Regional Differences in Diffusion Tensor Imaging Measurements: Assessment of Intrarater and Interrater Variability


AJNR Am J Neuroradiol 2008, 29 (6) 1124-1127
doi: https://doi.org/10.3174/ajnr.A0998
http://www.ajnr.org/content/29/6/1124
Regional Differences in Diffusion Tensor Imaging Measurements: Assessment of Intrarater and Interrater Variability

BACKGROUND AND PURPOSE: Diffusion tensor imaging (DTI) has become a valuable tool in both the research and clinical evaluation of subjects. We sought to quantify interobserver and intraobserver variability of diffusivity and diffusion anisotropy measurements with regard to specific regions of interest (ROIs).

MATERIALS AND METHODS: The subject group consisted of 5 healthy control subjects and 7 study subjects (all males; 16–19 years old; mean age = 17.5 years), as part of a protocol for closed head injury. Two whole-brain DTI scans were acquired on a 3T scanner for each subject. Analysis was performed using a ROI approach. Two independent observers analyzed the apparent diffusion coefficient (ADC) and fractional anisotropy (FA) indices in the corpus callosum, cortical spinal tract, internal capsules (ICs), basal ganglia, and centrum semiovale (CSO). Intraobserver and interobserver variability were calculated for the mean ADC, FA, and ordered eigenvalues of the diffusion tensor ($\lambda_1$, $\lambda_2$, and $\lambda_3$).

RESULTS: The overall $\kappa$ statistic for intraobserver variability for both observers showed slight-to-substantial agreement ($\kappa = 0.02–0.69$), however FA values in the CSO showed only slight agreement. Interobserver agreement was also slight to substantial for these DTI measurements with high variability in FA values in the IC and CSO.

CONCLUSIONS: When one is comparing 2 DTI measurements, it is important to assess intraobserver and interobserver variability. We recommend caution in the analysis of DTI contrasts in the IC and CSO, because we have found the widest range of variability in measurements within these structures.

Diffusion tensor imaging (DTI) has become well established as a research tool to investigate water diffusion properties in the central nervous system and is making inroads into clinical imaging. It is a safe, noninvasive in vivo method that allows a superior assessment (compared with conventional MR imaging) of white matter tracts by reconstructing their 3D shape and connectivity. Parameters derived from DTI can provide information about tissue organization, degree of myelination, and water mobility, enabling the study of white matter tracts and connectivity, integrity, and damage in the brain.1

Although DTI was initially used for anatomic purposes to understand human brain anatomy and to topographically depict the white matter tracts of the brain, the technique has increasingly been used to study changes in pathology by comparing quantifiable metrics of diffusion.

Two important parameters derived from DTI are the apparent diffusion coefficient (ADC) and fractional anisotropy (FA). The ADC and FA parameters characterize the average amount of diffusion and the diffusion anisotropy, respectively.2 These parameters are derived from the eigenvalues ($\lambda_1$, $\lambda_2$, and $\lambda_3$) computed from the diffusion tensor. The diffusivity measurements from the eigenvalues themselves can also be used to study tissue properties.3,4 Recent reviews5,6 outline the methodology and clinical applications of DTI.

In clinical practice, these parameters can be used for comparison between individual patients, for serial examinations in the same patient, and for the evaluation of maturation during childhood. The quantification of diffusion can be especially helpful, because it may allow earlier diagnosis of the presence and extent of pathology. Pathologic changes due to damage in the central nervous system start at the microstructural level. Previous studies have shown that changes in diffusion characteristics (ie, diffusivity and diffusion anisotropy) can be detected in stroke7 and multiple sclerosis8 before abnormalities can be detected with conventional MR imaging7,9,10.

The assessment of the variability of DTI measurements, both when the data are analyzed repeatedly by one reviewer or by different reviewers, is an essential and important step in evaluating the clinical utility of DTI and its strength as a quantitative measurement. We sought to provide measurements of the intrarater and interrater variability for DTI measurements and to assess the degree of difference that can be detected comparing 2 DTI studies.

