the amount of implausibly high FA values (Bayesian linear modelling: BF[#(voxels with FA > 1) ~ preprocessing pipeline] > $4.7*10^{10}$). The unringing tool reduced the amount of implausible FA values significantly (paired Bayesian t-test: BF[unfiltered ~ Kellner Method] > $8*10^{9}$, CoV=42.3%) whereas the quantity of implausible FA values did not change consistently after any other preprocessing tool (BF[unfiltered ~ denoising] = 0.7, CoV=104.7%), BF[unfiltered ~ Siemens low-pass window filtering] = 13, CoV=78.1%) (Figure 9). Further, there were no significant differences in the amount of physically implausible FA values between Verio and Skyra (Bayesian linear modelling, full/null model comparison: $BF\left[\frac{(voxelswithFA>1)\sim scanner*preprocessingpipeline}{(voxelswithFA>1)\sim preprocessingpipeline}\right]$ = 0.05±1.09%).

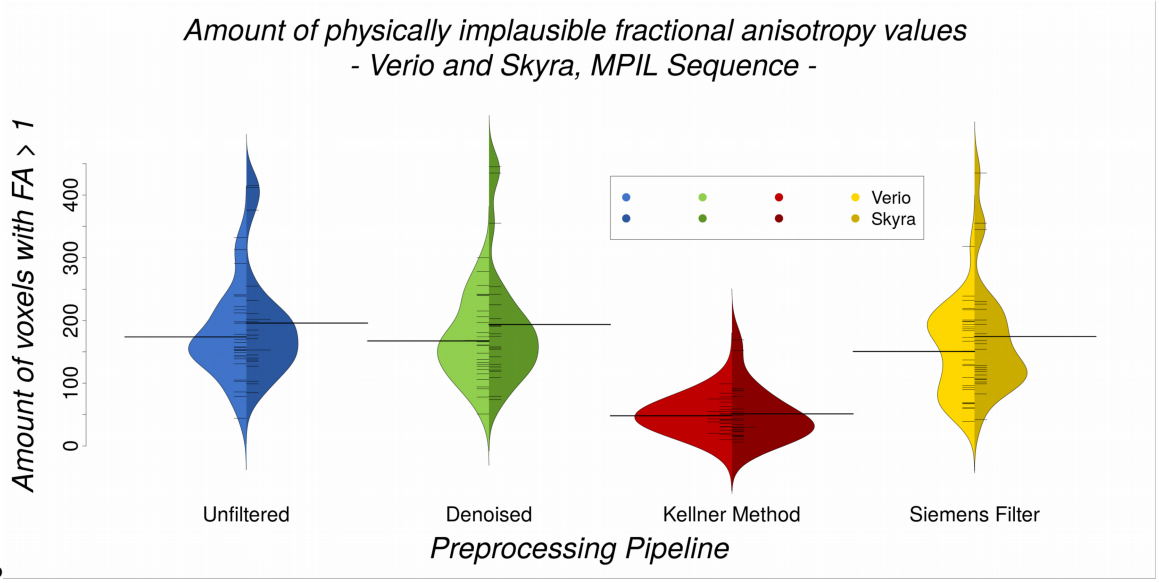


Figure 9: Amount of physically implausible FA values (Verio and Skyra, MPIL sequence, n=23 + 28) is most reduced after unringing (red) with the Kellner Method (paired Bayesian t-test: BF[unfiltered ~ Kellner Method] > $8*10^{9}$) but does not differ significantly between scanners neither before nor after differing preprocessing pipelines (Bayesian linear modelling, full/null model comparison: $BF\left[\frac{(voxelswithFA>1)\sim scanner*preprocessingpipeline}{(voxelswithFA>1)\sim preprocessingpipeline}\right]$ = 0.05±1.09%).

Motion effects

Comparing head motion quantified as mean frame-wise displacement (FD) values (in mm) between scanners (CMRR sequence, n=115), it became evident that the estimated motion effects differed significantly (Bayesian linear modelling, full/null model comparison: $BF\left[\frac{meanFDvalue\sim scanner*preprocessingpipeline}{meanFDvalue\sim preprocessingpipeline}\right]$ > $4.6*10^{70}$ ±1.96%). Regarding the significant differences between levels of preprocessing (Bayesian linear modelling, full/null model comparison: $BF\left[\frac{meanFDvalue\sim scanner*preprocessingpipeline}{meanFDvalue\sim scanner}\right]$ > $2.4*10^{11}$±1.93%) (Supplementary Figure S2), estimated motion effects could be attenuated significantly by applying denoising and unringing (see BF of post-hoc paired Bayesian t-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

tests in Table 3). Mean FD values are shown in Table 4. However, scanner differences remained significant even after further preprocessing (paired Bayesian t-test, n=115, all BF > 10^{18}).

Table 3: Results of statistical analysis of head motion values (FD) with and without preprocessing with paired Bayesian t-tests. Denoising and unringing reduce head motion artefacts significantly. The CoVs of the preprocessing pipeline differences in head motion differ largely.

Contrast of preprocessing pipelines	Bayes Factor of paired t-test on mean FD values (n=115)	CoV | preprocessing step – preprocessing step | (%)	
		Verio	Skyra
unfiltered ~ denoised	> 2 * 10^9	20.5	27.4
unfiltered ~ Kellner Method	> 5 * 10^6	27.8	40.2
denoised ~ Kellner Method	0.241	61.6	71.8

Table 4: Absolute head motion values per preprocessing pipeline and per scanner. Head motion estimated from the diffusion weighted images is significantly lower in Skyra than Verio independent of preprocessing pipeline (all BF >> 3). The CoV of the scanner difference in head motion is of comparable size between preprocessing pipelines.

Preprocessing pipeline	Mean FD value ± SD (mm)		CoV (%) | Verio - Skyra |
	Verio	Skyra	
unfiltered	0.417 ± 0.061	0.293 ± 0.064	41.9
denoised	0.355 ± 0.066	0.263 ± 0.066	47.3
Kellner Method	0.364 ± 0.067	0.269 ± 0.067	46.9

Physiological effects of interest

The age effect on whole brain WM skeleton mean FA equalled approximately -0.06% per year (CMRR sequence, after unringing). The effect was estimated with linear modelling (mean FA/MD value ~ age) and comparable between scanners (estimate ±std. error): mean FA: Verio: -3.203*10^{-4} ±1.237*10^{-4} and Skyra: -2.957*10^{-4} ±1.347*10^{-4} (Figure 10); mean MD: Verio: -2.420*10^{-7} ±1.654*10^{-7} and Skyra: -2.584*10^{-7} ±1.722*10^{-7}. Bayesian linear modelling confirmed the significance of the negative age effect on the FA skeleton in a full/null model comparison: $BF\left[\frac{meanFAvalue \sim age * scanner}{meanFAvalue \sim scanner}\right]$ = 5.49 ±1.39%. but a potential age effect on mean MD value could not be confirmed with Bayesian linear modelling: $BF\left[\frac{meanMDvalue \sim age * scanner}{meanMDvalue \sim scanner}\right]$ = 0.23 ±1.04%.

