Title: Longitudinal reproducibility of Neurite Orientation Dispersion and Density Imaging (NODDI) derived metrics in the white matter

Nico Lehmann*[,a,b], Norman Aye[a], Jörn Kaufmann[c], Hans-Jochen Heinze[c,d,e,f], Emrah Düzel[d,e,g,h], Gabriel Ziegler[d,g], & Marco Taubert[a,e]

[a] Faculty of Human Sciences, Institute III, Department of Sport Science, Otto von Guericke University, Zschokkestraße 32, 39104 Magdeburg, Germany

[b] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103 Leipzig, Germany

[c] Department of Neurology, Otto von Guericke University, Leipziger Straße 44, 39120 Magdeburg, Germany

[d] Germany German Center for Neurodegenerative Diseases (DZNE), Leipziger Straße 44, 39120 Magdeburg, Germany

[e] Center for Behavioral and Brain Science (CBBS), Otto von Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany

[f] Leibniz-Institute for Neurobiology (LIN), Brenneckestraße 6, 39118 Magdeburg, Germany

[g] Institute of Cognitive Neurology and Dementia Research, Otto von Guericke University, Leipziger Str. 44, 39120 Magdeburg, Germany

[h] Institute of Cognitive Neuroscience, University College London, Alexandra House, 17-19 Queen Square, Bloomsbury, London, UK, WC1N 3AZ

Available ORCID IDs

Lehmann, Nico: 0000-0002-3146-5084

Kaufmann, Jörn: 0000-0002-8513-7043

Düzel, Emrah: 0000-0002-0139-5388

Ziegler, Gabriel: 0000-0001-6589-6416

Corresponding author email address: nico1.lehmann@ovgu.de

Number of pages: 45

Number of figures: 6

Number of tables: 2

Abstract

Diffusion-weighted magnetic resonance imaging (DWI) is undergoing constant evolution with the ambitious goal of developing in-vivo histology of the brain. A recent methodological advancement is *Neurite Orientation Dispersion and Density Imaging (NODDI)*, a histologically validated multi-compartment model to yield microstructural features of brain tissue such as geometric complexity and neurite packing density, which are especially useful in imaging the white matter. Since NODDI is increasingly popular in clinical research and fields such as developmental neuroscience and neuroplasticity, it is of vast importance to characterize its reproducibility (or reliability). We acquired multi-shell DWI data in 29 healthy young subjects twice over a rescan interval of 4 weeks to assess the within-subject coefficient of variation (CV_{WS}), between-subject coefficient of variation (CV_{BS}) and the intraclass correlation coefficient (ICC), respectively. Using these metrics, we compared regional and voxel-by-voxel reproducibility of the most common image analysis approaches (tract-based spatial statistics [TBSS], voxel-based analysis with different extents of smoothing ["VBM-style"], ROI-based analysis). We observed high test-retest reproducibility for the "orientation dispersion index" (ODI) and slightly worse results for the "neurite density index" (NDI). Our findings also suggest that the choice of analysis approach might have significant consequences for the results of a study. Collectively, the voxel-based approach with Gaussian smoothing kernels of ≥ 4mm FWHM and ROI-averaging yielded the highest reproducibility across NDI and ODI maps (CV_{WS} mostly ≤ 3%, ICC mostly ≥ 0.8), respectively, whilst smaller kernels and TBSS performed consistently worse. Furthermore, we demonstrate that image quality (signal-to-noise ratio) is an important determinant of NODDI metric reproducibility. We discuss the implications of these results for longitudinal and cross-sectional research designs commonly employed in the neuroimaging field.

Keywords: Diffusion-weighted imaging, Neurite Orientation Dispersion and Density Imaging (NODDI), reproducibility, reliability, precision

Abbreviations

BBR – boundary-based registration

BMI – body mass index

CI – confidence interval

CSF – cerebrospinal fluid

CV – coefficient of variation

CV_{BS} – between-subject variation expressed as a CV

CV_{WS} – within-subject variation expressed as a CV

DW – diffusion-weighted

DWI – diffusion-weighted imaging

EPI – echo-planar imaging

FA – fractional anisotropy

FWHM – Full width at half maximum

FSL – FMRIB Software Library

GM – gray matter

ICC – intraclass correlation

ISO – isotropic volume fraction

MD – mean diffusivity

MRI – magnetic resonance imaging

NDI – neurite density index (aka intra-cellular volume fraction)

NODDI – Neurite Orientation Dispersion and Density Imaging

ODI – orientation dispersion index

RM-ANOVA – Repeated Measures Analysis of Variance

ROI – region of interest

SD – standard deviation

SNR – signal-to-noise ratio

TBSS – Tract-Based Spatial Statistics

T1w – T1-weighted

WM – white matter

Introduction

Although the cerebral cortex has a unique role in neural computation, about half of the human brain's volume is comprised of fibres, often summarized as white matter (WM) tissue (Walhovd et al., 2014). Within the brain, axons insulated by myelin sheaths connect remote cortical and subcortical gray matter (GM) regions, therefore enabling efficient neurotransmission subserving consciousness, emotions, cognition, and motor functions (Filley and Fields, 2016; Sampaio-Baptista and Johansen-Berg, 2017).

Non-invasive diffusion-weighted imaging (DWI) is a powerful magnetic resonance image technique which allows to infer microstructural features of WM such as axonal packing, membrane properties, and myelination *in vivo* by applying strong, directionally varying gradient fields (Alexander et al., 2019; Beaulieu, 2002; Jones et al., 2013). An increasing number of DWI studies have demonstrated that microstructural features of WM dynamically change throughout lifespan (Lebel et al., 2012; Mills et al., 2016) and are sensitive to modulations with disease, experience, lifestyle factors and learning (Bengtsson et al., 2005; Raja et al., 2019; Scholz et al., 2009; Voss et al., 2013).

To yield biologically interpretable variables of WM tissue, the measured DWI signal reflecting the motion of water molecules within the local tissue environment has to be described in a mathematically plausible way (Alexander et al., 2019; Jones et al., 2013; Novikov et al., 2018). Most commonly, DWI data are acquired with a single diffusion weighting constant (*b*-value) for at least six noncollinear gradient directions (Jones, 2004), and the measured signals along different axes are fitted to a diffusion ellipsoid or tensor (Basser et al., 1994; Basser and Pierpaoli, 1996). Thus, the tensor is a simplified representation of diffusion attenuation that makes no assumptions about the underlying biophysical tissue properties (Hutchinson et al., 2017; Novikov et al., 2018; Novikov et al., 2019). Notwithstanding that, popular tensor-derived metrics like fractional anisotropy (FA) and mean diffusivity (MD) have shown to be sensitive to tissue microstructure in health and disease, especially in WM regions of approximately parallel fiber bundles (Alexander et al., 2019; Chang et al., 2017). A common criticism of the diffusion tensor and the conventionally used low *b*-values (≤ 1,000 s/mm^2) are their limited ability in resolving complex fiber geometries such as crossing, fanning and kissing fibers, although it is estimated that roughly 90% of the brain's voxels exhibit such complex microstructure (Jones et al., 2013). Likewise, it has been emphasized that DWI with low *b*-values is rather insensitive to neurites (axons and dendrites) and neural tissue changes in gray and white matter (Fukutomi et al., 2019).

To overcome these issues, it has been suggested to acquire DWI data with multiple *b*-values (multi-shell) and a high number of gradient directions (angular sampling), along with the use of a theory-driven modeling framework to relate the *b*-value dependent diffusion signal to

more specific biophysical properties such as axon density and fiber dispersion (Alexander et al., 2019; Hutchinson et al., 2017; Novikov et al., 2019). One increasingly popular and clinically feasible biophysical model of diffusion is *Neurite Orientation Dispersion and Density Imaging* (NODDI; Zhang et al., 2012). The key parameters to emerge from this model are the neurite density index (NDI), a measure of axonal or neurite packing density, and the orientation dispersion index (ODI), a measure representing geometric complexity (angular variation) of neurite orientation and therefore reflecting tract disorganisation (Zhang et al., 2012). Importantly, these novel indices of brain microstructure were found to be associated with changes typically observed in neurodegeneration (reviewed by Lakhani et al., 2020; Sone, 2019) and were also validated against histological counterparts (Grussu et al., 2017; Jespersen et al., 2010; Mollink et al., 2017; Schilling et al., 2018; Sepehrband et al., 2015; Wang et al., 2019). Therefore, NODDI has potential to characterize the biological mechanisms underlying group differences, brain-behavior-correlations, or brain changes over time in unprecedented biological plausibility.

There is an increasing awareness of the importance of scientific quality standards such as accurate handling of effect sizes and scientific reporting (Loken and Gelman, 2017), appropriate sample sizes (Button et al., 2013; Szucs and Ioannidis, 2020), accounting for multiplicity of tests (Winkler et al., 2016) and using reliable measures (Poldrack et al., 2017; Zuo et al., 2019). Reliable measures are essential in neuroimaging research, as they are an important contributor to the sensitivity and specificity of the analysis in various statistical designs (Tofts, 2018b; Zuo et al., 2019). Yet surprisingly, our knowledge on NODDI reproducibility in the human brain's white matter is based on very limited data with sample sizes of $n \leq 10$ (Andica et al., 2020; Chung et al., 2016; Granberg et al., 2017; Tariq et al., 2013). Although the results of these studies are promising with reported within-subject coefficients of variation (CV_{WS}) ranging from approximately 1–7% (Andica et al., 2020; Chung et al., 2016; Granberg et al., 2017; Tariq et al., 2013) and retest-correlation values (Pearson's *r* or ICC) of ≥ 0.9 (Andica et al., 2020; Tariq et al., 2013), several important issues remain to be addressed. First, most previous studies focused on within-session (Andica et al., 2020; Chung et al., 2016; Granberg et al., 2017) rather than longitudinal reliability over longer time intervals (Tariq et al., 2013). Since repeated measurements in longitudinal neuroimaging studies are typically separated by weeks to several months (Lebel et al., 2012; Valkanova et al., 2014), reproducibility of NODDI maps with longer rescan intervals needs to be investigated. Second, a so far neglected area is the reproducibility of NODDI on the single-voxel level in stereotactic space, which is however the relevant quantity influencing popular voxel-by-voxel analysis methods (Cabeen et al., 2017; Snook et al., 2007; Vollmar et al., 2010). We consider this as the most important gap in knowledge, because many existing studies using NODDI carried out localized statistical testing based on a "VBM-style"

framework (e.g., Billiet et al., 2015; Broad et al., 2019; Churchill et al., 2019; Dowell et al., 2019; Kraguljac et al., 2019) or the tract-based spatial statistics (TBSS; Smith et al., 2006) approach (e.g., Kodiweera et al., 2016; Timmers et al., 2016; Zhang et al., 2018). A problem intrinsically tied to "VBM-style" approach is the choice of an adequate extent of spatial smoothing, since this particular postprocessing step impacts the results of later statistical tests (Jones et al., 2005; Smith et al., 2006).

Reproducibility of biophysical models of diffusion like NODDI is not only dependent on the choice of an adequate analysis approach, but also (and in the first place) on the quality of the underlying DWI data. DWI is an inherently signal-to-noise ratio (SNR) sensitive technique due to the loss of signal accompanying the application of strong diffusion gradients (Jones et al., 2013; Polders et al., 2011). In this respect, applying even stronger gradients – which is required to fit advanced biophysical models – leads to a further increase in signal attenuation, thus aggravating the SNR-problem (Chung et al., 2016; Hutchinson et al., 2017; Wang et al., 2019). In addition, factors like scanner hardware (e.g., field strength, gradient performance, receive coil sensitivity, scanner instabilities), geometric distortions (e.g., eddy current- and susceptibility-induced distortions), imaging protocol (e.g., number of diffusion-weighted and non-diffusion-weighted images, number of *b*-values, angular sampling, choice of the highest *b*-value, voxel size) and other measurement issues (e.g., subject placement, head motion, cardiac pulsation) may also contribute to imaging artifacts (Chen et al., 2015; Chung et al., 2016; Farrell et al., 2007; Hutchinson et al., 2017; Parvathaneni et al., 2018; Roalf et al., 2016; Vollmar et al., 2010; Wang et al., 2019). Determining how the combined effect of all these confounds affects NODDI (and DTI) metric reproducibility is a research problem yet to be addressed.

This paper examines the longitudinal reproducibility of NODDI metrics in white matter in a cohort of healthy adults at a magnetic field strength of 3T. We concentrate on a comparative evaluation of the region-of-interest (ROI) analysis approach – i.e. averaging voxels within atlas-derived WM tracts (cf. Froeling et al., 2016; Snook et al., 2007) – against the "VBM-style" approach with different extents of spatial smoothing (cf. Jones et al., 2005) and against TBSS (Smith et al., 2006). Furthermore, we address the impact of image quality on the reproducibility of NODDI metrics. We will therefore be able to evaluate NODDI's general reproducibility and to derive evidence-based recommendations regarding analysis approaches for future cross-sectional and longitudinal studies.

Materials and Methods

Participants and experimental design

In order to assess longitudinal reproducibility of NODDI microstructural maps, two MRI measurements separated by four weeks were acquired. This design was chosen because a between-scan interval of several weeks is frequently used in neuroimaging studies on plasticity (cf. Valkanova et al., 2014, for a review).

Twenty-nine cognitively healthy adults (24 ♂, 5 ♀; age: *M* = 23.07, *SD* = 3.98, range 19–35; BMI: *M* = 23.76, *SD* = 3.02, range 18.99–29.99) with no history of neurological, psychiatric or systemic diseases were included. Note that an age range of 18 to 35 y reflects a comparably homogenous phase of human ontogeny in terms of cognitive functions (Li et al., 2004) and structural brain development (Mills et al., 2016). The study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments. Approval was granted by the Ethics Committee of Otto von Guericke University Magdeburg. Written informed consent was obtained from all individual participants included in the study.

MR image acquisition

MRI data were acquired on a 3T MAGNETOM Prisma system (Siemens Healthcare, Erlangen, Germany) using a 64-channel head coil. We used the same protocol for each volunteer and each scanning session. Whenever possible, subjects were measured at approximately the same time of day during the study. The imaging protocol consisted of a series of MRI sequences, as outlined below. Subjects were asked to relax, keep their mind free of any thoughts, and to move as little as possible. A pillow was placed surrounding the sides and the back of the head to minimize head motion and within- as well as between-subject differences in positioning.

Anatomical images were acquired using a T1w three-dimensional magnetization-prepared rapid gradient echo sequence (Mugler & Brookeman, 1990) with 240 sagittal slices. The imaging parameters used were as follows: inversion time, TI = 1,100 ms; repetition time, TR = 2,600 ms; echo time, TE = 5.18 ms; readout pulse flip angle, α = 7°; parallel GRAPPA acceleration factor = 2; acquisition matrix = 320 × 320; field of view, FOV = 256 × 256 mm^2; nominal spatial resolution = 0.8 × 0.8 × 0.8 mm^3; scan duration = 7 min 25 s.