Materials and Methods

Subjects
The subjects were recruited as part of an ongoing study looking at closed-head injury in athletes performing contact sports. This study was approved by the institutional review board, and written informed consent was obtained from all of the adult subjects and from the parents of the minors before data acquisition. The subjects were all male and ranged in age from 16 to 19 years (mean age = 17.5 years). Five healthy control subjects with no history of neurologic disease and 7 research subjects participated in this study. All of the subjects were also evaluated...
for trauma-related changes in MR imaging by one observer (D.M.Y). There were no abnormal findings in MR images of subjects.

**MR Imaging**

All of the subjects were scanned using a 3T MR imaging scanner (Philips Medical Systems, Best, the Netherlands) capable of 60-mT/m magnetic field gradients, using the body coil for excitation and an 8-channel phased-array sensitivity-encoding coil for reception. Each DTI dataset was acquired with the following protocol. A multisection, single-shot echo-planar imaging, spin-echo sequence (TR/TE = 7897/84 ms; FOV = 212 × 212 mm) was used to acquire 60 transverse sections with no section gap and 2.2-mm nominal isotropic resolution (acquired matrix = 96 × 96, reconstructed to 256 × 256). Diffusion weighting was applied in 32 noncollinear directions with a b-value of 700 s/mm². A volume minimal diffusion weighting (b = 0 s/mm²) was also acquired. Two DTI datasets were acquired for each subject, and the total scan time for DTI was approximately 10 minutes.

**Image Processing**

DTI data were processed off-line using the Coregistration, Adjustment and Tensor-solving, a Nicely Automated Program (CATNAP; http://iacl.ece.jhu.edu/~bennett/catnap/, Johns Hopkins University School of Medicine), an in-house–designed data processing pipeline. CATNAP performs motion-correction with the Oxford Center for Functional Magnetic Resonance Imaging Linear Image Registration Tool (Oxford, United Kingdom), computes diffusion-weighted gradient tables adjusted for section angulation, and calculates the diffusion tensor, as well as the diffusivity (ADC, eigenvalues) and diffusion anisotropy (FA) metrics. Motion correction was performed by using a 6-degree of freedom rigid body model, and the 2 DTI datasets were concatenated and processed simultaneously in 1 diffusion tensor calculation. Image visualization and region of interest (ROI) placement were then performed using DTISTudio (Johns Hopkins University, Baltimore, Md).11

**ROI Placement**

The processed DTI contrasts were analyzed separately and independently by 2 observers (A.O. and A.D.S) who had been trained on the DTISTudio software package. To estimate the intraobserver and interobserver reliability of ROI-based DTI parameters, ROI placement was carried out by both observers on 2 separate occasions, 4–12 weeks apart, without the use of a template. The information about the first ROI placement was not available during the second assessment, and the investigator was blinded to the other observer’s evaluation. We selected 9 different ROIs in locations with different FA. The locations, which are commonly used in many clinical DTI studies, were easily visualized and delineated on DTI color maps. Predetermined circular single ROIs were manually placed on color maps at the following anatomic locations: corticospinal tract at the level of the pons, middle cerebellar peduncles (MCP), anterior limb of the internal capsule, posterior limb of the internal capsule, genu of the internal capsule, centrum semiovale (CSO), thalamus (Th), putamen (Pt), splenium, and genu of the corpus callosum (CCG). With the exception of the corpus callosum, all of the ROIs were positioned bilaterally. The ROIs were then propagated onto the DTI contrast images (i.e., FA, ADC, etc.), and the mean and SD over all voxels in the ROI were calculated. Statistical analysis (intrarater and interrater comparisons) were performed for the mean ADC, FA, $\lambda_1$, $\lambda_2$, and $\lambda_3$.

**Data Analysis**

To compare first and second measurements of 1 observer (intraobserver reliability) and to compare the first and the second measurements of 2 independent investigators (interobserver reliability), $\kappa$ statistics was used. (The $\kappa$ value of 0.11–0.20 is considered as “slight,” 0.21–0.40 as “fair,” 0.41–0.60 as “moderate,” 0.61–0.80 as “substantial,” and 0.81–1.00 as “almost perfect” agreement.) Finally, to determine the degree of differences between measurements as a percentage, we calculated the difference between 2 measurements divided by the mean of these 2 measurements for both evaluators for all of the ROIs.