The negative age effect can be further depicted by TBSS on the FA skeleton and is observable in Verio and Skyra. By extracting t-values of the t-maps from TBSS, we could confirm that - in line with the absolute FA value approach - the age effect was slightly stronger in Verio (mean t-value = 0.517) than in Skyra (mean t-value = 0.453) (CMRR sequence, after unringing; Supplementary Figure S3, images above, t-values averaged over all voxels). In the case of studies conducted on different scanners (simulated by randomly selected subjects from Skyra and Verio), the age effect size was of intermediate magnitude (Figure 10; Supplementary Figure S3, image below).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

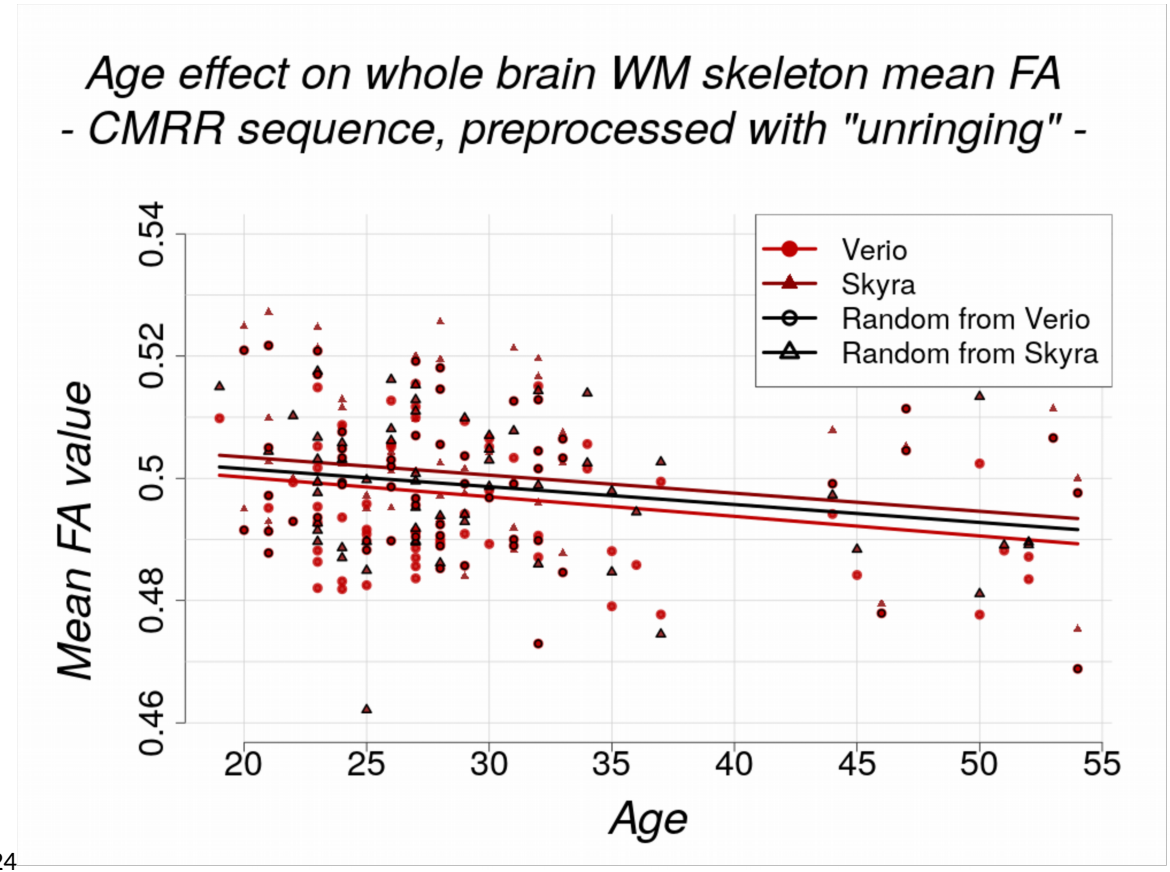


Figure 10: Negative effect of age on whole brain WM skeleton for both scanners separately (red circles: Verio, dark red triangles: Skyra, black framed circles and triangles: randomly selected to simulate pooled dataset from two different scanners). The effect modelled with linear modelling (mean FA value ~ age) is comparable between scanners (estimate ±std. error: Verio: Verio: $-3.203*10^{-4}$ $\pm 1.237*10^{-4}$ and Skyra: $-2.957*10^{-4}$ $\pm 1.347*10^{-4}$, random scanner: $-2.925*10^{-4}$ $\pm 1.265*10^{-4}$). Bayesian linear modelling delivered significant results for the negative age effect: $BF\left[\frac{meanFAvalue \sim age * scanner}{meanFAvalue \sim scanner}\right]$ = 5.49 ±1.39%. In the case of studies conducted on different scanners (simulated by randomly selected subjects from Skyra and Verio, black line), the age effect size was still present and of intermediate magnitude.

Harmonisation attempt

As suggested by Pohl et al. (2016), we calculated a whole brain correction factor in order to potentially harmonise the data and obtained $\frac{meanFA(Verio)}{meanFA(Skyra)} = cf = 0.9892$ (after denoising), which was comparable between preprocessing pipelines (Bayesian linear modelling, n=115: BF[mean FA ratio ~ preprocessing pipeline] = 0.6; Figure 11). Calculations of the correction factors for the selected ROIs (SCC, LSLF, LUF) however yielded correction factors which differed significantly (after denoising/after unringing with Kellner Method) depending on ROI: SCC = 1.055/1.0575, LSLF = 0.9885/0.9903, LUF = 0.9884/0.9881 (Bayesian linear modelling, n=115: $BF\left[\frac{meanFAratio \sim ROI * preprocessingpipeline}{meanFAratio \sim preprocessingpipeline}\right]$ >> $2*10^{196}$ ±1.15%.; Figure 11). The different levels of preprocessing did not play a major role here (Bayesian linear modelling, n=115: $BF\left[\frac{meanFA(Skyra) - meanFA(Verio)}{meanFA(Verio)}\right]$ < $7*10^{-6}$ ±1.52%).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2021 doi:10.20944/preprints202107.0134.v2

