Published in final edited form as:

Neuroimage. 2008 February 15; 39(4): 1693–1705.

Whole Brain Voxel-wise Analysis of Single-Subject Serial DTI by Permutation Testing

SungWon Chung^{a,b,c}, Daniel Pelletier^{b,d}, Michael Sdika^{b,d}, Ying Lu^{a,b,c,e}, Jeffrey I. Berman^{b,c}, and Roland G. Henry^{a,b,c,*}

aUCSF / UC Berkeley Joint Graduate Group in Bioengineering

bCenter for Molecular and Functional Imaging, San Francisco, CA

cDepartment of Radiology, University of California San Francisco, San Francisco, CA

dDepartment of Neurology, University of California San Francisco, San Francisco, CA

eDepartment of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA

Abstract

Diffusion tensor MRI (DTI) has been widely used to investigate brain microstructural changes in pathological conditions as well as for normal development and aging. In particular, longitudinal changes are vital to the understanding of progression but these studies are typically designed for specific regions of interest. To analyze changes in these regions traditional statistical methods are often employed to elucidate group differences which are measured against the variability found in a control cohort. However, in some cases, rather than collecting multiple subjects into two groups, it is necessary and more informative to analyze the data for individual subjects. There is also a need for understanding changes in a single subject without prior information regarding the spatial distribution of the pathology, but no formal statistical framework exists for these voxel-wise analyses of DTI. In this study, we present PERVADE (PERmutation Voxel-wise Analysis of Diffusion Estimates), a whole brain analysis method for detecting localized FA changes between two separate points in time of any given subject, without any prior hypothesis about where changes might occur. Exploiting the nature of DTI that it is calculated from multiple diffusion-weighted images of each region, permutation testing, a non-parametric hypothesis testing technique, was modified for the analysis of serial DTI data and implemented for voxel-wise hypothesis tests of diffusion metric changes, as well as for suprathreshold cluster analysis to correct for multiple comparisons. We describe PERVADE in detail and present results from Monte Carlo simulation supporting the validity of the technique as well as illustrative examples from a healthy subject and patients in the early stages of multiple sclerosis.

Keywords

diffusion tensor MRI; permutation; serial; longitudinal; voxel; cluster; fiber tracking; multiple sclerosis; white matter

* Address Correspondence to: Roland G. Henry, Ph.D., Center for Molecular and Functional Imaging, University of California, San Francisco, 185 Berry Street, Suite 350, Rm 339, Box 0946, San Francisco, CA 94107-0946, Email: roland.henry@radiology.ucsf.edu, Phone: 415-353-9406

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

In recent years, diffusion tensor magnetic resonance imaging (DTI) (Basser et al., 1994a; Basser et al., 1994b) has been widely used in studying normal development and aging, and diverse pathological conditions of the human brain due to its unique ability to identify microstructural abnormalities. Since DTI is a relatively new technique and is fundamentally different from other imaging techniques in that each voxel contains not a single value but a 3×3 matrix with 6 unique elements called the diffusion tensor, the statistical analysis of DTI is still under development. Multiple approaches exist for defining the target regions for statistical comparisons, including manually traced region of interest (ROI), fiber tracking defined ROI, whole brain histogram, and whole brain voxel-wise analysis by statistical parametric mapping SPM (Friston et al., 1995) or by tract-based spatial statistics (Smith et al., 2006). There are also diverse DTI-derived parameters to compare, such as mean diffusivity (D_{av}), parallel / transverse diffusivity, fractional anisotropy (FA) and the primary eigenvector (Schwartzman et al., 2005).

Most DTI studies with statistical testing performed in one or more of the aforementioned ways have one thing in common; they are all multiple-subject group comparison studies, usually one group of experimental subjects compared with a matched group of healthy control subjects. Group comparison studies are possible when the effects of interest are located in stereotypical anatomic structures, such as the same white matter tracts, across subjects. In certain conditions where the effects are expected to be focal (or multi-focal) with spatial distributions that are highly specific for individual subjects, finding and grouping subjects that share effects in similar anatomic locations may be difficult. Furthermore, group analyses demand that significant effects be larger than group variability and therefore may suffer from decreased sensitivity. A statistical analysis that can be performed in individual subjects is thus required. Due to the non-Gaussianity, DTI parameters such as anisotropy indices (Pajevic and Basser, 2003; Skare et al., 2000) makes parametric testing less optimal. Non-parametric resampling techniques have great potential in the statistical testing of these DTI data since the recent trend of oversampling DTI data makes it suitable for resampling. One of these techniques called bootstrap has been shown to be very useful in DTI (Chung et al., 2006a; Heim et al., 2004; Jones, 2003; Lazar and Alexander, 2005; Pajevic and Basser, 2003; Whitcher et al., 2007). These works have shown the ability of bootstrap to estimate the uncertainties of DTI parameters (descriptive statistics) but have not addressed the issue of statistical significance testing (inferential statistics). For statistical testing when comparing two groups, permutation testing (Edgington, 1995; Good, 2005; Manly, 2007), another non-parametric resampling technique, can be useful as well.

Permutation testing provides statistical significance testing of differences between groups, with the unique ability of directly estimating the null distribution of the statistic describing the difference, rather than assuming a null distribution of analytically known form (such as T-distribution). Multiple groups are required for resampling and permutation testing provides exact (or almost exact) p-values but does not estimate uncertainties of sample statistics (such as the standard error of mean). Bootstrap is mainly used to estimate the accuracy of sample statistics by resampling from one group (multiple groups are not required). Though it can be used not only in descriptive statistics (i.e. standard errors) but also in inferential statistics (i.e. significance testing), bootstrap testing is not able to estimate a data-driven null distribution and bootstrap-estimated p-value is only approximate.

We focus on permutation testing in this work. Permutation testing is well-established in the field of neuroimaging, especially functional MRI where numerous works advocate the strength of this approach in the last decade (for a review refer to (Nichols and Holmes, 2002)). Permutation based method 1) makes minimal assumptions and thus can be applied even in

situations where the assumptions of parametric approaches are not met or cannot be verified, 2) is conceptually simple and provides intuitive solutions to the multiple comparison problem, and 3) is easily applicable to any test statistics allowing the researchers to freely choose the statistic best suited for their studies. Unlike bootstrap, permutation testing can elegantly incorporate the whole statistical procedure of voxel-wise comparison and multiple-comparison correction in a completely non-parametric way due to the ability to estimate data-specific null distribution as described above, possibly with more accurate p-values.

