HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2017 July 20.

Published in final edited form as:

Neuroimage. 2015 October 01; 119: 406–416. doi:10.1016/j.neuroimage.2015.06.078.

Region of Interest Correction Factors Improve Reliability of Diffusion Imaging Measures Within and Across Scanners and Field Strengths

Vijay K Venkatraman[1], **Christopher E. Gonzalez**[1], **Bennett Landman**[2], **Joshua Goh**[1,3], **David A. Reiter**[1], **Yang An**[1], and **Susan M. Resnick**[1]

[1]Intramural Research Program, National Institute on Aging, NIH, Baltimore, MD, 21224, USA

[2]Institute of Imaging Science and Department of Electrical Engineering, Vanderbilt University, Nashville, TN 37235, USA

[3]Graduate Institute of Brain and Mind Sciences, National Taiwan University College of Medicine, Taipei, Taiwan

Abstract

Diffusion tensor imaging (DTI) measures are commonly used as imaging markers to investigate individual differences in relation to behavioral and health-related characteristics. However, the ability to detect reliable associations in cross-sectional or longitudinal studies is limited by the reliability of the diffusion measures. Several studies have examined reliability of diffusion measures within (i.e. intra-site) and across (i.e. inter-site) scanners with mixed results. Our study compares the test-retest reliability of diffusion measures within and across scanners and field strengths in cognitively normal older adults with a follow-up interval less than 2.25 years. Intra-class correlation (ICC) and coefficient of variation (CoV) of fractional anisotropy (FA) and mean diffusivity (MD) were evaluated in sixteen white matter and twenty-six gray matter bilateral regions. The ICC for intra-site reliability (0.32 to 0.96 for FA and 0.18 to 0.95 for MD in white matter regions; 0.27 to 0.89 for MD and 0.03 to 0.79 for FA in gray matter regions) and inter-site reliability (0.28 to 0.95 for FA in white matter regions, 0.02 to 0.86 for MD in gray matter regions) with longer follow-up intervals were similar to earlier studies using shorter follow-up intervals. The reliability of across field strengths comparisons was lower than intra- and inter-site reliability. Within and across scanner comparisons showed that diffusion measures were more stable in larger white matter regions (> 1500 mm^3). For gray matter regions, the MD measure showed stability in specific regions and was not dependent on region size. Linear correction factor estimated from cross-sectional or longitudinal data improved the reliability across field strengths. Our findings indicate that investigations relating diffusion measures to external variables must consider variable

Corresponding Author: Susan M. Resnick, Ph.D., Laboratory of Behavioral Neuroscience, BRC/NIA/NIH, 251 Bayview Blvd., Baltimore, MD, 21224, resnicks@mail.nih.gov.

8. Conflicts of Interest/Disclosure:
The authors declare no competing financial interests.

reliability across the distinct regions of interest and that correction factors can be used to improve consistency of measurement across field strengths. An important result of this work is that inter-scanner and field strength effects can be partially mitigated with linear correction factors specific to regions of interest. These data-driven linear correction techniques can be applied in cross-sectional or longitudinal studies.

Keywords

white matter; gray matter; DTI; Reliability; Correction Factor; Longitudinal

1. Introduction

Diffusion Tensor Imaging (DTI) is a magnetic-resonance (MR) based imaging approach that provides quantitative measurement of brain microstructure and can indicate abnormalities in white matter (Basser, 1995; Beaulieu, 2002; Marner et al., 2003) and gray matter (Nusbaum, 2002; Pfefferbaum et al., 2010). Two commonly used diffusion metrics, fractional anisotropy (FA) and mean diffusivity (MD), are obtained by fitting the diffusion tensor model to diffusion data (Basser et al., 1994; Mori and Zhang, 2006). DTI is widely used in neuroimaging studies (Le Bihan et al., 2001) and has been applied in studies of brain development (Cascio et al., 2007), aging (Charlton et al., 2010), Alzheimer's disease (Sexton et al., 2011), multiple sclerosis (Harrison et al., 2011), and traumatic brain injury (Farbota et al., 2012).

Previous studies have shown that the diffusion measures of FA and MD are affected by acquisition and analysis approaches, such as b-factor, signal-to-noise, image resolution, within-session averaging, co-registration methods, warping, reslicing, scanner performance, and segmentation approaches (Bisdas et al., 2008; Landman et al., 2007; Pagani et al., 2010; Papinutto et al., 2013). As quantitative MRI becomes more widely used in clinical practice and research studies, it is important to establish reproducibility and reliability of diffusion measures in the context of similar protocols from distinct scanners, as is commonly the case for longer term or multi-site studies. Such parameters are critical for interpreting DTI findings in patient populations, comparing results between studies, and evaluating whether and how data from multi-site cross-sectional or longitudinal studies can be pooled.

Several studies have investigated the reliability of diffusion measures within (i.e. intra-site) and across (i.e., inter-site) scanners, often using intra-class correlation (ICC) and coefficient of variation (CoV) to assess reliability. Findings from studies using 1.5T scanners and a variety of analysis approaches (Bonekamp et al., 2007; Zhu et al., 2011) indicate that intra-site comparisons generally show higher reliability than inter-site comparisons. For a single region analysis of the corpus callosum (Pfefferbaum et al., 2003), intra-site FA and trace were highly reproducible (mean CoVs for FA= 1.9% and trace=2.6%) and inter-site measures showed greater variability (mean CoVs for FA=4.5% and trace=7.5%). For 3T scanners, a study using a region of interest (ROI) approach on diffusion measures from two similar scanners (Vollmar et al., 2010) showed higher intra-site reliability (ICC: 0.90 to 0.99, CoV: 0.8% to 3% for FA) compared to inter-site reliability (ICC: 0.82 to 0.99, CoV: 1.0% to

4.1% for FA) for whole brain white matter and three white matter tracts. Another study using gray and white matter ROIs across five scanners from two manufacturers (Fox et al., 2012) showed that the average CoVs ranged from 6.8% to 9.1% for FA, 2.2% to 4.8% for MD, 4.2% to 6.6% for transverse diffusivity and 3.5% to 5.0% for longitudinal diffusivity measures. These prior reliability studies of diffusion measures have primarily investigated diffusion measures in relatively few ROIs, mostly in the white matter regions, over short follow-up intervals. If data from longitudinal multisite studies are to be combined, it is important to determine estimates of reliability over time. Stability estimates of FA and MD over longer intervals (>1 year) provide lower bound reliabilities, as they include both reliability of measurement and true biological change.

Few studies have explored the effect of MR field strength on reliability of diffusion measures and whether correction factors can allow continuity of measurement across different field strengths. In one study, FA values (Huisman et al., 2006) were statistically higher in 3T compared to 1.5T (percentage change between 4.04% to 11.15%) using twelve participants scanned within two hours, whereas apparent diffusion coefficient (ADC) for white and gray matter was significantly lower at 3T compared to 1.5T (-1.94% to −9.79%). Another study (Alexander et al., 2006) in four subjects using manually traced region of interests in white and gray matter regions showed that reproducibility of DTI measures (FA and MD) measured as noise estimates in 3T is better than 1.5T (reduced by 34% to 52%). These findings show clear differences in DTI measures across field strengths, and multi-field studies should be avoided if possible. However, retrospective aggregation of multi-site studies, institutional availability of MR scanners, and long-term studies can produce datasets of interest from a variety of scanners and field strengths.

Whether data from different field strengths can be combined (given that alternative study designs were not feasible) and how to best accomplish such an analysis remains an open question. Attempts to employ statistical approaches to harmonize DTI measures across scanners and field strengths have been limited. Use of a global scaling factor was shown to reduce the inter-site CoV to the range of intra-site CoV (Vollmar et al., 2010). Another study (Pagani et al., 2010) demonstrated that statistical adjustments for scanner manufacturer, field strength and number of diffusion-weighted directions was sufficient for discrimination of patients from healthy controls in a cross-sectional study. It has also been shown that weighing diffusion metrics based on their within scan variability, as evaluated by wild bootstrap analysis, can reduce the intra- and inter-site variability in phantom and human data (Zhu et al., 2011). However, it is unclear whether such statistical approaches can detect subtle differences such as longitudinal change or regional variations.

In this study, we characterized the reliability of diffusion measures in eighty-four gray and white matter regions in a large dataset of cognitively normal older adults using Philips scanners, highlighting regions of higher and lower reliability. We determined the reliability of FA and MD on the following comparisons to compare against existing studies with shorter follow-up intervals: (a) intra-site scanner differences on 1.5T and 3T scanners over a mean 1.4 years follow-up interval, (b) inter-site differences across two different 3T scanners conducted same day and mean 1.8 years follow-up interval and (c) across field strengths with mean 1.7 years follow-up interval. Most importantly, we evaluate a statistical approach

to improve the reliability across field strengths, using linear correction factors estimated for individual ROIs using cross-sectional and longitudinal datasets.

