# Longitudinal Reliability of Tract-Based Spatial Statistics in Diffusion Tensor Imaging

**Tara Madhyastha,[1]* Susan Mérillat,[2,3] Sarah Hirsiger,[2,3] Ladina Bezzola,[2,3,4] Franziskus Liem,[4] Thomas Grabowski,[1,5] and Lutz Jäncke[2,3,4]**

[1]*Department of Radiology and Integrated Brain Imaging Center (IBIC), University of Washington, Seattle, Washington*
[2]*International Normal Aging and Plasticity Imaging Center (INAPIC), University of Zurich, Zurich, Switzerland*
[3]*University Research Priority Program (URPP) "Dynamics of Healthy Aging", University of Zurich, Zurich, Switzerland*
[4]*University of Zurich, Institute of Psychology, Neuropsychology, Zurich, Switzerland*
[5]*Department of Neurology, University of Washington, Seattle, Washington*

**Abstract:** Relatively little is known about reliability of longitudinal diffusion-tensor imaging (DTI) measurements despite growing interest in using DTI to track change in white matter structure. The purpose of this study is to quantify within- and between session scan-rescan reliability of DTI-derived measures that are commonly used to describe the characteristics of neural white matter in the context of neural plasticity research. DTI data were acquired from 16 cognitively healthy older adults (mean age 68.4). We used the Tract-Based Spatial Statistics (TBSS) approach implemented in FSL, evaluating how different DTI preprocessing choices affect reliability indices. Test-Retest reliability, quantified as ICC averaged across the voxels of the TBSS skeleton, ranged from 0.524 to 0.798 depending on the specific DTI-derived measure and the applied preprocessing steps. The two main preprocessing steps that we found to improve TBSS reliability were (a) the use of a common individual template and (b) smoothing DTI data using a 1-voxel median filter. Overall our data indicate that small choices in the preprocessing pipeline have a significant effect on test-retest reliability, therefore influencing the power to detect change within a longitudinal study. Furthermore, differences in the data processing pipeline limit the comparability of results across studies. *Hum Brain Mapp 00:000–000, 2014.* © **2014 Wiley Periodicals, Inc.**

**Key words:** diffusion tensor imaging; tract-based spatial statistics; preprocessing; reliability; longitudinal

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Velux Stiftung; Contract grant number: project No. 369; Contract grant sponsor: Swiss National Science Foundation.

*Correspondence to: Tara Madhyastha, Department of Radiology and Integrated Brain Imaging Center (IBIC), University of Washington, Seattle, WA 98195, E-mail: madhyt@uw.edu

Received for publication 16 January 2014; Accepted 6 February 2014.

DOI 10.1002/hbm.22493
Published online 00 Month 2014 in Wiley Online Library (wileyonlinelibrary.com).

## INTRODUCTION

Interest in using neuroimaging to track longitudinal within-subject changes is increasing in the field of aging research. However, with this interest comes an even more pressing need to understand the reliability of neuroimaging measures in specific age groups to correctly interpret changes observed over time.

Diffusion tensor imaging (DTI), a magnetic resonance imaging (MRI) technique that allows the mapping of the diffusion of water molecules in the brain, is one of the most frequently used neuroimaging techniques to assess white matter structural brain changes associated with development and aging [Madden et al., 2012; Sullivan and Pfefferbaum, 2006]. In cerebral white matter (WM), the diffusion of water molecules follows the direction of the axon bundles, while the transverse movement is limited (as for example by cell membranes) resulting in an anisotropic diffusion pattern [Le Bihan et al., 2001; Mori and Zhang, 2006]. Several different DTI-derived parameters are used to describe different aspects of this diffusion pattern and, thus, allow inferences about WM microstructure. Considering one voxel as basic element of the MR image, DTI-derived diffusion information can be mathematically expressed as ellipsoid or so-called diffusion tensor. Each of the three eigenvalues (λ1, λ2, and λ3) that define the shape of the ellipsoid represents a measure of diffusivity along the corresponding primary axis of the ellipsoid (v1, v2, or v3). Diffusivity along the principal axis, λ1, is called axial diffusivity (AD) while radial diffusivity (RD) refers to the averaged diffusivity along the two minor axes [Madden et al., 2012; Mori and Zhang, 2006]. While earlier investigations associated AD/RD with white matter properties, such as axonal integrity and especially the degree of myelination [Song et al., 2003, 2005], more recent simulation work advocates caution when speculating about the underlying biophysical substrates of changes in AD and RD. It is known that these measures are sensitive to the eigenvalue sorting or the effects of noise and partial volume [Wheeler-Kingshott and Cercignani, 2009] in addition to the underlying fiber structure [Jones et al., 2013]. From the eigenvalues of the diffusion tensor, scalar DTI summary measures can be calculated. The fractional anisotropy (FA) coefficient is the most frequently used DTI summary measure of white matter integrity and represents the rate of orientation preference within a tissue [Le Bihan et al., 2001; Madden et al., 2012]. FA values range from 0 to 1 with higher values indicating increased directionality of diffusion, independent of the diffusion rate [Le Bihan et al., 2001]. Because the directionality of diffusion depends on the presence and density of physical barriers, FA is higher in WM where the diffusion of water molecules is restricted: for example, by myelin sheaths of axons. Mean diffusivity (MD) is a measure used to describe the local magnitude of diffusion regardless of direction [Assaf and Pasternak, 2008]. Higher values of MD are associated with low restrictions of the diffusion movement.

DTI studies consistently show an effect of age on FA and MD [Burzynska et al., 2010; Charlton et al., 2008; Sullivan and Pfefferbaum, 2006; Yoon et al., 2008; Ziegler et al., 2010], which is in line with the suggestion that these parameters represent potential of cerebral health that is—in terms of WM—defined e.g., by axonal integrity and intact myelination. Until now, our understanding of white matter properties in old age is mainly based on cross-sectional age group comparisons that do not allow one to draw conclusions about age-related change because of potential cohort effects and uncontrolled interindividual differences [Schaie, 2005]. For example, a typical concern in studies that compare different age groups cross-sectionally is selection bias, meaning that older subjects do not represent the same population as the younger subjects will when they reach the same age. Therefore, cross-sectional estimates of white matter integrity may under- or overestimate the true rate of longitudinal change [as demonstrated by Barrick et al., 2010], or misrepresent the shape of individual trajectories. Fortunately, the contributions of longitudinal studies are accumulating and there is beginning evidence for substantial age-related changes in white matter in older samples [Fjell and Walhovd, 2010; Teipel et al., 2010].