In this study, we describe how permutation testing can be properly implemented in voxel-wise analysis of single-subject serial DTI studies. We will pay special attention to not violating exchangeability assumption for permutation testing that if groups are not different then any re-grouping of the samples (permutation) are equally likely as the original grouping (observation). Then, we present a novel statistical analysis framework called PERVADE (PERmutation Voxel-wise Analysis of Diffusion Estimates) that is designed to localize subtle and local microstructural changes over time in the whole brain of a single subject without any prior hypothesis. PERVADE includes 1) non-linear registration between two time-points to account for any morphological changes over time, 2) voxel-wise calculation of p-values by permutation testing, and 3) suprathreshold cluster analysis with permutation testing to deal with the multiple comparison problem. Our preliminary results of local microstructural changes detected outside as well as inside the focal lesions of patients in the earliest stage of multiple sclerosis show the potential of this technique to provide additional information about microstructural white matter injury.

Methods

DTI Permutation Testing

Permutation Testing in Group Comparison—Before presenting the proposed voxel-wise permutation testing in a single-subject serial DTI study, we first examine how it would be done in a multiple-subject group comparison study (i.e. comparison of group-averaged metric between two cohorts), a simple and typical scenario of permutation testing. To test whether the observed difference in group-averaged FA, i.e. $\hat{\theta} = \bar{FA}_B - \bar{FA}_A$, is statistically significant, permutation testing can be performed as follows. All subjects are randomly assigned to either group A or B while maintaining the original number of subjects (thus re-defining groups A, B to A*, B*), and one permutation replication of $\hat{\theta}$ defined as $\hat{\theta}_i^* = \bar{FA}_{B^*} - \bar{FA}_{A^*}$ is calculated. If the null hypothesis H_0 that the FA values of groups A and B are not different holds, then this re-grouping (or re-labeling to either group in each subject) would be equally likely to happen as the original grouping, meaning that any possible $\hat{\theta}_i^*$ are equally likely as $\hat{\theta}$. Thus, the random variable $\hat{\theta}^*$ has the null distribution of $\hat{\theta}$ if H_0 is true, and it is called permutation distribution.

The statistical significance (p-value) in the permutation testing is defined as,

$$p = \text{Prob}_{\text{perm}} \left\{ \left| \hat{\theta}^* \right| \geq \left| \hat{\theta} \right| \right\} = \# \left\{ \left| \hat{\theta}^* \right| \geq \left| \hat{\theta} \right| \right\} / N, \quad [\text{Equation 1}]$$

where N is the total number of possible labelings. This simply quantifies the probability of observing group-averaged FA difference more extreme than $\hat{\theta}$ by random chance. If this probability is less than 5 percent ($p \leq 0.05$) then $\hat{\theta}$ is unlikely under H_0 , and we reject H_0 concluding that two groups are different in terms of FA. If N is too large to compute all possible labelings, the p-value can be approximated by using a randomly chosen subset of all possible labelings such that

$$p \approx \# \left\{ \left| \widehat{\theta}_i^* \right| \geq \left| \widehat{\theta} \right| \right\} / N' \quad (i = 1, 2, \dots, N'),$$
[Equation 2]

where $N' < N$. Typically $N' = 1000$ and $\widehat{\theta}_1^* = \widehat{\theta}$ (i.e. the first labeling of $\widehat{\theta}^*$ is just the observed statistic $\widehat{\theta}$).

Permutation Testing in Intra-subject Serial Comparison—On the other hand, if we want to test whether FA from two time points of a single subject is different, we can no longer permute the observed FA values since there are only two such values available, one from each time point. Instead, since the FA at one time point is calculated from the diffusion weighted images (DWIs, this refers to the raw images acquired in a single diffusion encoding direction for DTI in this study) at that time point (just like $\overline{FA}$ is calculated using the FA values from the individual subjects of the group), the DWIs can be permuted. Assuming that same DTI acquisition protocol was used for two time points, which is a typical scenario since it is always strongly encouraged to use the same protocol in the same scanner for follow-up MRI studies, the simplest approach satisfying exchangeability is to permute only the DWIs that have the same diffusion encoding (diffusion gradient direction and strength). This stratified permutation scheme shares similarity with permutation of scans only within an exchangeability block in functional neuroimaging (Nichols and Holmes, 2002), and with stratified (or repetition) DTI bootstrap (Chung et al., 2006a; Pajevic and Basser, 2003). While multiple repeated acquisitions are mandated for stratified DTI bootstrap, scans for stratified DTI permutation do not need to be repeated at each time point since pooling two time points naturally guarantees at least two repetitions. A caveat here, though, is that even with the same diffusion encodings, the DWIs may still not be theoretically exchangeable due to differences in the gain factors in the raw MR signal intensities and possible difference in head positioning (effectively changing the diffusion encoding directions relative to the brain tissue). Any drift in the gain factors of the MR system will result in changes of the DWI signal intensities. However, this difference can be easily remedied and exchangeability secured by simply estimating the ratio of the signal intensities from the images where we can assume that there are no biological changes, such as normal-appearing white matter (NAWM) in the non-diffusion-weighted images, simply called $b=0$ images. For the head positioning issue, if the relative angular locations of the brain to the MRI system are different between two time points, the brain will effectively be under different diffusion encodings even if the diffusion gradients are applied exactly the same way. Thus, we propose to permute the diffusion encoding gradients (rotated in the same degree as any registration performed in order to correct for the different head positioning) in conjunction with the DWIs themselves. While DWIs themselves might not be exchangeable in a strict sense, the complete set of permuted DWIs paired with properly adjusted diffusion encoding directions could be exchangeable.

The only assumption required to justify this approach is that slightly different configurations of diffusion encoding directions will not affect the probability distributions of DTI metrics. Our experience tells us that in typical serial MRI studies, the rotational components by different head angulations are usually a few degrees (axis of rotation either x, y, or z). After the permutations above, each of diffusion encoding direction will be ‘jittered’ around a few degrees, but the diffusion gradient scheme will be similar overall. Notice that even if we are permuting the pairs of DWIs and diffusion directions that are slightly jittered, we still opt for the stratified permutation since drastic variations of the set of diffusion gradients after permutation can distort the permutation distribution due to gradient configurations effects on the DTI parameter estimations (Jones, 2004). Limiting the permutation within the exchangeability block defined as the same physical diffusion encodings guarantees similar diffusion gradient sets. The permutation testing to calculate voxel-wise p-value is illustrated in figure 1 steps (a) to (e) and figure 2.