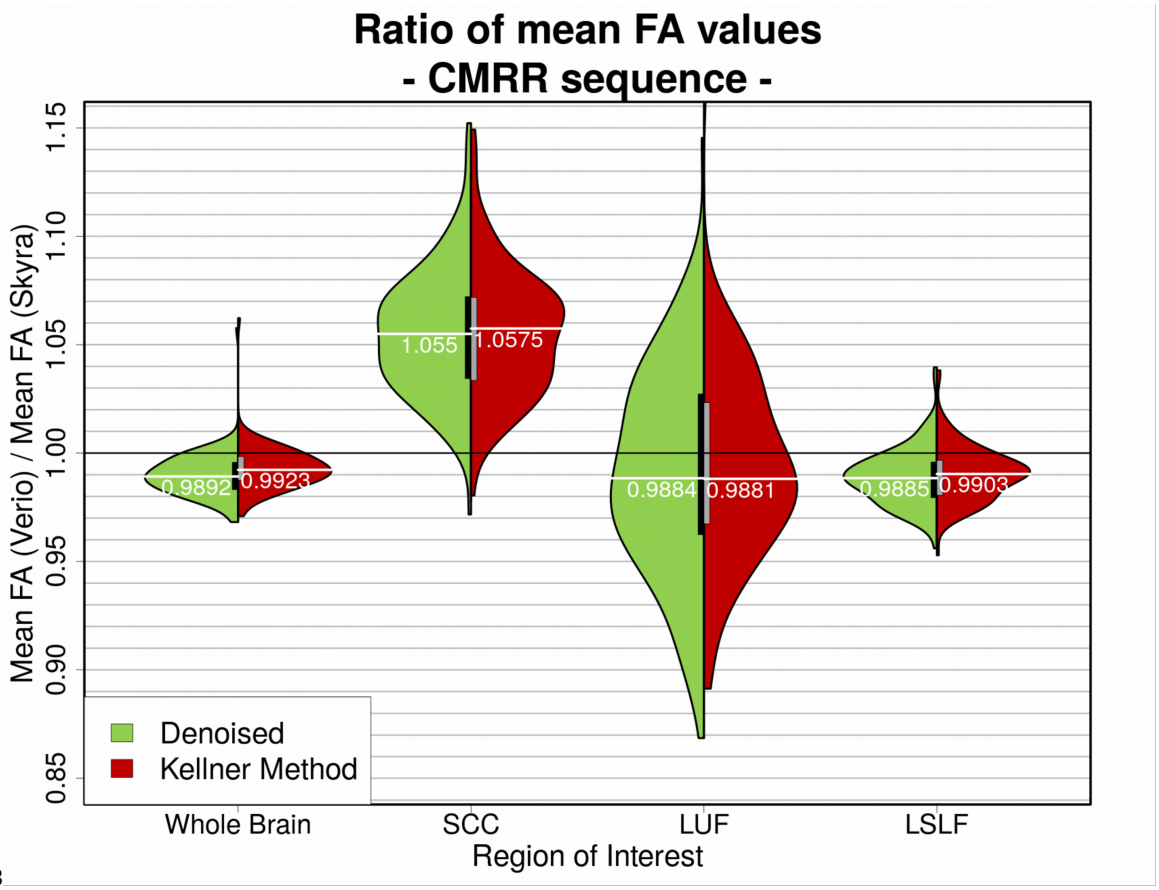


Figure 11: Ratio of whole brain mean FA values (Verio divided by Skyra) in different ROIs. For better visual clarity, unfiltered data is not shown. The left, green part of the violins represents denoised data; the right, red part shows the ratios after denoising and unringing. Bayesian linear modelling delivered significant differences between ROIs: $BF\left[\frac{meanFAratio \sim ROI * preprocessingpipeline}{meanFAratio \sim preprocessingpipeline}\right]$ >> 2*10196 ±1.15%. Mean FA ratios calculated after different preprocessing pipelines do not differ significantly: BF[mean FA ratio ~ preprocessing pipeline] = 0.6.

Discussion

Using a large sample size of healthy adults that underwent repeated MRI scanning at 3 Tesla with state-of-the-art acquisition and (pre)processing pipelines, we here report systematic global and regional differences in common DTI outcome measures between different scanners, sequences and pipelines. More specifically, we observed relative mean skeletonised FA value differences between scanners of up to 5% across brain regions and relative mean skeletonised MD value differences between scanner of up to 14%, which may well exceed potential effects of ageing (estimated to reach about 0.14%) or effects of disease on these measures. In addition, we found that the unringing tool from Kellner et al. (2016) reduced Gibbs ringing artefacts satisfactorily as opposed to other preprocessing approaches without unringing. Head motion quantified as mean frame-wise displacement (FD) values were consistently lower in the scanning sessions at Skyra compared to Verio, and motion-related artefacts were additionally reduced after preprocessing by denoising or unringing.

 doi:10.20944/preprints202107.0134.v2

Regional FA and MD variability due to different scanners and sequence parameters

Scanner differences in DTI outcome measures after "state of the art" preprocessing of DWI data with denoising range from about 1% globally to about 5% locally on the FA skeleton and from about 2% globally to about 14% locally on the MD skeleton. Our findings of these immense variations across distinct ROIs of the WM skeleton's mean FA and MD values have to our knowledge not yet been reported, and question recent attempts to harmonise multi-centre DWI data (as suggested by Pohl et al. (2016)): variations included not only large differences in magnitude but differed for mean FA also in direction (Skyra vs. Verio) mean FA: whole brain: ~+1%, SCC: ~ -5.2%, LUF: ~+1%, LSLF: ~+1.1%; mean MD: whole brain ~+2%, SCC: ~+14%, LUF: ~+3%, LSLF: ~ -0.2%. . We further calculated the coefficient of variance (CoV) of WM skeleton mean difference FA and MD values ($|\text{Verio} - \text{Skyra}|$ in %, single subject, voxel-based) after preprocessing with denoising on a whole brain level and in selected ROIs. This inter-scanner CoV for the mean difference FA ranged from ~7% globally to ~30% locally and is thereby locally much higher than inter-scanner CoVs from previous studies (1.0% (Vollmar et al., 2010) to 4.1% (Palacios et al., 2017) to 14.4% (Teipel et al., 2011)); similarly, the inter-scanner CoV for the mean difference MD ranged from ~11% globally to ~28% locally, also much higher than the according inter-vendor Siemens CoV from Prohl et al. (2019) of 4.4%. Our CoVs compared to past studies show that the mean FA and MD values might not be as robust to inter-scanner variations as previously assumed.

Pohl and colleagues (2016) harmonised scanner differences with correction factors for whole brain mean FA values regardless of variation across brain regions. This would lead to either fortified or attenuated local effects and is therefore not suitable for clinical diagnostics or studies focusing on regional effects.