Whole-brain DW images were obtained with a monopolar single-shot spin echo EPI sequence: TE = 74 ms; TR = 4970 ms; flip angle α = 90°; parallel GRAPPA acceleration factor = 2, matrix: 130 × 130; FOV = 208 × 208 mm^2; nominal spatial resolution = 1.6 × 1.6 × 1.6 mm^3; multiband acceleration factor = 2; phase-encoding direction: anterior >> posterior; 228 isotropically distributed diffusion sensitization directions (38 at b = 1,000 s/mm^2, 76 at b

= 2,000 s/mm^2, and 114 at b = 3,000 s/mm^2) and 14 b = 0 s/mm^2 images (interleaved throughout the acquisition) were collected. The sampling scheme was designed according to Caruyer and co-workers (http://www.emmanuelcaruyer.com/q-space-sampling.php; Caruyer et al., 2013). To generate appropriate fieldmaps to correct for susceptibility-induced distortions, nine b = 0 s/mm^2 images with reversed phase encoding (posterior >> anterior) were also acquired. The total scan duration was 22 min 31 s.

Processing of MR images

In accordance with the majority of existing NODDI papers we opted for preprocessing tools provided by the FMRIB Software Library ([FSL] Smith et al., 2004; see Fig. 1 for a graphical overview of the pipeline).

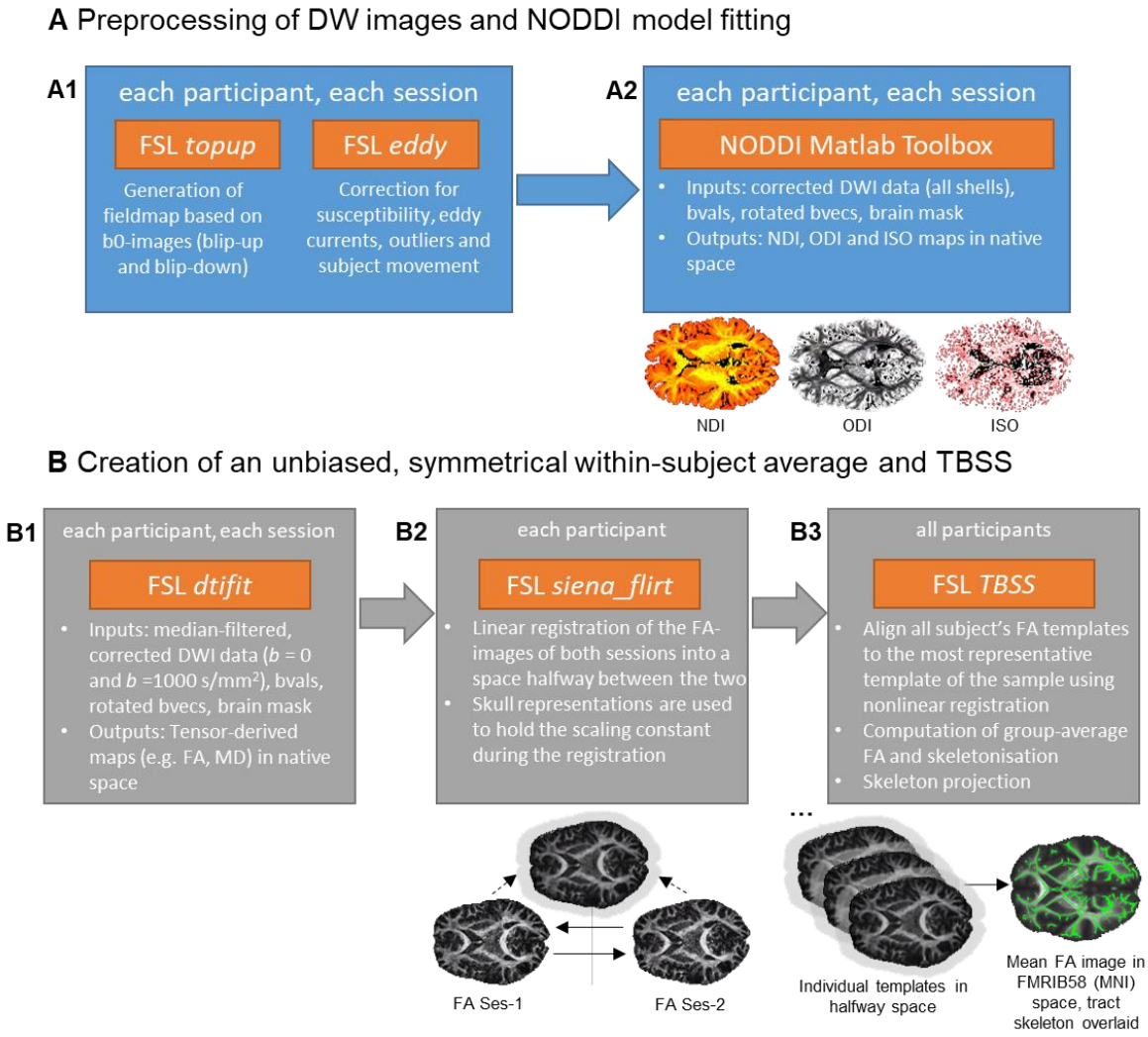

Figure 1: Graphical overview of the preprocessing pipeline (see text for details).

After visual quality assessment, preprocessing of DW images started with the creation of a fieldmap using *topup* (Andersson et al., 2003) for later correction of susceptibility-induced distortions (unwarping). The approach combines the b = 0 s/mm^2 images acquired with reversed phase-encoding as described in the previous section. Using the *eddy* tool (Andersson and Sotiropoulos, 2016), data sets were corrected for susceptibility (using the

fieldmap to emerge from *topup*), eddy current-induced distortions and head motion, and outlier slices were detected and corrected (Andersson et al., 2016). Realignment of images in the course of motion correction was accompanied by appropriate correction of gradient directions (Leemans and Jones, 2009).

NODDI parameter maps were estimated from corrected multishell DW images images (b = 0 s/mm^2, b = 1,000 s/mm^2, b = 2,000 s/mm^2, and b = 3,000 s/mm^2) using the NODDI Matlab Toolbox v1.0.1 (http://nitrc.org/projects/noddi_toolbox, default settings), implementing the model formulation of Zhang et al. (2012). In brief, NODDI models the diffusion signal in each voxel as contribution from three compartments: intraneurite signal, referring to the space bounded by the membrane of neurites, extraneurite signal, referring to the space around the neurites (glial cells, cell bodies), and CSF signal, referring to the space occupied by CSF. In the mathematical formulation of the model, intraneurite signal is represented by a set of zero-radius sticks following a Watson distribution, extraneurite signal is represented by a cylindrically symmetric tensor and CSF is modeled as isotropic Gaussian diffusion. The full normalized signal A can be written as

$$A = (1-v_{iso})(v_{ic}A_{ic} + (1-v_{ic})A_{ec}) + v_{iso}A_{iso}, \quad (1)$$

where A_{ic} and v_{ic} refer to the normalized signal and volume fraction of the intra-cellular compartment; A_{ec} is the normalized signal of the extracellular compartment; and A_{iso} and v_{iso} are the normalized signal and volume fraction of the CSF compartment, respectively (Zhang et al., 2012). Microstructural maps to emerge from the model are NDI (synonymous to v_{ic}), the fraction of tissue that comprises axons or dendrites, ODI, a measure of spatial configuration of the neurite structures and therefore tract disorganisation, and ISO, representing the freely diffusing water (i.e., CSF).

To analyse the region-based and voxel-wise reproducibility of these maps, it is paramount to have accurate registrations from native space to a standard template space. To this end, we used an established longitudinal TBSS-based pipeline (Smith et al., 2006; see Fig. 1) which has been evaluated in terms of reliability (Madhyastha et al., 2014) and has shown sensitivity to neuroplastic changes in longitudinal studies (e.g., Engvig et al., 2012; Lehmann et al., 2020). As a first step, a diffusion tensor (Basser et al., 1994; Basser and Pierpaoli, 1996) was fitted at each voxel of the preprocessed images (b = 0 s/mm^2 and b = 1,000 s/mm^2 shells; cf. Barrio-Arranz et al., 2015; Hutchinson et al., 2017; Novikov et al., 2018) using FSL's *dtifit*. Second, diffusion indices such as fractional anisotropy (FA) and mean diffusivity (MD) were computed from the eigenvalues of the diffusion tensor with the respective formulas (Pierpaoli and Basser, 1996). Third, an unbiased halfway space between the two FA images of each participant was determined using the *siena_flirt* tool (Smith et al., 2002),

as described by Engvig et al. (2012) and Madhyastha et al. (2014). Fourth, the original FA images were linearly registered to the computed halfway point and subsequently averaged to generate a subject-wise FA halfway template (Engvig et al., 2012; Madhyastha et al., 2014). Fifth, each subject-wise FA template in midpoint space was nonlinearly (Andersson et al., 2007) aligned to every other one in order to identify the most representative template of the sample (Rueckert et al., 1999; Smith et al., 2006). Sixth, after warping each subject's template to the target, images were registered to MNI152 space (FMRIB58 1mm template) using affine transformation (Jenkinson et al., 2002). Seventh, a group-average FA image was computed and thinned/binarized with an FA-value of > 0.25 (skeletonization). Eighth, the previously created warp fields were applied to all midpoint-space registered NODDI/DTI maps of both measurement points, and the aligned NODDI/DTI data was projected onto the skeleton. In sum, the aforementioned procedures yield linear transformations from native diffusion space to each subject's individual template, and from each individual template to MNI152 space (see Fig. 1, C and D). Inversion of these transformations and warps allowed us to register atlas ROIs to each subject's native diffusion space.

For the "VBM-style" approach, registered maps were smoothed with Gaussian kernels of different sizes (0mm [unsmoothed], 2mm, 4mm, 6mm, 8mm and 10mm FWHM) based on previously reported settings applied in the literature (Billiet et al., 2015; Broad et al., 2019; Churchill et al., 2019; Dowell et al., 2019; Jones et al., 2005; Kraguljac et al., 2019). To minimize potential problems with mixing of tissue types ("partial voluming", Smith et al., 2006), only voxels within the group-specific, conservative white matter mask (see "Regions-of-interest and masks") were subjected to smoothing.

Quality Assurance

Because preprocessing pipelines cannot fully compensate for the effects of potential confounding factors (see "Introduction"), it is important to assure that these confounders did not exert an unsystematic influence on the pre- and post-measurements. For this purpose, we calculated a recently proposed index of DW image quality – the temporal signal-to-noise ratio (tSNR; Roalf et al., 2016) – from the preprocessed and brain-extracted DWI data. Average tSNR is estimated by first calculating the mean and standard deviation of each voxel's intensity over time, and then averaging the resulting values across all brain voxels to yield a single metric of image quality (Roalf et al., 2016). Note that these calculations were performed separately for each subject, session, and *b*-shell.

As emphasized by Roalf et al. (2016), a note of caution should be sounded concerning the application of tSNR to diffusion-weighted images ($b \geq 0$ s/mm^2), because the latter have varying signal intensity and noise profiles. However, this potential drawback should be

bearable because significantly more diffusion-weighted than non-diffusion-weighted images were collected, thus providing a robust estimate of SNR (Roalf et al., 2016).

Besides the question whether image artifacts are comparable between pre- and post-test, it is also important to check the data for the presence of outliers. To this end, we calculated the absolute deviation around the median (MAD; see Leys et al., 2013, for details) as a measure of dispersion separately for each *b*-shell (data from both sessions merged) and defined a moderately conservative rejection criterion of 2.5 times the MAD below the median (Leys et al., 2013). In other words, individual DWI data were categorized as outliers if their tSNR fell outside the predefined rejection criterion of 2.5*MAD below the median.

Reliability metrics

Reproducibility (or *reliability)* generally refers to the "degree to which multiple assessments of a subject agree" (Bartko, 1991, p. 483). Here, we analyse agreement both in terms of measurement precision (Sullivan et al., 2015) and in terms of consistent ranking of individuals (Bartko, 1991).

Starting point of reliability analysis according to classical test theory is the decomposition of observed scores into between-subject variability ("true score") and within-subject variability (Bartko, 1991; Hopkins, 2000; Tofts, 2018b). *Within-subject variability* describes the inconsistency (or dispersion) of observations when repeatedly measuring a single individual, thus representing the amount of random error or noise contributing to the measure. To calculate within-subject variability, the two measurements (or replicates) of each subject were first transformed according to

$$\tilde{y}_i = 100 \cdot \log y_i \quad (2)$$

for scan (i=1) and re-scan (i=2) where log refers to the natural logarithm (Hopkins, 2000). This procedure was chosen because the standard deviation of observations often increases with mean value in brain measures (Tofts, 2018b), and quantities derived from log-transformed data vary less with mean value (i.e., residuals are more uniform). Next, signed differences of the log-transformed data were calculated:

$$\Delta = \tilde{y}_2 - \tilde{y}_1. \quad (3)$$

Subsequently, the within-subject standard deviation in absolute units was computed according to the formula

$$SD_\Delta = \frac{\sigma(\Delta)}{\sqrt{2}} \quad (4)$$

where σ refers to the standard deviation (Hopkins, 2000; Tofts, 2018b). Finally, SD_Δ was converted to a coefficient of variation (CV) using the formula

$$CV_{WS} = 100 \cdot \left(e^{SD_{\Delta}/100} - 1\right) \quad (5)$$

where *e* is the base of the (natural) exponential function (Hopkins, 2000). The CV_{WS} is equivalent to the standard deviation of replicate measures for a subject, expressed as a percent of the subject's mean value. For example, a CV_{WS} of 10% reflects that the variation about the mean value is typically 1/1.1 to 1.1 times the mean, or ≈ 0.91 to 1.1 (Hopkins, 2000). The interpretation of this measure is straightforward: the smaller the CV_{WS}, the better the reproducibility.

The other source of variability is arising from differences between subjects (*between-subject variability)* therefore representing an indicator of sample heterogeneity (Bartko, 1991; Tofts, 2018b; Zuo et al., 2019). Between-subject variability was computed by first averaging the two measurements within-subject, second calculating the mean ($\bar{x}$) and standard deviation (σ) of the resulting scores across subjects, and third calculating the CV_{BS} according to the formula

$$CV_{BS} = \frac{\sigma}{\bar{\bar{x}}} \cdot 100. \quad (6)$$

In line with most reliability studies, we also report the intraclass correlation coefficient (ICC; Bartko, 1991; Shrout and Fleiss, 1979), a ratio measure between the previously introduced sources of variability. Conceptually, the ICC reflects the fraction of observed test score variance that is attributed to between-subject variability (Tofts, 2018b). If within-subject variability is small compared to between-subject variability, ICC approaches 1. The ICC was computed using the two-way mixed model where agreement is defined in terms of consistency. According to the Shrout and Fleiss (1979) convention, this type of ICC is termed ICC(3,1), where the "3" refers to the two-way mixed model (i.e., participants are treated as random effect, sessions as fixed effect), while the "1" refers to the reliability of single repeated measurements (instead of the mean of several measurements). Assuming that the data is arranged in a convenient matrix with subjects in rows and repeated measurements in columns, ICC(3,1) is calculated according to

$$ICC(3,1) = \frac{MSS_R - MSS_E}{MSS_R + (k-1)MSS_E} \quad (7)$$

where *k* refers to the number of measurements/scans, MSS_R refers to the mean sum of squares (i.e., between-subject) and MSS_E refers to mean sum of squares of errors (within-subject) (Shrout and Fleiss, 1979).