Permutation Testing for Multiple Comparisons—The permutation scheme can be used not only for voxel-wise p-values but also to deal with the multiple comparison problem arising from testing multiple voxels simultaneously. Specifically, we chose the suprathreshold cluster analysis (STCA) method that is known to provide a good balance between the statistical power and the ability to localize the changes (Bullmore et al., 1999; Friston et al., 1996; Nichols and Holmes, 2002; Poline et al., 1997; Roland et al., 1993). In STCA scheme, the statistic map is initially thresholded at some arbitrary single voxel level to define clusters (spatially contiguous voxels with contiguous defined as sharing a side). The multiple comparison corrected p-values for clusters are then calculated based on the distribution of maximum cluster sizes identified in each permutation originally done for the voxel-wise analysis; i.e., the same set of permutations are used for all voxels in the analysis domain. Equation 2 is valid here as well except that taking the absolute values of $\hat{\theta}_i^*$ and $\hat{\theta}$ is not necessary since cluster sizes are always positive (thus it is one-tailed testing, not two-tailed) and the equation is shown below,

$$p \approx \# \{ \hat{\theta}_i^* \geq \hat{\theta} \} / N' \quad (i = 1, 2, \dots, N'), \quad [\text{Equation 3}]$$

where $\hat{\theta} = \max_{\{p\}} (\{\text{cluster size}\})$ and $\hat{\theta}_i^* = \max_{\{p_i^*\}} (\{\text{cluster size}\})$. $\{p\}$ and p_i^* denotes the observed p-value map and p-value map from i^{th} permutation each, in the domain of whole white matter.

The implemented permutation-based STCA is similar to that described previously (Holmes, 1994; Holmes et al., 1996; Locascio et al., 1997), except that we use the p-value maps created from voxel-wise permutation tests to define clusters instead of some other statistic maps (like t-statistic). During the voxel-wise permutation testing, the whole image of DWIs are permuted together instead of permuting individual voxels independently. While not affecting the voxel-wise permutation testing in any way, this helps to create permutation replicas of the statistic maps that are equally likely in terms of spatial correlation structures as well as the distribution of p-values. Each cluster in the observed p-value map are assigned the cluster-wise p-value by comparing the cluster size with the maximal cluster permutation distribution $\hat{\theta}^*$, i.e. the collection of image-wise maximal cluster size from all the permutation replicas. The cluster-wise p-values estimated as such are by definition multiple-comparison corrected since the permutation distribution describes the probability of having any cluster(s) above certain size in the image by random chance. It has been shown that a strong control of family-wise error is achieved in this way, meaning that we can localize significantly changing regions by rejecting the null hypothesis for the clusters with adjusted p-values smaller than 0.05, rather than simply knowing that changes are occurring somewhere in the brain (Holmes et al., 1996). The permutation testing to calculate multiple-comparison-corrected cluster-wise p-values is illustrated in figure 1 steps (f) to (h).

Identification of significant clusters was performed with a multi-step iterative algorithm called the step-down test (Holmes, 1994). After rejecting the null hypothesis for clusters exceeding the 95th percentile of the permutation distribution $\hat{\theta}^* = \max_{\{p^*\}} (\{\text{cluster size}\})$, a new white matter domain that does not include these clusters was defined. $\hat{\theta}^*$ was then re-estimated from this new domain using the same p-value maps, and any previously undetected clusters were searched. These steps were repeated until no more clusters were rejected. Then, $\hat{\theta}^*$ in the final step was considered the most appropriate null distribution least affected by any real changes, and all the clusters in the whole white matter domain were tested against this distribution. Actually, this approach of re-estimating $\hat{\theta}^*$ after excluding all rejected tests simultaneously rather than only the most significant test is called the jump-down variant of the step-down approach.

Jump-down testing can substantially decrease the computation time due to fewer steps.

Non-linear Registration

As is the case for most voxel-wise analysis techniques, PERVADE depends on the accuracy of registration techniques. It is designed for detecting changes over time in the same subject, and it is easier to register data from the same subject than across subjects since there is much less intra-individual than inter-individual biological variability. This makes PERVADE less vulnerable to misregistration than group comparison methods. Still, any morphological changes over time such as atrophy should be accounted for, and, if not, they may be detected as significant changes. While acknowledging that detecting atrophy of specific white matter structures based on the unique ability of DTI to create rich contrast within white matter is also a subject of interest, the focus of PERVADE for now is to detect ‘microstructural’ changes over time, parametrized by DTI scalar indices such as FA, rather than ‘macrostructural’ volumetric changes associated with atrophy. Thus we attempted to eliminate volume changes by non-linear registration but remain sensitive to microstructural related intensity changes due to the atrophy process.

A high-order non-linear registration algorithm (Sdika, 2007) was used for registration of the diffusion data between time points. The transformation was parametrized using cubic B-splines, making it intrinsically smooth, local, compact and fast to compute. The local support property of B-spline is particularly important as it ensures that a deformation in a given region of the brain has only a local effect on the transformation. An empirically chosen node spacing of 12 voxels in x, y, and z directions was used. To prevent folding in the transformation, the registration was constrained. These constraints made the Jacobian of the transformation greater than a small positive threshold on the pixels. Negative Jacobians between the pixels were penalized by making Jacobian derivatives small on pixels. The constrained optimization problem was solved using a combination of the multipliers method and the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method (Liu and Nocedal, 1989) with a non-monotonic line search. The cost function was the squared difference between the reference and the deformed version of the floating image, suitable for intramodal registration of quantitative images such as FA.

Monte Carlo Simulation

Monte Carlo simulation studies were performed to investigate the effects of rotated diffusion encoding directions and different gain factors on the permutation exchangeability between different time points. DTI acquisition in each time point was assumed to be six diffusion directions (dual gradient scheme, $b=1000 \text{ s/mm}^2$) in addition to one $b=0 \text{ s/mm}^2$ acquisition, repeated three times (total 21 measurements). The noise-free diffusion tensor $\mathbf{D}$ was defined with $\text{FA}=0.5$, $D_{\text{av}} (= \text{Tr}(\mathbf{D})/3) = 0.7 \times 10^{-3} \text{ mm}^2/\text{s}$, and $\lambda_2=\lambda_3$.