2. Methods

2.1. Participants

This study uses MRI data from 545 participants and 800 visits (mean age (SD) = 68.5 (12.7) years) from participants enrolled in the Baltimore Longitudinal Study of Aging (BLSA) (Shock et al., 1984). Measures of reliability for the current study were calculated from a sub-sample (mean age = 78 ± 8 years) of this larger dataset. The sub-sample was chosen with follow-up intervals less than 2.25 years for intra-site and inter-site datasets as shown in Table. 1. All participants were cognitively normal at all MRI assessments. Those with a history of stroke were also excluded from the study. Diagnoses of dementia and Alzheimer's disease were determined by the Diagnostic and Statistical Manual (DSM)-III-R and the National Institute of Neurological and Communication Disorders Alzheimer's Disease and Related Disorders Association criteria (McKhann et al., 1984). The local Institutional Review Board approved this study, and all participants gave written informed consent at each visit. Demographic characteristics of the sample are shown in Table. 1.

2.2. Image acquisition

Data were acquired on a single 1.5 Tesla Philips Intera scanner (scanner A) and three different 3 Tesla Philips Achieva scanners (scanner B and C at the Kennedy Krieger Institute, and scanner D at the National Institute on Aging). Scanners B and C used the same platform and protocol, and the data were combined after verifying comparability of the diffusion measures as explained in Section 2.4. Our study compared the reliability of diffusion measures within 1.5T (Dataset 1; Table. 1), within 3T (Dataset 2; Table. 1), and across field strengths (Dataset 3; Table. 1), with all follow-up intervals less than 2.25 years. Additionally, we calculated reliability measures for fifteen participants scanned on the same day (Dataset 4; Table. 1) and thirteen participants scanned with follow-up intervals less than 2.25 years (Dataset 5; Table. 1) within two different 3T scanners to study inter-site reliability and effect of follow-up interval.

Each participant underwent a magnetization-prepared rapid gradient-recalled echo (MPRAGE) scan and two DTI scans at each visit. The scanning protocol is presented in detail in Table. 2. Each DTI acquisition had two b0 images, which were averaged in k-space to reduce bias (Henkelman, 1985). On all three systems, two separate DTI acquisitions each with NSA = 1 were obtained and then combined offline (as explained in Section 2.3) for an effective NSA = 2 to improve signal-to-noise ratio.

2.3. Image processing

DTI processing followed standard practice for tensor fitting and quality assessment (Lauzon et al., 2013). Briefly, individual diffusion weighted volumes were affine co-registered to a minimally weighted (b0) target to compensate for eddy current effects and physiological motion. The gradient tables were corrected for the identified rotational component using finite strain (Alexander et al., 2001). At each voxel, the RESTORE algorithm (Chang et al.,

2005) was used to fit a tensor while simultaneously excluding outliers with noise locally estimated (Landman et al., 2009). To combine two DTI sessions with different (unknown) intensity normalization constants, each diffusion-weighted image was normalized by its own reference image prior to tensor fitting.

To segment gray matter regions, multi-atlas registration using Non-local STAPLE (Asman and Landman, 2013, 2012) was performed using the Advanced Normalization Toolkit (ANTs) with the SyN image similarity criteria (Avants et al., 2011) and 35 manually labeled atlases from NeuroMorphometrics with the BrainCOLOR protocol (Klein et al., 2010). To segment the white matter, the Eve White Matter atlas (Lim et al., 2013) was combined with corresponding labels from multi-atlas segmentation, and an FA mapped MRI. Briefly, the Eve atlas' T1-weighted image and FA maps were non-rigidly registered to the T1-weighted and FA of a target subject using ANTS multi-modal registration (Avants et al., 2011). The white matter labels within Eve are intersected with the regions defined as white matter from the multi-atlas segmentation. The resulting labels were iteratively grown to fill the remaining white matter space from the multi-atlas. Abbreviations of white and gray matter regions are provided in Table. 3. The white and gray matter ROI labels obtained from the T1-weighted image for each visit were affine registered to the FA image and used to extract region-specific FA and MD measures.

Note that EPI distortion correction was not specifically applied. Reverse gradient DTI acquisitions were not acquired, and so correction such as FSL's TOPUP (Andersson et al., 2003; Smith et al., 2004) was not possible. B0 maps were not acquired for all subjects, so that consistent field map correction such as FSL's FUGUE (Andersson et al., 2003; Smith et al., 2004) was not possible. The multi-channel image-based non-rigid image registration was used to compensate for geometric difference between the atlases and the subjects. However, application of non-rigid image registration to correct distortions between minimally weighted diffusion images ("b0", distorted) and the T1 structural images (non-distorted) (Wu et al., 2008) was visually judged to introduce distortion artifacts within the central white matter regions (which are of high interest for DTI) at a cost of slightly improved peripheral alignment (which are of lower interest for DTI). Therefore, affine registration was applied between the b0 and T1 image spaces to correct for eddy current and first order EPI distortion; findings in the prefrontal region bordering sinus and temporal lobe, surrounding the acoustic meatuses should be interpreted in the context of known uncorrected distortion, which is commonly seen in DTI.

For quality control (QC) of our imaging data, we reviewed several summary statistics computed for each combined DTI scan by our processing pipeline (Lauzon et al., 2013). Specifically, we plotted the distributions of FA Bias from SIMEX, a wild-bootstrap experimental variance of FA, and median FA in each scanner and performed QC. After reviewing these distributions, twenty-seven scans were excluded because of either excessive motion or images that had globally high FA bias to yield the MR dataset used in this study.

2.4. Reliability Statistics

We used the ICC, CoV and Pearson correlation (r) as our main statistics to assess reliability intra-site, inter-site and across field strengths. For ICC, total variance was partitioned into

within- and between-subject variance by fitting the data to a linear mixed-effects model for each diffusion measure and region, and ICC was calculated as the proportion of between-subject variance to total variance. The dependent variable for these models was FA or MD for each person on two separate visits, fixed effects included baseline age and follow-up interval between two visits, and subject-specific intercept was included as a random effect. CoV was calculated by dividing the within-subject standard deviation (Bland and Altman, 1996) by the mean across all visits. We calculated Pearson and Spearman's correlation coefficients (not shown in results) across visits and the mean and standard deviation of the diffusion measures at each visit for each region. All statistical analyses were performed using R version 3.1.0 and mixed effects models were conducted using lme4 version 1.1–6. Intra-site reliability measures were calculated for 16 participants with two visits on scanner A for 1.5T (Dataset 1; Table. 1) and for 99 participants with two visits on scanner D for 3T with follow-up intervals 1.2 and 1.6 years on average, respectively (Dataset 2; Table. 1). Inter-site across field strength reliability were calculated on 29 participants with a scan on scanner A followed by a scan on scanners B or C separated by 1.7 years on average (Dataset 3; Table. 1). Inter-site same field strength scanner reliability measures were calculated on 13 participants with a scan on scanner C followed by a scan on scanner D with mean 1.8 years follow-up interval (Dataset 5; Table. 1). Finally, inter-site same field strength scanner reliability was also assessed for 15 participants with same day acquisitions on scanners C and D (Dataset 4; Table. 1). Furthermore, we compared the differences in FA in white matter regions and MD in gray matter regions between scanners B and C using mixed effects models, controlling for baseline age and time interval after the baseline scan. In a sensitivity analysis excluding Scanner B data (N=20 from Dataset 3, Table.1), we calculated the across-field strength reliability estimates and re-calculated the reliability after applying the linear correction factor.

2.5. Linear Correction Factor

To determine if we could improve across field strength reliability (i.e. ICC measure) for FA in white matter regions and MD in gray matter regions, we used cross-sectional and longitudinal training datasets to estimate the scanner differences for each diffusion measure and each ROI. We used the across field strength reliability sample (Dataset 3; Table. 1) as the test dataset. The training datasets contained no visits from the test dataset.

For the longitudinal training dataset, we used a linear mixed effects model to estimate the difference between scanner A (1.5T) scans and scanner B/C (3T) scans ($\beta 2$). We included additional fixed effects for baseline age (age at first visit; $\beta 1 Age_i$), follow-up interval (time-variant; $\beta 3 Interval_{ij}$), and scanner (coded as a categorical variable with three levels: A, B/C, and D; $\beta 2 Scanner_i$) and random effects with subject-specific intercepts (b_{0i}).

$$y_{ij} = \beta 0 + \beta 1\, Age_i + \beta 2\, Scanner_i + \beta 3\, Interval_{ij} + b_{0i} + \varepsilon_{ij}$$

The longitudinal training dataset comprised of 742 visits (544 participants; 140 participants with at least two visits) after excluding the visits from the test dataset (58 visits; Dataset 3; Table. 1) from the complete dataset (n=545, 800 visits; Table. 1). Of these 742 visits, 48

visits were from scanner A, 35 visits were from scanner B/C, and 659 visits were from scanner D. The estimated difference due to the fixed effect of scanner was used as a linear correction term, which was calculated separately in bilateral white matter regions for FA and bilateral gray matter regions for MD. The ROI-specific correction factors were applied to the original FA or MD values for the testing dataset (Dataset 3; Table. 1) on scanner A (1.5T) scans. The ICC and CoV were recalculated using adjusted diffusion measures for the testing dataset. To verify there were no unexpected relationships between certain scanners and demographic information that influenced our linear estimates, we selected the largest gray and white matter regions and checked for interactions between scanner and baseline age, sex, and follow-up interval (time-variant).