As a sanity check that all formulas were rightly implemented, we calculated the ratio of CV_{WS}/CV_{BS} for each mask and for all analysis approaches and correlated the resulting scores with the ICC. Expectedly, the CV_{WS}/CV_{BS} ratio correlated very well with the ICC (R^2 = 0.98, p < .001). The small difference to a perfect correlation can be explained by the fact that in the

present paper within- and between-subject variability were not calculated in exactly the same way as in the ICC formula.

Regions-of-interest and masks

Based on the most commonly used analysis strategies in the literature, we analyzed the reliability of NODDI maps on the ROI level as well as on the single-voxel level ("VBM-style" approach and TBSS). To comparatively evaluate the approaches, the probabilistic JHU white-matter tractography atlas (Hua et al., 2008; Wakana et al., 2007; thresholded at 25% probability) was used, which contains masks of twenty major white matter fiber tracts (Fig. 2). We focused on reproducibility of the maps in WM tracts, because the default fixed values for the compartment diffusivities in the original NODDI model are suboptimal for gray matter (Fukutomi et al., 2019; Guerrero et al., 2019). In case of the ROI-based analysis approach, atlas regions were transformed to each subject's native diffusion space by inverting the previously created warp fields (Fig. 1, D).

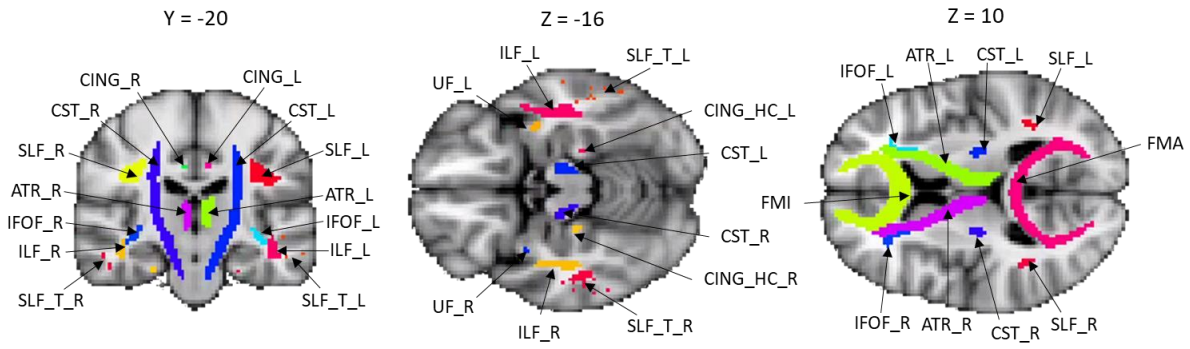

Figure 2: WM tract ROIs as derived from the probabilistic JHU tractography atlas (Hua et al., 2008; Wakana et al., 2007) in MNI152 space. Note that SLF_T_L and SLF_T_R were extracted from the unthresholded JHU atlas due to their small spatial extent in the 25% probability atlas.

Abbreviations: ATR_L/ATR_R – anterior thalamic radiation left/right, CING_HC_L/CING_HC_R – hippocampal part of the cingulum bundle left/right, CING_L/CING_R – cingulum bundle left/right, CST_L/CST_R – corticospinal tract left/right, FMA – forceps major, FMI – forceps minor, IFOF_L/IFOF_R – inferior fronto-occipital fasciculus left/right, ILF_L/ILF_R – inferior longitudinal fasciculus left/right, SLF_L/SLF_R – superior longitudinal fasciculus left/right, SLF_T_L/SLF_T_R – superior longitudinal fasciculus, temporal part left/right, UF_L/UF_R – uncinate fasciculus left/right.

Additionally, we also evaluated and compared the reproducibility of the "VBM-style" and TBSS approaches in a voxel-by-voxel fashion. In this respect, we aimed to restrict the analyses mainly to WM voxels, whilst not excluding all voxels in adjacency to the cortex. To this end, we created a group-specific white matter mask in standard space with the following steps. First, each subject's T1w-image was processed using the *fsl_anat* tool, including inhomogeneity correction, segmentation (Zhang et al., 2001) and brain extraction (Smith, 2002). Subsequently, WM partial volume maps were extracted from each subject's segmentation summary image, in which each voxel is assigned to the tissue class with the greatest partial volume fraction. Afterwards, intra-subject inter-modal registration (diffusion-to-T1w) was performed using FSL's *epi_reg*, which makes use of the white-matter

boundaries from the segmented T1w image and the grey-white intensity contrast in a b = 0 s/mm^2 image from the corrected diffusion data (Greve and Fischl, 2009; Jenkinson et al., 2002). Next, the WM partial volume maps of each subject and session were registered to MNI152 space by concatenating the transformations from structural to native diffusion space, from native diffusion space to midpoint, and from midpoint to standard space (see "Processing of MR images"). Finally, normalized partial volume maps were summed across subjects and sessions and subsequently binarized at 2/3 of the total number of images. Therefore, the resulting group-specific white matter mask contains only voxels in which WM has the greatest partial volume in at least 2/3 of the sample.

Reproducibility of image analysis approaches

To investigate the reproducibility of the ROI-based analysis approach, we averaged voxel values of NODDI (DTI) maps within a respective region of the JHU atlas for each participant and session. Note that within-ROI average voxel values were derived from unsmoothed maps (Froeling et al., 2016). ICC(3,1), CV_{WS} and CV_{BS} were calculated according to the abovementioned formulas using *R* (R Development Core Team, 2013).

With respect to reproducibility analysis in a voxel-by-voxel fashion in standard space ("VBM-style", TBSS), we used Matlab (Mathworks, Sherborn, MA) and bash scripts (based on *fslmaths* functions) to calculate the ICC(3,1), CV_{WS} and CV_{BS}, respectively (Fig. 3). Computation of reproducibility metrics was restricted to the group-specific white matter mask. Likewise, in case of the TBSS approach, an intersection mask of the tract skeleton and the white matter mask was used.

To globally summarize the results across voxels constituting an atlas region, we focus on the median (50th percentile) of the reliability metrics. For example, if the median of the ICC in a given atlas region has a value of 0.8, this means that 50 % of the voxels in this region have an ICC of 0.8 or higher. With respect to the TBSS approach, the summary statistic is exclusively based on voxels located a) on the group skeleton and b) inside a given atlas region, i.e. voxels outside the skeleton are ignored.

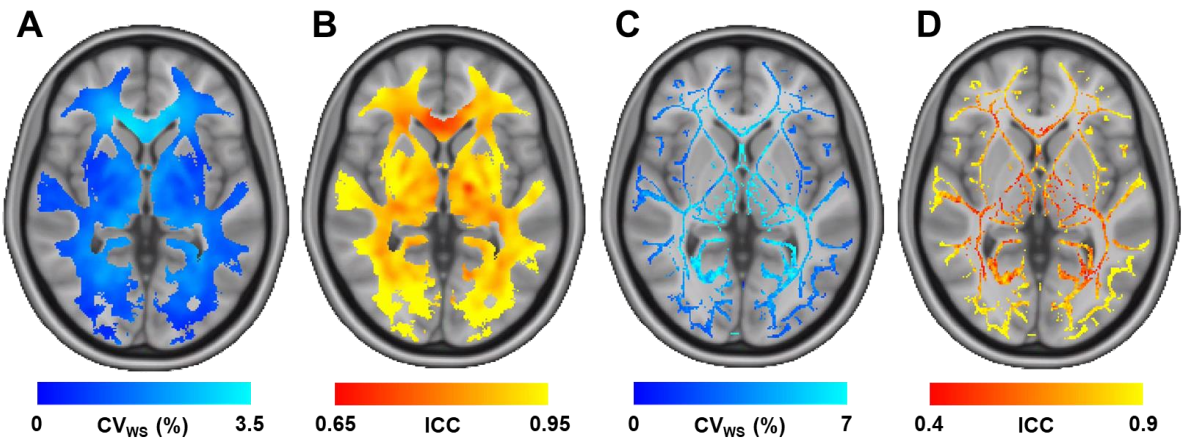

Figure 3: Exemplary depiction of voxel-based reproducibility maps in a horizontal section. A and B show CV_{WS} and ICC of NDI with 6mm smoothing within the group-specific white matter mask ("VBM-style" approach). C and D show the same reproducibility indices calculated based on the skeletonized NDI map (TBSS approach). To ensure fair comparison, the TBSS skeleton was masked with the group-specific white matter mask before reproducibility metrics were extracted.

Descriptive and inferential statistical analysis

Descriptive data underlying the comparison of the ROI-based approach against the voxel-by-voxel approaches (different levels of smoothing and TBSS) were visualized using boxplots. The latter were created using the packages *beeswarm* (Eklund, 2016) and *ggplot2* (Wickham, 2016) running in an *R* v3.5.1 environment (R Development Core Team, 2013). With respect to inferential statistical analysis, we assumed that the analysis approaches reflect related groups. Since the normality assumption was not tenable in all cases, we used Friedman tests with reliability metrics (point estimates in case of ROI approach, median values in case of voxel-based and TBSS-approach) of the 20 tracts of the JHU atlas as dependent variables. Post-hoc comparisons between analysis approaches were conducted by means of Wilcoxon signed-rank tests. We report uncorrected results of all follow-up tests, but additionally indicate if comparisons would survive Bonferroni correction. Effect sizes of follow-up tests are reported as matched-pairs rank biserial correlation coefficient (King et al., 2011). Rules of thumb for interpreting correlation-based effect sizes are $|r| < 0.30$ "small", $0.30 \leq |r| < 0.50$ "medium", and $|r| \geq 0.50$ "large" effects, respectively (Cohen, 1988).

Not least, we also compared the "VBM-style" approach against TBSS in terms of reproducibility on the single-voxel level. To this end, we extracted the intensity of each voxel of the reproducibility maps (ICC, CV_{WS}, CV_{BS}) located within the group-specific white matter mask (or skeleton mask) using *fslmeants*. Distributions of reproducibility values within the masks were visualized as a function of analysis approach using boxplots. Because the assumption of normality was violated across imaging modalities and reproducibility metrics (positively skewed distributions) and there was an unequal number of cases (the TBSS skeleton contains fewer voxels than the white matter mask), Kruskal-Wallis tests were used for statistical inference. Post-hoc comparisons between analysis approaches were conducted

by means of Mann–Whitney *U*-tests, and the respective effect sizes are reported as Cliff's delta (Cliff, 1996). The magnitude of the effect can be classified as follows: $0.147 \leq |d| < 0.33$ "small", $0.33 \leq |d| < 0.474$ "medium", $|d| \geq 0.50$ "large" effect (Torchiano, 2016).

Inferential statistical analyses as described above were calculated using *R*'s standard library and the packages *rcompanion* (Mangiafico, 2020) and *effsize* (Torchiano, 2016).

Impact of data quality on NODDI reproducibility

An increasing number of studies have suggested that data quality exerts a significant impact on NODDI reproducibility (Chung et al., 2016; Hutchinson et al., 2017; Parvathaneni et al., 2018; Wang et al., 2019). In the present study, we make use of the observation that SNR in DWI images typically varies across distinct white matter regions (Chen et al., 2015; Choi et al., 2011; Chung et al., 2016; Farrell et al., 2007; Marenco et al., 2006; Polders et al., 2011) to ask whether regional differences in tSNR contribute to regional differences in NODDI/DTI reproducibility. This phenomenon has been attributed to factors like regionally varying T2 relaxation, distance of the region from the receive coil elements as well as tissue type (gray matter versus white matter) and complexity of local fiber organization (Chen et al., 2015; Choi et al., 2011; Chung et al., 2016; Farrell et al., 2007; Marenco et al., 2006).

To address this question, we first standardised (*z*-transformed) the tSNR maps of each subject and session. This step yields maps in which positive (negative) voxel intensities indicate that the local tSNR is higher (lower) compared to an individual's white matter grand mean. Next, the *z*-transformed tSNR maps in native diffusion space were registered to MNI152 space according to the previously described procedure (see "Processing of MR Images"). Afterwards, all tSNR maps in standard space were averaged across subjects and sessions. The underlying idea of this procedure was to assign positive (negative) values to voxels where tSNR is inherently – i.e. in the average subject – high (low).

Next, we tested whether local tSNR affects image reproducibility by means of regression models calculated using *R*'s standard library (v3.5.1) and the package *yhat* (Nimon and Oswald, 2013). To this end, within the 20 regions of the JHU white matter tractography atlas, we extracted the median intensities a) of the tSNR summary maps of all *b*-shells and b) of the unsmoothed ICC maps (NDI and ODI), respectively. Based on these data, multiple linear regression models with $tSNR_{b0}$, $tSNR_{b1000}$, $tSNR_{b2000}$ and $tSNR_{b3000}$ as predictors of ICC (separate models for NDI and ODI, respectively) were fitted. Although regression models were calculated based on only 20 cases, results from a recent simulation study suggest that two cases per predictor variable are sufficient for an adequate estimation of regression coefficients and associated standard errors (Austin and Steyerberg, 2015). Also note that, since we focused on within-ROI median values of tSNR and ICC, perfect voxel-to-voxel

correspondence between maps is not mandatory, such that the potential influence of registration imperfection on the tSNR-ICC regression results should be negligible.

With respect to multiple regression, it must be assumed that the extent of shared variance between the predictors is high (high collinearity; cf. Nimon and Oswald, 2013), which complicates inferences about the relative importance of tSNR of each *b*-shell on NODDI reproducibility. Therefore, instead of reporting standardized regression weights, we used a method that decomposes multiple R^2 into contributions from the individual regressors ("relative weights"/RLW; c.f. Fabbris, 1980; Nimon and Oswald, 2013). In a nutshell, RLW creates a set of regressors that is as highly correlated as possible with the original set of regressors but orthogonal (uncorrelated) to each other. Therefore, the RLW reflects the contribution of a predictor to variance of the dependent variable, considering the predictor's unique contribution as well as its common contribution with the other predictors in the model (see Fabbris, 1980 and Nimon and Oswald, 2013, for statistical details).