TBSS on the FA skeleton presented a pattern with higher FA values for Verio data in more superficial white matter and higher FA values for Skyra data in rather deep WM areas. We observed a bias similar to this pattern when comparing anatomical MR images from Verio and Skyra (Medawar et al., 2020), namely Skyra exhibiting higher cortical thickness and larger GM volumes in medial frontal and central regions and Verio showing higher cortical thickness in lateral and occipital regions. This pattern could be caused by scaling differences between scanners that were observed on the anatomical images and could affect diffusion image processing due to registration on the individual anatomical images. TBSS on the MD skeleton showed that values from Skyra are almost uniformly higher throughout the whole brain. MD values might be more sensitive to the possible difference in the homogeneity of the magnetic field which leads to a locally variable signal-to-noise ratio between the different scanners. Therefore, higher MD values throughout the whole brain might reflect that DWI data from Skyra is less noisy then DWI data from Verio.

These findings emphasise that a retrospective correction for scanning at different imaging sites is hardly possible. Therefore "imaging site" should always be considered as a covariate in statistical analyses. Possible sources for such large differences between scanners could lie in hardware (e.g. radio frequency transmission, receiver coil sensitivity or signal processing elements) or software (reconstruction algorithms, data processing) differences.

We also showed that DWI data from sequences with harmonised but not identical parameters collected at the same scanner present region-dependent differences in TBSS, which is in line with earlier studies suggesting a strong sensitivity of DWI and its outcome measures to sequence parameters (Holdsworth & Bammer, 2008). The only difference between the CMRR and MPIL protocol identifiable with the scanner software was the amount of k-space reconstruction (partial Fourier). For EPI sequences, a large k-space coverage is necessary to reduce the EPI readout time and therefore increase the image quality, and thus 6/8 for MPIL and 7/8 for CMRR of k-space lines were acquired. Nevertheless, we cannot exclude that this sampling difference could cause slight image quality differences due to different k-space coverage. Yet, those are expected to be global differences in resolution, e.g. more blurriness for lower coverage, but not regionally specific effects.

Of note, a negative influence of age on whole brain WM coherence (represented by skeletonised mean FA values) as physiological effect of interest is much smaller— 0.06% reduction per year in this relatively young sample and 0.14%

reduction per year in the additionally analysed older cohort of the LIFE Adult Study— than the differences introduced by multi-site (and also partly by multi-sequence) data collection.

Despite the relatively young cohort (29.9±8.2 y.o.), we confirmed the negative effect of ageing on WM coherence (estimated cross-sectionally by FA) in analyses within-scanners. However, when pooling data from the two imaging sites, the physiological effect of age detected with TBSS was not extinguished but attenuated and changed in regional extent. We failed to detect an effect of ageing on the WM skeleton MD values possibly because FA may be a more specific measures of age-related changes in WM. Nevertheless, we conclude that pooling datasets from different imaging sites might fail to detect small effect sizes and/or may deliver regionally inconsistent patterns of the effect of interest if during analysis it is not accounted for the different imaging sites. This is especially crucial in clinical diagnostics if patients are scanned at different imaging sites. In this case, pathological changes could be attenuated or masked and therefore be missed.

As we did not assess intra-scanner variability, by e.g. repeating the same imaging protocol on the same scanner, the observed differences might be partly due to intra-scanner variability. Yet, a previous study showed high reliability of intra-scanner DTI metrics which was similar to intra-session differences and mainly influenced by the applied preprocessing steps (Madhyastha et al., 2014). Therefore, our finding of systematic differences between scanners is likely to be driven mainly by inter-scanner variability, largely independent of intra-scanner variability. Nevertheless, future studies should incorporate a test/retest intra-scanner acquisition as to quantify the contributions of the different sources of variability.

To account for a spatial heterogeneity of scanner differences, Fortin and colleagues (2017) suggested ComBat as a tool to harmonise FA and MD maps. ComBat is a batch effect correction tool used in genomics (Johnson et al., 2007) which aims to remove site effects from DTI maps and seems to preserve biological phenotype such as age. Yet, the locally largely differing CoVs as well as their divergence from the whole brain CoV indicate that the extent of scanner differences is not consistent across regions and subjects, rendering retrospective correction difficult. Future analyses need to test if applying ComBat in multi-site DWI effectively reduces between-scanner variance.

Taken together, our finding of gross regional differences in skeletonised FA and MD values between scanners and sequences strongly argue to keep imaging parameters stable if possible and to remain with data collection at one imaging site, or to increase sample sizes dramatically in multi-site studies to adjust for the reduction of statistical power. In the clinical context, we recommend to rescan a patient at the same MRI machine.

Gibbs ringing and motion artefacts

Regarding attempts to reduce common artefacts such as the Gibbs ringing (GR), we compared three different preprocessing approaches plus data without additional filtering, and demonstrated that visually and quantitatively the unringing tool from Kellner et al. (2016) reduced the GR artefact most efficiently. To the best of our knowledge, quantitative assessment of the GR has so far not yet been established with an easy, ready-to-apply method which is why we introduced the amount of voxels with an implausible FA value (FA>1) as an approximation of the amount of GR. We confirmed by visual inspection that the implausible FA values in the selected voxels (FA>1) were caused by GR and not by other artefacts. GR can of course affect FA values without causing them to exceed 1, especially in areas with lower FA values, so that the amount of implausible FA values cannot be seen as an absolute measure of GR but rather as a conservative estimation of the number of voxels clearly affected by GR. Even though other measurement noise could affect FA to exceed 1 by e.g. causing negative eigenvalues (Koay et al., 2006), other preprocessing tools such as the Siemens low-pass window filtering supposed to address this noise or the denoising tool from MRtrix did not attenuate the GR visually or the amount of FA>1 quantitatively. Additionally, eddy current and vibration artefacts could lead to systematic patterns of artificially high FA values with a particular spatial pattern. In our experiment,

eddy currents were successfully compensated by the twice refocused pulse-sequence and we did not observe vibration artefacts in any of the measurements. We therefore conclude that FA>1 is an appropriate lower approximation of GR in our data. We suggest to include the unringing tool from Kellner et al. (2016) in DW preprocessing in order to increase data quality and to possibly mitigate differences between data from different scanners before pooling them in a multi-site study. Promising future steps towards automatic GR artefact detection and reduction besides the Kellner tool might be the application of convolutional neural networks as suggested and experimentally verified by Zhang et al. (2019), Zhao et al. (2020) and Muckley et al. (2021).