However, the statistical power to detect within-person individual change in longitudinal studies is highly dependent upon the reliability of the measurement. Reliability is a quantification of the true variance of a measurement relative to the total variance of a measurement. Thus, a perfect measurement has a reliability of 1. A measurement that has variability caused by error has a reliability that is less than 1. Unfortunately, there are a wide range of issues in DTI acquisition, preprocessing, and analysis that contribute to reliability and interpretation [Jones and Cercignani, 2010]. For example, raw data must be corrected for eddy currents (distortions produced by local electric fields) that cannot be completely eliminated through acquisition parameters, and for subject motion. Other preprocessing steps to reduce noise in DTI, such as smoothing, may be performed before tensor estimation. Finally, there are several popular analysis methods that have been used in previous studies to obtain comparable DTI statistics: (1) Voxel wise analyses, (2) Region-of-Interest (ROI) analyses, (3) probabilistic tractography (fiber tracking), and (4) TBSS (Tract-Based Spatial Statistics, [Smith et al., 2006]). There is remarkably little information on test-retest reliability of measures derived from state-of-the-art DTI protocols on 3.0T MRI machines [Fox et al., 2012; Jones et al., 2007; Magnotta et al., 2012; Vollmar et al., 2010; Wang et al., 2012]. To the best of our knowledge only one previous study, in which the primary interest was to assess the test–retest multicenter reliability, reported test–retest reliability of the default TBSS analysis pipeline [Vollmar et al., 2010]. Overall, the studies consistently found very high intrasite reliability of FA, MD, AD, and RD for ROI analyses and voxel wise analysis using the same preprocessing pipeline. However, Jones et al.

[2007] found that voxel wise analysis results were extremely sensitive to the specific choices made for smoothing, image coregistration, etc., making it difficult to compare results from different research groups. Reliability of probabilistic tractography was shown to be significantly lower [Vollmar et al., 2010]. Specifically, we can conclude from recent literature that (a) intrasite reliability is higher than intersite reliability [Magnotta et al., 2012; Vollmar et al., 2010], (b) reliability on 3.0 T systems is higher compared to 1.5 T systems [Vollmar et al., 2010], (c) reliability is affected by the approach chosen for image coregistration [Magnotta et al., 2012; Vollmar et al., 2010], and (d) reliability of DTI-derived measures is generally worse as compared to for example reliability of gray matter parameters derived from T1-weighted structural brain images (such as cortical thickness values calculated in Freesurfer [Hirsiger et al., n.d.; Morey et al., 2010]. Regarding the latter, there is certainly room for improvement of reliability, e.g., by means of methodological adaptations.

TBSS has gained popularity for its ease of use and improved alignment of white matter as compared to voxel wise analysis methods [Jones and Cercignani, 2010] and for its ability to conduct whole brain analysis. TBSS uses a carefully optimized nonlinear registration procedure [FNIRT: Andersson et al., 2007a,b] to warp subjects to a common space, and then projects all subjects' FA onto a mean FA tract skeleton that includes skeletonized representations of the major fiber tracts. The projection is achieved by searching perpendicular to the local skeleton structure for the maximum FA value in the subject's FA image. After creation of the FA skeleton, other DTI scalar measures can be aligned to the same space for analysis using voxel-wise statistics within the skeleton. The data do not need to be smoothed to correct for small alignment errors, and the number of voxels used for comparison is reduced to those in the skeleton [Smith et al., 2006]. Because of TBSS's applicability to the examination of whole-brain change, it has the potential to be useful in longitudinal study of white matter integrity.

In the TBSS pipeline, large potential causes of error are scanner noise and registration differences that cause measurements of scalar DTI statistics within a single voxel to be different even when the underlying "true" value is identical. Reliability is related to the effect size ($d$), usually defined as:

$$d = \frac{\Delta}{\sigma} \tag{1}$$

where Δ is the mean difference in the statistic of interest and σ is the standard deviation of the sample. Improving reliability reduces the standard deviation ($\sigma$), making the same Δ translate into a larger effect, making the test more sensitive. Finally, effect size is related to statistical power and the rate of Type 1 (false positive) and Type II (false negative) such that if we keep the number of subjects and Type 1 and Type 2 error rate identical, increasing reliability will decrease the magnitude of change (Δ) that can be identified (without changing the false negative rate). Alternatively, increasing the reliability without changing (Δ) will increase the power of the test by decreasing the Type II error rate [see Rice, 2006 for *derivation of formulas, chapter 11.2.2*].

In this study, we assessed within- and across-session within-subject reliability of different common DTI measures calculated by TBSS in a sample of older subjects (65+). In the context of longitudinal DTI data analysis, where within-subject changes are of potentially greater interest than group differences, we can consider the effect of different processing alternatives and of error introduced by the scanner itself on the reliability of the TBSS results (and subsequent ability to detect significant change). To this end, we tested to what degree reliability coefficients are affected by different TBSS-based processing pipelines including different sets of coregistration approaches, smoothing, and preprocessing options.

## MATERIAL AND METHODS

### Research Participants

Participants were 16 older adults (mean age 68.4 ± 2.45 years, 8 women) selected from a larger longitudinal study with seniors (longitudinal Healthy Aging Brain (l-HAB) project) conducted at the International Normal Aging and Plasticity Imaging Center (INAPIC) at the University of Zurich [Zöllig et al., 2011]. All participants were cognitively healthy, had no history of neurological or psychiatric disorder and gave written informed consent and the local ethics committee approved the present study, in compliance with the Helsinki Declaration.

### Study Design

Participants were scanned on two occasions (time 1, time 2), one week apart. Each occasion had two sessions that were separated by a break of 15 minutes that subjects spent outside the scanner. At time 1, two T1-weighted images, one Diffusion-weighted tensor (DTI) image and a resting-state fMRI image were obtained. After re-positioning the participants in the scanner bore after the break, at the second session of time 1, one T1-weighted and one DTI image were obtained. The scan protocol was the same for time 2. Scan order was randomized across time points and subjects. With this study design we were able to explore (1) within-day reliability, comparing images obtained during the two sessions of one occasion and (2) across-day reliability, comparing images obtained during the two occasions. Reliability analysis described in this article was limited to the DTI images.

### Imaging Acquisition

Data acquisition was performed on a 3.0 T Philips Achieva whole body scanner (Philips Medical Systems, Best, The Netherlands) equipped with a transmit-receive body coil and a commercial 32-element sensitivity encoding (SENSE) head coil array. Diffusion-weighted single-shot spin echo EPI sequence scans were obtained with a measured spatial resolution of 2 × 2 × 2 mm (acquisition matrix 112 × 112 pixels, 75 slices, no gap). Further imaging parameters were: Field of view FOV = 224 × 224 mm, echo-time TE = 55 ms, repetition time TR = 13.24 s, flip-angle = 90°, SENSE factor $R = 2.0$. Diffusion-weighted scans were performed along 32 noncollinear directions with a maximum b-factor of 1000 s/mm$^2$, complemented by one scan with $b = 0$ s/mm$^2$ (reference volume). Total acquisition time for the DTI image was 8 min 50 s.