In the methodological literature on DW image processing, it has been suggested that spatial smoothing does not only compensate for registration misalignments, but also that it mitigates the effects of low SNR (Jones et al., 2005; van Hecke et al., 2010). If this applies, the benefit of smoothing on reproducibility should be higher (lower) in regions with inherently low (high) SNR. To directly test this assumption, we calculated regression models with the same regressors as described above, but with smoothing-induced percent change in ICC as dependent variable. ICC percent change maps were calculated between unsmoothed and 6mm isotropically smoothed maps (Billiet et al., 2015; Broad et al., 2019; Churchill et al., 2019).

Reproducibility of clinically feasible NODDI

Following recent recommendations in the literature, the analyses of NODDI's reproducibility in this paper are based on a multi-shell resolution protocol with three equally spaced *b*-values (Sotiropoulos et al., 2013; Wang et al., 2019), high angular resolution (228 diffusion sensitization directions), and a comparably high outer shell *b*-value of 3,000 s/mm^2 (Hutchinson et al., 2017; Parvathaneni et al., 2018). Due to time constraints in settings like the clinic, however, it might be necessary to use a scan protocol with a shorter acquisition time. We therefore compared NODDI reproducibility of the "full" scan protocol against a "standard" protocol with 30 and 60 directions at b = 1,000 s/mm^2 and b = 2,000 s/mm^2, respectively (Zhang et al. 2012).

For the original set of 38 (at b = 1,000 s/mm^2) and 76 (at b = 2,000 s/mm^2) gradient directions, we used Camino's "subsetpoints" tool (Cook et al., 2006) to search for an ordering that minimizes the electrostatic energy for the desired subsets with 30 (at b = 1,000 s/mm^2)

and 60 directions (at b = 2,000 s/mm^2), respectively. This ensures an as evenly spread of gradient directions over the sphere as possible. NODDI parameter maps were then re-estimated from the subsampled data. Reliability metrics on the ROI- and single-voxel-levels were calculated as described before.

Finally, the reproducibilities of the “full” vs. “standard” protocols were compared using a robust mixed ANOVA based on 20% trimmed means (cf. Wilcox, 2017) as implemented in the *WRS* package (Wilcox and Schönbrodt, 2019) running in *R*. “Protocol” (“rich” vs. “standard”) was defined as between-subjects factor and “analysis approach” as within-subjects factor (levels: ROI-based approach, “VBM-style” approach without smoothing, “VBM-style” approach with 6mm isotropic smoothing, TBSS). As described before, ICC in the 20 ROIs of the JHU tractography atlas was used as dependent variable. Between-protocol comparisons at each level of the within-subjects factor were conducted using post-hoc Wilcoxon signed-rank tests.

Results

For the sake of clarity and to avoid overloading the exposition, we focus on the presentation of region-based NODDI reproducibility in the following. All analyses as conducted below were also performed on tensor-derived metrics (FA, MD), and the respective results were included in the Supplementary Material (Supplementary Figures 1–2, 6–7; Supplementary Tables 10–15, 25–31). Likewise, from time to time we refer to comparisons of the "VBM-style" approach against TBSS on the single-voxel level, whose results can be found in the Supplementary Material, too (Supplementary Figures 3–5; Supplementary Tables 16–24).

Quality assurance

We started our analyses by comparing sessions for differences in tSNR and checking for the presence of outliers. There were no significant pre-post differences in tSNR for neither *b*-shell as assessed with Wilcoxon signed-rank tests (all *p*'s ≥ .31; Table 1). Likewise, subjects' individual tSNR values exceeded the predefined rejection criterion of 2.5*MAD below the median across all *b*-shells (Leys et al., 2013), thus indicating the absence of extreme outliers in the sample.

Table 1: Between-session comparison and reproducibility (CV_{WS}, CV_{BS}) of tSNR in all b-shells. Descriptive statistics refer to median and interquartile range (25^{th} and 75^{th} percentile).

	tSNR ses-1	tSNR ses-2	Wilcoxon test (*p*)	CV_{WS} (%)	CV_{BS} (%)
b0	16.36 (16.02,17.37)	16.91 (16.18,17.59)	.31	3.30	4.70
b1000	6.25 (6.15,6.42)	6.33 (6.24,6.41)	.65	1.66	2.41
b2000	3.68 (3.63,3.78)	3.72 (3.68,3.77)	.39	1.87	2.17
b3000	2.79 (2.72,2.83)	2.80 (2.76,2.85)	.31	1.55	2.33

Longitudinal Reproducibility of NDI

Fig. 4 visualizes scan–rescan CV_{WS} and ICC's of NDI as a function of the analysis approach. Friedman tests indicate that analysis approach has a significant impact on CV_{WS}, $\chi^2(7) = 132.2$, $p < .001$, CV_{BS}, $\chi^2(7) = 131.88$, $p < .001$, and ICC, $\chi^2(7) = 108.98$, $p < .001$, respectively (for follow-up Wilcoxon signed-rank tests, see Supplementary Tables 1–3).

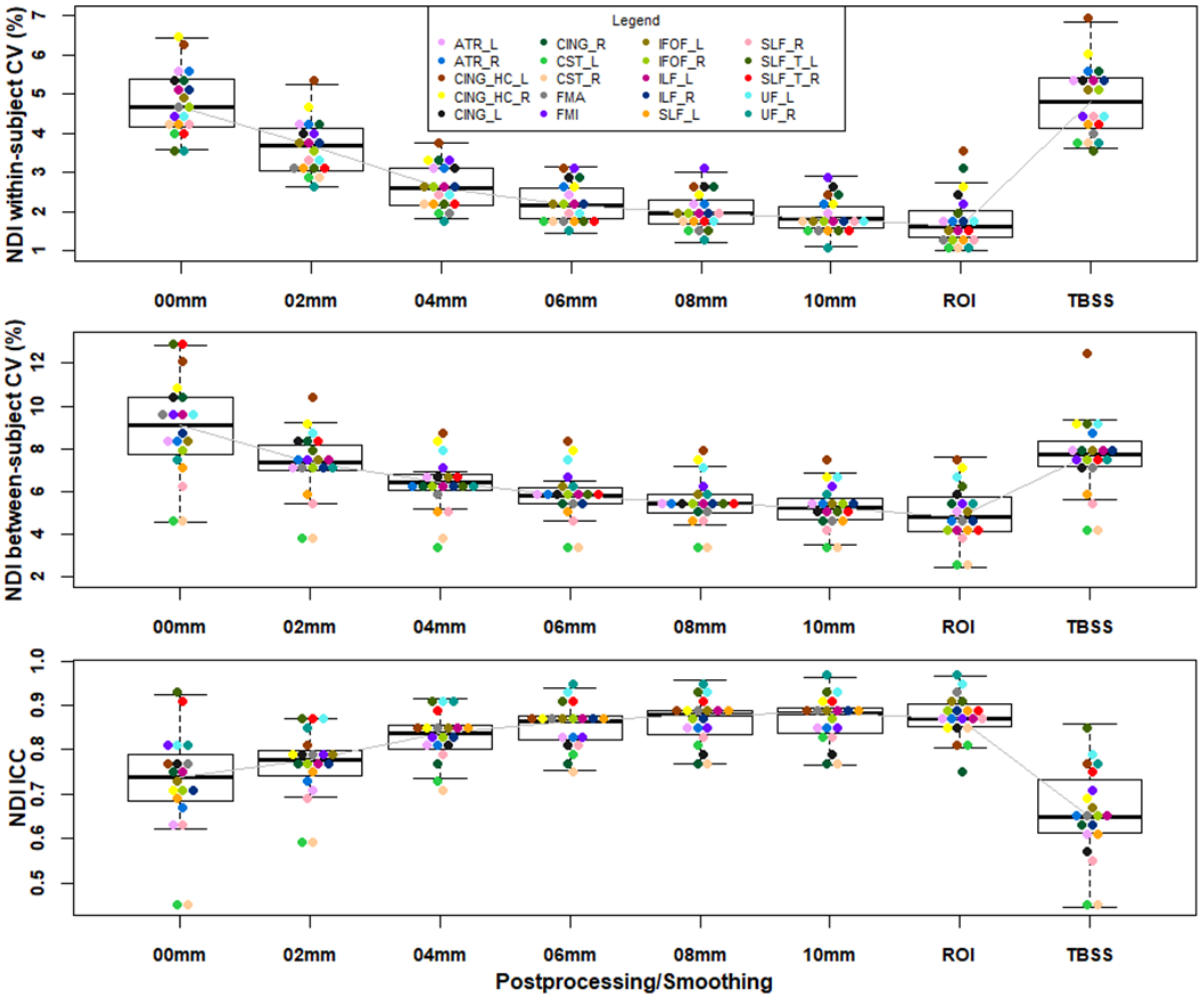

Figure 4: Reproducibility boxplots (CV_{WS}, CV_{BS} and ICC) of NDI in 20 ROIs of the JHU tractography atlas as a function of image analysis approach. Colored dots represent reproducibility of the median voxel within an ROI (in case of voxel-by-voxel approaches) or point estimates of reliability (in case of the ROI-based approach), respectively. Legend shown on top applies to all plots, tract abbreviations as in Figure 2.

The median values across 20 tracts show a global trend towards decreasing CV_{WS} and CV_{BS} and increasing ICC with larger extents of spatial smoothing. This pattern of results can be explained by a stronger decline of CV_{WS} compared to CV_{BS} with increasing smoothing extent. Smoothing with Gaussian kernels of ≥ 4mm FWHM yielded excellent scan–rescan reproducibility (CV_{WS} boxplot *Mdn* ≤ 2.6%, ICC boxplot *Mdn* ≥ 0.84), as did ROI-averaging (CV_{WS} boxplot *Mdn* = 1.6%, ICC boxplot *Mdn* = 0.87). Remarkably, TBSS performs consistently worse than ≥ 2mm FWHM smoothing and the ROI approach, respectively, regarding both CV_{WS} and ICC (Supplementary Tables 1–3).

The same pattern of results emerges when comparing TBSS against the "VBM-style" approach with different smoothing kernels on the level of single voxels (Supplementary Figure 3, Supplementary Tables 16–18). Of note, these analyses revealed that TBSS and unsmoothed maps did not meaningfully differ with regard to CV_{WS} (Cliff's delta = 0.032, negligible effect), while there was a trend for higher CV_{BS} in unsmoothed maps (Cliff's delta = 0.179, small effect).

Longitudinal Reproducibility of ODI

Fig. 5 shows the scan–rescan CV_{WS} and ICC of ODI. As with NDI, CV_{WS}, $\chi^2(7) = 124.82$, $p < .001$, CV_{BS}, $\chi^2(7) = 128.92$, $p < .001$, and ICC, $\chi^2(7) = 78.6$, $p < .001$ were significantly affected by analysis approach (for follow-up Wilcoxon signed-rank tests, see Supplementary Tables 4–6).

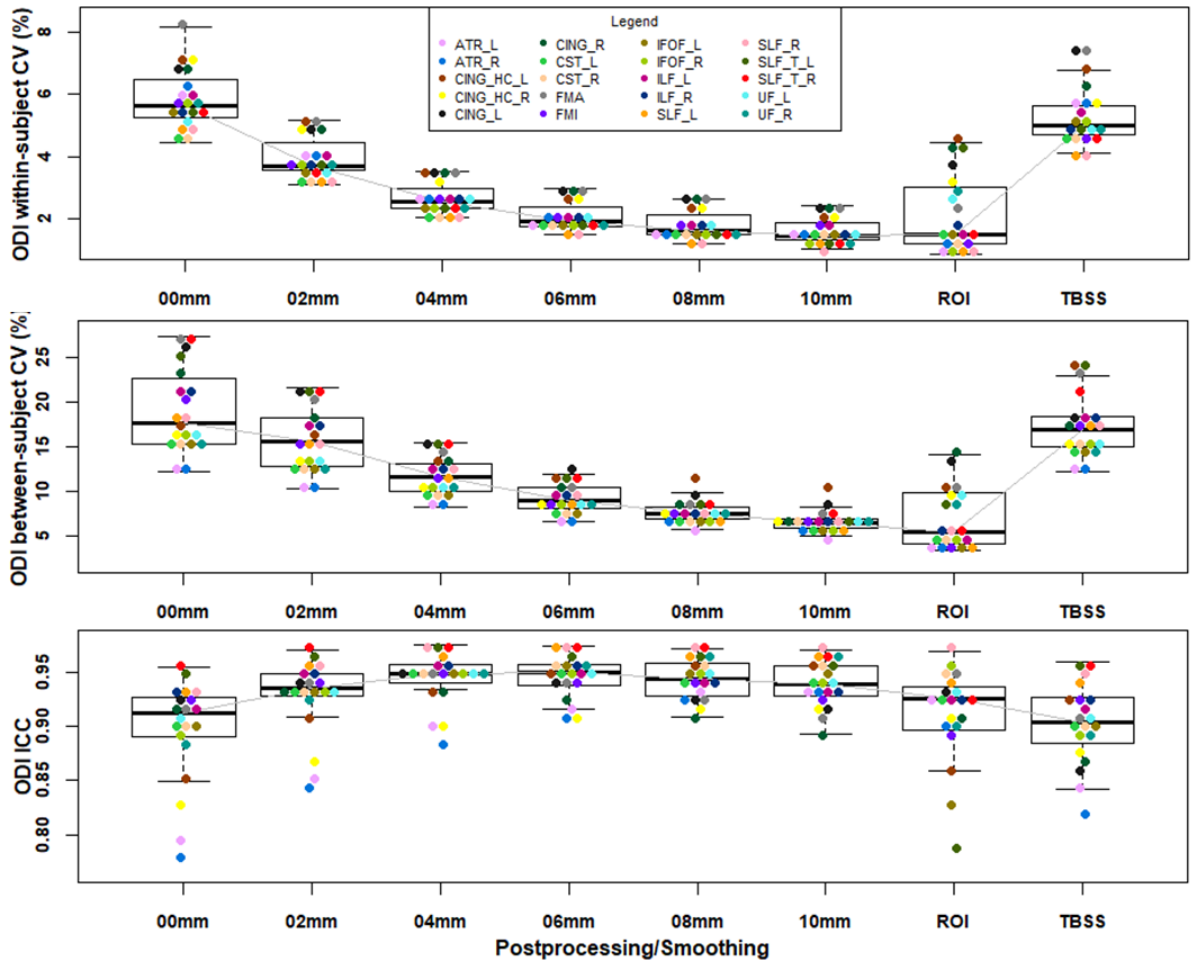

Figure 5: Reproducibility boxplots (CV_{WS}, CV_{BS} and ICC) of ODI in 20 ROIs of the JHU tractography atlas as a function of image analysis approach. Colored dots represent reproducibility of the median voxel within an ROI (in case of voxel-by-voxel approaches) or point estimates of reliability (in case of the ROI-based approach), respectively. Legend shown on top applies to all plots, tract abbreviations as in Figure 2.