Regarding head motion, mean FD values were estimated consistently lower in the scanning sessions at Skyra. Even though, all participants underwent their second scan at Skyra, most participants are very MRI-experienced and therefore the chronology of the scanning sessions unlikely explains the considerable attenuation in head motion. While speculative, we suppose DWI is that demanding for the hardware such as the gradient coils that the wearing off during the years of use (Verio in use since 2008) compared to Skyra (upgraded in 2016) might have an effect on the increased estimated motion effects in Verio. The scanners might show a slight difference in the non-compensated eddy currents or a scanner drift which might result in a difference in the estimated head motion parameters by the FSL eddy tool. This tool estimates the eddy currents and head motion at the same time and both estimations are not independent. FD values could be reduced by including denoising or unringing in the preprocessing pipeline. This reduction in apparent head motion can be explained by an improved image quality introduced through these filtering techniques and therefore improved motion estimation. Nevertheless, differences in head motion were used as covariate in statistical analyses and did not influence scanner, sequence or pipeline differences significantly.

Limitations and strengths

Our study includes two scanner systems of the same manufacturer and two diffusion-weighted sequences – a main limitation therefore is that it does not reflect the whole range of most commonly used MRI systems in clinics and research neither all of the most commonly implemented DWI sequences. However, considering that the two very similar scanner systems and two harmonised sequences exhibit considerable differences in DTI outcome measures, it can be speculated that more differing scanners and sequences exhibit more substantial differences.

More detailed limitations include that we did not correct for gradient non-linearities which could affect diffusion tensor metrics in a similar way as the apparent diffusion coefficient as shown by Fedeli et al. (2018) and Tan et al. (2013) and thereby account for some of the differences we found on the mean FA and MD skeletons. However the small non-linearity in the gradients of the used clinical MR systems, the relative small field-of-view (only the brain) and the comparably low b-value (b=1000 s/mm²) minimize the effect of gradient non-linearity on image distortions and b-value variation compared to other studies where such corrections are needed (e.g. Human Connectome Project, van Essen et al., 2013, Jones 2010). To further control for a correct application of the diffusion gradients, the positioning of the participants followed a standardised protocol to position the brain in the isocenter of the gradient coil which presents the smallest non-liberalities. Additionally, in a parallel comparison of the anatomical images from Verio and Skyra, gradient non-linearity correction— conducted with the gradunwarp implementation [https://github.com/Washington-University/gradunwarp] in Python 2.7.)— did not substantially reduce the detected differences (Medawar et al., 2020). This is why we estimate that gradient non-linearities might only be a small share of the sources leading to the differences in the WM skeletons between the scanners.

Concerning the quantification of the GR artefact, the linear least squares method employed by Lipsia to estimate the diffusion tensor in our preprocessing pipeline, also used in widely established DWI preprocessing software such as FSL's dtifit, comes with the negative eigenvalue problem. Negative eigenvalues can be caused by measurement noise and lead to FA values larger than 1. This is e.g. the case for voxel presenting the GR artefact where the signal of the

 doi:10.20944/preprints202107.0134.v2

diffusion weighted images is locally increased and might be higher that the non-diffusion weighted (b0) signal in the same voxel. The linear tensor fit leads to physically implausible negative eigenvalues (and therefore FA>1) in those voxels. Koay et al. (2006) showed in simulations that the constrained non-linear least squares method is, in terms of mean squared error for estimating trace and FA, the most effective method for correcting negative eigenvalues. Studies focussing on FA and areas with high anisotropy such as the corpus callosum should therefore reconsider the approach to estimate the diffusion tensor in order to ensure data quality. In our case, measurement noise might have been different in areas with high anisotropy leading to inflated differences in the FA skeleton between scanners and sequences, especially in the splenium of the corpus callosum (SCC).

Additionally, as scanner order could not be randomised in this project due to scheduling issues, all participants underwent the first MRI at Verio and the second at Skyra which may have led to effects of scanning order we could not account for in our analysis. Further, we did not include a retest measurement on the same scanner to discriminate between within- and between-scanner effects. Lastly, we did not monitor hydration state and time of day at scanning, factors which could also affect measures of brain microstructure (Streitbuerger et al., 2012).

Nevertheless, this work excels with its large amount of participants and longitudinal design with closely timed acquisitions, rendering true effects of seasonal or age-related changes practically unlikely. Including two MRI systems which are usually linked by an upgrade, namely 3T Siemens Magnetom Skyra^{fit} as upgraded version of Verio, the consequences of such scanner upgrades on DTI outcome measures can be directly inferred from our study. Such across-upgrade investigations have been to date very rare (Zhan et al., 2014; Fox et al., 2011).

Conclusions

In summary, based on two widely used Siemens MRI systems of the same field strength and two established DWI acquisition sequences we demonstrate that the reproducibility of DTI outcome measures strongly depends on imaging site, software and brain region. This is an alarming finding considering the importance of replicability of MRI assessments in the clinical context and increasing availability and diverseness of research-oriented MRI assessments on a large scale. It also underlines the necessity to carefully document, correct and adjust for different modifications of imaging parameters and applied data analysis pipelines. If not controlled for, such variations lead to much larger sample sizes which compensate the loss of statistical power. Our findings further support the use of the Gibbs ringing correction tool from Kellner et al. (2016), encourage to adhere to one imaging system, scanning protocol and preprocessing pipeline and to conscientiously document every change in the aforementioned steps. Moreover, physiological effects such as ageing reflected in the decrease of FA were found to be robust against scanner differences and may be traceable despite variation in DWI data collection and processing, however, by the cost of a reduced effect signal and regional specificity. Regarding clinical applications, the potential impact of these variations on pathological changes should be kept in mind when assessing DWI data. Future studies need to further develop novel strategies to harmonise data acquisition and retrospective correction of hardware- and software-introduced differences in common MRI outcome measures to augment neuroimaging data reliability and replicability.

Supplementary Material

Figure S1: ROIs selected for ROI-based analyses.

Figure S2: Motion effects quantified as mean frame-wise displacement (Verio and Skyra, CMRR sequence, n=115) differ between preprocessing pipelines.

Figure S3: TBSS results of the negative age effect on the whole brain WM skeleton compared between scanners.

Funding

This work was supported by the German Research Foundation (DFG), contract grant numbers WI 3342/3-1 and 209,933,838 – SFB 1052.

Acknowledgements

We are thankful to all contributors of the LIFE-Upgrade Study for helping with data collection, namely C. Barth, T. Ballerini, N. Dermody, J. Grothe, M. Lammert, S. Huhn, E. Medawar, M. Polyakova, J. Sacher, L. Schaare, U. Scharrer, M. Schroeter, B. Sehm and K. Thomas. I further want to thank André Pampel for advising with, matching and setting up the scanning protocol.