We further investigated if across field strength reliability could be improved using linear correction terms estimated from a cross-sectional training dataset. We fit the cross-sectional training data to a general linear model that had a scanner covariate ($\beta 2 Scanner_i$) with the same coding as in the longitudinal data analysis and also controlled for age ($\beta 1 Age_i$).

$$y_i = \beta 0 + \beta 1\ Age_i + \beta 2\ Scanner_i + \varepsilon_i$$

The cross-sectional training dataset comprised of 27 scans from scanner A (1.5T), 27 from scanners B/C (3T), and 490 from scanner D (3T) by selecting a single scan from each subject in a way that maximized the number of scans from scanner A and B/C scanners. Again, we added the estimated difference between scanners A and B/C for each diffusion measure and each ROI to the original scanner A (1.5T) scans and recalculated ICC and CoV in the test dataset (Dataset 3; Table. 1).

To determine if our sample size was large enough to detect a reliable correction factor, we examined the variability of the corrected ICC as a function of sample size. We restricted this analysis to FA in white matter for the 54 subjects in the cross-sectional training dataset; half of whom had scans on scanner A (1.5T) and the other half on the scanners B/C (3T). We bootstrapped this training data for each region with varying sample sizes (n= 10, 15, 20, 25, 30, 35, 40, 45, 50, 54) for 500 iterations each, with each iteration estimating a scanner difference. We then applied these correction factors to scanner A (1.5T) scans from the reliability analysis (the "test" data) and calculated an ICC measure for each iteration. This resulted in an ICC distribution for each sample size, for each white matter region.

2.6. Effect of ROI definition and motion parameters on reliability measures

Because each diffusion measure was calculated within the native space for each visit (visit-specific ROI - ROI segmentation generated at each visit), we computed Dice-Similarity Coefficients (DSCs) for white and gray matter regions to determine the overlap in ROI segmentations across visits for participants used in the reliability comparisons. For each participant, the T1-weighted image of the second visit was rigid-body registered to the first visit and then the transformation was applied to the white and gray matter ROI labels to compute the DSC. Average DSCs were calculated for intra-site (Dataset 1 and Dataset 2; Table. 1) and across field strength (Dataset 3; Table. 1) reliability comparisons. We also

determined how using subject-specific ROIs instead of visit-specific ROIs (ROI segmentation generated at each visit was applied) affected the reliability measures. For subject-specific ROIs, each subject's last visit ROI segmentation was applied to all earlier visits for that subject. We compared the ICC calculated using subject- and visit-specific white and gray matter ROIs for the intra-site reliability (Dataset 2; Table. 1) comparison. We also determined how motion parameters affected reliability measures. The ICC measure was recalculated for Dataset 2 (99 subjects) adjusting for summary measures of three dimensional motion parameters (mean of root sum squared for translation and peak angular rotation) for each visit.

3. Results

Overall, the reliability measures from this study (ICC, CoV, r) provide complementary information. To evaluate the reliability measures we used cut-offs based on previous literature, where acceptable reliability is defined as ICC > 0.6; CoV < 10 and r > 0.6. The raw FA and MD values of bilateral regions for all scanners (1.5T and 3T) are presented in Tables S1a and S1b. In comparison of scanners B and C, there were no significant differences in FA in white matter regions and only a few gray matter regions (PHG, Amy, SPL, Palli and Ent) showed minor differences for MD between scanners using the mixed effects model. Furthermore, our sensitivity analysis excluding scanner B data showed no differences in trends of across field strength reliability compared to intra-site 1.5T and 3T reliability. There were also no differences in trends of the overall improvement in reliability across brain regions with correction factor for FA and MD. The average reliability measures of left and right hemisphere for each region are presented, as there were no significant hemispheric differences in reliability across white or gray matter regions.

3.1. Intra-site reliability for 1.5T and 3T

The intra-site reliability measures in 1.5T (Dataset 1; Table. 1) and 3T (Dataset 2; Table. 1) for FA and MD were acceptable in most of the white matter regions (Figures. 1 and S2a, Table. S2a). Certain regions showed poor reliability for FA and MD in at least two reliability indices. These included FX, UF, Scc and SS in 1.5T and CH and UF in 3T. Across all white matter regions, mean reliability of FA at 3T (ICC=0.76; CoV= 5.04; r=0.79) was slightly better than at 1.5T (ICC=0.72; CoV= 4.90; r=0.73), but the difference was not significant for any of the reliability indices. Mean reliability across white matter regions for MD in 1.5T (ICC= 0.71; CoV=3.56; r= 0.75) was slightly higher than in 3T (ICC= 0.66; CoV=5.77; r= 0.72;), however the difference was only significant for CoV (p= 0.0041; Paired T-test, one-tailed). As expected, in the gray matter regions, FA was not reliable within 1.5T or 3T for most regions (Figures. 1 and S3b, Table. S2b). The mean reliability of MD across gray matter regions showed that 1.5T (ICC = 0.61; CoV=5.80; r= 0.72) was equivalent to 3T (ICC = 0.61; CoV=6.23; r= 0.74) with respect to reliability measures. However, reliability was higher in 1.5T than 3T for the largest gray matter regions (MFG, SFG, PrG, SPL, PoG).

3.2. Inter-site reliability at 3T

Next we compared inter-site reliability (Dataset 5; Table. 1) with intra-site reliability (Dataset 2; Table. 1) at 3T, over follow-up intervals less than 2.25 years. Mean inter-site

reliability measures across white matter regions for FA (inter-site: ICC= 0.71; CoV= 5.98; r= 0.74) were significantly lower ($p<0.05$; Paired T-test, one-tailed) compared to intra-site (Figure. 2a) for all reliability indices. Similarly for MD in gray matter regions (Figure. 2b), the reliability for intra-site was higher than inter-site (mean values: ICC= 0.59; CoV= 7.19; r= 0.64), where the reliability measures (CoV and r) were significantly different ($p<0.01$; Paired T-test, one-tailed). For data acquired the same day at two different sites (Dataset 4; Table. 1), the mean inter-site reliability measures (FA in white matter regions, MD in in gray matter regions) were mean ICC= (0.78, 0.61); mean CoV= (4.57, 6.03); mean r= (0.83, 0.69). Next, we compared the inter-site reliability of scans acquired the same day (Dataset 4; Table. 1) with inter-site reliability with follow-up interval (Dataset 5; Table. 1). ICC reliability measures were not statistically different, but CoV was significantly different ($p<0.05$; Paired T-test, two-tailed) for white and gray matter regions, and r ($p<0.05$; Paired T-test, two-tailed) differed for white matter regions in the comparison of inter-site reliability measures for same day and follow-up acquisitions.

3.3. Reliability across field strengths

The reliability across field strengths (Dataset 3; Table. 1) for FA (mean values of ICC =0.35; CoV = 8.59; r =0.63) and MD (mean values of ICC= 0.31; CoV= 8.20; r= 0.50) were significantly lower ($p<0.01$; Paired T-test, one-tailed) than reliability measures of within scanner strengths at 1.5T and 3T for white matter regions (Figures. 1 and S3a, Table. S2a). The reliability measures for gray matter regions (Figures. 1 and S3b, Table. S2b) were also significantly lower ($p<0.01$; Paired T-test, one-tailed) across field strengths for FA (mean values of ICC= 0.14; CoV= 19.06; r= 0.28) and MD (mean values of ICC= 0.32; CoV= 11.92; r= 0.64). Some regions had very low ICC or were even computed as zero, this indicates the within-subject variation was much larger than between-subject variation in those regions.

3.4. Linear correction term

Although across field strength reliability was significantly lower than within field strength, we examined whether applying a linear correction factor could improve reliability estimates to the level seen within field strength. We used a larger pool of data to estimate the difference between scanner A and scanner B/C to maximize the accuracy of the correction factor. Linear correction terms were estimated and applied for left and right hemispheres of each region and each diffusion measure separately. The average ICC across left and right hemisphere was used to evaluate the effect of the correction factors. The examination of the utility of these statistical correction factors was limited to FA in white matter regions and MD in gray matter regions. Figures 3a and S4a, Table S3a show that use of the linear correction factors improved ICC in every white matter region compared to the uncorrected ICC values after applying the linear estimate from cross-sectional (ICC change ranged from −0.01 to 0.56) and longitudinal (ICC change ranged from 0.03 to 0.58) data. Figures 3b and S4b, Table S3b indicate ICC in most gray matter regions improved with the correction term estimated from cross-sectional data (ICC change ranged from −0.06 to 0.64) or longitudinal data (ICC change ranged from −0.05 to 0.70). Overall the improvement of ICC in white matter ($p<0.019$; Paired T-test, one-tailed) and gray matter ($p<0.001$; Paired T-test, one-tailed) regions was statistically higher when estimated using longitudinal compared to cross-

sectional data. Our sensitivity analysis to determine the sample size needed to reliably estimate correction factor showed a decrease in variance with increasing sample size for each region. The decrease in variance seemed to plateau around N=40 for the majority of regions. This suggests that our training dataset of n=54 was sufficient for estimation of regional correction factors. Figure S5 shows the plot of variance by sample size for left SCR.