## Image Processing

### Tract-based spatial statistics

We focus on the reliability of tract-based spatial statistics (TBSS) introduced by Smith et al. [2006], which is part of FSL 5.0 (http://www.fmrib.ox.ac.uk/fsl) [Smith et al., 2003]. After warping subjects to a common space, all subjects' FA maps are projected onto a mean FA tract skeleton. The skeleton is thresholded at an FA value of 0.2 (the default recommended threshold, which was appropriate for our data). After creation of the FA skeleton, other DTI scalar statistics can be aligned to the same space for analysis using voxel-wise statistics within the skeleton. In this study, we examine the reliability of commonly used scalar DTI indices: FA, MD, AD, and RD.

### Subject-specific template

One difficulty with the default process of nonlinear registration to common space is that, if a different nonlinear warp is used for the same subject at multiple time points, within-subject longitudinal differences may be removed. For this reason, the use of a nonbiased individual subject template has been recommended [Engvig et al., 2011]. We evaluated the use of a subject-specific template, created by registering the FA map from time 1 to time 2 and from time 2 to time 1, computing the halfway point between the two images, creating a mean FA map of the two time points and registering each time point to that halfway template. Note that this subject-specific template is different from the study template, which may be formed in TBSS by selecting "the most representative subject" or by using the standard space FA template.

### Motion correction, eddy current correction, and tensor calculation

Motion correction, eddy current correction, and tensor calculation was performed either using the FSL pipeline or the DTIPrep v1.1.1 pipeline. The FSL pipeline consists of the programs eddycorrect to perform motion and eddy current correction by using an affine registration to the first image, and dtifit to estimate the diffusion tensor and calculate the scalar DTI statistics. The DTIPrep pipeline performs motion correction and eddy-correction using DTIPrep [Liu et al., 2010] and estimation of the diffusion tensor using the guided tensor restored anatomical connectivity tractography (GTRACT) software [Cheng et al., 2006].

### Noise removal

Diffusion images are sensitive to a number of artifacts besides eddy currents, including vibrational artifacts, venetian blind artifacts, and noise that selectively affects a subset of gradients. We investigated the effect of using DTIPrep to perform automatic quality control on DTI scans, removing noise [Liu et al., 2010]. The DTIPrep pipeline performs quality control checks on image information (against the acquisition protocol), the *b*-values, slice-wise checks for intensity-related artifacts, interlace-wise checking for Venetian blind artifacts, baseline averaging of non-diffusion weighted images, eddy current, and head motion artifacts correction, and gradient-wise checking. Gradients that fail checks for intensity-related artifacts are removed. To perform this check, DTIPrep computes the correlations between successive slices in each gradient and examines them at each slice position across all the diffusion gradient volumes. Intensity artifacts are detected as large deviations from the mean of all the gradients [Liu et al., 2010].

### Smoothing

We evaluated the effect of smoothing the DTI image using a 1 voxel median filter, applied by GTRACT before computation of the DTI scalar indices, as recommended by Magnotta et al. [2012] to improve intersite reliability. A median filter replaces each entry with the median of neighboring entries, increasing the signal to noise ratio while being robust to outliers and preserving structure [Tukey, 1977]. An alternative to median smoothing with GTRACT is to use fslmaths with a 1 voxel box kernel and the–fmedian flag, which we also evaluated.

## Statistical Analysis

There is no standard method for reporting reliability in DTI studies. To ensure comparability of metrics with other studies, we assessed reliability by calculating voxel-wise statistics on the final TBSS skeleton using Pearson's correlation coefficient ($r$) and the Intraclass Correlation (ICC) coefficient. Regarding the latter, there are many different statistical ICC models that are appropriate for specific experimental designs [McGraw and Wong, 1996]. In our study, we used a two-way mixed effects model with absolute agreement. This model assumes that the dependent variable (DTI scalar statistic) is assessed by the same "rater" (the scanner at the first occasion and the scanner at the second occasion), and that this "rater effect" should be

TABLE I. Overview of the tested TBSS-based data processing pipelines

| | QA check/No ise removal | Eddy/Motion Correction | Smoothing | Tensor Estimation/ Scalar Estimation | Subject-specific registration template | TBSS |
|---|---|---|---|---|---|---|
| Intersession, individual template (*FSL Subject-Specific template + smoothing*) | – | 1 | 5 | 2 | Yes | Yes |
| Intersession, individual template (*DTIPrep + noise removal + smoothing*) | 3 | 3 | 4 | 4 | Yes | Yes |
| Intersession, individual template (*DTIPrep + noise removal*) | 3 | 3 | – | 4 | Yes | Yes |
| Intersession, individual template (*DTIPrep without noise removal*) | – | 3 | – | 4 | Yes | Yes |
| Intersession, individual template (*FSL Subject-Specific template*) | – | 1 | – | 2 | Yes | Yes |
| Intersession (*FSL default*) | – | 1 | – | 2 | – | Yes |
| Intraession | – | 1 | – | 2 | – | Yes |

Numbers represent the following software-specific processing tools: (1) eddycorrect (FSL), (2) dtifit (FSL), (3) DTIprep, (4) GTRACT, and (5) fslmaths (FSL).

treated as a fixed effect. Differences between individuals are treated as random effects. Furthermore, the model is sensitive to systematic differences in the data between the two scan occasions [i.e., it models absolute agreement vs. consistency). The formula for this model may be found by McGraw and Wong, 1996]. Statistical analysis of reliability was conducted using R version 2.15.1. The different processing pipelines tested are listed in Table I. We determined whether there was an overall significant difference in reliability between voxels in a skeleton by using a Mann–Whitney *U*-test, a nonparametric test of the null hypothesis that the distributions of the reliability coefficients in the skeleton are identical for two different processing pipelines. The Mann–Whitney *U*-test is appropriate in this case because the distributions of reliability measurements are not normal (they are positively skewed), and because the resulting skeletons do not have a voxel-to-voxel correspondence (specifically, different preprocessing pipelines generate skeletons of different sizes). The lack of voxel-to-voxel correspondence is also the reason why differences between processing pipelines are not reported on a voxel-by-voxel basis and displayed accordingly.

## RESULTS

### Summary of Reliability of Methods Tested

Figure 1 shows the output FA maps produced by three different processing streams (FSL default, DTIPrep noise removal + smoothing, FSL default + smoothing) for an example subject. We can see that the default FSL eddy correction introduces smoothing as a result of affine registration, but this level of smoothing is clearly lower compared with the pipelines that include smoothing with a median filter. Table II shows the reliability obtained for each of the tested pipelines for commonly used scalar DTI statistics (FA, MD, RD, and AD). We note that reliability coefficients follow the same rank order for all statistics, with “FSL subject-specific template + smoothing” showing the highest and “Intersession” showing the lowest reliability. Table III shows the differences in reliability between the tested processing streams. All differences are significant at Bonferroni-corrected *P*-value of $P < 0.05$ except for the DTIPrep pipelines with and without noise removal.

Focusing on FA, we find that the mean ICC calculated for intrasession reliability using the standard TBSS pipeline (two measurements on the same day) is almost identical to the mean ICC calculated for across-session reliability (two measurements one week apart) using the standard TBSS processing pipeline. In other words, the cost of reliability that is attributed solely to scanner state and subject position within the scanner is relatively small. We obtain a much larger increase in reliability by introducing adaptations to the basic intersession FSL TBSS pipeline (e.g., using the subject-specific template approach, see Fig. 2).