In general, ODI shows consistently high reproducibility across all analysis approaches. Specifically, like in NDI, there is a trend for decreasing CV_{WS} with increased extent of smoothing. While we found no significant difference between the ROI approach on the one hand and smoothing in a range from 4mm to 10mm on the other hand, smoothing with Gaussian kernels of ≥ 2mm FWHM yielded significantly lower CV_{WS} than TBSS (Supplementary Table 4).

Compared to NDI, a different pattern of results emerged when focusing on the ICC as a function of analysis approach. Here, the results indicate that ICC's across smoothing kernels in a range from 4mm to 8mm FWHM (ICC boxplot *Mdn* ≥ 0.948) have higher ICC compared

to the ROI (ICC boxplot *Mdn* = 0.925) approach (Supplementary Table 6), and a large, but not significant effect was registered for the 4mm smoothing vs ROI comparison (r = -0.71). These results can be explained by the presence of tipping points after which CV_{BS} falls disproportionately compared to CV_{WS}. TBSS (ICC boxplot *Mdn* = 0.90) showed consistently lower ICC's compared to Gaussian smoothing with ≥ 2mm FWHM.

Again, the results from regional analysis are paralleled by voxel-by-voxel reproducibilities (Supplementary Figure 4, Supplementary Tables 19–21). Unsmoothed maps and TBSS reproducibility metrics were by and large comparable, only CV_{WS} tended to be lower in TBSS (Cliff's delta = 0.177, small effect). Equivalent results were registered for isotropic smoothing in a range from 4mm to 10mm in terms of the ICC (all Cliff's deltas ≤ |0.084|, negligible effects).

Longitudinal Reproducibility of ISO

Fig. 6 shows the scan–rescan reproducibilities of ISO. Again, CV_{WS}, $\chi^2(7) = 133.48$, $p < .001$, CV_{BS}, $\chi^2(7) = 135.15$, $p < .001$, and ICC, $\chi^2(7) = 80.85$, $p < .001$ were significantly affected by analysis approach (for follow-up Wilcoxon signed-rank tests, see Supplementary Tables 7–9).

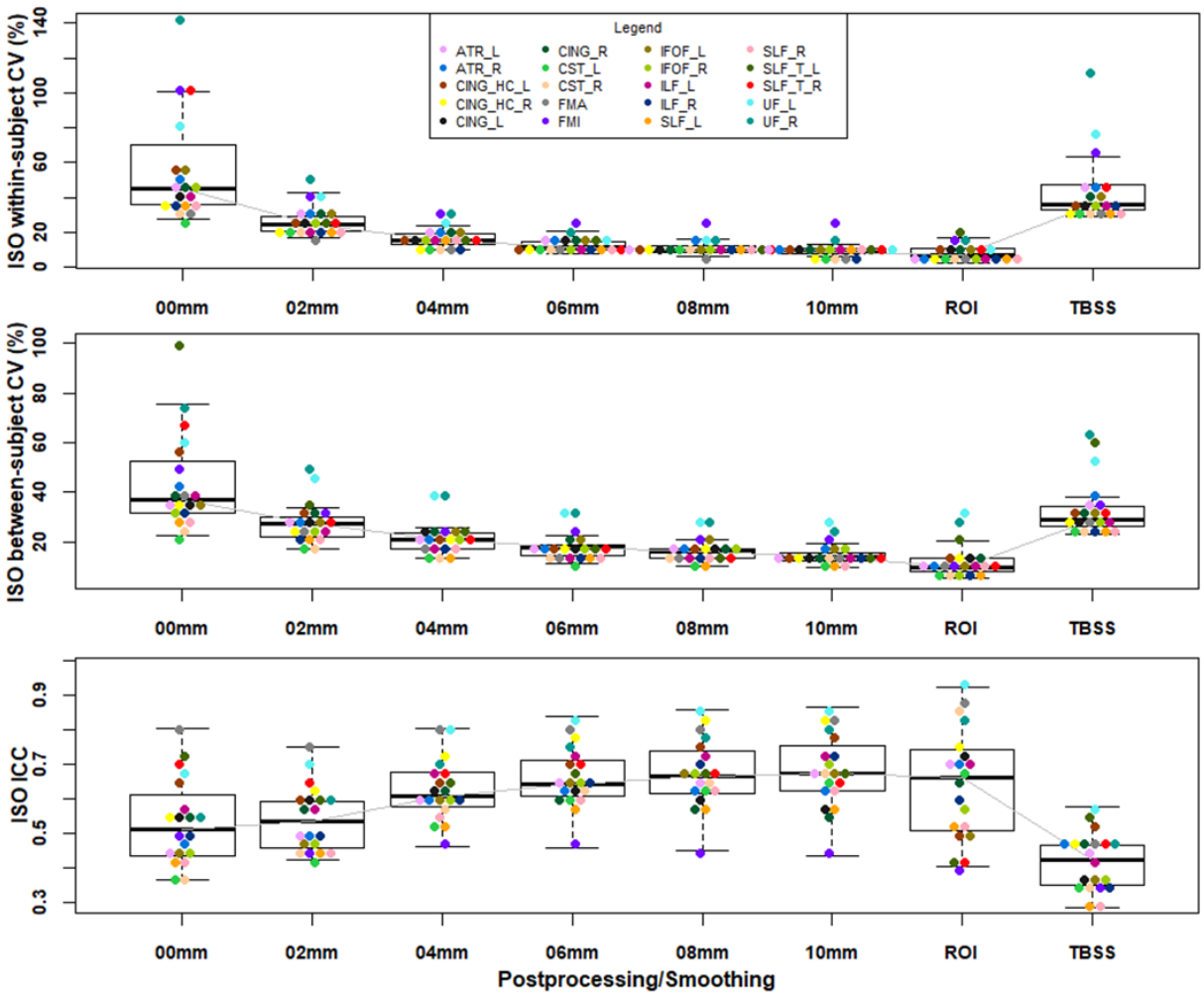

Figure 6: Reproducibility boxplots (CV_{WS}, CV_{BS} and ICC) of ISO in 20 ROIs of the JHU tractography atlas as a function of image analysis approach. Colored dots represent reproducibility of the median voxel within an ROI (in case of voxel-by-voxel approaches) or point estimates of reliability (in case of the ROI-based approach), respectively. Legend shown on top applies to all plots, tract abbreviations as in Figure 2.

Expectedly (cf. Andica et al., 2020; Chung et al., 2016), ISO showed consistently the poorest scan–rescan reproducibility of all NODDI maps. This is corroborated by the analysis of ISO reproducibility in a voxel-by-voxel fashion (Supplementary Figure 5, Supplementary Tables 22–24). The overall results pattern resembled the one observed in the NDI maps revealing a tendency for decreased CV_{WS} and increased ICC with larger extents of smoothing. Note, however, that reproducibility in some regions benefits from ROI averaging, whilst in others it does not. Again, this behavior can be explained by different effects of smoothing/averaging on CV_{WS} and CV_{BS}, respectively (Fig. 6).

Impact of data quality on NODDI reproducibility

Results emerging from relative weights analyses (Table 2) show that reproducibility of NDI and ODI can be well accounted for by data quality of the underlying DW images. Note that all predictors were positively related to the respective dependent variables (NDI_ICC, ODI_ICC); the higher the intrinsic tSNR of a tract, the higher the regional ICC and therefore

reproducibility. Likewise, intrinsic regional tSNR statistically accounted for smoothing-induced change of ICC (NDI_Δ_ICC, ODI_Δ_ICC). Here, the relationship was inverse: the lower (higher) regional intrinsic tSNR, the higher (lower) the smoothing-induced increase in ICC.

Note that similar results were obtained for FA and MD predicted by tSNR of the b = 0 s/mm^2 and b = 1,000 s/mm^2 shells (Supplementary Table 31).

Table 2: Relative weights analysis of regional tSNR (b = 0 s/mm², b = 1,000 s/mm², b = 2,000 s/mm² and b = 3,000 s/mm² shells) as predictor of regional NDI and ODI in 20 ROIs of the JHU tractography atlas. In statistical models denoted with "Δ", percent change of ICC between unsmoothed maps and 6mm isotropic smoothing was used as dependent variable. Ninety-five percent bias-corrected and accelerated bootstrap confidence intervals (95% BCa CI) are based on 10,000 bootstrap samples. Note that relative weights (RLW) of all predictors sum up to the multiple R^2 of the respective multiple linear regression model.

	Multiple R^2	RLW_tSNR_b0	RLW_tSNR_b1000	RLW_tSNR_b2000	RLW_tSNR_b3000
NDI_ICC	R^2 = 0.76, p < .001	0.164 (0.063,0.293)	0.269 (0.211,0.374)	0.197 (0.168,0.258)	0.129 (0.113,0.219)
ODI_ICC	R^2 = 0.74, p < .001	0.402 (0.214,0.556)	0.128 (0.086,0.199)	0.126 (0.077,0.209)	0.086 (0.052,0.189)
NDI_Δ_ICC	R^2 = 0.55, p = .014	0.068(0.023,0.186)	0.207(0.139,0.356)	0.155(0.121,0.261)	0.115(0.089,0.327)
ODI_Δ_ICC	R^2 = 0.71, p < .001	0.366 (0.131,0.611)	0.126 (0.080,0.194)	0.126 (0.063,0.212)	0.097 (0.036,0.235)

Comparison of "full" vs. "standard" DWI protocol

TSNR was expectedly lower in the “standard” compared to the “full” protocol for both the *b* = 1,000 s/mm^2 (pre-test: -2.13%, $p < .001$; post-test: -3.01%, $p < .001$) and the *b* = 2,000

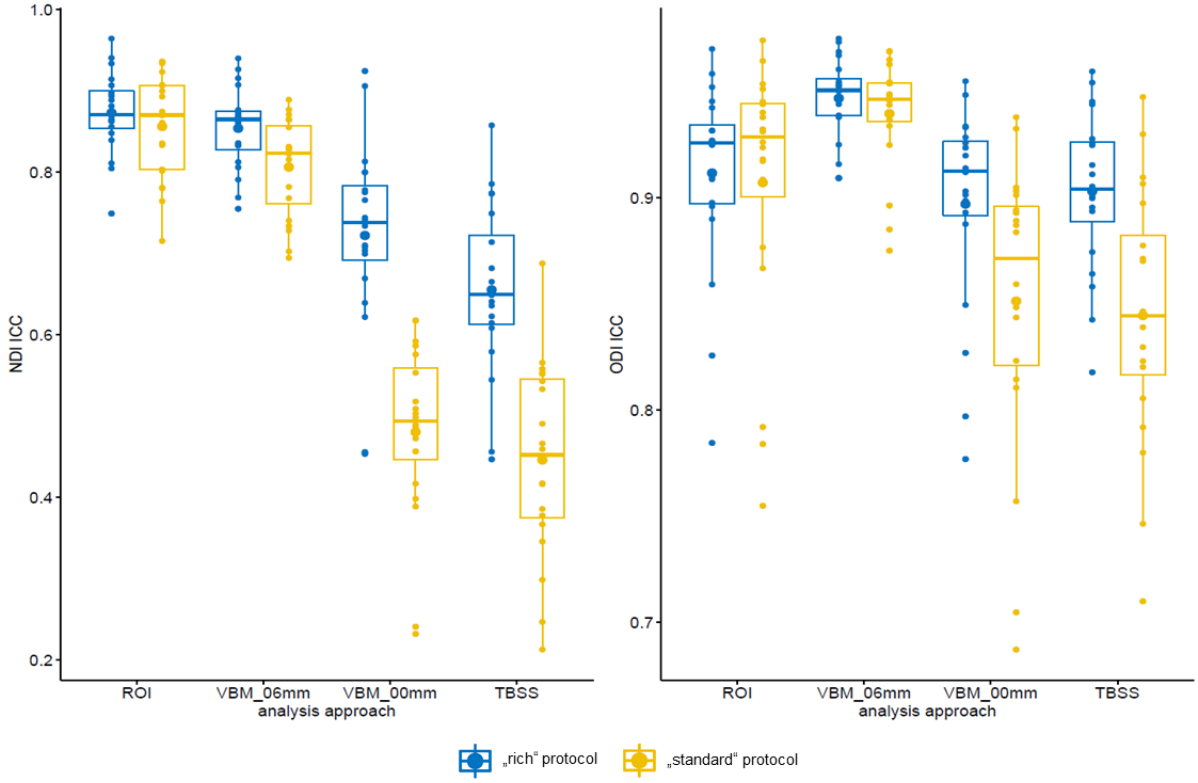

Figure 7: Graph visualizing the interaction between scan protocol and analysis approaches (left side: NDI, right side: ODI).

s/mm^2 shell (pre-test: -3.05%, $p < .001$; post-test: -2.97%, $p < .001$). Mixed ANOVAs based on 20% trimmed means (Table 3 and Figure 7) yielded a significant main effect for the between-subjects factor “protocol” in both modalities (NDI and ODI), revealing that the “rich” three-shell protocol generally had higher reproducibility compared to the “standard” two-shell protocol. The main effect of the within-subjects factor in both modalities is well in line with the results reported in the previous sections, reinforcing that the choice of analysis approach significantly influences reproducibility. Highly significant interaction effects indicate that the reproducibility differences between protocols vary dependent on the level of the within-subjects factor. Post-hoc Wilcoxon-tests to break down this interaction reveal that between-protocol differences were most pronounced in analysis approaches without some kind of smoothing or voxel value averaging (i.e., “VBM-style” approach and TBSS; Table 3 and Figure 7). ICC differences between the protocols are weaker when considering the “VBM-style” approach with 6mm FWHM smoothing, but the respective effect sizes are still “large” ($p \leq .02$, $|r| \geq 0.6$). If averaging of voxel values over an entire fiber tract is applied (“ROI-based approach”), between-protocol differences are not significant anymore ($p \geq 0.165$).

Table 3: Results from mixed ANOVAs based on 20% trimmed means with between-subjects factor “protocol”, within-subjects factor “analysis approach”, and ICC as dependent variable. Follow-up Wilcoxon tests were conducted to compare “protocol” at each factor level of “analysis approach”. Note that W = 0 and W = 210 means that values consistently differed between

samples, therefore indicating complete dominance of one protocol over the other. Effect sizes of Wilcoxon-tests are reported as matched-pairs rank biserial correlation coefficient (King et al., 2011).