Author Contributions

Conceptualization, Frauke Beyer and A. Veronica Witte; Data curation, Ronja Thieleking and Alfred Anwander; Formal analysis, Ronja Thieleking and Rui Zhang; Funding acquisition, Arno Villringer and A. Veronica Witte; Investigation, Ronja Thieleking and Maria Paerisch; Methodology, Ronja Thieleking, Rui Zhang, Alfred Anwander, Frauke Beyer, Arno Villringer and A. Veronica Witte; Project administration, Maria Paerisch, Kerstin Wirkner and A. Veronica Witte; Supervision, Frauke Beyer and A. Veronica Witte; Writing – original draft, Ronja Thieleking; Writing – review & editing, Ronja Thieleking, Rui Zhang, Maria Paerisch, Kerstin Wirkner, Alfred Anwander, Frauke Beyer, Arno Villringer and A. Veronica Witte.

Ethics Statement

This study involving human participants adhered to the Human Subjects Guidelines of the Declaration of Helsinki and the Ethics Committee of the Medical Faculty of the University of Leipzig raised no obejctions (no. 289-15-13072015 and no. 263-2009-14122009). The participants provided their written informed consent to participate in this study.

Data Availability Statement

Data (such as FA and FD values, ROIs, scanning protocols and data for age correlation) are stored at the Open Science Framework and openly available (https://osf.io/vnuqp/). Imaging data (anatomical and diffusion-weighted) are only available from subjects who signed additional consent (n=57) and can be shared upon request (please contact witte@cbs.mpg.de).

Conflict of Interest Statement

The authors declare no conflict of interest.

References

Alexander, D. C., & Barker, G. J. (2005). Optimal imaging parameters for fibre-orientation estimation in diffusion MRI. NeuroImage, 27(2), 357–367. https://doi.org/10.1016/j.neuroimage.2005.04.008

Andersson, J. L. R., & Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. NeuroImage, 125, 1063–1078. https://doi.org/10.1016/j.neuroimage.2015.10.019

Assaf, Y., & Pasternak, O. (2008). Diffusion Tensor Imaging (DTI)-based White Matter Mapping in Brain Research: A Review. Journal of Molecular Neuroscience, 34(1), 51–61. https://doi.org/10.1007/s12031-007-0029-0

Basser, P. J., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. Biophysical Journal, 66(1), 259–267. https://doi.org/10.1016/S0006-3495(94)80775-1

Basser, P. J., & Jones, D. K. (2002). Diffusion-tensor MRI: theory, experimental design and data analysis - a technical review. NMR in Biomedicine, 15(7–8), 456–467. https://doi.org/10.1002/nbm.783

Belli, G., Busoni, S., Ciccarone, A., Coniglio, A., Esposito, M., Giannelli, M., … Gobbi, G. (2016). Quality assurance multicenter comparison of different MR scanners for quantitative diffusion-weighted imaging. Journal of Magnetic Resonance Imaging, 43(1), 213–219. https://doi.org/10.1002/jmri.24956

Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., Lampe, L., Luck, T., Riedel-Heller, S. G., … Witte, A. V. (2017). Higher body mass index is associated with reduced posterior default mode connectivity in older adults. Human Brain Mapping, 38(7), 3502–3515. https://doi.org/10.1002/hbm.23605

Blumenfeld-Katzir, T., Pasternak, O., Dagan, M., & Assaf, Y. (2011). Diffusion MRI of Structural Brain Plasticity Induced by a Learning and Memory Task. PLoS ONE, 6(6), e20678. https://doi.org/10.1371/journal.pone.0020678

Branzoli, F., Ercan, E., Valabrègue, R., Wood, E. T., Buijs, M., Webb, A., & Ronen, I. (2016). Differentiating between axonal damage and demyelination in healthy aging by combining diffusion-tensor imaging and diffusion-weighted spectroscopy in the human corpus callosum at 7 T. Neurobiology of Aging, 47, 210–217. https://doi.org/10.1016/j.neurobiolaging.2016.07.022

Buchanan, C. R., Pernet, C. R., Gorgolewski, K. J., Storkey, A. J., & Bastin, M. E. (2014). Test–retest reliability of structural brain networks from diffusion MRI. NeuroImage, 86, 231–243. https://doi.org/10.1016/j.neuroimage.2013.09.054

Constable, R. T., & Henkelman, R. M. (1991). Data extrapolation for truncation artifact removal. Magnetic Resonance in Medicine, 17(1), 108–118. https://doi.org/10.1002/mrm.1910170115

de Groot, M., Cremers, L. G. M., Ikram, M. A., Hofman, A., Krestin, G. P., van der Lugt, A., … Vernooij, M. W. (2016). White Matter Degeneration with Aging: Longitudinal Diffusion MR Imaging Analysis. Radiology, 279(2), 532–541. https://doi.org/10.1148/radiol.2015150103

de Groot, M., Ikram, M. A., Akoudad, S., Krestin, G. P., Hofman, A., Van Der Lugt, A., … Vernooij, M. W. (2015). Tract-specific white matter degeneration in aging: The Rotterdam Study. Alzheimer's and Dementia, 11(3), 321–330. https://doi.org/10.1016/j.jalz.2014.06.011

Fedeli, L., Belli, G., Ciccarone, A., Coniglio, A., Esposito, M., Giannelli, M., … Busoni, S. (2018). Dependence of apparent diffusion coefficient measurement on diffusion gradient direction and spatial position – A quality assurance intercomparison study of forty-four scanners for quantitative diffusion-weighted imaging. Physica Medica, 55(July), 135–141. https://doi.org/10.1016/j.ejmp.2018.09.007

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., … Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. NeuroImage, 161(August), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Fox, R. J., Sakaie, K., Lee, J.-C., Debbins, J. P., Liu, Y., Arnold, D. L., … Fisher, E. (2012). A validation study of multicenter diffusion tensor imaging: reliability of fractional anisotropy and diffusivity values. AJNR. American Journal of Neuroradiology, 33(4), 695–700. https://doi.org/10.3174/ajnr.A2844

Gelb, A., & Archibald, R. (2002). Reducing the Gibbs ringing artifact in MRI scans while maintaining tissue boundary integrity. In Proceedings IEEE International Symposium on Biomedical Imaging (pp. 923–926). IEEE. https://doi.org/10.1109/ISBI.2002.1029412

Giannelli, M., Cosottini, M., Michelassi, M. C., Lazzarotti, G., Belmonte, G., Bartolozzi, C., & Lazzeri, M. (2010). Dependence of brain DTI maps of fractional anisotropy and mean diffusivity on the number of diffusion weighting directions. Journal of Applied Clinical Medical Physics, 11(1), 176–190. https://doi.org/10.1120/jacmp.v11i1.2927