The DTIPrep pipeline offers an alternative method for performing eddy correction, motion correction, and calculating the tensor and scalar statistics. The baseline reliability of DTIPrep as compared to the equivalent FSL pipeline was very slightly lower (ICC difference = −0.006, $P < 0.05$, corrected). However, using DTIPrep to remove noisy gradients from the data had the effect of reducing reliability beyond that. For the first occasion of imaging, the mean number of gradients retained was 30.56 (SD = 1.15, min = 28, max = 32). On the second occasion, the mean number of gradients retained was 30.06 (SD = 1.69,

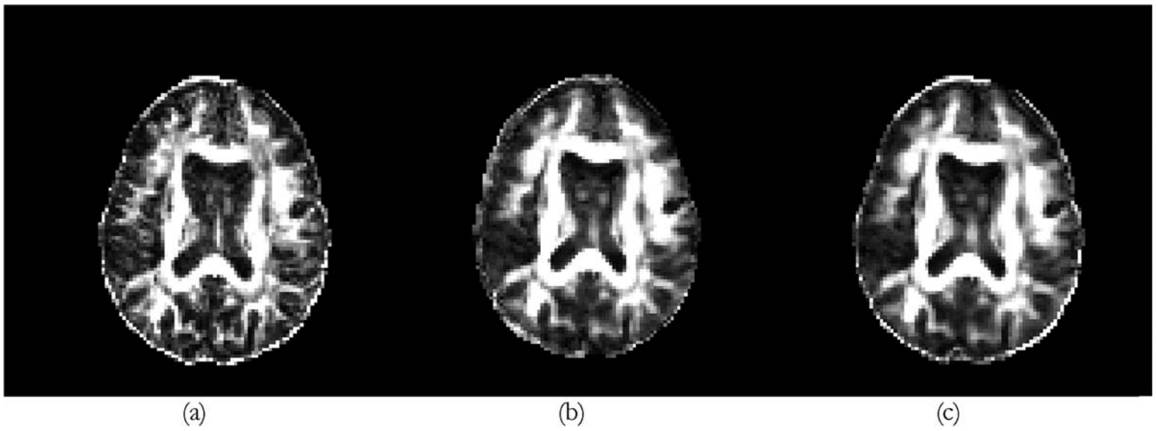

**Figure 1.**
FA maps produced by different processing streams for a single subject. (**a**) FSL default, (**b**) DTI-Prep + noise removal + smoothing, (**c**) FSL default + median filtering smoothing with fslmaths.

min = 27, max = 32). The largest benefits to the FSL and DTIPrep pipelines came from median smoothing of the motion and eddy corrected image (Table II). The distribution of the ICC reliability coefficients across voxels in the TBSS skeleton is displayed in Figure 3. The blue and green lines in Figure 3 clearly reflect the improvement in reliability obtained by applying median filter smoothing.

## Smoothing Improves Reliability by Improving Alignment

To test whether the increase in reliability seen for pipelines that include some higher degree of smoothing is because smoothing removes noise or because smoothing adjusts for minute differences in registration, we reasoned that if we created a synthetic data set with perfect within-

**TABLE II. Reliability (ICC and Pearson correlation) for evaluated methods**

| | No. of voxels in skeleton | % skeletal voxels significant ICC[a] | $\sigma$[b] | Relative power[c] | FA | | MD | | RD | | AD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ICC | *R* | ICC | *r* | ICC | *r* | ICC | *r* |
| Intersession, individual template (*FSL Subject-Specific template + smoothing*) | 102,385 | 0.947 | 0.047 | 1.358 | 0.788 | 0.798 | 0.760 | 0.772 | 0.786 | 0.797 | 0.775 | 0.786 |
| Intersession, individual template (*DTIPrep + noise removal + smoothing*) | 104,612 | 0.942 | 0.048 | 1.325 | 0.776 | 0.787 | 0.751 | 0.764 | 0.778 | 0.790 | 0.763 | 0.775 |
| Intersession, Individual template (*DTIPrep + noise removal*) | 138,133 | 0.891 | 0.064 | .996 | 0.702 | 0.714 | 0.608 | 0.626 | 0.658 | 0.675 | 0.639 | 0.654 |
| Intersession, individual template (*DTIPrep without noise removal*) | 137,340 | 0.895 | 0.063 | 1.002 | 0.705 | 0.717 | 0.610 | 0.628 | 0.660 | 0.677 | 0.642 | 0.657 |
| Intersession, individual template (*FSL Subject-Specific template*) | 130,736 | 0.896 | 0.062 | 1.023 | 0.711 | 0.723 | 0.624 | 0.640 | 0.668 | 0.684 | 0.655 | 0.669 |
| Intersession (*FSL default*) | 132,911 | 0.847 | 0.063 | 1.000 | 0.650 | 0.663 | 0.524 | 0.543 | 0.577 | 0.595 | 0.565 | 0.581 |
| Intrasession | 132,911 | 0.867 | 0.063 | 1.010 | 0.661 | 0.676 | 0.535 | 0.555 | 0.590 | 0.608 | 0.572 | 0.590 |

[a] Significant ICC at $P < 0.05$.
[b] Standard deviation is computed at each voxel as the square root of the average of the squared standard deviations of the FA values for subjects at each occasion, averaged across all voxels in the skeleton.
[c] Relative power compared to FSL default (Inter-Session) reliability assuming $\Delta$ is kept constant ($\sigma$ Inter-session/$\sigma$ of alternative pipeline).

**TABLE III. Difference in reliability in FA (cells contain the mean difference in reliability between the column and row)**

| | Inter-Session, Individual template (*FSL + smoothing*) | Inter-Session, Individual template (*DTIPrep + noise removal + smoothing*) | Inter-Session, Individual template (*DTIPrep + noise removal*) | Inter-Session, Individual template (*DTIPrep without noise removal*) | Inter-Session, Individual template (*FSL Subject-Specific Template*) | Inter-Session (*FSL Default*) | Intra-Session |
|---|---|---|---|---|---|---|---|
| Intersession, individual template (*FSL + smoothing*) | – | −0.011 | −0.083 | −0.080 | −0.075 | −0.133 | −0.123 |
| Intersession, individual template (*DTIPrep + noise removal + smoothing*) | 0.011 | – | −0.072 | −0.069 | −0.064 | −0.123 | −0.112 |
| Intersession, individual template (*DTIPrep + noise removal*) | 0.083 | 0.072 | – | *0.003* | 0.008 | −0.050 | −0.040 |
| Intersession, individual template (*DTIPrep without noise removal*) | 0.080 | 0.069 | *−0.003* | – | 0.005 | −0.053 | −0.043 |
| Intersession, individual template (*FSL Subject-Specific Template*) | 0.075 | 0.064 | −0.008 | −0.005 | – | −0.058 | −0.048 |
| Intersession *(FSL Default)* | 0.133 | 0.123 | 0.050 | 0.053 | 0.058 | – | 0.010 |
| Intrasession | 0.123 | 0.112 | 0.040 | 0.043 | 0.048 | −0.010 | – |

All differences are significant at Bonferroni corrected p < .05 except for those values in italics. Similar tables provided for MD, RD, AD (supplemental tables).

subject registration but additional noise, smoothing should result in a major improvement in quality if it operated largely on noise rather than registration. We used the eddy-corrected, unsmoothed data from GTRACT as the reference input set and added Gaussian noise to decrease the SNR by 20%, creating a synthetic data set of reduced quality. Because the only difference in the synthetic data set was the addition of noise, the halfway registration matrices were the identity matrix. The mean ICC without smoothing was 0.32, and the mean ICC with smoothing was 0.33.