	NDI_ICC	ODI_ICC
Between-subjects factor (protocol)	$F(1, 35.84) = 415.17$, $p < .001$	$F(1, 34.49) = 58.06$, $p < .001$
Within-subjects factor (analysis approach)	$F(3, 23.68) = 67.29$, $p < .001$	$F(3, 21.06) = 22.76$, $p < .001$
Interaction (protocol * analysis approach)	$F(3, 23.15) = 120.33$, $p < .001$	$F(3, 21.41) = 20.88$, $p < .001$
Post-hoc ROI	$W = 143$, $p = .165$, $r = 0.36$	$W = 124$, $p = .50$, $r = 0.18$
Post-hoc VBM_6mm	$W = 208$, $p < .001$, $r = 0.98$	$W = 168$, $p = .02$, $r = 0.6$
Post-hoc VBM_0mm	$W = 210$, $p < .001$, $r = 1$	$W = 210$, $p < .001$, $r = 1$
Post-hoc TBSS	$W = 210$, $p < .001$, $r = 1$	$W = 210$, $p < .001$, $r = 1$

Discussion

Using multi-shell high-angular DWI along with theory-driven biophysical models of diffusion like NODDI (Zhang et al., 2012) allows to characterize the brain mechanisms underlying disease, development and plasticity in unprecedented biological plausibility. Here we assessed the longitudinal reproducibility of NODDI metrics in white matter of healthy subjects with special emphasis on the most frequently used image analysis approaches. Reproducibility of NDI and ODI was high during a comparably long time interval of four weeks. The voxel-based approach with Gaussian smoothing kernels of ≥ 4mm FWHM and ROI-averaging yielded the highest reproducibilities (CV_{WS} mostly ≤ 3%, ICC mostly ≥ 0.8), whilst unsmoothed maps and TBSS had consistently higher CV_{WS} and lower ICC. The need to apply some kind of smoothing or averaging of voxel values is even more important if a standard two-shell scan protocol is used. Descriptive statistics suggest that NODDI metrics had comparable (NDI) or even better (ODI) reproducibility than tensor-derived metrics (cf. Andica et al., 2020; Chung et al., 2016), although it must be kept in mind that the latter were computed only based on single-shell DWI data (Barrio-Arranz et al., 2015). Not least, our results indicate that data quality (SNR) is an important determinant of NODDI and DTI metric reproducibility.

NODDI reproducibility and image analysis approach

Collectively, our findings (CV_{WS}, ICC) were coarsely consistent with one previous report on NODDI's between-session reproducibility (Tariq et al., 2013). It is not surprising that between-session reproducibility as reported in the present study tended to be slightly lower than previous within-session reproducibility studies with (Andica et al., 2020) and without subject repositioning (Chung et al., 2016). Of note, the aforementioned studies exclusively assessed reproducibility of what we termed the ROI-based analysis approach (Froeling et al., 2016; Snook et al., 2007), i.e. after voxel values within certain regions of the brain were averaged. In this study, we additionally explored reproducibility of two alternative analysis approaches commonly used in neuroimaging, namely "VBM-style" analysis (Billiet et al., 2015; Broad et al., 2019; Churchill et al., 2019; Dowell et al., 2019; Kraguljac et al., 2019) and the TBSS framework (Kodiweera et al., 2016; Timmers et al., 2016; Zhang et al., 2018).

As a general trend, we observed that with increasing levels of smoothing or averaging voxels within an atlas region, numerical NODDI voxel values across the sample tend to become more homogeneous (decreasing CV_{BS}), as do the repeated measurements of a subject (decreasing CV_{WS}). Therefore, our results align well with previous DTI scan-rescan studies showing that the ROI-based approach (Cabeen et al., 2017; Farrell et al., 2007; Luque Laguna et al., 2020; Vollmar et al., 2010) and smoothing (Cabeen et al., 2017) increase

precision (lower CV_{WS}) compared to (unsmoothed) voxel-wise measures ("VBM-style" analysis/TBSS). However, a less noticed feature of ROI-averaging/smoothing is the concomitant reduction of interindividual variability (lower CV_{BS}), which must be considered an undesired effect under the tacit assumption that CV_{BS} (mainly) reflects true biological variation (Seghier and Price, 2018; Zuo et al., 2019). This indicates that there is a tradeoff among analysis approaches regarding precision and the preservation of sample heterogeneity, which has important consequences for statistical testing (as discussed in the next section). However, ICC – a ratio measure that essentially places the "noise" (CV_{WS}) in the context of biological variation between subjects (CV_{BS}) (Bartko, 1991; Tofts, 2018b) – also indicates that ROI-averaging and smoothing at least tend to outperform approaches without smoothing (i.e., unsmoothed "VBM-style"/TBSS).

Implications for statistical testing

If we now turn to the implications of NODDI reproducibility for statistical testing, it makes sense to differentiate between three common statistical designs, namely *correlation analysis*, *group comparisons* and *analysis of within-subject changes over time* (within or between groups).

The ability to detect correlations with other constructs is crucially dependent on the ability of the variables to discriminate between individuals, or in other words on high ICC (Seghier and Price, 2018; Zuo et al., 2019). With respect to NDI data, researchers can be confident to detect correlations with other (reliable) constructs when using the "VBM-style" framework with ≥ 4mm FWHM smoothing or the ROI approach, yielding ICC's of ≥ 0.8 in most regions/voxels of the white matter. Conversely, based on the ICC results of our study, smoothing extents of < 4mm FWHM and TBSS are clearly less well suited for correlational research using NDI data. We emphasize, however, that it is possible that TBSS outweighs its comparably worse ICC by markedly relaxing the necessary corrections for multiple comparisons across space, thus increasing statistical power (Bach et al., 2014). Regarding ODI maps, all analysis approaches can be used with confidence, as ICC's are collectively in the very good to excellent range. The main reason for this seems to be rooted in the fact that voxel values of ODI maps do generally not cluster in a narrow numerical range, resulting in comparably high between-subject variation (cf. Chung et al., 2016). Interestingly, smoothing of ODI maps with Gaussian kernels of > 8mm FWHM and ROI-averaging leads to a disproportionate decrease of CV_{BS} compared to CV_{WS}, such that ICC tends to fall. Collectively, researchers have the highest chance to detect correlations between ODI and other constructs when opting for a "VBM-style" framework with smoothing kernels between

4mm and 8mm FWHM, but the other analysis approaches investigated here can also be used with confidence.

Cross-sectional group comparisons and longitudinal statistical designs (within or between groups) have in common that variability within subjects/groups is assumed to reflect random noise (Hopkins, 2000; Tofts, 2018b; Zimmerman and Zumbo, 2015), such that the intuitive recommendation for researchers would be to choose an analysis approach with as low as possible CV_{WS}. However, as repeatedly stressed in the literature (Cabeen et al., 2017; Chung et al., 2016; Snook et al., 2007; Vollmar et al., 2010), it is possible that high measurement precision (low CV_{WS}) is at the expense of reduced sensitivity to "true" biological differences or changes. For example, adopting the ROI-based approach might reduce a map's sensitivity if data from a comparably extended region are averaged, whilst the effect is only present in a small part of that region (Snook et al., 2007; Tofts, 2018a). The problem of averaging out true differences or changes might be less pronounced if spatial smoothing is applied, but this approach comes with the cost of an increased risk of partial volume effects (Cabeen et al., 2017). In the present study, this problem was addressed by applying spatial smoothing exclusively within a rather conservative group white matter mask. However, to address the issue of an "optimal" analysis approach in an unbiased way, sensitivity analysis on a ground truth is urgently needed (van Hecke et al., 2009). For the time being, it seems to be a safe choice for researchers to follow the abovementioned recommendations for correlational studies. Interestingly, the majority of previous NODDI studies using "VBM-style" analysis opted for isotropic Gaussian smoothing with kernels ranging from 6mm (Billiet et al., 2015; Broad et al., 2019; Churchill et al., 2019) to 8mm (Dowell et al., 2019), which ensures – according to our data – a reasonable balance between CV_{WS} and CV_{BS}.

Data quality as important determinant of NODDI reproducibility

The impact of analytical and random biological variation on biomarkers of interest is unavoidable in the biological and medical sciences (Fraser and Fogarty, 1989; Hopkins, 2000). To name only a few examples related to NODDI, previous studies have shown that factors like magnetic field strength (Chung et al., 2016), added noise (Hutchinson et al., 2017) and DWI sampling scheme (Hutchinson et al., 2017; Parvathaneni et al., 2018; Wang et al., 2019) affect SNR and NODDI map reproducibility. Our results are generally in line with previous research indicating that regionally varying image quality (quantified via tSNR; Roalf et al., 2016) is highly correlated with regionally varying NODDI and DTI reproducibility. In line with this observation, we also demonstrate that the use of a clinically feasible two-shell scan protocol reduces tSNR and therefore the reproducibility of NODDI maps compared to a three-shell protocol with higher angular resolution. Both of the aforementioned results underline that high SNR, influenced for example by adequate hardware, good imaging

protocols and other measurement issues (see “Introduction”), constitutes a basic requirement for precise and reliable quantification of NODDI and DTI metrics (Chen et al., 2015; Farrell et al., 2007; Hutchinson et al., 2017; Jones, 2004; Wang et al., 2019).

Moreover, we also show a significant inverse relationship between regional tSNR and smoothing-induced changes in ICC, therefore aligning with the notion that smoothing and ROI-averaging aid to mitigate the adverse effects of low SNR (Jones et al., 2005; Snook et al., 2007; van Hecke et al., 2010). Since it has been suggested that smoothing also suppresses the effect of registration inaccuracies (Jones et al., 2005; van Hecke et al., 2010), it is possible that improved within- and between-subject image registration and skeleton projection (de Groot et al., 2013; Zalesky, 2011; see also the “Limitations” section) would reduce the observed reproducibility differences between analysis approaches with and without smoothing. However, given the high correlations between tSNR and DTI/NODDI reproducibility, we hypothesize that the beneficial effects of smoothing (ROI-averaging) are valid even if alternative methods of image registration/skeletonization would be used.

Applications for monitoring individuals and sample size planning

Reproducibility estimates of NODDI metrics (region- or voxel-based) as reported here can in principle be used for evidence-based monitoring of individuals and for sample size planning. We emphasize, however, that these estimates should only be regarded as an approximate order of magnitude due to possible effects of site, vendor, and many others (see “Limitations” section).

If biomarkers are used for monitoring an individual (e.g. patient) over time, the major challenge that physicians/researchers face is to decide whether changes observed in the biomarker are meaningful or whether they are simply caused by analytical and random biological variation (Fraser and Fogarty, 1989; Hopkins, 2000). Monitoring of individuals can also be of high value in the research setting, for example if the aim is to test whether theoretical predictions are also manifested at the individual participant level (Smith and Little, 2018). In terms of monitoring individuals, Hopkins (2000) has proposed that an observed change in an individual’s values exceeding (or deceeding) 1.5 to 2.0 times the CV_{WS} would indicate that a real change has likely occurred (corresponding odds of a real change 6:1 to 12:1).

In the neuroimaging field, the influence of measurement reliability on statistical power (Kanyongo et al., 2007; Zimmerman and Zumbo, 2015) has at best indirectly been considered in sample size planning (Szucs and Ioannidis, 2020; Zuo et al., 2019). A solution for the most common statistical models (independent and dependent samples *t*-test and nonparametric equivalents, one-way ANOVA) suggests to multiply the expected population

effect size by the square root of retest reliability/ICC (Kanyongo et al., 2007; Zuo et al., 2019). Since the composition of the sample of the present study was comparably homogeneous (healthy, young adults, narrow age range), ICC values reported here can be used for studies with assumingly similar or higher between-subject variation (Hopkins, 2000; see also “Limitations” section).

Limitations

Our results show that reproducible characterization of brain microstructure can be achieved using a contemporary 3T scanner, a multi-shell high-angular resolution imaging protocol with whole brain coverage and feasible acquisition time (≈22 min), and reasonable pre- and postprocessing steps. Nevertheless, we are aware that several potential limitations of the present study need to be considered. In the following, we separate the CV_{WS} into its components biological and analytical variation (Fraser and Fogarty, 1989) to discuss these limitations in a structured way. Note that it is in principle possible to estimate the effect of different sources of error with more complex study designs (Brandmaier et al., 2018), but such an approach was beyond the scope of the present study.

Regarding biological variation, we opted for a comparably long interval between measurements, consistent with frequently used study designs in the fields of developmental neuroscience and neuroplasticity (Lebel et al., 2012; Valkanova et al., 2014). However, unlike the case with simulated data or repeatedly measuring a phantom, the degree of “true” biological variation during this interval (e.g., due to developmental changes or cyclical rhythms) is unknown (van Hecke et al., 2009). This should be kept in mind against the background that scan-rescan experiments in humans are based on the tacit assumption that the subjects’ brains are unchanging during the study (Tofts, 2018b). In this vein, it is possible that the trajectory of short-term microstructural changes varies e.g. as a function of age and sex (Kodiweera, 2016, Lebel et al., 2018; Lawrence et al., 2020), although we are not aware of previous studies demonstrating such effects during time intervals as short as four weeks.

Analytical variation is mainly influenced by measurement imprecision (see “Introduction”) as well as the applied pre- and postprocessing steps. Regarding the former, we used a contemporary 3T MRI scanner, but generalizability of results is limited by potential effects of imaging site and vendor (Andica et al., 2020). Furthermore, in line with recommendations in the literature, we used a multi-shell high-angular resolution protocol with three equally spaced *b*-values (Sotiropoulos et al., 2013; Wang et al., 2019) and a comparably high outer shell *b*-value of 3,000 s/mm^2 (Hutchinson et al., 2017; Parvathaneni et al., 2018). When interpreting the results of the present study, it should be kept in mind that factors like the number of shells, choice of *b*-values, angular sampling (number of volumes/ diffusion-encoding directions) and voxel resolution influence both image quality (and reproducibility)

and sensitivity to complex and heterogeneous neurobiological features (Chen et al., 2015; Farrell et al., 2007; Fukutomi et al., 2019; Hutchinson et al., 2017; Jones et al., 2013; Wang et al., 2019). The use of a high outer shell *b*-value, to name just one example, has shown to be beneficial in terms of precise quantification of NDI (Hutchinson et al., 2017; Parvathaneni et al., 2018), but comes at the cost of increased signal variance and rectified noise floor (Hutchinson et al., 2017), which lowers SNR. Conversely, ODI appears to be more sensitive to the number of diffusion-sensitized directions than to the use of high *b*-values (Parvathaneni et al., 2018). In this paper the importance of the aforementioned factors was exemplarily demonstrated by the comparison of a three-shell against a two-shell protocol.