3.5. Effect of ROI definition and motion parameters on reliability measures

Within field strength comparisons showed that almost all gray and white matter regions had high DSC (>0.8), although DSC showed some regional variability (Figure. S1). Some white matter regions, such as FX and UF, and gray matter regions, such as MFC, showed poor overlap in ROI segmentation across visits. In the white matter regions, the DSCs in smaller ROIs were much lower than larger ROIs, and DSC was correlated with size of the region ($r = 0.71$ for 1.5T; $r = 0.73$ for 3T). The DSCs ($r = 0.46$ for 1.5T; $r = 0.60$ for 3T) and size of the region ($r = 0.50$ for 1.5T; $r = 0.59$ for 3T) were correlated with reliability measure (ICC) for FA. The variability of DSC was higher in 3T compared to 1.5T in most of white and gray matter regions. In the gray matter regions, DSC was significantly greater in 1.5T compared to 3T ($p<0.001$; Paired T-test, one-tailed), and there was no relation to size of the region. However, the relationship of DSC and reliability measure (ICC) after controlling for size of the region was not significant across white matter regions for FA ($r = -0.08$ for 1.5T; $r = 0.42$ for 3T) and gray matter regions for MD ($r = -0.28$ for 1.5T; $r = 0.36$ for 3T). The across field strength DSCs were much lower, with many gray and white matter regions ranging from 0.6–0.8 and some regions lower than 0.6.

Differences of reliability measures between visit-specific and subject-specific ROIs were statistically significant in white and gray matter regions using Dataset 2 (Table. S4). FA showed slight improvement in reliability due to subject-specific ROI across most white matter regions (mean values of change in ICC, CoV and r = 0.06, −0.74, 0.06). In the gray matter regions, the MD showed slight improvement due to subject-specific ROI (mean values of change in ICC, CoV and r = 0.11, −0.66, 0.07).

In comparison of reliability measures estimated with and without accounting for motion in white and gray matter regions using Dataset 2, FA and MD showed some improvement in reliability after accounting for motion parameters in most white matter regions (mean values of change in ICC for FA and MD = 0.007, 0.008). In the gray matter regions, the MD also showed slight improvement after accounting for motion parameters (mean values of change in ICC = 0.011). These differences in ICC were statistically significant for white (p-value for FA and MD = 0.005, 0.009) and gray (p-value for MD = 0.004) matter regions using a one-tailed paired t-test.

4. Discussion

This study investigated a large dataset acquired on a variety of Philips MR scanners to characterize regional variability in reliability measures within and across field strength, over a follow-up interval of <2.25 years. The results highlight the importance of considering the reliability of diffusion measures and their regional variability when designing a study relating these measures to clinical, behavioral or other external variables (Section 4.1). The

proposed data-driven approach to estimate regional correction factors improved the comparability across scanners for a large number of white and gray matter regions and could be applied in multisite studies for which single scanner/single protocol studies are not possible (Section 4.2). Future studies are needed to explore the reliability of other diffusion measures and to expand the development of correction approaches that take regional variability into account (Section 4.3).

4.1 Intra- and Inter-site Reproducibility

The intra-site reliability analyses replicate findings in previous studies with shorter follow-up intervals and identify regional variation in reliability that further depends on field strength. The results are consistent with an earlier study at 1.5T with shorter follow-up, which showed CoV for FA was less than 9.5% (Bonekamp et al., 2007) in white matter regions. Similar reliabilities for very short and moderate (<2.25 years) term consistency addresses potential concerns with regard to the feasibility of estimating lower-bound reliability in our study design.

The intra-site reliability estimates at 1.5T and 3T for ICC, CoV and r ranged from 0.32 to 0.96, 2.06% to 12.11% and 0.39 to 0.96 for FA in white matter regions. Earlier studies reported a smaller range for ICC (0.90–0.99) when limiting analysis to a small number of white matter regions (Vollmar et al., 2010). MD in white matter regions in our study showed mean CoV= 3.56% at 1.5T and mean CoV = 5.07% at 3T, consistent with mean CoV = 4.5% using a 1.5T scanner reported in an earlier study with small number of regions (Pfefferbaum et al., 2003). In addition to capturing biological change, our slightly lower ICC are likely associated with our fine ROI parcellation and corresponding intrinsic variability with label generation. Previous studies have not explored in detail the intra-site reliability in gray matter regions, looking only at a few gray matter regions (Fox et al., 2012; Papinutto et al., 2013). In our study, ICC, CoV and r ranged from 0.18 to 0.89, 2.64% to 14.59% and 0.30 to 0.93, respectively, for MD in gray matter regions for intra-site reliability at 1.5T and 3T. These estimates were consistent with those reported previously for specific regions and mean reliability measures. The sub-cortical regions such as hippocampus showed higher reliability compared to gyral ROIs even though the DSCs were similar. As expected, the FA in most of the gray matter regions were not reliable, as the fitted tensors in gray matter regions lack a large principal eigenvector guiding local diffusion compared to white matter regions.

Our study also demonstrated higher intra-site than inter-site reliability for 3T in white matter regions (intra-site ICC = 0.76 versus inter-site ICC= 0.71 for FA). These results were similar to previous studies (Vollmar et al., 2010) with shorter follow-up intervals. In gray matter regions, the inter- and intra- site reliability was not significantly different for ICC (intra-site ICC = 0.61 versus inter-site ICC = 0.59 for MD) but was for CoV and r measures, a finding not previously reported. The trends remained the same after excluding regions of poor reliability (ICC<0.4). In addition, FA had higher reliability at 3T for most white matter regions compared to 1.5T and MD had higher reliability at 1.5T for large gray matter regions compared to 3T. Both our intra- and inter-site reliability measures at 1.5T and 3T validate reliability measures from previous studies with shorter follow-up intervals (with ranges of

days to months). Furthermore, we did not see a significant difference in inter-site ICC values for visits scanned across two different 3T scanners on the same day (Dataset 4) compared with visits scanned over a mean follow-up interval of 1.8 years on the same two 3T scanners (Dataset 5).

Our results across a large number of white and gray matter regions suggest that there is regional variability in reliability of diffusion measures, consistent with earlier findings (Marenco et al., 2006). Our study also indicates a slight difference in 1.5T versus 3T regional variability in intra-site reliability measures for white and gray matter regions. Since intra-site reliability is used as the gold standard, it is important to take regional variability and possible differences in reliability due to field strengths into account in diffusion studies.

In across field strength comparisons, reliability estimates for both white and gray matter regions were much lower than intra- and inter- site reliabilities. Earlier studies (Huisman et al., 2006) have shown that FA in white matter regions was significantly higher at 3T compared with 1.5T (4% to 11.15%). Our study also showed significantly higher FA at 3T compared with 1.5T with average increase of 6.2% across regions. In the present study, we show that compared to within field strength reliability, the across field strength measures performed poorly in white (mean ICC= 0.35 for FA) and gray (mean ICC = 0.32 for MD) matter regions.

4.2 Correction Factors to Improve Inter-site Reproducibility

We investigated the potential of linear correction factors to reduce variation due to field strength differences between scanners. This approach takes regional variability into account and is data driven compared to other approaches (Vollmar et al., 2010). Our results suggest that sample sizes as small as n=40 may be sufficient for application of this regional correction factor approach, ideally with half of the sample from one scanner and half from the other. However there is some regional variability in the decrease in variance with increasing sample size, and larger training datasets could improve regions, especially in smaller regions, that showed little to no improvement. Using correction factors at the individual ROI level, we found improvement in reliability estimates with greater improvement for ICC using correction factors estimated from longitudinal over cross-sectional data. We suspect this is because modeling within-subject variability, especially for subjects that have data over time on different scanners, provides improved estimates of scanner contributions to the diffusion measures. While the use of correction factors improves ICC for most white and gray matter regions, for some regions the ICC remains substantially lower than within field strength reliability. Because high ICC depends on large between-subject variability, and because our correction approach uses the same estimate of scanner differences for each subject, regions that have small between-subject variability may experience little benefit. Adding a subject-specific estimate for scanner difference (the random effect deviation) may lower the within-subject variability or raise the between-subject variability and further improve the reliability. Although we have only applied this approach to repeated measures across field strengths, it could also be used to improve intra- and inter-site measurements as well.