Goveas, J., O'Dwyer, L., Mascalchi, M., Cosottini, M., Diciotti, S., De Santis, S., … Giannelli, M. (2015). Diffusion-MRI in neurodegenerative disorders. Magnetic Resonance Imaging, 33(7), 853–876. https://doi.org/10.1016/j.mri.2015.04.006

Hasan, K. M., Parker, D. L., & Alexander, A. L. (2001). Comparison of gradient encoding schemes for diffusion-tensor MRI. Journal of Magnetic Resonance Imaging, 13(5), 769–780. https://doi.org/10.1002/jmri.1107

Holdsworth, S., & Bammer, R. (2008). Magnetic Resonance Imaging Techniques: fMRI, DWI, and PWI. Seminars in Neurology, 28(04), 395–406. https://doi.org/10.1055/s-0028-1083697

Horsfield, M. A., & Jones, D. K. (2002). Applications of diffusion-weighted and diffusion tensor MRI to white matter diseases - a review. NMR in Biomedicine, 15(7–8), 570–577. https://doi.org/10.1002/nbm.787

Jellison, B. J., Field, A. S., Medow, J., Lazar, M., Salamat, M. S., & Alexander, A. L. (2004). Diffusion tensor imaging of cerebral white matter: a pictorial review of physics, fiber tract anatomy, and tumor imaging patterns. AJNR. American Journal of Neuroradiology, 25(3), 356–369. Retrieved from http://www.ajnr.org/content/25/3/356

Johansen-Berg, H., & Behrens, T. E. J. (2014). Diffusion MRI : from quantitative measurement to in-vivo neuroanatomy. Elsevier Science.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Jovicich, J., Marizzoni, M., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., … Frisoni, G. B. (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

Kellner, E., Dhital, B., Kiselev, V. G., & Reisert, M. (2016). Gibbs-ringing artifact removal based on local subvoxel-shifts. Magnetic Resonance in Medicine, 76(5), 1574–1581. https://doi.org/10.1002/mrm.26054

Koay, C. G., Carew, J. D., Alexander, A. L., Basser, P. J., & Meyerand, M. E. (2006). Investigation of anomalous estimates of tensor-derived quantities in diffusion tensor imaging. Magnetic Resonance in Medicine, 55(4), 930–936. https://doi.org/10.1002/mrm.20832

Kuhn, T., Gullett, J. M., Nguyen, P., Boutzoukas, A. E., Ford, A., Colon-Perez, L. M., … Bauer, R. M. (2016). Test-retest reliability of high angular resolution diffusion imaging acquisition within medial temporal lobe connections assessed via tract based spatial statistics, probabilistic tractography and a novel graph theory metric. Brain Imaging and Behavior, 10(2), 533–547. https://doi.org/10.1007/s11682-015-9425-1

Leemans, A. (2010). Theory and applications of diffusion MRI. In 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (pp. 628–631). IEEE. https://doi.org/10.1109/ISBI.2010.5490100

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods, 264, 47–56. https://doi.org/10.1016/j.jneumeth.2016.03.001

Loeffler, M., Engel, C., Ahnert, P., Alfermann, D., Arelin, K., Baber, R., … Thiery, J. (2015). The LIFE-Adult-Study: Objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. BMC Public Health. https://doi.org/10.1186/s12889-015-1983-z

Lohmann, G., Müller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., … von Cramon, D. Y. (2001). Lipsia—a new software system for the evaluation of functional magnetic resonance images of the human brain. Computerized Medical Imaging and Graphics, 25(6), 449–457. https://doi.org/10.1016/S0895-6111(01)00008-8

Madhyastha, T., Mérillat, S., Hirsiger, S., Bezzola, L., Liem, F., Grabowski, T., & Jäncke, L. (2014). Longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging. Human Brain Mapping, 35(9), 4544–4555. https://doi.org/10.1002/hbm.22493

Malyarenko, D., Galbán, C. J., Londy, F. J., Meyer, C. R., Johnson, T. D., Rehemtulla, A., … Chenevert, T. L. (2013). Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. Journal of Magnetic Resonance Imaging : JMRI, 37(5), 1238–1246. https://doi.org/10.1002/jmri.23825

Medawar, E., Thieleking, R., Manuilova, I., Villringer, A., Witte, A. V., & Beyer, F. (2020). Estimating the effect of a scanner upgrade on measures of grey matter structure for longitudinal designs. BioRxiv. https://doi.org/10.1101/2020.08.28.271296

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nature Neuroscience, 19(11), 1523–1536. https://doi.org/10.1038/nn.4393

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., … Rathi, Y. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. NeuroImage, 135, 311–323. https://doi.org/10.1016/j.neuroimage.2016.04.041

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain FMRI. Magnetic Resonance in Medicine, 63(5), 1144–1153. https://doi.org/10.1002/mrm.22361

Muckley, M. J., Ades-Aron, B., Papaioannou, A., Lemberskiy, G., Solomon, E., Lui, Y. W., … Knoll, F. (2021). Training a neural network for Gibbs and noise removal in diffusion MRI. Magnetic Resonance in Medicine, 85(1), 413–428. https://doi.org/10.1002/mrm.28395

Ni, H., Kavcic, V., Zhu, T., Ekholm, S., & Zhong, J. (2006). Effects of number of diffusion gradient directions on derived diffusion tensor imaging indices in human brain. AJNR. American Journal of Neuroradiology, 27(8), 1776–1781. Retrieved from http://www.ajnr.org/content/27/8/1776

Palacios, E. M., Martin, A. J., Boss, M. A., Ezekiel, F., Chang, Y. S., Yuh, E. L., … Mukherjee, P. (2017). Toward Precision and Reproducibility of Diffusion Tensor Imaging: A Multicenter Diffusion Phantom and Traveling Volunteer Study. American Journal of Neuroradiology, 38(3), 537–545. https://doi.org/10.3174/AJNR.A5025

Pan, C. (2001). Gibbs phenomenon removal and digital filtering directly through the fast Fourier transform. IEEE Transactions on Signal Processing, 49(2), 444–448. https://doi.org/10.1109/78.902128

Perrone, D., Aelterman, J., Pižurica, A., Jeurissen, B., Philips, W., & Leemans, A. (2015). The effect of Gibbs ringing artifacts on measures derived from diffusion MRI. NeuroImage, 120, 441–455. https://doi.org/10.1016/j.neuroimage.2015.06.068

Pfefferbaum, A., Adalsteinsson, E., & Sullivan, E. V. (2003). Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. Journal of Magnetic Resonance Imaging, 18(4), 427–433. https://doi.org/10.1002/jmri.10377