The simulation of noisy diffusion weighted signals and DTI processing was done in a similar manner as described elsewhere (Chung et al., 2006a). From $\mathbf{D}$, noise-free diffusion weighted signals were calculated by $S(\mathbf{g}) = S_0 \exp(-b\mathbf{g}^T\mathbf{D}\mathbf{g})$ where S_0 is the signal without diffusion weighting, b is the diffusion weighting factor, and $\mathbf{g}$ is 3×1 unit vector of the diffusion-encoding gradient direction. For time point B, in addition to using the same gradient set and scaling factors, we also tested scenarios where $\mathbf{g}$ was replaced by $\mathbf{g}'=\mathbf{R}\mathbf{g}$ ($\mathbf{R}$ is a product of three 3×3 matrices, each describing rotations around x, y, and z axis each) or S_0 was replaced by S_0' in order to simulate the rotated diffusion directions or drift in gain factors. Next, the noise modeled as complex random number with real and imaginary parts following Gaussian distribution of zero mean and standard deviation $\sigma(=S_0/\text{SNR})$ was added to $S(\mathbf{g})$ and the magnitude of the noisy signal was calculated. The magnitude of the added noise ($\text{SNR}=25$ in S_0 or S_0') was constant for both time points, and for all the simulation studies.

From a complete set of noisy signals, the noisy diffusion tensor, FA, and finally the statistic of interest $\hat{\theta} = FA_B - FA_A$ were calculated. Then, the stratified DTI permutation algorithm was performed on a pool of 42 signals (7×3 from each time point) 999 times, with and without the 'corrected' permutation scheme that ensures exchangeability by permuting $\mathbf{g}$ ($\mathbf{g}'$ for B) in conjunction with $S(\mathbf{g})$ (scaled in B to match A). Using the statistic from the original labeling $\hat{\theta}$ and the values from the 999 permuted labelings $\hat{\theta}_i^*$, the p-value of $\hat{\theta}$ was calculated by equation 2. This experiment of estimating a p-value from a noisy dataset was repeated 10,000 times to derive a distribution of p-values. We assumed that the distribution of FA at both time points was equivalent and therefore the distribution of p-values should have a uniform probability between 0 and 1. This fundamental statistical property was used to evaluate our methods. The gold standard (a uniform distribution) and derived p-value distributions were compared using quantile-quantile plots (Q-Q plots).

Human Brain Data

Subjects / Acquisitions—Clinically isolated syndrome (CIS) is a diagnosis given to patients who experience for the first time neurological symptoms indicative of central nervous system dysfunction (visual, brain/brainstem, or spinal cord) thought to be due to primary demyelinating disease. It is considered by clinicians to be the earliest stage of multiple sclerosis (MS), even before a definite diagnosis can be confirmed by subsequent clinical symptoms or new lesions seen on MRI scans. Three CIS subjects at risk of developing MS and one healthy volunteer were included in this study. Patients were part of an ongoing follow-up study of CIS patients, with scans at baseline (within 3 month of initial clinical presentation) and every 3 to 6 months afterward for two years. A 27-year-old healthy male subject was scanned four times within two months in order to validate the technique when no biological changes over time can be assumed. All subjects gave informed consent and the study was approved by our institution's committee on human research.

All images were acquired on a 1.5T GE Signa system (General Electric, Milwaukee, WI) equipped with 4G/cm gradients and a standard quadrature head coil. The DTI scans were performed with a single-shot multislice axial spin echo EPI sequence (TR/TE = 7s/105ms, 9 NEX) at 1.7×1.7×2.1 mm voxel resolution using six diffusion-encoding directions (dual gradient scheme) at $b=2000 \text{ s/mm}^2$ in addition to one $b=0 \text{ s/mm}^2$ image. The brain was covered from the level of midbrain to the vertex with 40 slices. Data averaging in the scanner was done up to 3 NEX, which was acquired three times to come up with 3 repetitions of 3 NEX data.

Analysis—The overall procedure of PERVADE can be split into 5 steps, and the details of each step are given below.

Step 1: Initial DTI processing at each time point: After motion correction by linear image registration (FLIRT) (Jenkinson and Smith, 2001), the DWIs were slightly smoothed using a 3×3×3 voxel Gaussian kernel (FWHM 1.5 voxel) to increase the SNR and were processed to calculate FA, as described elsewhere (Chung et al., 2006a). Brain masks were created from the $b=0$ images using the brain extraction tool (BET) (Smith, 2002) in order to mask out the artificially high FA regions outside the brain so that these will not affect the inter-time-point registration of FA maps in the next step.

Step 2: Inter-time-point registration: Non-linear registration was done after the initial linear registrations by FLIRT. For both linear and non-linear registrations, FA maps from all the time points were matched to a reference FA map that was created as follows. First, a linear transformation between FA maps from two arbitrarily chosen time points was calculated. Then, the FA map from one of these time points was half-transformed to the other time point such that the resulting FA map will be spatially in the middle in between these two FA maps. This

procedure ensured that none of the time points benefited from having no registration / interpolation error that can arise in other time points. Also, sinc interpolation was used for all the registrations in this study. The DWIs were resampled based on the transformation between the FA maps.

Step 3: Masking Out Unnecessary Regions: We masked out brain regions where subsequent permutation tests were not to be performed. Not only does this save computation time, but this also increases the power of detecting statistically significant clusters by reducing the number of multiple comparisons. By default, white matter from the whole brain was defined as the domain for permutation testing. The whole brain white matter was defined as voxels within the brain that had FA values larger than 0.2 from the average of the FA maps from the two time points, and that were not contaminated by CSF (defined by segmenting $b=0$ image using an automated segmentation tool (FAST) (Zhang et al., 2001)).

Step 4: Permutation testing for voxel-wise p-values: The DWIs that were resampled to the common space by step 2 were smoothed using the same smoothing kernel of step 1 before voxel-wise permutation testing. Voxels outside the white matter were not included in the smoothing kernel. The first time point (baseline) was compared to subsequent time points. For each time point, the directions of the diffusion gradients were rotated to the same degree as determined from the linear transformation in step 2. In order to be exact, the degree of rotation should include the additional components of the non-linear transformation, but this was deemed minor in comparison to rotation described by linear transformation and thus was neglected. The scaling factor correcting for the drift in the gain factors over time was estimated from the median of ratio of signal intensities from NAWM (lesion masked out from white matter defined in step 3) and then multiplied back to DWIs. Permutation testing was performed with the observed labeling plus a randomly chosen set of 999 labelings, and the voxel-wise p-values were estimated as described above.