Regarding the pre- and postprocessing, we opted for default settings of FSL's FDT diffusion and TBSS pipelines to keep the processing in line with frequent use in the neurodevelopment and neuroplasticity literature. According to evidence-based recommendations, we made use of unbiased individual halfway templates and applied only one nonlinear warp per subject to standard space (Engvig et al., 2012; Keihaninejad et al., 2013; Madhyastha et al., 2014; Ridgway et al., 2015). Note, however, that a vast number of alternative data processing options exist whose application might yield similar or even better NODDI metric reproducibility than reported here. A thorough treatment of these options would be of high value but was beyond the scope of this paper. For example, although FA maps are frequently used for registration of NODDI maps within- and between-subjects (e.g., Alfaro-Almagro et al., 2018; Andica et al., 2020; Broad et al., 2019; Kodiweera et al., 2016), tensor-based registration algorithms have also been evaluated with success (Keihaninejad et al., 2013; Liu et al., 2014). Moreover, alternative approaches to construct within-subject and group-specific templates have been proposed (Reuter et al., 2012; Zhang et al., 2007; Zhang and Arfanakis, 2018). Furthermore, it has been suggested that the normalisation, skeletonization and skeleton projection steps of the most recent version of TBSS (v1.2) can be further optimized (Bach et al., 2014; de Groot et al., 2013; Leming et al., 2016; Schwarz et al., 2014; Zalesky, 2011). Our results are also somewhat limited with respect to the use of smoothing. For example, instead of using isotropic smoothing after map generation, it might have an influence to apply some kind of adaptive smoothing before (Tabelow et al., 2008) or after (van Hecke et al., 2010) computation of NODDI/DTI maps. Finally, for fitting the NODDI model, we used the non-linear routines introduced in the seminal Zhang et al. (2012) paper and implemented in the NODDI Matlab toolbox. Daducci et al. (2015) have proposed a re-formulation of these routines as linear systems, which reduces the computational burden of fitting the model and renders *Accelerated microstructure imaging via convex optimization (AMICO)* an increasingly popular alternative to the non-linear model. Since extensive validation work by Daducci et al. (2015) revealed that differences between the linear and

non-linear model formulation are negligible, readers planning to use AMICO in their research should be able to use the reproducibility estimates reported here as benchmarks.

Readers should also be aware of specific limitations when interpreting ICC and CV_{WS}, respectively. Although the sample of the present study was comparably homogeneous including only young, healthy adults, we emphasize that the ICC crucially depends on the between-subject variation of a sample, which is generally not known a priori (Hopkins, 2000). Consequently, ICC does only generalize to individuals similar to those in the investigated sample (Hopkins, 2000). We therefore expect that the ICC of NODDI maps would be higher in more heterogeneous populations (e.g., wider age range, inclusion of diseased subjects etc.), and lower in more homogeneous populations (e.g., narrower age range, only one sex, similar expertise level etc.). In contrary, a remarkable property of the CV_{WS} is that it can be estimated from a sample of individuals that is not particularly representative for the population, but nevertheless applies to most individuals in the population (Hopkins, 2000).

To sum up, the evidence from this study suggests that NODDI maps generally possess sufficient measurement precision and remarkable properties to discriminate between individuals based on white matter microstructural features, thus rendering NODDI a suitable modality for the most common cross-sectional and longitudinal research designs. Related to the ability to mitigate the detrimental effects of low SNR (Jones et al., 2005; van Hecke et al., 2010), we demonstrate that the voxel-based approach with Gaussian smoothing kernels of ≥ 4mm FWHM as well as ROI-averaging yielded the highest reproducibility across NODDI metrics. Finally, our results underline the importance of data quality for precise and reliable quantification of NODDI metrics, such that researchers are well-advised to ensure as high as possible SNR in their studies (Chen et al., 2015; Farrell et al., 2007; Hutchinson et al., 2017; Jones et al., 2013; Wang et al., 2019).

Declarations of interest: none

Data and code availability statement:

Data availability: The datasets generated during and/or analysed during the current study will be available on request from the corresponding author [NL] without undue reservation.

Code availability: All previously unpublished computer code used to generate results that are reported in the paper will be available on request from the corresponding author [NL] without undue reservation.

Funding: Norman Aye was funded by the innovation fund of the Otto von Guericke University Magdeburg (2039236003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments: The authors thank Arturo Cardenas-Blanco for the kind support in preparing the imaging protocols and Claus Tempelmann for critical suggestions and discussion.

Author's contributions:

Nico Lehmann: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - original draft, Visualization

Norman Aye: Conceptualization, Methodology, Investigation, Project administration, Writing - Review & Editing

Jörn Kaufmann: Methodology, Investigation, Data Curation, Writing - Review & Editing

Hans-Jochen Heinze: Resources

Emrah Düzel: Conceptualization, Resources

Gabriel Ziegler: Conceptualization, Methodology, Formal analysis, Writing - Review & Editing

Marco Taubert: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

REFERENCES

Alexander DC, Dyrby TB, Nilsson M, Zhang H (2019) Imaging brain microstructure with diffusion MRI: practicality and applications. NMR Biomed 32:e3841.

Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, Sotiropoulos SN, Jbabdi S, Hernandez-Fernandez M, Vallee E, Vidaurre D, Webster M, McCarthy P, Rorden C, Daducci A, Alexander DC, Zhang H, Dragonu I, Matthews PM, Miller KL, Smith SM (2018) Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166:400–424.

Andersson JLR, Graham MS, Zsoldos E, Sotiropoulos SN (2016) Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. Neuroimage 141:556–572.

Andersson JLR, Jenkinson M, Smith SM (2007) Non-linear registration, aka spatial normalisation.

Andersson JLR, Skare S, Ashburner J (2003) How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20:870–888.

Andersson JLR, Sotiropoulos SN (2016) An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 125:1063–1078.

Andica C, Kamagata K, Hayashi T, Hagiwara A, Uchida W, Saito Y, Kamiya K, Fujita S, Akashi T, Wada A, Abe M, Kusahara H, Hori M, Aoki S (2020) Scan-rescan and inter-vendor reproducibility of neurite orientation dispersion and density imaging metrics. Neuroradiology 62:483–494.

Austin PC, Steyerberg EW (2015) The number of subjects per variable required in linear regression analyses. J Clin Epidemiol 68:627–636.

Bach M, Laun FB, Leemans A, Tax CMW, Biessels GJ, Stieltjes B, Maier-Hein KH (2014) Methodological considerations on tract-based spatial statistics (TBSS). Neuroimage 100:358–369.

Barrio-Arranz G, Luis-García R de, Tristán-Vega A, Martín-Fernández M, Aja-Fernández S (2015) Impact of MR Acquisition Parameters on DTI Scalar Indexes: A Tractography Based Approach. PLoS ONE 10:e0137905.

Bartko JJ (1991) Measurement and reliability: statistical thinking considerations. Schizophr Bull 17:483–489.

Basser PJ, Mattiello J, LeBihan D (1994) MR diffusion tensor spectroscopy and imaging. Biophysical Journal 66:259–267.

Basser PJ, Pierpaoli C (1996) Microstructural and Physiological Features of Tissues Elucidated by Quantitative-Diffusion-Tensor MRI. J Magn Reson B 111:209–219.

Beaulieu C (2002) The basis of anisotropic water diffusion in the nervous system - a technical review. NMR Biomed 15:435–455.

Bengtsson SL, Nagy Z, Skare S, Forsman L, Forssberg H, Ullén F (2005) Extensive piano practicing has regionally specific effects on white matter development. Nat Neurosci 8:1148–1150.

Billiet T, Vandenbulcke M, Mädler B, Peeters R, Dhollander T, Zhang H, Deprez S, van den Bergh BRH, Sunaert S, Emsell L (2015) Age-related microstructural differences quantified using myelin water imaging and advanced diffusion MRI. Neurobiol Aging 36:2107–2121.

Brandmaier AM, Wenger E, Bodammer NC, Kühn S, Raz N, Lindenberger U (2018) Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). Elife 7.

Broad RJ, Gabel MC, Dowell NG, Schwartzman DJ, Seth AK, Zhang H, Alexander DC, Cercignani M, Leigh PN (2019) Neurite orientation and dispersion density imaging (NODDI) detects cortical and corticospinal tract degeneration in ALS. J Neurol Neurosurg Psychiatry 90:404–411.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Cabeen RP, Bastin ME, Laidlaw DH (2017) A Comparative evaluation of voxel-based spatial mapping in diffusion tensor imaging. Neuroimage 146:100–112.

Caruyer E, Lenglet C, Sapiro G, Deriche R (2013) Design of multishell sampling schemes with uniform coverage in diffusion MRI. Magn Reson Med 69:1534–1540.

Chang EH, Argyelan M, Aggarwal M, Chandon T-SS, Karlsgodt KH, Mori S, Malhotra AK (2017) The role of myelination in measures of white matter integrity: Combination of diffusion tensor imaging and two-photon microscopy of CLARITY intact brains. Neuroimage 147:253–261.

Chen Y, Tymofiyeva O, Hess CP, Xu D (2015) Effects of rejecting diffusion directions on tensor-derived parameters. Neuroimage 109:160–170.

Choi S, Cunningham DT, Aguila F, Corrigan JD, Bogner J, Mysiw WJ, Knopp MV, Schmalbrock P (2011) DTI at 7 and 3 T: systematic comparison of SNR and its influence on quantitative metrics. Magn Reson Imaging 29:739–751.

Chung AW, Seunarine KK, Clark CA (2016) NODDI reproducibility and variability with magnetic field strength: A comparison between 1.5 T and 3 T. Hum Brain Mapp 37:4550–4565.

Churchill NW, Caverzasi E, Graham SJ, Hutchison MG, Schweizer TA (2019) White matter during concussion recovery: Comparing diffusion tensor imaging (DTI) and neurite orientation dispersion and density imaging (NODDI). Hum Brain Mapp 40:1908–1918.

Cliff N (1996) Ordinal methods for behavioral data analysis. Mahwah, NJ: Erlbaum.

Cohen J (1988) Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cook PA, Bai Y, Nedjati-Gilani S, Seunarine KK, Hall MG, Parker GJ, Alexander DC (2006) Camino: Open-Source Diffusion-MRI Reconstruction and Processing. 14th Scientific Meeting of the International Society for Magnetic Resonance in Medicine, Seattle, WA, USA, p. 2759.

Daducci A, Canales-Rodríguez EJ, Zhang H, Dyrby TB, Alexander DC, Thiran J-P (2015) Accelerated Microstructure Imaging via Convex Optimization (AMICO) from diffusion MRI data. Neuroimage 105:32–44.

de Groot M, Vernooij MW, Klein S, Ikram MA, Vos FM, Smith SM, Niessen WJ, Andersson JLR (2013) Improving alignment in Tract-based spatial statistics: evaluation and optimization of image registration. Neuroimage 76:400–411.

Dowell NG, Bouyagoub S, Tibble J, Voon V, Cercignani M, Harrison NA (2019) Interferon-alpha-Induced Changes in NODDI Predispose to the Development of Fatigue. Neuroscience 403:111–117.

Eklund A (2016) beeswarm. The Bee Swarm Plot, an Alternative to Stripchart. R package version 0.2.3. https://CRAN.R-project.org/package=beeswarm.

Engvig A, Fjell AM, Westlye LT, Moberget T, Sundseth Ø, Larsen VA, Walhovd KB (2012) Memory training impacts short-term changes in aging white matter: a longitudinal diffusion tensor imaging study. Hum Brain Mapp 33:2390–2406.

Fabbris L (1980) Measures of predictor variable importance in multiple regression: An additional suggestion. Qual Quant 14:787–792.

Farrell JAD, Landman BA, Jones CK, Smith SA, Prince JL, van Zijl PCM, Mori S (2007) Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T. J Magn Reson Imaging 26:756–767.

Filley CM, Fields RD (2016) White matter and cognition: making the connection. J Neurophysiol 116:2093–2104.

Fraser CG, Fogarty Y (1989) Interpreting laboratory results. BMJ 298:1659–1660.

Froeling M, Pullens P, Leemans A (2016) DTI Analysis Methods: Region of Interest Analysis. In: Diffusion Tensor Imaging (van Hecke W, Emsell L, Sunaert S, eds), pp 175–182. New York, NY: Springer.

Fukutomi H, Glasser MF, Murata K, Akasaka T, Fujimoto K, Yamamoto T, Autio JA, Okada T, Togashi K, Zhang H, van Essen DC, Hayashi T (2019) Diffusion Tensor Model links to Neurite Orientation Dispersion and Density Imaging at high b-value in Cerebral Cortical Gray Matter. Sci Rep 9:12246.

Granberg T, Fan Q, Treaba CA, Ouellette R, Herranz E, Mangeat G, Louapre C, Cohen-Adad J, Klawiter EC, Sloane JA, Mainero C (2017) In vivo characterization of cortical and white matter neuroaxonal pathology in early multiple sclerosis. Brain 140:2912–2926.

Greve DN, Fischl B (2009) Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63–72.

Grussu F, Schneider T, Tur C, Yates RL, Tachrount M, Ianuş A, Yiannakas MC, Newcombe J, Zhang H, Alexander DC, DeLuca GC, Gandini Wheeler-Kingshott CAM (2017) Neurite dispersion: a new marker of multiple sclerosis spinal cord pathology? Ann Clin Transl Neurol 4:663–679.

Guerrero JM, Adluru N, Bendlin BB, Goldsmith HH, Schaefer SM, Davidson RJ, Kecskemeti SR, Zhang H, Alexander AL (2019) Optimizing the intrinsic parallel diffusivity in NODDI: An extensive empirical evaluation. PLoS ONE 14:e0217118.

Hopkins WG (2000) Measures of reliability in sports medicine and science. Sports Med 30:1–15.

Hua K, Zhang J, Wakana S, Jiang H, Li X, Reich DS, Calabresi PA, Pekar JJ, van Zijl PCM, Mori S (2008) Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. Neuroimage 39:336–347.

Hutchinson EB, Avram AV, Irfanoglu MO, Koay CG, Barnett AS, Komlosh ME, Özarslan E, Schwerin SC, Juliano SL, Pierpaoli C (2017) Analysis of the effects of noise, DWI sampling, and value of assumed parameters in diffusion MRI models. Magn Reson Med 78:1767–1780.

Jenkinson M, Bannister PR, Brady M, Smith SM (2002) Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. Neuroimage 17:825–841.

Jespersen SN, Bjarkam CR, Nyengaard JR, Chakravarty MM, Hansen B, Vosegaard T, Østergaard L, Yablonskiy D, Nielsen NC, Vestergaard-Poulsen P (2010) Neurite density from magnetic resonance diffusion measurements at ultrahigh field: comparison with light microscopy and electron microscopy. Neuroimage 49:205–216.

Jones DK (2004) The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study. Magn Reson Med 51:807–815.

Jones DK, Knösche TR, Turner R (2013) White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. Neuroimage 73:239–254.

Jones DK, Symms MR, Cercignani M, Howard RJ (2005) The effect of filter size on VBM analyses of DT-MRI data. Neuroimage 26:546–554.

Kanyongo GY, Brook GP, Kyei-Blankson L, Gocmen G (2007) Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Nonparametric Statistics. J Mod App Stat Meth 6:81–90.

Keihaninejad S, Zhang H, Ryan NS, Malone IB, Modat M, Cardoso MJ, Cash DM, Fox NC, Ourselin S (2013) An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to Alzheimer's disease. Neuroimage 72:153–163.