The implementation of longitudinal DTI described by Engvig et al. [2011] included smoothing of the halfway-registered maps with a 2 mm FWHM Gaussian kernel to correct for small alignment differences; we implemented this approach (see Inline Supporting Information Table I and Supporting Information Fig. 1) but results for FA reliability were slightly worse than median filtering using fslmaths, and comparable for other scalar DTI parameters.

Smoothing has systematic effects upon the size of the skeleton obtained, the reliability of the voxels within the skeleton, the mean FA values. Mean FA values obtained using median smoothing are lower than those obtained without. Table IV shows descriptive statistics for the scalar DTI indices (FA, MD, AD, and RD) across all individuals within the skeleton. For DTIPrep + noise removal, the mean FA within the skeleton was 0.42 (SD = 0.15). With smoothing, the mean FA dropped to 0.37 (SD = 0.12). The lower FA values result in a smaller number of voxels in the skeleton that meet the default TBSS threshold. Table II shows that smoothing methods result in an over 20% reduction in the skeleton size. However, an additional 10% of the skeleton voxels have a significant ($P < 0.05$) ICCs.

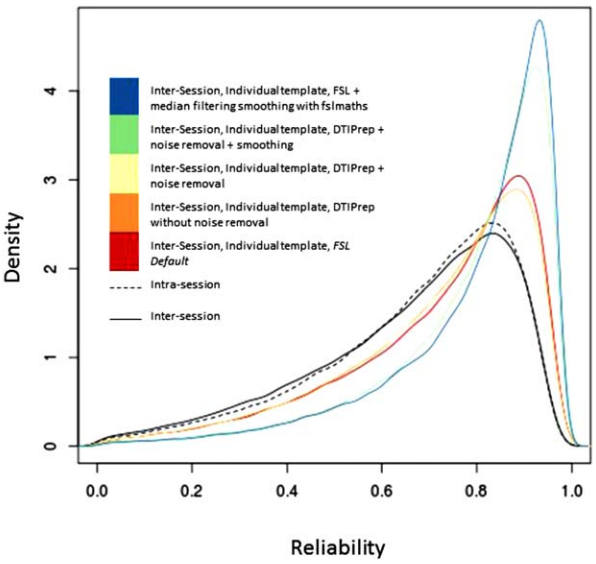

**Figure 2.**

Density plot of reliability (ICC) for each method tested. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

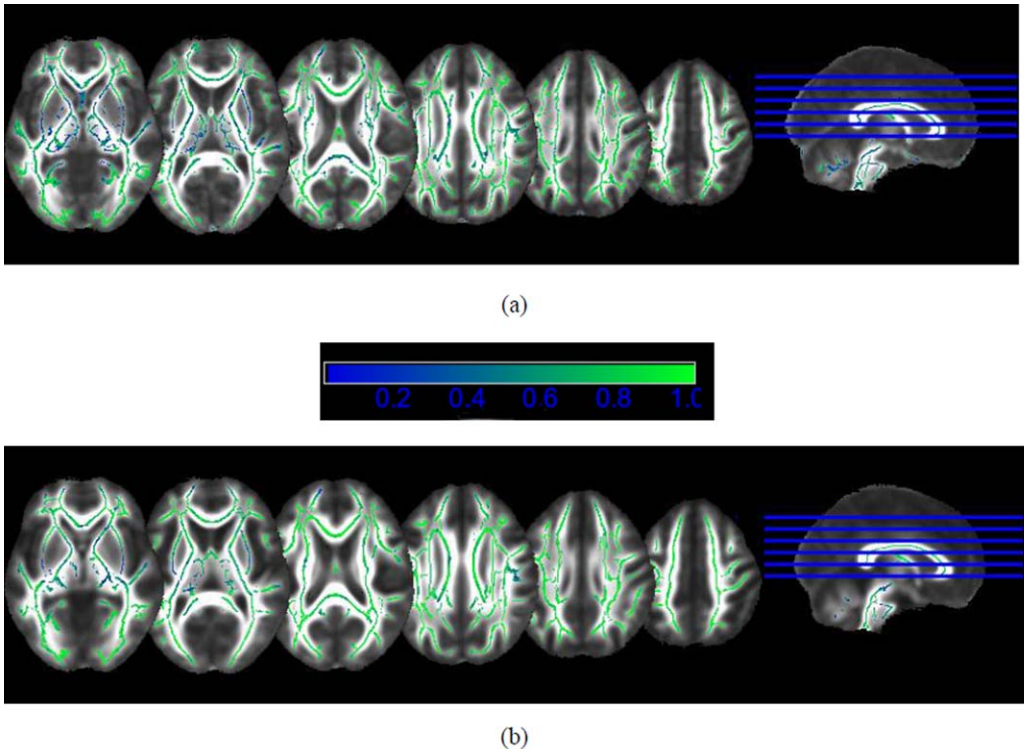

**Figure 3.**
ICC skeletons showing reliability for (**a**) Intersession, individual template (FSL subject-specific template) and (**b**) intersession, individual template (FSL + median filtering smoothing with fslmaths). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

### Higher FA Thresholds for TBSS Skeleton Do Not Significantly Improve Reliability

We note from Table II that the skeleton sizes for methods with higher reliability are smaller, raising the possibility that reliability could be improved by using a higher FA threshold in skeleton formation (tbss_4_prestats). To test this possibility, we recomputed the skeleton for the Intersession Reliability, thresholding at a mean FA value of 0.265 to obtain a skeleton size of 102,229 voxels, just slightly smaller than that of the skeleton for the FSL Subject-Specific template + smoothing pipeline. FA reliability within this restricted skeleton increased a little, but not very much (from 0.650 to 0.657). Similarly, we recomputed the skeleton for the FSL Subject-