Step 5: Permutation testing for suprathreshold cluster analysis (STCA): In step 4, the permutation of the DWIs were done at the image-level, thus each permutation could generate a null p-value map maintaining the inter-voxel spatial correlation structures. For each of the 1000 p-value maps, the maximum cluster size found after thresholding the voxel-wise p-values at $p=0.01$ was entered as a data point in the cluster null distribution. Actual definition of clusters was not only based on the voxel-wise p-values but also the sign of the changes; in order to be a cluster, all the voxels in the cluster have to be changing in the same direction (increased or decreased FA values). This eliminated the chance of defining clusters of mixed changes (which is difficult to interpret biologically) and increased the power of detecting clusters with unidirectional changes. Significant clusters identified by the jump-down method were overlaid on the averaged FA maps for better visualization. In some cases, fiber tracking was performed to reveal the spatial connectivity between a focal lesion and significant clusters in the NAWM. The tracking algorithm was based on FACT (Mori et al., 1999), with termination criteria of $FA < 0.2$ and angles of primary eigenvectors between connecting voxels $> 50^\circ$.

Results

Monte Carlo Simulation

Comparison of the distribution of observed p-values from the simulation of permutation testing and the expected p-values are shown (Figure 3). Fig 3a shows that when the diffusion directions were rotated by 20 degrees each around x, y and z axes for one time point, and if the permutation was carried out ignoring the diffusion gradient directions, the observed p-values were deviated from the ideal distribution, with the trend of over-estimating p-values. This is not surprising

since increased variability due to the mismatched diffusion gradient directions will lead to more dispersed permutation null distributions that will eventually inflate the estimated p-values.

Permuting rotated gradients in pair with the signals led to observed p-values that closely followed the expected distribution, demonstrating the effectiveness of the modified permutation strategy. Note that the p-values hit the floor of 0.001 since using 1000 labelings does not allow the p-values to be lower than 0.001. This is not an important issue for this study since the voxel-wise p-values are to be thresholded at 0.01 for the cluster analysis. However, if estimating smaller p-values are desired, more permutation labelings would be necessary at the cost of increased computation time.

Results for smaller degrees of rotation of 10 degrees show that the deviation from the expected distribution without permuting diffusion gradients was much less. Since the degrees of rotation for real follow-up studies were typically within the range of 5 degrees (see below), we expect that the problem of over-estimating p-values will not be a serious issue even if diffusion gradients were not permuted, and any small effects will be eliminated by our proposed correction algorithm.

The differences in signal gain factors can bias the p-values in the similar way as diffusion gradient rotation (Figure 3c,d), and this effect was eliminated by scaling the intensities as described above.

Healthy Volunteer

Figure 4 shows voxel-wise p-value maps from the observed Δ FA estimated by permutation testing from three pairs (0 month versus 0 month (head rotated), 1 month, and 2 months) for a healthy subject. We can appreciate the smoothed random field pattern with few noticeable non-random structures. Further evidence of the validity of PERVADE in human data is demonstrated by the Q-Q plot showing excellent agreement of single-voxel p-values from the whole brain white matter with the expected distribution. The permutation distribution

$\hat{\theta}^* = \max_{\{p^*\}} (\{\text{cluster size}\})$ indicates that the probability of having any clusters larger than (or equal to) 11 voxels is 0.05, thus the critical value for significant clusters (for 0 versus 2 months) could be set as 11. Overall, only one significant cluster (slightly larger than the critical value) was identified in the three pairs of image-wise cluster analysis (data not shown).

CIS Patient #1

This patient had a significant thinning of the splenium of the corpus callosum over time that could be easily identified visually. For this patient, we compared the result from linear and non-linear registrations to investigate the effects of atrophy (tissue loss) with or without the non-linear registration. The baseline FA map registered to the FA map 10 months later, either by linear alone and by linear + non-linear transformation, are shown (Figure 5). The corpus callosum in the registered baseline FA map was 'thicker' by about a half voxel than the reference image with linear registration only, while addition of the non-linear registration brought a better match in a qualitative sense. With the linear registration, significant clusters were identified at the border of the corpus callosum by PERVADE and those were not detected after non-linear registration. Even after non-linear registration, p-value maps of baseline versus 16 months showed the trend of general FA decrease in the corpus callosum (most voxels are in 'blue' color) though they were not detected by PERVADE as clusters.

CIS Patient #2

This patient developed three new lesions in and around the splenium of the corpus callosum at 9 months after the baseline scan. While a few clusters (most of them attributable to lesions

visible in conventional MRI as well) were detected in the analysis of baseline versus follow-up scans before 9 months (data not shown), the analysis of baseline versus 9 months revealed numerous clusters with $p < 0.05$, not only within lesions but also in the NAWM (Figure 6). In the later time points, there was a tendency for the FA clusters to remain longer than T2 hyperintensities in some lesions. Using DTI fiber tracking, part of the forceps major (a subset of corpus callosum fibers connecting occipital lobes on both sides) was delineated that intersects one of the new lesions at 9 months (the lesion within the splenium of corpus callosum close to the midline). Significant NAWM clusters were located within or near the tracks, suggesting that the detected NAWM changes might be related to the new lesion.

CIS Patient #3

This patient had a lesion within the right pyramidal tract at the superior level of the internal capsule and inferior aspect of the corona radiata. The degree of T2 hyperintensity within the lesion lessened after 3 month compared to the baseline scan, possibly indicating edema resolution. Results from the serial DTI permutation tests highlighted two significant clusters, one in the lesion with increasing FA over time, and another in the NAWM at the superior aspect of the corona radiata with decreasing FA over time (figures 7). Tracks (part of pyramidal tract) seeded from the lesion showed that the distal cluster of decreased FA was connected to the lesion, suggesting that they are possibly related.

Gradient tables rotation and gain factors

For all the human data analyzed by PERVADE, the average difference of angulation between the two time points were 2.52, 1.55 and 1.30 degrees around x-, y- and z-axes, respectively. The average difference in the gain factors was 5.6%.

Discussion

In this paper, we have proposed a novel non-parametric statistical framework for detecting subtle and local diffusion MRI changes over time in a single subject. To our knowledge, this is the first study to analyze single-subject serial DTI data at the voxel / cluster level. This was possible with the ability of DTI permutation testing to calculate voxel-wise statistic p-values and to correct for multiple comparisons. We have demonstrated that we can ensure exchangeability of the DWIs which justified our implementation of permutations for testing two DTI datasets. We have also demonstrated that voxel-wise p-values from the DTI permutation tests follow the expected null distribution in a control subject, and that we can detect statistically significant clusters in lesions and normal-appearing white matter of patients at the earliest stage of MS.