Pohl, K. M., Sullivan, E. V, Rohlfing, T., Chu, W., Kwon, D., Nichols, B. N., … Pfefferbaum, A. (2016). Harmonizing DTI measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the NCANDA study. NeuroImage, 130, 194–213. https://doi.org/10.1016/j.neuroimage.2016.01.061

Prohl, A. K., Scherrer, B., Tomas-Fernandez, X., Filip-Dhima, R., Kapur, K., Velasco-Annis, C., … the TACERN Study Group. (2019). Reproducibility of structural and diffusion tensor imaging in the TACERN multi-center study. Frontiers in Integrative Neuroscience, 13. https://doi.org/10.3389/fnint.2019.00024

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage, 59(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Sampaio-Baptista, C., & Johansen-Berg, H. (2017). White Matter Plasticity in the Adult Brain. Neuron, 96(6), 1239–1251. https://doi.org/10.1016/j.neuron.2017.11.026

Schilling, K. G., Nath, V., Blaber, J., Harrigan, R. L., Ding, Z., Anderson, A. W., & Landman, B. A. (2017). Effects of b-Value and Number of Gradient Directions on Diffusion MRI Measures Obtained with Q-ball Imaging. Proceedings of SPIE--the International Society for Optical Engineering, 10133. https://doi.org/10.1117/12.2254545

Schlett, C., Hendel, T., Weckbach, S., Reiser, M., Kauczor, H., Nikolaou, K., … Bamberg, F. (2016). Population-Based Imaging and Radiomics: Rationale and Perspective of the German National Cohort MRI Study. RöFo - Fortschritte Auf Dem Gebiet Der Röntgenstrahlen Und Der Bildgebenden Verfahren, 188(07), 652–661. https://doi.org/10.1055/s-0042-104510

Scholz, J., Klein, M. C., Behrens, T. E. J., & Johansen-Berg, H. (2009). Training induces changes in white-matter architecture. Nature Neuroscience, 12(11), 1370–1371. https://doi.org/10.1038/nn.2412

Schwartz, D. L., Tagge, I., Powers, K., Ahn, S., Bakshi, R., Calabresi, P. A., … Rooney, W. D. (2019). Multisite reliability and repeatability of an advanced brain MRI protocol. Journal of Magnetic Resonance Imaging, 50(3), 878–888. https://doi.org/10.1002/jmri.26652

Smith, S. M. (2002). Fast robust automated brain extraction. Human Brain Mapping, 17(3), 143–155. https://doi.org/10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., … Behrens, T. E. J. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. NeuroImage, 31(4), 1487–1505. https://doi.org/10.1016/j.neuroimage.2006.02.024

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., … Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage, 23, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Streitbuerger, D.-P., Möller, H. E., Tittgemeyer, M., Hund-Georgiadis, M., Schroeter, M. L., & Mueller, K. (2012). Investigating Structural Brain Changes of Dehydration Using Voxel-Based Morphometry. PLoS ONE, 7(8), e44195. https://doi.org/10.1371/journal.pone.0044195

Tan, E. T., Marinelli, L., Slavens, Z. W., King, K. F., & Hardy, C. J. (2013). Improved correction for gradient nonlinearity effects in diffusion-weighted imaging. Journal of Magnetic Resonance Imaging, 38(2), 448–453. https://doi.org/10.1002/jmri.23942

Tax, C. M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., John Evans, C., … Veraart, J. (2019). Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. NeuroImage, 195(February), 285–299. https://doi.org/10.1016/j.neuroimage.2019.01.077

Teipel, S. J., Reuter, S., Stieltjes, B., Acosta-Cabronero, J., Ernemann, U., Fellgiebel, A., … Hampel, H. (2011). Multicenter stability of diffusion tensor imaging measures: A European clinical and physical phantom study. Psychiatry Research: Neuroimaging, 194(3), 363–371. https://doi.org/10.1016/j.pscychresns.2011.05.012

Tournier, J.-D., Mori, S., & Leemans, A. (2011). Diffusion tensor imaging and beyond. Magnetic Resonance in Medicine, 65(6), 1532–1556. https://doi.org/10.1002/mrm.22924

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium, for the W.-M. H. (2013). The WU-Minn Human Connectome Project: an overview. NeuroImage, 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Veraart, J., Fieremans, E., Jelescu, I. O., Knoll, F., & Novikov, D. S. (2016). Gibbs ringing in diffusion MRI. Magnetic Resonance in Medicine, 76(1), 301–314. https://doi.org/10.1002/mrm.25866

Veraart, J., Fieremans, E., & Novikov, D. S. (2016). Diffusion MRI noise mapping using random matrix theory. Magnetic Resonance in Medicine, 76(5), 1582–1593. https://doi.org/10.1002/mrm.26059

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., … Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. NeuroImage, 51(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., … Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. NeuroImage, 45(1), S173–S186. https://doi.org/10.1016/j.neuroimage.2008.10.055

Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. Nature Neuroscience, 15(4), 528–536. https://doi.org/10.1038/nn.3045

Zavaliangos-Petropulu, A., Nir, T. M., Thomopoulos, S. I., Reid, R. I., Bernstein, M. A., Borowski, B., … Thompson, P. M. (2019). Diffusion MRI indices and their relation to cognitive impairment in brain aging: The updated multi-protocol approach in ADNI3. Frontiers in Neuroinformatics, 13, 2. https://doi.org/10.3389/fninf.2019.00002

Zhan, L., Jahanshad, N., Jin, Y., Nir, T. M., Leonardo, C. D., Bernstein, M. A., … Thompson, P. M. (2014). Understanding scanner upgrade effects on brain integrity & connectivity measures. 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), 234–237. https://doi.org/10.1109/isbi.2014.6867852

Zhang, Q., Ruan, G., Yang, W., Liu, Y., Zhao, K., Feng, Q., … Feng, Y. (2019). MRI Gibbs-ringing artifact reduction by means of machine learning using convolutional neural networks. Magnetic Resonance in Medicine, (October 2018), 1–13. https://doi.org/10.1002/mrm.27894

Zhang, R., Beyer, F., Lampe, L., Luck, T., Riedel-Heller, S. G., Loeffler, M., … Witte, A. V. (2018). White matter microstructural variability mediates the relation between obesity and cognition in healthy adults. NeuroImage, 172, 239–249. https://doi.org/10.1016/j.neuroimage.2018.01.028

Zhao, X., Zhang, H., Zhou, Y., Bian, W., Zhang, T., & Zou, X. (2020). Gibbs-ringing artifact suppression with knowledge transfer from natural images to MR images. Multimedia Tools and Applications, 79(45–46), 33711–33733.https://doi.org/10.1007/s11042-019-08143-6