4.3 Study Limitations and Lessons Learned

Earlier studies have not explored the effects of ROI definition in relation to reliability measures. Our results show that variation in ROI definition will impact reliability of diffusion measures and should be considered in cross-sectional and longitudinal diffusion studies. In this study, detailed analysis of ROI definition could explain poor reliability in certain regions, such as UF and FX, which showed suboptimal overlap across repeated scans (Figure. S1). Similar findings of low ICC values for structures such as uncinate fasciculus and fornix were reported (Bach et al., 2014) using the TBSS approach mainly due to influence of partial volume and skeleton shape. When comparing subject-specific ROI versus visit-specific ROI definitions, we demonstrated an improvement in intra-site reliability in 3T for MD in gray matter regions (mean ICC change = 0.11) and FA in white matter regions (mean ICC change = 0.06) when using subject-specific ROI. Our findings demonstrate that variability in ROI definitions have important effects on reliability estimates, including those based on longitudinal follow-ups and its regional differences. Our results also show that motion parameters have a significant effect on reliability measures for white and gray matter regions. These results are very interesting as we excluded scans based on very stringent parameters mentioned in Section 2.3, and as Dataset 2 used for this comparison had very minimal motion. These results demonstrate the importance of accounting for motion parameters in diffusion measures and show that the impact varies regionally.

Tract-based spatial statistics (TBSS) is a widely used skeletonized approach for comparing diffusion measures. Recent studies have investigated the reliability of TBSS approach (Bach et al., 2014; Madhyastha et al., 2014) and several studies (Edden and Jones, 2011; Zalesky, 2011) have evaluated the TBSS approach. As future work, in-depth comparison of ROI-based and TBSS approaches are warranted.

A limitation of our study is that we restricted our analysis to reliability measured as ICC, r and CoV on select diffusion measures (i.e., FA and MD). In addition, our scanning protocols had subtle differences across scanners, which do not seem to affect the reliability estimates. We employed a ROI-based approach (Lauzon et al., 2013), similar to earlier studies (Fox et al., 2012), however some specific procedures used for image processing in this study might have some impact on reliability measures, such as the effects of ROI definition and motion parameters. Another limitation to consider is that some of the smaller gray and white matter regions could be affected by partial volume effects, which may vary over time. The potential errors due to registration method used in ROI definition and calculation of DSC were not explored in this study. In this study we have used the RESTORE technique for the tensor fitting given its strong performance (Walker et al., 2011) especially in aging cohorts. For broader application, further comparisons with other tensor fitting approaches such as FSL dtifit would be beneficial. The registration was performed using structural images, and the effects of EPI geometric distortion and susceptibility artifacts on white and gray matter ROI segmentation and registration was low when judged visually and analyses were performed excluding sinuses and auditory canals, above steps were taken to minimize the effects however we did not explicitly explore its effect on reliability separately. Our evaluation of the utility of linear correction factors was tested in relation to ICC only as it is the most

commonly used reliability measure. As shown by previous studies (Alexander et al., 2006), the head coils can contribute to variation in diffusion data collected across scanners. The current study used the product head coil for all the scanners and was not evaluated as a source of variability.

5. Conclusion

Unlike biological measures that naturally fluctuate over time, we expect indices of brain structure to be relatively stable. Combining data across scanners for the same subject adds noise to measuring this structural stability, and here we characterize the effect of combining different scanners or different field strengths in a longitudinal setting. This study highlights the importance of considering the impact of regional variation on reliability and the interpretation of images obtained on different scanners as well as the application of diffusion imaging to clinical trials. The quantitative ROI variability provided in the supplementary material enables specific power calculations for study planning and design, especially in cases where same-scanner studies are not feasible. Within a single manufacturer and well-controlled protocols, substantively “the same” acquisitions are possible. However, even substantively similar scanners from the same manufacture deployed years apart at different sites yield systematic differences in DTI metrics. Note that smaller, more tortuous white mater tracks are less reliable (FX, UF, Scc and SS in 1.5T and CH and UF at 3T) and should be interpreted with great caution. The linear correction factors greatly increase reliability across scanners (Table S3), but do not fully correct the systematic differences between scanners. Statistical significance and effect sizes derived from multi-site studies should be evaluated within the context of scanner variability to assess whether such effects could be due to data acquisition as opposed to subject variation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the BLSA participants and staff for their dedication to these studies and the staff of the Kennedy Krieger Institute and National Institute on Aging MRI facilities for their assistance.

7. Source of Funding: This research was supported by the Intramural Research Program of the NIH and National Institute on Aging and part by Research and Development Contract N01-AG-3-2124.

References

Alexander A, Lee J, Wu Y, Field A. Comparison of diffusion tensor imaging measurements at 3.0 T versus 1.5 T with and without parallel imaging. Neuroimaging Clin North …. 2006; 16:299–309. DOI: 10.1016/j.nic.2006.02.006

Alexander DC, Pierpaoli C, Basser PJ, Gee JC. Spatial transformations of diffusion tensor magnetic resonance images. IEEE Trans Med Imaging. 2001; 20:1131–1139. DOI: 10.1109/42.963816 [PubMed: 11700739]

Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. Neuroimage. 2003; 20:870–888. DOI: 10.1016/S1053-8119(03)00336-7 [PubMed: 14568458]

Asman AJ, Landman BA. Non-local STAPLE: an intensity-driven multi-atlas rater model. Med Image Comput Comput Assist Interv. 2012; 15:426–434. [PubMed: 23286159]

Asman AJ, Landman BA. Non-local statistical label fusion for multi-atlas segmentation. Med Image Anal. 2013; 17:194–208. DOI: 10.1016/j.media.2012.10.002 [PubMed: 23265798]

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011; 54:2033–2044. DOI: 10.1016/j.neuroimage.2010.09.025 [PubMed: 20851191]

Bach M, Laun FB, Leemans A, Tax CMW, Biessels GJ, Stieltjes B, Maier-hein KH. NeuroImage Methodological considerations on tract-based spatial statistics (TBSS). Neuroimage. 2014; 100:358–369. DOI: 10.1016/j.neuroimage.2014.06.021 [PubMed: 24945661]

Basser PJ. Inferring microstructural features and the physiological state of tissues from diffusion-weighted images. NMR Biomed. 1995; 8:333–344. [PubMed: 8739270]

Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophys J. 1994; 66:259–267. DOI: 10.1016/S0006-3495(94)80775-1 [PubMed: 8130344]

Beaulieu C. The basis of anisotropic water diffusion in the nervous system - a technical review. NMR Biomed. 2002; 15:435–455. DOI: 10.1002/nbm.782 [PubMed: 12489094]

Bisdas S, Bohning DE, Besenski N, Nicholas JS, Rumboldt Z. Reproducibility, interrater agreement, and age-related changes of fractional anisotropy measures at 3T in healthy subjects: effect of the applied b-value. AJNR Am J Neuroradiol. 2008; 29:1128–1133. DOI: 10.3174/ajnr.A1044 [PubMed: 18372415]

Bland JM, Altman DG. Statistics notes : Measurement Error. BMJ. 1996; 313:744. [PubMed: 8819450]

Bonekamp D, Nagae LM, Degaonkar M, Matson M, Abdalla WM, Barker PB, Mori S, Horska A. Diffusion tensor imaging in children and adolescents: reproducibility, hemispheric, and age-related differences. Neuroimage. 2007; 34:733–742. DOI: 10.1016/j.neuroimage.2006.09.020 [PubMed: 17092743]

Cascio CJ, Gerig G, Piven J. Diffusion tensor imaging: Application to the study of the developing brain. J Am Acad Child Adolesc Psychiatry. 2007; 46:213–223. DOI: 10.1097/01.chi.0000246064.93200.e8 [PubMed: 17242625]

Chang LC, Jones DK, Pierpaoli C. RESTORE: robust estimation of tensors by outlier rejection. Magn Reson Med. 2005; 53:1088–1095. DOI: 10.1002/mrm.20426 [PubMed: 15844157]

Charlton RA, Barrick TR, Lawes IN, Markus HS, Morris RG. White matter pathways associated with working memory in normal aging. Cortex. 2010; 46:474–489. DOI: 10.1016/j.cortex.2009.07.005 [PubMed: 19666169]

Edden, Ra, Jones, DK. Spatial and orientational heterogeneity in the statistical sensitivity of skeleton-based analyses of diffusion tensor MR imaging data. J Neurosci Methods. 2011; 201:213–9. DOI: 10.1016/j.jneumeth.2011.07.025 [PubMed: 21835201]

Farbota KD, Bendlin BB, Alexander AL, Rowley HA, Dempsey RJ, Johnson SC. Longitudinal diffusion tensor imaging and neuropsychological correlates in traumatic brain injury patients. Front Hum Neurosci. 2012; 6:160.doi: 10.3389/fnhum.2012.00160 [PubMed: 22723773]