Diffusion MRI, including three-directional diffusion-weighted MRI and DTI, has been widely used in the study of MS (Rovaris et al., 2005). Numerous studies have shown that diffusion MRI is sensitive to subtle damage in the normal appearing brain tissue as well as lesions and that it might even provide insights about pathological processes such as inflammation, demyelination and axonal injury (Henry et al., 2003). Serial studies can be especially valuable in studying disease evolution with or without treatment that can lead to a better understanding of pathophysiology and drug response, although they are rarely reported, probably due to more involved recruitment of patients. Four serial studies used ROI analysis of apparent diffusion coefficient (ADC) on either lesion or NAWM (Caramia et al., 2002; Castriota-Scanderbeg et al., 2002; Rocca et al., 2000; Werring et al., 2000) and one study used histogram analysis of D_{av} and FA (Oreja-Guevara et al., 2005). PERVADE is a completely different approach of analyzing serial DTI that can complement the techniques described above.

Before further discussing the merits and challenges of the technique presented here, we would like to point out that there have been numerous efforts in the MRI literature to use voxel-wise techniques to detect single-subject structural changes over time. Since the demonstration of detecting subtle brain changes by subvoxel registration and subtraction of serially acquired conventional MRIs (Hajnal et al., 1995a; Hajnal et al., 1995b), image subtraction techniques have been widely applied in detecting intensity changes (such as lesions in MS or tumor evolution), morphological changes (such as atrophy) or both. For localization and quantification of atrophy, image subtraction schemes have either focused on the structural boundary shift (Freeborough and Fox, 1997; Smith et al., 2001; Wang and Doddrell, 2002) or Jacobian matrix (Crum et al., 2001; Freeborough and Fox, 1998; Studholme et al., 2003; Thompson et al., 2003). For multiple sclerosis, the major interest was in the analysis of lesion evolution (Gerig et al., 2000; Lee et al., 1998; Meier and Guttman, 2003; Rey et al., 2002).

The techniques listed above have used conventional MRIs, and usually do not provide the statistical significance for the difference detected since conventional MRIs yield a simple one-image datum that does not allow uncertainties to be calculated in each voxel. One study, though, has attempted statistical testing with conventional MRIs (Lemieux et al., 1998). In DTI, a study utilized subtraction of FA maps over two time points to visualize the spatial distribution of FA changes (Qiu et al., 2006), though no formal voxel-wise statistical tests were attempted. DTI is unique in that numerous images are acquired, and this allows sufficient number of labelings to be used for permutation. We took advantage of this property to create statistical procedures for detection, rather than simple subtraction.

The most important strength of PERVADE is that it can lead to an almost exact statistical testing with minimal assumptions, while parametric approaches will have to make numerous compromises to reach the same goal. For instance, if the classical statistical framework of SPM is attempted for this study, we would first need to estimate the standard error of FA by an analytical formula (Anderson, 2001; Chang et al., 2007) with approximations from matrix perturbations and assumptions of known SNR or non-parametrically (Chung et al., 2006b). Then, we would need to create t-statistic maps assuming that FA probabilistic distribution can be approximated by a Gaussian distribution. Finally, for multiple comparison corrections, Gaussian random field theory (Friston et al., 1991; Worsley et al., 1992) can be used to investigate the distribution of the maximal statistic (largest t or cluster size) within the image, with numerous assumptions (Petersson et al., 1999) that might not be satisfied in our case (such as heavy smoothing of FWHM of at least 2-3 times the voxel size). The feasibility of a parametric statistical framework for serial DTI voxel-wise analysis would be a subject of future study.

While we have presented detection of significant FA clusters, the flexibility of the DTI permutation in the test statistic θ to be tested ($\theta = \text{FA}_B - \text{FA}_A$ in this study) allows it to be easily applied to different types of problems, as long as θ can be formulated. One simple example would be the comparison of FA averaged within a ROI. A more complicated scenario would be to test whether the connectivity between two predefined ROIs was changing over time in a subject. Connectivity testing are neither whole-brain voxel-wise nor ROI based tests, and certainly not well-defined in terms of parametric statistics since these problems are very specific to DTI analysis. For permutation testing, however, tests can be carried out in exactly the same way once θ is defined. Let's assume that $\theta = N_B - N_A$ where N is the number of reconstructed tracks connecting two regions, at time point A or B. If the permutation distribution of θ is created by tracking and counting the number tracks in each of the relabeled DTI datasets, we can calculate the p-value of observing the difference in the track counts by random chance. Last, permutation testing may be used in testing statistics derived from high angular resolution diffusion-weighted MRI or q-space imaging such as difference in

generalized fractional anisotropy (GFA) or peaks of diffusion orientation diffusion function (ODF) (Tuch, 2004).

In order to properly implement PERVADE, there are a few requirements to be met. First, it is designed to be suitable for serial DTI with the same DTI protocol. If comparing data from different sets of diffusion gradients is desired, stratified permutation of DWIs might not be optimal and the permutation distribution would include the additional variability from the effect of differing gradient configuration. Permuting residuals from fitting DWIs to the diffusion tensor model, similar to the way that residual bootstrap was implemented (Chung et al., 2006a), might be an alternative that can be investigated in the future. Next, the ability of the permutation test to estimate small p-values depends on how many different labelings are possible. Suppose we have seven images (minimum for DTI) from each time point. Then, $({}_2C_1)^7=128$ different labelings exist and the smallest possible p-value is $1/128=0.0078$ (or 0.0156 for two-tailed tests due to symmetry of labeling), which is not small enough if a p-value threshold of 0.01 is desired, for instance. Fortunately, most of the recent DTI studies acquire many more than seven images, and the number of possible labelings increase dramatically with more DWIs, such as $({}_2C_1)^{14}=16,384$ for 14 images with no repeated acquisitions, or $({}_6C_3)^7 \approx 1.28 \times 10^9$ for the acquisition scheme used in this study.