**TABLE IV. Descriptive statistics for DTI scalar values within skeleton for evaluated methods**

| | FA | | MD | | RD | | AD | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | M | SD | M | SD |
| Intersession, individual template *(FSL default + smoothing)* | 0.365 | 0.122 | 5.66 | 1.45 | 7.76 | 1.29 | 11.9 | 2.45 |
| Intersession, individual template *(DTIPrep + noise removal + smoothing)* | 0.368 | 0.123 | 6.13 | 1.29 | 7.77 | 1.28 | 11.0 | 2.30 |
| Intersession, individual template *(DTIPrep + noise removal)* | 0.422 | 0.148 | 5.83 | 1.83 | 7.72 | 1.76 | 11.5 | 2.82 |
| Intersession, individual template *(DTIPrep without noise removal)* | 0.422 | 0.148 | 5.82 | 1.79 | 7.71 | 1.73 | 11.5 | 2.79 |
| Intersession, individual template *(FSL Default)* | 0.416 | 0.146 | 5.81 | 1.75 | 7.65 | 1.69 | 11.3 | 2.72 |
| Intersession | 0.419 | 0.146 | 5.77 | 1.69 | 7.62 | 1.63 | 11.3 | 2.71 |
| Intrasession | 0.419 | 0.146 | 5.77 | 1.69 | 7.62 | 1.63 | 11.3 | 2.71 |

MD, RD, and AD are in units of $10^{-4}$ mm$^2$/s.

specific template pipeline, thresholding at a mean FA value of 0.261, to obtain a skeleton size of 102,230 voxels. The mean FA reliability within this more strictly thresholded skeleton was 0.717 (vs. 0.711). This clarifies the point that the gains in reliability that we observe cannot be achieved simply by restricting the skeleton to a subset of voxels with higher FA values.

### Improvements in Reliability Come from Improved Skeleton Alignment

A main source of variability in the TBSS pipeline is alignment of the skeleton to equivalent voxels within a subject. We assess this by backpropagating the skeleton to the individual subject FA, nonlinearly aligned to standard space (using tbss_deproject). We can then quantify, for each subject, what percentage of skeleton voxels at each timepoint overlap as a fraction of the voxels in the skeleton. FA skeleton voxels that do not map to the same voxel in the same subject after registration to standard space are less likely to be measuring the same underlying FA value than those that do overlap, leading to lower reliability. For intersession FSL default, the mean percentage of overlapping voxels was 55.9% (53.4–58.4%, SD = 1.5%). For intrasession, the mean percentage of overlapping voxels was 56.0% (53.1–58.7%, SD = 1.7%). The use of a subject-specific template reduced variability caused by different within-subject nonlinear warps. For the subject-specific template pipeline, the mean percentage of voxels that overlap was 65.6% (61.7–68.8%, SD = 2.0%). Smoothing additionally improves voxel alignment. For the FSL subject-specific template + smoothing, the mean percentage of voxels that overlap was 73.9% (71.9–76.0%, SD = 1.2%). This increase in skeleton overlap means that the steps of subject-specific template creation and smoothing reduce variability in the TBSS projection of individual aligned FA values to the skeleton.

### Reducing Variability Increases Statistical Power

Finally, we consider the effect of improving reliability upon our statistical power to detect an effect within the TBSS skeleton. Table II shows the pooled standard deviation ($\sigma$), averaged across skeleton voxels, for each method. We note that reducing $\sigma$ means that it is possible to detect a smaller longitudinal change $\Delta$ without affecting the rates of Type 1 or Type 2 error. Alternatively, the statistical power to detect the same magnitude of change increases with a reduction in $\sigma$ (Table II column Relative Power).

## DISCUSSION

The main goal of the present study was to quantify within- and between session scan-rescan reliability of DTI-derived parameters that are commonly used to describe the characteristics of neural white matter in the context of neural plasticity research. To do so, we applied the TBSS approach implemented in FSL, which is a frequently used analysis tool. Moreover, we were interested in testing how different DTI preprocessing choices affect reliability indices, with the goal of identifying a processing pipeline that maximizes scan-rescan consistency and is therefore adequate in the context of longitudinal studies. The two main preprocessing steps that we found to improve TBSS reliability were (a) the use of a common individual template and (b) smoothing DTI data using a 1-voxel median filter. Both of these steps systematically increased the number of voxels in the skeleton that overlapped within an individual subject, increasing reliability by improving within-subject skeleton alignment. Scripts that we used to implement these steps are available at https://github.com/fliem/longitudinalTBSS.

Regarding the image registration strategy, we implemented the longitudinal registration method proposed by Smith et al. [2002] for use in SIENA as an adaptation of the standard TBSS processing pipeline [Engvig et al., 2011]. This halfway-space registration strategy ensures that the images that will be compared (to detect change) undergo equivalent processing steps and that there is no registration bias towards one of the time points of data acquisition. Importantly, after linear registration of the time point 1 and time point 2 images to the common individual template, the nonlinear registration of the individual FA images to the study template (standard space) is then identical for all time points for each individual. This approach reduces the magnitude of individual differences caused by nonlinear registration. Furthermore, it avoids the problem that separate nonlinear registration might obscure real change by trying to “compensate” for it. Taken together this approach improves reliability and, in turn, improves the ability of TBSS to detect individual change. An improved registration procedure [e.g., de Groot et al., 2013] will likely address this aspect of variability with TBSS.

With respect to smoothing, we acknowledge that one of the advantages of TBSS is that it does not require smoothing. Instead, the mean FA skeleton represents the centers of all FA tracts common to the group of subjects included in a given analysis. However, we found that the use of a 1-voxel median filter greatly improved the reliability of TBSS, which is in line with a recent study on multi-site reliability by Magnotta et al. [2012] who also showed that median filter smoothing significantly improved the reliability of the DTI scalar measures. Filtering of the raw DWI images should improve estimation of the diffusion tensors, which improves estimation of scalar DTI parameters. Median filtering is theoretically preferable to application of a Gaussian smoothing kernel [as in Engvig et al., 2011] because median filtering better preserves structure in the de-noising process [Welk et al., 2007]. Although more sophisticated filtering strategies have been applied to DWI data [Descoteaux et al., 2008; Wiest-Daesslé et al., 2008], median filtering is a reasonable choice because it effectively preserves structure and removes noise.

Alternatively, it is possible that it improves FA reliability simply by smoothing the data and thus reducing the effect of small registration differences. To test this hypothesis, we simulated the introduction of noise while maintaining perfect registration. The difference in the mean ICC was small, lending support to the idea that median filtering fosters FA reliability in TBSS by reducing the effect of small registration differences. We compared the results of smoothing using a small Gaussian smoothing kernel to that of the use of the median filter on the raw data and found the latter to have slightly better reliability for FA values, and to be comparable on the other scalar parameters (see Inline Supporting Information Table I and Supporting Information Fig. 1).

In line with Magnotta et al. [2012], we also observed that the median filter considerably decreases mean FA values. Application of the median filter increases RD and decreases AD, thus decreasing FA. A corollary of this is that the skeleton obtained from methods with greater levels of smoothing is smaller in size (if computed using the same threshold value for FA, as we have done here). Therefore, it is very important to take this "side effect" into account when comparing FA values for data on which median filtering has been applied to previously reported data processed with default TBSS routines.