**Results**

**Intrarater and Interrater Agreement**

The overall $\kappa$ statistic for intraobserver variability for both observers showed slight to substantial agreement ($\kappa = 0.02–0.69$) for FA, ADC, and eigenvalues (On-line Table 1); however, FA values in the CSO showed only slight agreement ($\kappa < 0.20$). The interobserver agreement was also slight to substantial for these DTI measurements ($\kappa = 0.02–0.69$; On-line Table 1) with high variability in FA values in the internal capsule and CSO. Eigenvalues also showed slight agreement in internal capsules, CSO, and MCPs for interobserver agreement. $\lambda_1$ was more reliable than the $\lambda_2$ and $\lambda_3$ (On-line Table 2).

**Regional Distribution of Measurements**

When the individual ROIs were analyzed to determine whether the reliability was lower in some location than others, the overall span of variability of interobserver readings showed the lowest agreement within the internal capsule and CSO, whereas the splenium and CCG showed the highest agreement (On-line Tables 1 and 2). FA measurements are least reliable in the CSO and gray matter ROIs (On-line Table 2). The percentage of variability (as measured by the difference between the 2 readings divided by the mean of the 2 readings) was obtained for all of the DTI measures for each reader (On-line Table 2).

**Discussion**

**The Potential Clinical Use of DTI Measurements**

The potential use of DTI assessments in clinical practice is encouraging. ADC and FA measurements are frequently used as potential biomarkers of the degree of tissue injury in brain diseases. Some authors have used DTI in the assessment of brain disorders and have shown abnormal hemispheric fiber connections in acquired disease or congenital abnormalities.13

In a similar vein, the degree of white matter disruption due to MS or diffuse axonal injury may be quantitatively assessed with DTI metrics, as opposed to merely depicting the tracts as smaller in size or finer in quality. Although the interpretation of diffusion changes measured by DTI is not straightforward, measures of the severity of demyelination and axonal damage would be of great clinical relevance.8,14-19 Ptak et al20 have used DTI indices to propose a cerebral FA score to serve as an index of white matter injury due to trauma that successfully correlates with outcome and predictor variables.21

DTI is a valuable tool to assess the impact of neoplasms on the white matter tracts. Another widespread white matter disorder, such as adrenoleukodystrophy, could be assessed
quantitatively with DTI as the dietary therapy is implemented. In this way, an improvement from baseline can be demonstrated eloquently in a quantitative manner.22,23 Other clinical applications include the use of DTI to investigate the affects of small vessel ischemic disease on white matter integrity, determining thresholds by which the patient becomes symptomatic. More severe white matter disease, such as cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy orBinswanger disease, which both tend to affect the white matter more so than the gray matter,24 could also be investigated. Neurophysiologic disorders have also been investigated with DTI, and some researchers propose that errant white matter connectivity may be the intrinsic problem in schizophrenia, autism, or dyslexia.25 From the clinical perspective, it is hoped that the sensitivity of DTI contrasts to changes in microstructure may permit the detection of recovery (as a function of time or medical treatment), which could be correlated with standard clinical outcome measures.

Variability in DTI Measurements

The acceptable range of variability of a quantifiable index in medicine cannot be arbitrarily set. Nonetheless, it is valuable to measure this variability if one hopes to be able to make a claim of abnormality when a value deviates from "the norm." DTI will increasingly be used to access pathology in the brain, and, therefore, knowing the interobserver and intraobserver variability is critical. Calculating DTI indices requires an intimate knowledge of the anatomy of the brain so that the ROIs can be placed appropriately. The variability from one observer to the other largely lies in the placement of the ROIs and, when tractography is performed, is also influenced by the subjective determination of the fiber tracking termination criteria (lower bound FA and turning angle thresholds). The variability produced by ROI placement is due, in part, to the neuroanatomic training of the observer, as well as the selection of the proper location for analysis as visualized on DTI contrasts.