It is also important to understand the limitations of PERVADE. This technique will be susceptible to artifacts that dominate the random noise, such as incomplete fat saturation. In a group comparison, the effects of artifacts will tend to be diluted due to the nature of having multiple subjects, but in a single subject study with only two datasets, image artifacts will play a more important role. For all the results presented here, we made sure that detected clusters were not the result of any possible unsuppressed fat signals at one of the time points. Another major artifact is due to cardiac pulsation, but this is believed to be less of an issue since the two time points will be affected to a similar degree. However, the cardiac artifacts could still have different effects at different time points. The fact that we do not tend to see these artifactual effects in our analyses may arise from the permutation method, which to some degree will increase the variability in the null distribution in response to time-wise random effects during the individual DTI acquisitions.

Perhaps a more challenging issue is the effect of atrophy, as described in the registration method section. Even with a perfect registration, a complete distinction of microscopic and macroscopic changes may not be easy because of the following reasons. Technically, it is conceivable that changes in the microstructure will lead to slightly different definition of the 'boundary' of specific white matter structures that guides the registration, since no imaging technique other than DTI itself provides such rich contrast within white matter for the registration. Biologically, any irreversible damage in the axon / myelin structures will likely be followed by atrophy, thus it is not surprising that the two conditions often co-exist. Detecting these effects are both of much interest, but at some point we may have to accept that in trying to detect one of these two, the other may often exist to some degree. This difficulty does not invalidate the technique of detecting biologic changes, but rather adds some uncertainty to the interpretation in terms of pure microscopic versus macroscopic changes.

In general, the statistical power of detecting biologically-driven significant clusters is somewhat unclear at this moment. Not only do we have little prior information about the effect size we want to detect, but this is made more complicated by the fact that FA in the different ranges have different uncertainties even with the same SNR of the images. Also, it is not known how similar degrees of injuries will affect FA in white matter with different magnitudes of FA and whether more than one fiber population exists within a voxel. Thus the power to detect a similar degree of injury in different white matter structures can be variable. With more information available on how pathological changes affect DTI metrics, we will be able to

estimate the requirements for the SNR of acquisition and more generally the reliability of this technique. In summary, we presented a novel statistical framework for whole brain voxel / cluster-wise analysis of single-subject serial DTI based on non-parametric permutation testing. This completely automated technique called PERVADE can be useful in detecting subtle and local DTI changes over time in a specific subject, especially in disease processes such as MS or traumatic brain injury that have heterogeneous regional distributions across different patients.

Acknowledgements

We are grateful to Daniel Handwerker and Pratik Mukherjee for helpful discussions. This study was supported by RG3240A1 from the National Multiple Sclerosis Society.

References

- Anderson AW. Theoretical analysis of the effects of noise on diffusion tensor imaging. *Magn Reson Med* 2001;46:1174–88. [PubMed: 11746585]
- Basser PJ, Mattiello J, LeBihan D. Estimation of the effective self-diffusion tensor from the NMR spin echo. *J Magn Reson B* 1994a;103:247–54. [PubMed: 8019776]
- Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. *Biophys J* 1994b; 66:259–67. [PubMed: 8130344]
- Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans Med Imaging* 1999;18:32–42. [PubMed: 10193695]
- Caramia F, Pantano P, Di Legge S, Piattella MC, Lenzi D, Paolillo A, et al. A longitudinal study of MR diffusion changes in normal appearing white matter of patients with early multiple sclerosis. *Magn Reson Imaging* 2002;20:383–8. [PubMed: 12206862]
- Castriota-Scanderbeg A, Sabatini U, Fasano F, Floris R, Fracacci L, Mario MD, et al. Diffusion of water in large demyelinating lesions: a follow-up study. *Neuroradiology* 2002;44:764–7. [PubMed: 12221449]
- Chang LC, Koay CG, Pierpaoli C, Basser PJ. Variance of estimated DTI-derived parameters via first-order perturbation methods. *Magn Reson Med* 2007;57:141–9. [PubMed: 17191228]
- Chung S, Lu Y, Henry RG. Comparison of bootstrap approaches for estimation of uncertainties of DTI parameters. *Neuroimage* 2006a;33:531–41. [PubMed: 16938472]
- Chung, S.; Pelletier, D.; Berman, JI.; Metcalf, MA.; Henry, RG. Bootstrap-based longitudinal analysis of DTI estimates (BLADE): applications in clinically isolated syndrome patients; 14th Annual Meeting of ISMRM; Seattle, WA: ISMRM. 2006b; p. 275
- Crum WR, Scahill RI, Fox NC. Automated hippocampal segmentation by regional fluid registration of serial MRI: validation and application in Alzheimer's disease. *Neuroimage* 2001;13:847–55. [PubMed: 11304081]
- Edgington, ES. Randomization tests. Marcel Dekker; New York: 1995.
- Freeborough PA, Fox NC. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Trans Med Imaging* 1997;16:623–9. [PubMed: 9368118]
- Freeborough PA, Fox NC. Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images. *J Comput Assist Tomogr* 1998;22:838–43. [PubMed: 9754126]
- Friston KJ, Frith CD, Liddle PF, Frackowiak RS. Comparing functional (PET) images: the assessment of significant change. *J Cereb Blood Flow Metab* 1991;11:690–9. [PubMed: 2050758]
- Friston KJ, Holmes A, Poline JB, Price CJ, Frith CD. Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage* 1996;4:223–35. [PubMed: 9345513]
- Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp* 1995;2:189–210.
- Grig G, Welti D, Guttmann CR, Colchester AC, Szekely G. Exploring the discrimination power of the time domain for segmentation and characterization of active lesions in serial MR data. *Med Image Anal* 2000;4:31–42. [PubMed: 10972319]
- Good, PI. Permutation, parametric and bootstrap tests of hypotheses. Springer; New York: 2005.