We did not observe a benefit to removal of bad gradients, also in line with findings by Magnotta et al. [2012]. This may be because the overall quality of our acquisition was good. However, removal of gradients had a small effect of reducing reliability. One may choose to avoid removal of noisy gradients in analysis of longitudinal data, or at least use the number of gradients removed at each time point as a covariate in analysis.

Given the importance of within-subject skeleton alignment to reliability, it is not surprising that we observe the greatest reliability in straight, wide regions of the skeleton with high FA values. This finding is specific to the TBSS method of registration and skeleton projection. Because our measurements of reliability are based on measures from healthy elderly, and FA values tend to decrease with age, reliability in a younger adult sample might be higher for all methods tested. Similarly, any changes to the scanner protocol, which change the quality of acquisition, and consequently, within-subject alignment, will also systematically change overall measures of reliability. However, unless within-subject registration and TBSS skeleton formation is perfect, we would expect the same general pattern of findings (that improvements to within-subject registration improve reliability) to be valid. We did not find that noise removal (implemented by DTIPrep) improved reliability. This specific finding is dependent upon scanner and sequence and possibly subject motion and may not by true for all acquisitions.

We observe that reducing extraneous sources of variability improves the statistical power to detect longitudinal change. However, the methods that reduce variability result in smaller skeletons, reducing the number of voxels in which one can potentially detect change. This is a characteristic of the TBSS method. The skeleton is limited to regions of relatively high FA that are common across individuals, and at best we observed that only 73.9% of skeleton voxels overlapped within a single individual. This suggests that even within the skeleton, there are a large percentage of voxels that do not correspond to each other, making it difficult to find reliable change in these areas. An ROI-based method might be more suitable for testing hypotheses about longitudinal change in FA in very specific regions.

Given the different parameters previously used to quantify the reliability of neuroimaging data (e.g., ICC: Intraclass Correlation Coefficient, r: Pearsons correlation coefficient, CV: coefficient of variation, Lin concordance correlation, standard deviation of the measurement error) and given the differences and ambiguity regarding the appropriate statistical model in the specific case of the ICC, we decided to provide some more details on this matter. We based our decision about the ICC model used in this article on the methodological paper written by McGraw and Wong, which is in large part based on the seminal work of Shrout and Fleiss [1979] and McGraw and Wong, [1996]. We chose the case 3 model ICC(A,1) [McGraw and Wong, 1996, page 35, Table IV] for the following reasons. This model is a two-way mixed effect model with the row variable (subjects) treated as random effect. The model further assumes that the dependent variable (DTI scalar statistic) is assessed by the same rater (the scanner at the first occasion, and the scanner at the second occasion). Thus, the column variable (measurement time points, "rater effect") is treated as fixed effect. Because measurements were all acquired on the same scanner, we used the ICC model for single (instead of average) measurements. Lastly, we decided to use an ICC model that measures correlation using an absolute agreement definition, since this model considers systematic errors in the data set (e.g., time point 1 data differ in absolute values from time point 2 data) and views them as disagreement.

The main focus of our study was to examine reliability of TBSS given different preprocessing alternatives. Because this approach is fairly well standardized and keeps the subjective choices during analysis to a minimum, our results and recommendations are relevant to a wide group of neuroimaging scientists. As reliability also is a pivotal prerequisite for cross-sectional studies, the scope of our results is not restricted to longitudinal study designs. We did not explore reliability of tractography methods that may be more appropriate for studying longitudinal change of specific tracts, in contrast to a whole-brain approach. However, the relatively small sample size in our study limits our power to place confidence intervals on reliability of specific tracts, and thus to statistically compare the reliability of one method to another. Further work with large samples is required to examine the effect of processing alternatives on the reliability of tractography.

In summary, our conclusions are as follows. First, the small details of a processing pipeline have a profound effect upon both the reliability of the data and on the ability to compare findings across studies. Second, the use of an individual template with linear registration greatly increases reliability of longitudinal TBSS studies. Finally, smoothing implemented before calculation of the DTI tensor helps to compensate for small alignment errors and further improves reliability of TBSS results.

## ACKNOWLEDEMENTS

The sponsors had no involvement in this research or in the preparation of this manuscript. Furthermore, the authors would like to thank Michel Philipp for his helpful advice on statistical issues.

## REFERENCES

Andersson J, Jenkinson M, Smith S (2007a): Non-linear optimisation. FMRIB technical report TR07JA1 from www.fmrib.ox.ac.uk/analysis/techrep.

Andersson J, Jenkinson M, Smith S (2007b): Non-linear registration, aka Spatial normalization. FMRIB technical report TR07JA2 from www.fmrib.ox.ac.uk/analysis/techrep.

Assaf Y, Pasternak O (2008): Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. J Mol Neurosci 34:51–61. doi:10.1007/s12031-007-0029-0.

Barrick TR, Charlton RA, Clark CA, Markus HS (2010): White matter structural decline in normal ageing: A prospective longitudinal study using tract-based spatial statistics. NeuroImage 51:13–13. doi:10.1016/j.neuroimage.2010.02.033.

Burzynska AZ, Preuschhof C, Bäckman L, Nyberg L, Li SC, Lindenberger U, Heekeren HR (2010): Age-related differences in white matter microstructure: Region-specific patterns of diffusivity. NeuroImage 49:2104–2112. doi:10.1016/j.neuroimage.2009.09.041.

Charlton RA, Landau S, Schiavone F, Barrick TR, Clark CA, Markus HS, Morris RG (2008): A structural equation modeling investigation of age-related variance in executive function and DTI measured white matter damage. Neurobiol Aging 29: 1547–1555. doi:10.1016/j.neurobiolaging.2007.03.017.

Cheng P, Magnotta VA, Wu D, Nopoulos P, Moser DJ, Paulsen J, Jorge R, Andreasen NC (2006): Evaluation of the GTRACT diffusion tensor tractography algorithm: A validation and reliability study. NeuroImage 31:11–11. doi:10.1016/j.neuroimage.2006.01.028.

de Groot M, Vernooij MW, Klein S, Ikram MA, Vos FM, Smith SM, Niessen WJ, Andersson JL (2013): Improving alignment in Tract-based spatial statistics: Evaluation and optimization of image registration. NeuroImage 76:400–411. doi:10.1016/j.neuroimage.2013.03.015.

Descoteaux M, Wiest-Daesslé N, Prima S, Barillot C, Deriche R. (2008): Impact of Rician adapted non-local means filtering on HARDI. Medical image computing and computer-assisted intervention: MICCAI . . . International Conference on Medical Image Computing and Computer-Assisted Intervention. 11: 122–130.

Engvig A, Fjell AM, Westlye LT, Moberget T, Sundseth Ø, Larsen VA, Walhovd KB. (2011): Memory training impacts short-term changes in aging white matter: A longitudinal diffusion tensor imaging study. Hum Brain Mapp 33:2390–2406. doi:10.1002/hbm.21370.

Fjell AM, Walhovd KB (2010): Structural brain changes in aging: courses, causes and cognitive consequences. Rev Neurosci 21: 187–221.