Fox RJ, Sakaie K, Lee JC, Debbins JP, Liu Y, Arnold DL, Melhem ER, Smith CH, Philips MD, Lowe M, Fisher E. A validation study of multicenter diffusion tensor imaging: reliability of fractional anisotropy and diffusivity values. AJNR Am J Neuroradiol. 2012; 33:695–700. DOI: 10.3174/ajnr.A2844 [PubMed: 22173748]

Harrison DM, Caffo BS, Shiee N, Farrell JA, Bazin PL, Farrell SK, Ratchford JN, Calabresi PA, Reich DS. Longitudinal changes in diffusion tensor-based quantitative MRI in multiple sclerosis. Neurology. 2011; 76:179–186. DOI: 10.1212/WNL.0b013e318206ca61 [PubMed: 21220722]

Henkelman RM. Measurement of signal intensities in the presence of noise in MR images. Med Phys. 1985; 12:232.doi: 10.1118/1.595711 [PubMed: 4000083]

Huisman TA, Loenneker T, Barta G, Bellemann ME, Hennig J, Fischer JE, Il'yasov KA. Quantitative diffusion tensor MR imaging of the brain: field strength related variance of apparent diffusion coefficient (ADC) and fractional anisotropy (FA) scalars. Eur Radiol. 2006; 16:1651–1658. DOI: 10.1007/s00330-006-0175-8 [PubMed: 16532356]

Klein, A., Canton, TD., Ghosh, SS., Landman, BA., Lee, J., Worth, A. Open labels: online feedback for a public resource of manually labeled brain images. 16th Annu. Meet. Organ. Hum. Brain Mapp; 2010.

Landman BA, Bazin PL, Prince JL. Estimation and application of spatially variable noise fields in diffusion tensor imaging. Magn Reson Imaging. 2009; 27:741–751. DOI: 10.1016/j.mri.2009.01.001 [PubMed: 19250784]

Landman BA, Farrell JA, Jones CK, Smith SA, Prince JL, Mori S. Effects of diffusion weighting schemes on the reproducibility of DTI-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5T. Neuroimage. 2007; 36:1123–1138. DOI: 10.1016/j.neuroimage.2007.02.056 [PubMed: 17532649]

Lauzon CB, Asman AJ, Esparza ML, Burns SS, Fan Q, Gao Y, Anderson AW, Davis N, Cutting LE, Landman BA. Simultaneous analysis and quality assurance for diffusion tensor imaging. PLoS One. 2013; 8:e61737.doi: 10.1371/journal.pone.0061737 [PubMed: 23637895]

Le Bihan D, Mangin JF, Poupon C, Clark CA, Pappata S, Molko N, Chabriat H. Diffusion tensor imaging: concepts and applications. J Magn Reson Imaging. 2001; 13:534–546. [PubMed: 11276097]

Lim IA, Faria AV, Li X, Hsu JT, Airan RD, Mori S, van Zijl PC. Human brain atlas for automated region of interest selection in quantitative susceptibility mapping: application to determine iron content in deep gray matter structures. Neuroimage. 2013; 82:449–469. DOI: 10.1016/j.neuroimage.2013.05.127 [PubMed: 23769915]

Madhyastha T, Mérillat S, Hirsiger S, Bezzola L, Liem F, Grabowski T, Jäncke L. Longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging. Hum Brain Mapp. 2014; 35:4544–4555. DOI: 10.1002/hbm.22493 [PubMed: 24700773]

Marenco S, Rawlings R, Rohde GK, Barnett AS, Honea RA, Pierpaoli C, Weinberger DR. Regional distribution of measurement error in diffusion tensor imaging. Psychiatry Res. 2006; 147:69–78. DOI: 10.1016/j.pscychresns.2006.01.008 [PubMed: 16797169]

Marner L, Nyengaard JR, Tang Y, Pakkenberg B. Marked loss of myelinated nerve fibers in the human brain with age. J Comp Neurol. 2003; 462:144–152. DOI: 10.1002/cne.10714 [PubMed: 12794739]

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984; 34:939–944. [PubMed: 6610841]

Mori S, Zhang J. Principles of diffusion tensor imaging and its applications to basic neuroscience research. Neuron. 2006; 51:527–539. DOI: 10.1016/j.neuron.2006.08.012 [PubMed: 16950152]

Nusbaum AO. Diffusion tensor MR imaging of gray matter in different multiple sclerosis phenotypes. AJNR Am J Neuroradiol. 2002; 23:899–900. [PubMed: 12063212]

Pagani E, Hirsch JG, Pouwels PJ, Horsfield MA, Perego E, Gass A, Roosendaal SD, Barkhof F, Agosta F, Rovaris M, Caputo D, Giorgio A, Palace J, Marino S, De Stefano N, Ropele S, Fazekas F, Filippi M. Intercenter differences in diffusion tensor MRI acquisition. J Magn Reson Imaging. 2010; 31:1458–1468. DOI: 10.1002/jmri.22186 [PubMed: 20512899]

Papinutto ND, Maule F, Jovicich J. Reproducibility and biases in high field brain diffusion MRI: An evaluation of acquisition and analysis variables. Magn Reson Imaging. 2013; 31:827–839. DOI: 10.1016/j.mri.2013.03.004 [PubMed: 23623031]

Pfefferbaum A, Adalsteinsson E, Rohlfing T, Sullivan EV. Diffusion tensor imaging of deep gray matter brain structures: effects of age and iron concentration. Neurobiol Aging. 2010; 31:482–493. DOI: 10.1016/j.neurobiolaging.2008.04.013 [PubMed: 18513834]

Pfefferbaum A, Adalsteinsson E, Sullivan EV. Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. J Magn Reson Imaging. 2003; 18:427–433. DOI: 10.1002/jmri.10377 [PubMed: 14508779]

Sexton CE, Kalu UG, Filippini N, Mackay CE, Ebmeier KP. A meta-analysis of diffusion tensor imaging in mild cognitive impairment and Alzheimer's disease. Neurobiol Aging. 2011; 32:2322e5–18. DOI: 10.1016/j.neurobiolaging.2010.05.019

Shock N, Greulich R, Andres R, Arenberg D, Costa P, Lakatta E, Tobin J. Normal human aging: The Baltimore longitudinal study of aging. 1984 NIH Publ. No. 84–245.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, 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. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 2004; 23:208–219. DOI: 10.1016/j.neuroimage.2004.07.051

Vollmar C, O'Muircheartaigh J, Barker GJ, Symms MR, Thompson P, Kumari V, Duncan JS, Richardson MP, Koepp MJ. Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. Neuroimage. 2010; 51:1384–1394. DOI: 10.1016/j.neuroimage.2010.03.046 [PubMed: 20338248]

Walker L, Chang LC, Koay CG, Sharma N, Cohen L, Verma R, Pierpaoli C. Effects of physiological noise in population analysis of diffusion tensor MRI data. Neuroimage. 2011; 54:1168–1177. DOI: 10.1016/j.neuroimage.2010.08.048 [PubMed: 20804850]

Wu M, Chang LC, Walker L, Lemaitre H, Barnett aS, Marenco S, Pierpaoli C. Comparison of EPI distortion correction methods in diffusion tensor MRI using a novel framework. Med Image Comput Comput Assist Interv. 2008; 11:321–329. DOI: 10.1007/978-3-540-85990-1-39 [PubMed: 18982621]

Zalesky A. Moderating registration misalignment in voxelwise comparisons of DTI data: a performance evaluation of skeleton projection. Magn Reson Imaging. 2011; 29:111–25. DOI: 10.1016/j.mri.2010.06.027 [PubMed: 20933352]

Zhu T, Hu R, Qiu X, Taylor M, Tso Y, Yiannoutsos C, Navia B, Mori S, Ekholm S, Schifitto G, Zhong J. Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study. Neuroimage. 2011; 56:1398–1411. DOI: 10.1016/j.neuroimage.2011.02.010 [PubMed: 21316471]

Highlights

- Assessed reliability for intra and inter-site diffusion measures with long follow-up intervals
- Across field strength reliability improved with region-specific correction factor
- Correction factor improves more for longitudinal data than cross-sectional data
- Characterized regional variability in reliability of diffusion measures