According to our results, measurements of FA were most reliable in the CCG and least reliable in the Pt, Th, and CSO. For ADC, the percentage of variability was highest in Th and MCP and lowest in CSO and CCG. Reliability can be affected by a number of factors, including scanner performance, initial signal intensity-to-noise and acquisition resolution, positioning, segmentation, alignment, warping, and resectioning.26 Pfefferbaum et al26 were the first to address reproducibility of FA and ADC images in detail. Many studies have compared measured ADC and FA values in healthy children and adults on 1.5T MR imaging units.4,10,27-29 Characterizing regional variation in measurement error for this method is important for the understanding of the results of group comparisons and longitudinal studies. This is particularly true in the case of some diseases where differences from controls may be quite subtle.30 The possible explanation of higher variability in the internal capsule and CSO may be due to section shifts of the ROI location leading to addition or subtraction of a group of pixels within the section which would have a large effect on the k value.

The pattern of high variability in the internal capsule and in other white matter structures with different FA might indicate a combination of effects due to noise, partial volume effects, and complex fiber architectures within a pixel, which could easily vary from one section to another section. Image noise produces errors in the calculated tensor and, hence, in its eigenvalues and eigenvectors. Our results showed that λ1 was more reliable than λ2 and λ3 in several regions, such as the corpus callosum and CSO and corticospinal tract at the pons level. Overall percentage of variability was higher for λ3. Random variations in these quantities complicate the analysis and interpretation of DTI experiments. It is known that, in anisotropic systems, the expectation value of the largest eigenvalue is overestimated, and the lowest eigenvalue is underestimated.31,32

ADC values of the healthy brain are very similar in gray matter and white matter; however, FA shows distinct mean values between these structures. This may explain inaccuracies of FA values due to partial volume effects and ROI outlines of fiber tracts close to their borders.33 Overall, a lower percentage of variability of our ADC measurements is consistent with other studies reported in the literature (On-line Table 2).34,35

Rater performance is also an important source of variability and is one of the limiting factors in the detection of variance in both cross-sectional and longitudinal studies. Standardized rater training is increasingly used to improve the quality of the investigated outcome parameters.35 It is possible that, while 2 observers may have equivalent training for anatomic localization and use of software packages for DTI analysis, delineation of ROIs without the assistance of a template or reliance on single ROIs that may be subject to section shift errors may introduce sources of error and contribute to data variability. In a similar fashion to methods used in fiber tracking, the confidence level of validity could be improved by the use of anatomic limitations using multiple ROIs.36

One of the purposes of this article was to look at the degree of variability and to determine at what level a difference in DTI values would be significant. If one uses 2 SDs of variability as a marker for 95% accuracy corresponding to a typical p value of .05, it would suggest that identifying a difference more than twice that of the variability would be reasonable to compare a single subject versus a population norm. For example, given our variability of 2.6% for FA of the CCG, a difference of more than twice that, or 5.2% from either baseline or from a matched control would be a reasonable standard to use. For each DTI parameter, for each location, there are different thresholds. Overall, if one reviews on-line Table 2, these threshold values would range from 6.6% for the ADC of the CSO to as high as 25.0% for the ADC of the MCP. For comparative group studies, the empirical variability for each region may guide appropriate power analyses.

The limitations of this study include the relatively small number of subjects evaluated. One would expect nonetheless that the degree of variability would not change significantly with increasing numbers of subjects. The results herein reflect that of a 3T protocol, which may not apply to the more widely used 1.5T magnets. It would also be of benefit to have more than 2 observers for every location and to have evaluated additional locations with additional ROIs, which may be used for various analyses.

DTI measurements are sensitive to differences in hardware, acquisition parameters, analysis software, and data processing strategies. We suggest that standardizing and using schemes of ROIs should allow reduction of interobserver and intraobserver variability. The variability is greatest when slight differ-
References

2. Sigmund EE, Song YQ. Multiple echo diffusion tensor acquisition technique. Magn Reson Imaging 2006;24:7–18