Fox RJ, Sakaie K, Lee J-C, Debbins JP, Liu Y, Arnold DL, Melhem ER, Smith CH, Philips MD, Lowe M, Fisher E (2012): A validation study of multicenter diffusion tensor imaging: reliability of fractional anisotropy and diffusivity values. AJNR Am J Neuroradiol 33:695–700. doi:10.3174/ajnr.A2844.

Hirsiger S, Liem F, Bezzola L, Madhyastha TM, Martin M, Mérillat S, Jäncke L (2013): Test-retest reliability of cortical thickness in an older sample. 19th Annual Meeting of the Organization for Human Brain Mapping, Seattle, WA.

Jones DK, Cercignani M. (2010): Twenty-five pitfalls in the analysis of diffusion MRI data. NMR Biomed 23:803–820. doi:10.1002/nbm.1543.

Jones DK, Chitnis XA, Job D, Khong PL (2007): What happens when nine different groups analyze the same DT-MRI data set using voxel-based methods. Presented at the Proceedings of the International Society for Magnetic Resonance in Imaging, 15th annual meeting, Seattle, WA.

Jones DK, Knösche TR, Turner R (2013): White matter integrity, fiber count, and other fallacies: the do "s and dont's of diffusion MRI. NeuroImage 73:239–254. doi:10.1016/j.neuroimage.2012.06.081.

Le Bihan D, Mangin JF, Poupon C, Clark CA, Pappata S, Molko N, Chabriat H (2001): Diffusion tensor imaging: Concepts and applications. J Magn Reson Imaging 13:534–546.

Liu Z, Wang Y, Gerig G, Gouttard S, Tao R, Fletcher T, Styner M (2010): Quality control of diffusion weighted images. In: Liu BJ, Boonn WW, editors. Presented at the SPIE Medical Imaging, SPIE, San Diego, CA, Vol. 7628, pp. 76280J–76280J–9. doi:10.1117/12.844748.

Madden DJ, Bennett IJ, Burzynska A, Potter GG, Chen N-K, Song AW (2012): Diffusion tensor imaging of cerebral white matter integrity in cognitive aging. Biochim Biophys Acta 1822:386–400. doi:10.1016/j.bbadis.2011.08.003.

Magnotta VA, Matsui JT, Liu D, Johnson HJ, Long JD, Bolster BD Jr, Mueller BA, Lim K, Mori S, Helmer KG, Turner JA, Reading S, Lowe MJ, Aylward E, Flashman LA, Bonett G, Paulsen JS (2012): MultiCenter reliability of diffusion tensor imaging. Brain Connectivity 2:345–355. doi:10.1089/brain.2012.0112.

McGraw KO, Wong SP (1996): Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30.

Morey RA, Selgrade ES, Wagner HR II, Huettel SA, Wang L, McCarthy G (2010): Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Hum Brain Mapp NA–NA. doi:10.1002/hbm.20973.

Mori S, Zhang J (2006): Principles of diffusion tensor imaging and its applications to basic neuroscience research. Neuron 51:527–539. doi:10.1016/j.neuron.2006.08.012.

Rice J (2006): Power. In Mathematical Statistics and Data Analysis. Cengage Learning, Stamford, Connecticut. p 433–435.

Schaie KW (2005): What can we learn from longitudinal studies of adult development? Res Hum Develop 2:133–158. doi:10.1207/s15427617rhd0203_4.

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

Smith SM, Zhang YY, Jenkinson M, Chen J, Matthews PM, Federico A, De Stefano N (2002): Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17:479–489. doi:10.1006/nimg.2002.1040.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2003): Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23(Suppl 1):S208–S219. doi:10.1016/j.neuroimage.2004.07.051.

Smith SM, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols TE, Mackay CE, Watkins KE, Ciccarelli O, Cader MZ, Matthews PM, Behrens TE (2006): Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. NeuroImage 31:19–19. doi:10.1016/j.neuroimage.2006.02.024.

Song S-K, Sun S-W, Ju W-K, Lin S-J, Cross AH, Neufeld AH (2003): Diffusion tensor imaging detects and differentiates axon and myelin degeneration in mouse optic nerve after retinal ischemia. NeuroImage 20:1714–1722.

Song S-K, Yoshino J, Le TQ, Lin S-J, Sun S-W, Cross AH, Armstrong RC. (2005): Demyelination increases radial diffusivity in corpus callosum of mouse brain. NeuroImage 26:132–140. doi:10.1016/j.neuroimage.2005.01.028.

Sullivan EV, Pfefferbaum A (2006): Diffusion tensor imaging and aging. Neurosci Biobehav Rev 30:749–761. doi:10.1016/j.neubiorev.2006.06.002.

Teipel SJ, Meindl T, Wagner M, Stieltjes B, Reuter S, Hauenstein KH, Filippi M, Ernemann U, Reiser MF, Hampel H (2010): Longitudinal changes in fiber tract integrity in healthy aging and mild cognitive impairment: A DTI follow-up study. J Alzheimer's Dis 22:507–522. doi:10.3233/JAD-2010-100234.

Tukey JW (1977): Exploratory Data Analysis. Addison-Wesley, Reading, Mass.

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. doi:10.1016/j.neuroimage.2010.03.046.

Wang JY, Abdi H, Bakhadirov K, Diaz-Arrastia R, Devous MD (2012): A comprehensive reliability assessment of quantitative diffusion tensor tractography. NeuroImage 60:1127–1138. doi:10.1016/j.neuroimage.2011.12.062.

Weickert W, Schnorr B, Burgeth F (2007): Median and related local filters for tensor-valued images. Signal Process 87:18–18. doi:10.1016/j.sigpro.2005.12.013.

Wheeler-Kingshott CAM, Cercignani M (2009): About "axial" and "radial" diffusivities. Magn Reson Med 61:1255–1260. doi:10.1002/mrm.21965.

Wiest-Daesslé N, Prima S, Coupé P, Morrissey SP, Barillot C (2008): Rician noise removal by non-local means filtering for low signal-to-noise ratio MRI: applications to DT-MRI. Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention. 11:171–179.

Yoon B, Shim Y-S, Lee K-S, Shon Y-M, Yang D-W (2008): Region-specific changes of cerebral white matter during normal aging: A diffusion-tensor analysis. Arch Gerontol Geriatr 47:10. doi:10.1016/j.archger.2007.07.004.

Ziegler DA, Piguet O, Salat DH, Prince K, Connally E, Corkin S (2010): Cognition in healthy aging is related to regional white matter integrity, but not cortical thickness. Neurobiol Aging 31:1912–1926. doi:10.1016/j.neurobiolaging.2008.10.015.

Zöllig J, Mérillat S, Eschen A, Röcke C, Martin M, Jäncke L (2011): Plasticity and imaging research in healthy aging: Core ideas and profile of the international normal aging and plasticity imaging center (INAPIC). Gerontology 57:190–192. doi:10.1159/000324307.