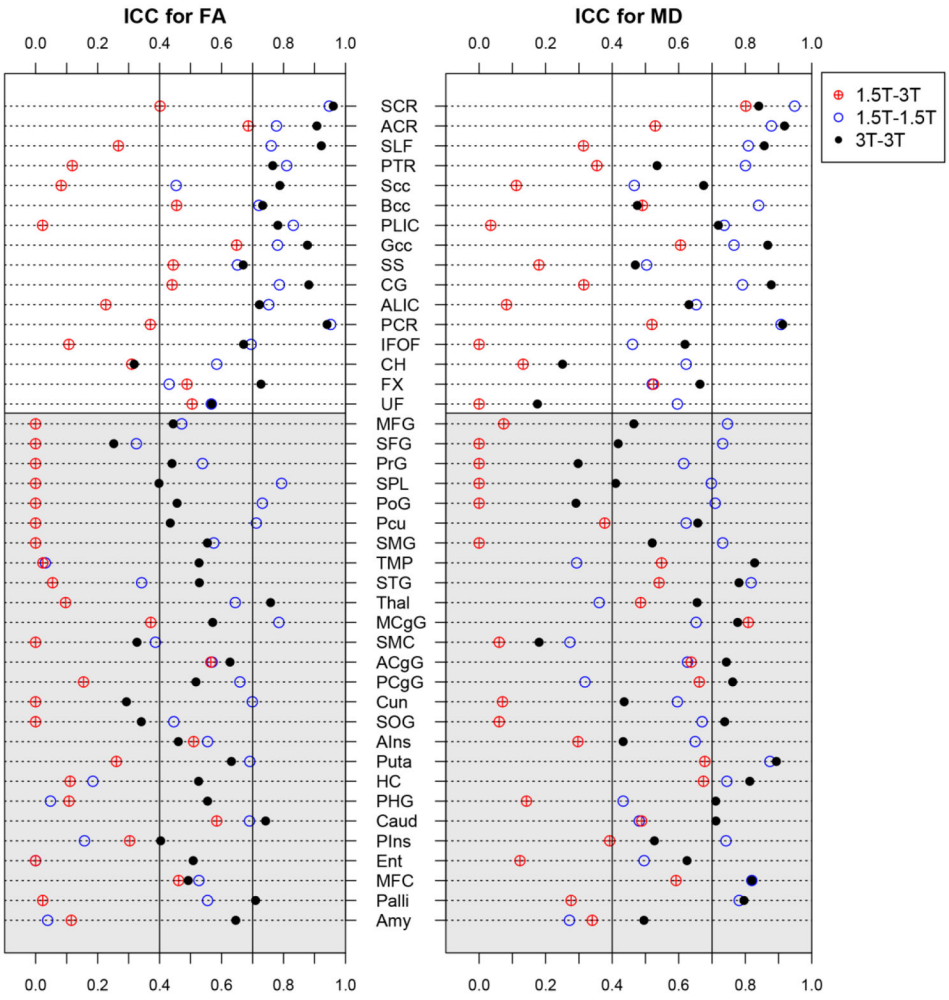


Figure 1.

Intra-site and across field strength reliability measures with follow-up interval for FA and MD in white and gray matter regions, shown as average of left and right hemispheres.

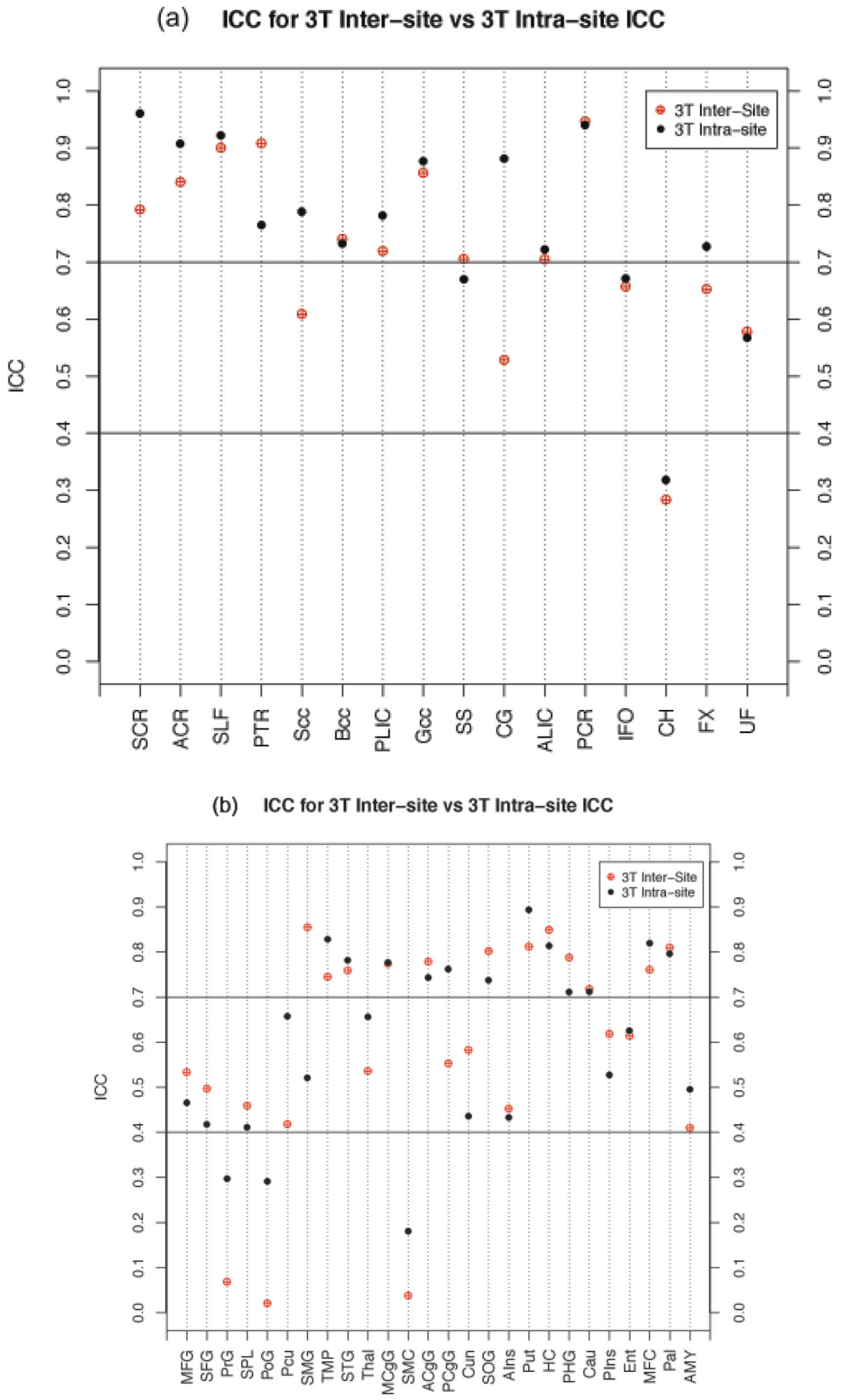
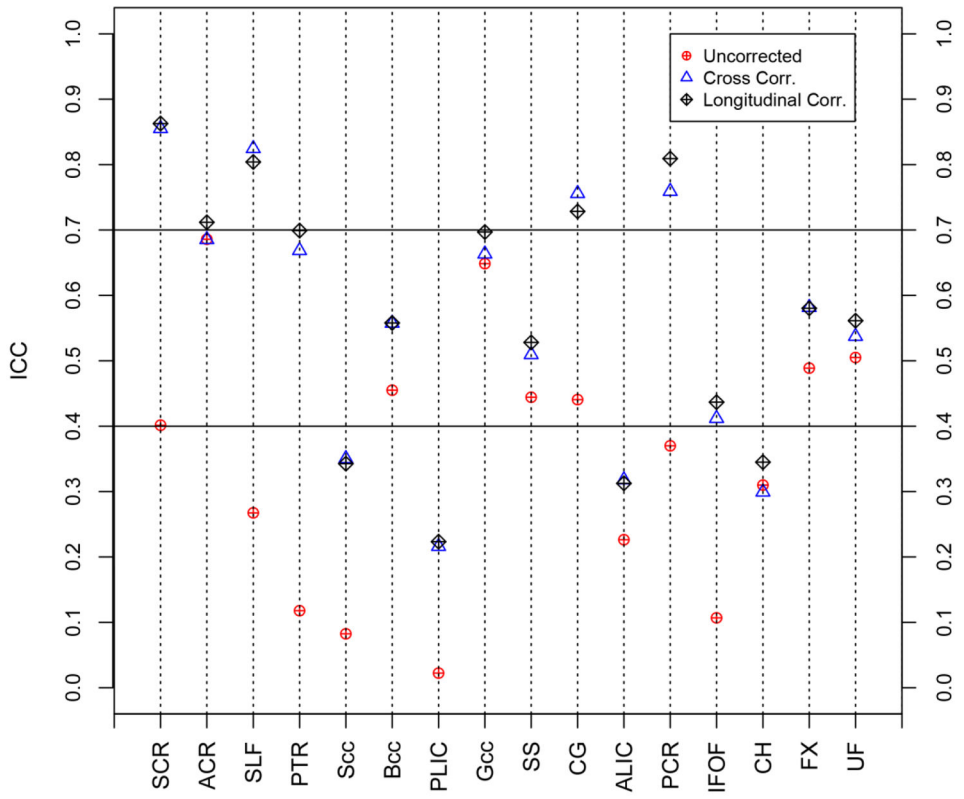


Figure 2.