King BM, Rosopa P, Minium EW (2011) Statistical reasoning in the behavioral sciences. Hoboken, NJ: John Wiley.

Kodiweera C, Alexander AL, Harezlak J, McAllister TW, Wu Y-C (2016) Age effects and sex differences in human brain white matter of young to middle-aged adults: A DTI, NODDI, and q-space study. Neuroimage 128:180–192.

Kraguljac NV, Anthony T, Monroe WS, Skidmore FM, Morgan CJ, White DM, Patel N, Lahti AC (2019) A longitudinal neurite and free water imaging study in patients with a schizophrenia spectrum disorder. Neuropsychopharmacology 44:1932–1939.

Lakhani DA, Schilling KG, Xu J, Bagnato F (2020) Advanced Multicompartment Diffusion MRI Models and Their Application in Multiple Sclerosis. AJNR Am J Neuroradiol 41:751–757.

Lawrence KE, Nabulsi L, Santhalingam V, Abaryan Z, Villalon-Reina JE, Nir TM, Ba Gari I, Zhu AH, Haddad E, Muir AM, Jahanshad N, Thompson PM Advanced diffusion-weighted MRI metrics detect sex differences in aging among 15,000 adults in the UK Biobank. In: The 16th International Symposium on Medical Information Processing and Analysis, 2020, Lima, Peru (Brieva J, Lepore N, Romero Castro E, Linguraru MG, eds), p 28: SPIE Digital Library.

Lebel C, Gee M, Camicioli R, Wieler M, Martin W, Beaulieu C (2012) Diffusion tensor imaging of white matter tract evolution over the lifespan. Neuroimage 60:340–352.

Lebel C, Treit S, Beaulieu C (2019) A review of diffusion MRI of typical white matter development from early childhood to young adulthood. NMR Biomed 32:e3778.

Leemans A, Jones DK (2009) The B-matrix must be rotated when correcting for subject motion in DTI data. Magn Reson Med 61:1336–1349.

Lehmann N, Villringer A, Taubert M (2020) Colocalized White Matter Plasticity and Increased Cerebral Blood Flow Mediate the Beneficial Effect of Cardiovascular Exercise on Long-Term Motor Learning. J Neurosci 40:2416–2429.

Leming M, Steiner R, Styner M (2016) A framework for incorporating DTI Atlas Builder registration into Tract-Based Spatial Statistics and a simulated comparison to standard TBSS. Proc SPIE Int Soc Opt Eng 9788.

Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 49:764–766.

Li S-C, Lindenberger U, Hommel B, Aschersleben G, Prinz W, Baltes PB (2004) Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychol Sci 15:155–163.

Liu X, Yang Y, Sun J, Yu G, Xu J, Niu C, Tian H, Lin P (2014) Reproducibility of diffusion tensor imaging in normal subjects: an evaluation of different gradient sampling schemes and registration algorithm. Neuroradiology 56:497–510.

Loken E, Gelman A (2017) Measurement error and the replication crisis. Science 355:584–585.

Luque Laguna PA, Combes AJE, Streffer J, Einstein S, Timmers M, Williams SCR, Dell'Acqua F (2020) Reproducibility, reliability and variability of FA and MD in the older healthy population: A test-retest multiparametric analysis. Neuroimage Clin 26:102168.

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. Hum Brain Mapp 35:4544–4555.

Mangiafico S (2020) rcompanion: Functions to Support Extension Education Program Evaluation. R package version 2.3.25. https://CRAN.R-project.org/package=rcompanion.

Marenco 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.

Mills KL, Goddings A-L, Herting MM, Meuwese R, Blakemore S-J, Crone EA, Dahl RE, Güroğlu B, Raznahan A, Sowell ER, Tamnes CK (2016) Structural brain development between childhood and adulthood: Convergence across four longitudinal samples. Neuroimage 141:273–281.

Mollink J, Kleinnijenhuis M, van Cappellen Walsum A-M, Sotiropoulos SN, Cottaar M, Mirfin C, Heinrich MP, Jenkinson M, Pallebage-Gamarallage M, Ansorge O, Jbabdi S, Miller KL (2017) Evaluating fibre orientation dispersion in white matter: Comparison of diffusion MRI, histology and polarized light imaging. Neuroimage 157:561–574.

Nimon KF, Oswald FL (2013) Understanding the Results of Multiple Linear Regression. Organ Res Methods 16:650–674.

Novikov DS, Fieremans E, Jespersen SN, Kiselev VG (2019) Quantifying brain microstructure with diffusion MRI: Theory and parameter estimation. NMR Biomed 32:e3998.

Novikov DS, Kiselev VG, Jespersen SN (2018) On modeling. Magn Reson Med 79:3172–3193.

Parvathaneni P, Nath V, Blaber JA, Schilling KG, Hainline AE, Mojahed E, Anderson AW, Landman BA (2018) Empirical reproducibility, sensitivity, and optimization of acquisition protocol, for Neurite Orientation Dispersion and Density Imaging using AMICO. Magn Reson Imaging 50:96–109.

Pierpaoli C, Basser PJ (1996) Toward a quantitative assessment of diffusion anisotropy. Magn Reson Med 36:893–906.

Polders DL, Leemans A, Hendrikse J, Donahue MJ, Luijten PR, Hoogduin JM (2011) Signal to noise ratio and uncertainty in diffusion tensor imaging at 1.5, 3.0, and 7.0 Tesla. J Magn Reson Imaging 33:1456–1463.

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline J-B, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18:115–126.

R Development Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Raja R, Rosenberg G, Caprihan A (2019) Review of diffusion MRI studies in chronic white matter diseases. Neurosci Lett 694:198–207.

Reuter M, Schmansky NJ, Rosas HD, Fischl B (2012) Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61:1402–1418.

Ridgway GR, Leung KK, Ashburner J (2015) Computing Brain Change over Time. In: Brain Mapping: An Encyclopedic Reference (Toga AW, ed), pp 417–428. London, UK: Academic Press.

Roalf DR, Quarmley M, Elliott MA, Satterthwaite TD, Vandekar SN, Ruparel K, Gennatas ED, Calkins ME, Moore TM, Hopson R, Prabhakaran K, Jackson CT, Verma R, Hakonarson H, Gur RC, Gur RE (2016) The impact of quality assurance assessment on diffusion tensor imaging outcomes in a large-scale population-based cohort. Neuroimage 125:903–919.

Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ (1999) Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging 18:712–721.

Sampaio-Baptista C, Johansen-Berg H (2017) White Matter Plasticity in the Adult Brain. Neuron 96:1239–1251.

Schilling KG, Janve V, Gao Y, Stepniewska I, Landman BA, Anderson AW (2018) Histological validation of diffusion MRI fiber orientation distributions and dispersion. Neuroimage 165:200–221.

Scholz J, Klein MC, Behrens TEJ, Johansen-Berg H (2009) Training induces changes in white-matter architecture. Nat Neurosci 12:1370–1371.

Schwarz CG, Reid RI, Gunter JL, Senjem ML, Przybelski SA, Zuk SM, Whitwell JL, Vemuri P, Josephs KA, Kantarci K, Thompson PM, Petersen RC, Jack CR (2014) Improved DTI registration allows voxel-based analysis that outperforms tract-based spatial statistics. Neuroimage 94:65–78.

Seghier ML, Price CJ (2018) Interpreting and Utilising Intersubject Variability in Brain Function. Trends Cogn Sci 22:517–530.

Sepehrband F, Clark KA, Ullmann JFP, Kurniawan ND, Leanage G, Reutens DC, Yang Z (2015) Brain tissue compartment density estimated using diffusion-weighted MRI yields tissue parameters consistent with histology. Hum Brain Mapp 36:3687–3702.

Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428.

Smith PL, Little DR (2018) Small is beautiful: In defense of the small-N design. Psychon Bull Rev 25:2083–2101.

Smith SM (2002) Fast robust automated brain extraction. Hum Brain Mapp 17:143–155.

Smith SM, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols TE, Mackay CE, Watkins KE, Ciccarelli O, Cader MZ, Matthews PM, Behrens TEJ (2006) Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31:1487–1505.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, Luca M de, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, Stefano N de, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1:S208-19.

Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, Stefano N de (2002) Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 17:479–489.

Snook L, Plewes C, Beaulieu C (2007) Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. Neuroimage 34:243–252.

Sone D (2019) Neurite orientation and dispersion density imaging: clinical utility, efficacy, and role in therapy. Rep Medical Imaging 12:17–29.

Sotiropoulos SN, Jbabdi S, Xu J, Andersson JL, Moeller S, Auerbach EJ, Glasser MF, Hernandez M, Sapiro G, Jenkinson M, Feinberg DA, Yacoub E, Lenglet C, van Essen DC, Ugurbil K, Behrens TEJ (2013) Advances in diffusion MRI acquisition and processing in the Human Connectome Project. Neuroimage 80:125–143.

Sullivan DC, Obuchowski NA, Kessler LG, Raunig DL, Gatsonis C, Huang EP, Kondratovich M, McShane LM, Reeves AP, Barboriak DP, Guimaraes AR, Wahl RL (2015) Metrology Standards for Quantitative Imaging Biomarkers. Radiology 277:813–825.

Szucs D, Ioannidis JP (2020) Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. Neuroimage:117164.

Tabelow K, Polzehl J, Spokoiny V, Voss HU (2008) Diffusion tensor imaging: structural adaptive smoothing. Neuroimage 39:1763–1773.

Tariq M, Schneider T, Alexander DC, Wheeler-Kingshott CAM, Zhang H (2013) Assessing scan-rescan reproducibility of the parameter estimates from NODDI. In: Proceedings of the 21st Annual Meeting of the ISMRM, Salt Lake City, Utah, USA, 20-26 April 2013 (Gold GE, ed), p 3187. Salt Lake City, UT: International Society for Magnetic Resonance in Medicine.

Timmers I, Roebroeck A, Bastiani M, Jansma B, Rubio-Gozalbo E, Zhang H (2016) Assessing Microstructural Substrates of White Matter Abnormalities: A Comparative Study Using DTI and NODDI. PLoS ONE 11:e0167884.

Tofts PS (2018a) Measurement Process: MR Data Collection and Image Analysis. In: Quantitative MRI of the Brain: Principles of Physical Measurement (Cercignani M, Dowell NG, Tofts PS, eds), pp 13–31. Milton: CRC Press.

Tofts PS (2018b) Quality Assurance: Accuracy, Precision, Controls and Phantoms. In: Quantitative MRI of the Brain: Principles of Physical Measurement (Cercignani M, Dowell NG, Tofts PS, eds), pp 33–53. Milton: CRC Press.

Torchiano M (2016) Effsize - A Package For Efficient Effect Size Computation. R package version 0.8.0. https://CRAN.R-project.org/package=effsize.

Valkanova V, Eguia Rodriguez R, Ebmeier KP (2014) Mind over matter—what do we know about neuroplasticity in adults? Int Psychogeriatr 26:891–909.

van Hecke W, Leemans A, Backer S de, Jeurissen B, Parizel PM, Sijbers J (2010) Comparing isotropic and anisotropic smoothing for voxel-based DTI analyses: A simulation study. Hum Brain Mapp 31:98–114.

van Hecke W, Sijbers J, Backer S de, Poot D, Parizel PM, Leemans A (2009) On the construction of a ground truth framework for evaluating voxel-based diffusion tensor MRI analysis methods. Neuroimage 46:692–707.

Vollmar C, O'Muircheartaigh J, Barker GJ, Symms MR, Thompson P, Kumari V, Duncan JS, Richardson MP, Koepp MJ (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.

Voss MW, Heo S, Prakash RS, Erickson KI, Alves H, Chaddock L, Szabo AN, Mailey EL, Wójcicki TR, White SM, Gothe N, McAuley E, Sutton BP, Kramer AF (2013) The influence of aerobic fitness on cerebral white matter integrity and cognitive function in older adults: results of a one-year exercise intervention. Hum Brain Mapp 34:2972–2985.

Wakana S, Caprihan A, Panzenboeck MM, Fallon JH, Perry M, Gollub RL, Hua K, Zhang J, Jiang H, Dubey P, Blitz A, van Zijl PCM, Mori S (2007) Reproducibility of quantitative tractography methods applied to cerebral white matter. Neuroimage 36:630–644.

Walhovd KB, Johansen-Berg H, Káradóttir RT (2014) Unraveling the secrets of white matter—bridging the gap between cellular, animal and human imaging studies. Neuroscience 276:2–13.

Wang N, Zhang J, Cofer G, Qi Y, Anderson RJ, White LE, Johnson GA (2019) Neurite orientation dispersion and density imaging of mouse brain microstructure. Brain Struct Funct 224:1797–1813.

Wheeler-Kingshott CAM, Cercignani M (2009) About "axial" and "radial" diffusivities. Magn Reson Med 61:1255–1260.

Wickham H (2016) ggplot2: Elegant graphics for data analysis. Cham: Springer.

Wilcox RR (2017) Introduction to robust estimation and hypothesis testing. Burlington: Elsevier.

Wilcox RR, Schönbrodt FD (2019) WRS: A package of R.R. Wilcox' robust statistics functions. R package version 0.36. https://github.com/nicebread/WRS.

Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM, Nichols TE (2016) Non-parametric combination and related permutation tests for neuroimaging. Hum Brain Mapp 37:1486–1511.

Zalesky A (2011) Moderating registration misalignment in voxelwise comparisons of DTI data: a performance evaluation of skeleton projection. Magn Reson Imaging 29:111–125.

Zhang H, Schneider T, Wheeler-Kingshott CAM, Alexander DC (2012) NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 61:1000–1016.

Zhang H, Yushkevich PA, Rueckert D, Gee JC (2007) Unbiased white matter atlas construction using diffusion tensor images. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2007: 10th International Conference, Brisbane, Australia, October 29 - November 2, 2007, Proceedings, Part II, vol. 4792 (Ayache N, Ourselin S, Maeder A, eds), pp 211–218. Berlin, Heidelberg: Springer.

Zhang J, Gregory S, Scahill RI, Durr A, Thomas DL, Lehericy S, Rees G, Tabrizi SJ, Zhang H (2018) In vivo characterization of white matter pathology in premanifest huntington's disease. Ann Neurol 84:497–504.

Zhang S, Arfanakis K (2018) Evaluation of standardized and study-specific diffusion tensor imaging templates of the adult human brain: Template characteristics, spatial normalization accuracy, and detection of small inter-group FA differences. Neuroimage 172:40–50.

Zhang Y, Brady M, Smith SM (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20:45–57.

Zimmerman DW, Zumbo BD (2015) Resolving the Issue of How Reliability is Related to Statistical Power: Adhering to Mathematical Definitions. J Mod App Stat Meth 14:9–26.

Zuo X-N, Xu T, Milham MP (2019) Harnessing reliability for neuroscience research. Nat Hum Behav 3:768–771.