Inter-site reliability measures with follow-up interval (a) white matter regions, (B) gray matter regions.

(a)



(b)

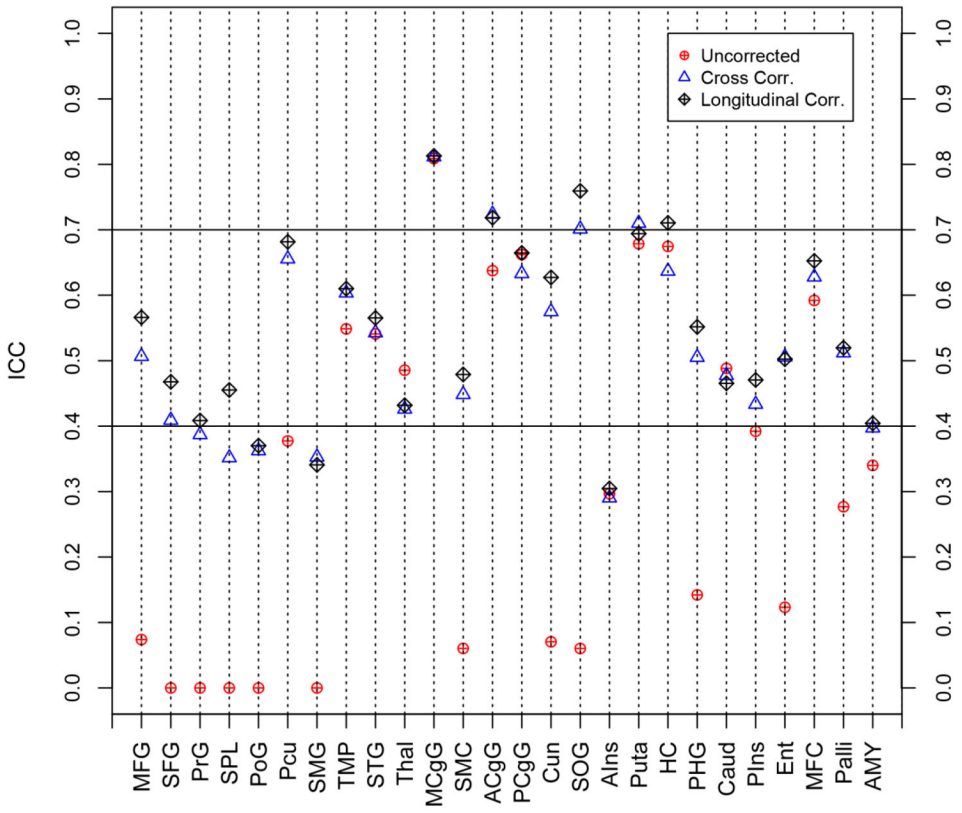


Figure 3. Improvement in across field strength reliability measure due to cross-sectional and longitudinal linear correction factors on (a) white matter regions in FA, (b) gray matter regions in MD

(a) Across field strength ICC for FA measure in white matter regions with and without correction

(b) Across field strength ICC for MD measure in gray matter regions with and without correction

Table 1

Demographics information for the study

	Complete Dataset	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Scanners	A, B/C, D	A-A	D-D	A-B/C	C-D (Same day)	C-D (Follow-up)
Field Strength (Tesla)	1.5, 3T, 3T	1.5T- 1.5T	3T- 3T	1.5T-3T	3T-3T	3T-3T
Subjects	545	16	99	29	15	13
Baseline age (years)	68.5 (12.7)	82.1 (5.7)	77.9 (8.5)	78.9 (6.7)	81.4 (4.6)	81.6 (8.6)
Follow-up interval (years)	0.77 (1.5)	1.2 (0.4)	1.6 (0.5)	1.7 (0.5)	-	1.8 (0.5)
Sex (% Male)	44.8	56.3	43.4	44.8	53.3	53.9
Race (% White)	64.4	87.5	70.7	79.3	80	92.3
Baseline Mini-Mental State Examination	28.6 (1.5)	29.4 (0.93)	28.5 (1.7)	29.2 (0.94)	29 (0.68)	29 (2.1)

Standard deviation shown in parentheses.

Table 2

Scanning protocol for the study

	Scanner A		Scanners B/C		Scanner D	
	MPRAGE	**DTI**	**MPRAGE**	**DTI**	**MPRAGE**	**DTI**
Head Coil	Philips 8-ch	Philips 8-ch	Philips 8-ch	Philips 8-ch	Philips 8-ch	Philips 8-ch
Scan Time (mins: secs)	3:58	3:56	10:52	3:58	10:52	4:20
Number of Gradients	-	30	-	32	-	32
Number of b0 images	-	1	-	1	-	1
Max b-factor (s/mm^2)	-	700	-	700	-	700
No. of Signal Averages (NSA)	-	1	-	1	-	1
Diffusion gradient timing DELTA/delta (msec)	-	39.2/15.1	-	36.3/16	-	36.3/13.5
Slice Thickness (mm)	1.5	2.5	1.2	2.2	1.2	2.2
Number of Slices	124	50	170	65	170	70
Flip angle (deg)	8	90	8	90	8	90
TR/TE (msec)	6.6/3.3	6210/80	6.8/3.1	6801/75	6.5/3.1	7454/75
Field of View (mm)	240*240	240*240	256*240	212*212	256*240	260*260
Acquisition Matrix	208*208	96*96	256*240	96*95	256*240	116*115
Reconstruction Matrix	256*256	256*256	256*256	256*256	256*256	320*320
Reconstructed Voxel Size (mm)	0.94*0.94	0.94*0.94	1.00*1.00	0.83*0.83	1.00*1.00	0.81*0.81

Table 3

Abbreviations and volume of gray and white matter regions

(a) White matter regions using EVE labels

Abbreviation	Region Name	Volume (cm^3)*
SCR	Superior Corona Radiata	7.82 ± 1.11
ACR	Anterior Corona Radiata	7.30 ± 1.15
SLF	Superior Longitudinal Fasciculus	5.70 ± 0.85
PTR	Posterior Thalamic Radiation (Include Optic Radiation)	5.48 ± 0.95
Scc	Splenium of Corpus Callosum	5.10 ± 1.08
Bcc	Body of Corpus Callosum	4.16 ± 0.63
PLIC	Posterior Limb of Internal Capsule	3.05 ± 0.43
Gcc	Genu of Corpus Callosum Sagittal stratum (Include Inferior	2.98 ± 0.61
SS	Longitudinal fasciculus and Inferior fronto occipital fasciculus)	2.85 ± 0.44
CG	Cingulum of Cingulate Gyrus	2.66 ± 0.55
ALIC	Anterior Limb of Internal Capsule	2.53 ± 0.40
PCR	Posterior Corona Radiata	2.38 ± 0.44
IFOF	Inferior Fronto Occipital Fasciculus	2.14 ± 0.53
CH	Cingulum of Hippocampus	1.45 ± 1.42
FX	Fornix (body and column)	0.47 ± 0.15
UF	Uncinate Fasciculus	0.24 ± 0.10

(b) Gray matter regions using Brain Color labels

Abbreviation	Region Name	Volume (cm^3)*
MFG	Middle Frontal Gyrus	20.75 ± 3.05
SFG	Superior Frontal Gyrus	15.54 ± 2.03
PrG	Precentral Gyrus	14.37 ± 1.87
SPL	Superior Parietal Lobule	11.81 ± 1.73
PoG	Postcentral Gyrus	11.66 ± 1.66
PCu	Precuneus	11.36 ± 1.62
SMG	Supramarginal Gyrus	8.50 ± 1.49
TMP	Temporal Pole	8.42 ± 1.22
STG	Superior Temporal Gyrus	6.98 ± 0.98
Thal	Thalamus	6.70 ± 0.78
MCgG	Middle Cingulate Gyrus	5.68 ± 0.87
SMC	Supplementary Motor Cortex	5.60 ± 0.97
ACgG	Anterior Cingulate Gyrus	5.22 ± 1.16
PCgG	Posterior Cingulate Gyrus	5.07 ± 0.81
Cun	Cuneus	4.99 ± 0.80
SOG	Superior Occipital Gyrus	4.48 ± 0.77
AIns	Anterior Insula	4.40 ± 0.59
Puta	Putamen	3.86 ± 0.62
HC	Hippocampus	3.53 ± 0.44

(b) Gray matter regions using Brain Color labels

Abbreviation	Region Name	Volume (cm^3)*
PHG	Parahippocampal gyrus	3.44 ± 0.52
Caud	Caudate	3.01 ± 0.53
PIns	Posterior Insula	2.54 ± 0.36
Ent	Entorhinal Area	2.28 ± 0.37
MFC	Medial Frontal Cortex	1.70 ± 0.37
Palli	Pallidum	1.28 ± 0.19
Amy	Amygdala	1.00 ± 0.16

*Volume was calculated using the segmented regions of interest and represented as the average and standard deviation of bilateral volume across subjects