- Hajnal JV, Saeed N, Oatridge A, Williams EJ, Young IR, Bydder GM. Detection of Subtle Brain Changes Using Subvoxel Registration and Subtraction of Serial Mr-Images. *Journal of Computer Assisted Tomography* 1995a;19:677–691. [PubMed: 7560311]
- Hajnal JV, Saeed N, Soar EJ, Oatridge A, Young IR, Bydder GM. A Registration and Interpolation Procedure for Subvoxel Matching of Serially Acquired Mr-Images. *Journal of Computer Assisted Tomography* 1995b;19:289–296. [PubMed: 7890857]
- Heim S, Hahn K, Samann PG, Fahrmeir L, Auer DP. Assessing DTI data quality using bootstrap analysis. *Magn Reson Med* 2004;52:582–9. [PubMed: 15334578]
- Henry RG, Oh J, Nelson SJ, Pelletier D. Directional diffusion in relapsing-remitting multiple sclerosis: a possible in vivo signature of Wallerian degeneration. *J Magn Reson Imaging* 2003;18:420–6. [PubMed: 14508778]
- Holmes, AP. Statistical issues in functional brain mapping. University of Glasgow; 1994. Vol PhD
- Holmes AP, Blair RC, Watson JD, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. *J Cereb Blood Flow Metab* 1996;16:7–22. [PubMed: 8530558]
- Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Med Image Anal* 2001;5:143–56. [PubMed: 11516708]
- Jones DK. Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magn Reson Med* 2003;49:7–12. [PubMed: 12509814]
- Jones DK. The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study. *Magn Reson Med* 2004;51:807–15. [PubMed: 15065255]
- Lazar M, Alexander AL. Bootstrap white matter tractography (BOOT-TRAC). *Neuroimage* 2005;24:524–32. [PubMed: 15627594]
- Lee MA, Smith S, Palace J, Matthews PM. Defining multiple sclerosis disease activity using MRI T2-weighted difference imaging. *Brain* 1998;121(Pt 11):2095–102. [PubMed: 9827769]
- Lemieux L, Wieshmann UC, Moran NF, Fish DR, Shorvon SD. The detection and significance of subtle changes in mixed-signal brain lesions by serial MRI scan matching and spatial normalization. *Med Image Anal* 1998;2:227–42. [PubMed: 9873901]
- Liu DC, Nocedal J. On the Limited Memory Bfgs Method for Large-Scale Optimization. *Mathematical Programming* 1989;45:503–528.
- Locascio JJ, Jennings PJ, Moore CI, Corkin S. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping* 1997;5:168–193.
- Manly, BfJ. Randomization, bootstrap, and Monte Carlo methods in biology. Chapman & Hall/CRC; Boca Raton, FL: 2007.
- Meier DS, Guttman CR. Time-series analysis of MRI intensity patterns in multiple sclerosis. *Neuroimage* 2003;20:1193–209. [PubMed: 14568488]
- Mori S, Crain BJ, Chacko VP, van Zijl PC. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Ann Neurol* 1999;45:265–9. [PubMed: 9989633]
- Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 2002;15:1–25. [PubMed: 11747097]
- Oreja-Guevara C, Rovaris M, Iannucci G, Valsasina P, Caputo D, Cavarretta R, et al. Progressive gray matter damage in patients with relapsing-remitting multiple sclerosis: a longitudinal diffusion tensor magnetic resonance imaging study. *Arch Neurol* 2005;62:578–84. [PubMed: 15824256]
- Pajevic S, Basser PJ. Parametric and non-parametric statistical analysis of DT-MRI data. *J Magn Reson* 2003;161:1–14. [PubMed: 12660106]
- Petersson KM, Nichols TE, Poline JB, Holmes AP. Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. *Philos Trans R Soc Lond B Biol Sci* 1999;354:1261–81. [PubMed: 10466150]
- Poline JB, Worsley KJ, Evans AC, Friston KJ. Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage* 1997;5:83–96. [PubMed: 9345540]
- Qiu D, Leung LH, Kwong DL, Chan GC, Khong PL. Mapping radiation dose distribution on the fractional anisotropy map: applications in the assessment of treatment-induced white matter injury. *Neuroimage* 2006;31:109–15. [PubMed: 16448821]

- Rey D, Subsol G, Delingette H, Ayache N. Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis. *Med Image Anal* 2002;6:163–79. [PubMed: 12045002]
- Rocca MA, Cercignani M, Iannucci G, Comi G, Filippi M. Weekly diffusion-weighted imaging of normal-appearing white matter in MS. *Neurology* 2000;55:882–4. [PubMed: 10994017]
- Roland PE, Levin B, Kawashima R, Akerman S. Three-dimensional analysis of clustered voxels in O-butanol brain activation images. *Human Brain Mapping* 1993;1:3–19.
- Rovaris M, Gass A, Bammer R, Hickman SJ, Ciccarelli O, Miller DH, et al. Diffusion MRI in multiple sclerosis. *Neurology* 2005;65:1526–32. [PubMed: 16301477]
- Schwartzman A, Dougherty RF, Taylor JE. Cross-subject comparison of principal diffusion direction maps. *Magn Reson Med* 2005;53:1423–31. [PubMed: 15906307]
- Sdika M. A fast non rigid image registration with constraints on the Jacobian using large scale constrained optimization. Submitted to IEEE on Medical Imaging. 2007
- Skare S, Li T, Nordell B, Ingvar M. Noise considerations in the determination of diffusion tensor anisotropy. *Magn Reson Imaging* 2000;18:659–69. [PubMed: 10930775]
- Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp* 2002;17:143–55. [PubMed: 12391568]
- Smith SM, De Stefano N, Jenkinson M, Matthews PM. Normalized accurate measurement of longitudinal brain change. *J Comput Assist Tomogr* 2001;25:466–75. [PubMed: 11351200]
- Smith SM, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols TE, Mackay CE, et al. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* 2006;31:1487–505. [PubMed: 16624579]
- Studholme C, Cardenas V, Maudsley A, Weiner M. An intensity consistent filtering approach to the analysis of deformation tensor derived maps of brain shape. *Neuroimage* 2003;19:1638–49. [PubMed: 12948718]
- Thompson PM, Hayashi KM, de Zubicaray G, Janke AL, Rose SE, Semple J, et al. Dynamics of gray matter loss in Alzheimer's disease. *J Neurosci* 2003;23:994–1005. [PubMed: 12574429]
- Tuch DS. Q-ball imaging. *Magn Reson Med* 2004;52:1358–72. [PubMed: 15562495]
- Wang D, Doddrell DM. MR image-based measurement of rates of change in volumes of brain structures. Part I: method and validation. *Magn Reson Imaging* 2002;20:27–40. [PubMed: 11973027]
- Werring DJ, Brassat D, Droogan AG, Clark CA, Symms MR, Barker GJ, et al. The pathogenesis of lesions and normal-appearing white matter changes in multiple sclerosis: a serial diffusion MRI study. *Brain* 2000;123(Pt 8):1667–76. [PubMed: 10908196]
- Whitcher B, Tuch DS, Wisco JJ, Sorensen AG, Wang L. Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. *Hum Brain Mapp*. 2007
- Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab* 1992;12:900–18. [PubMed: 1400644]
- Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 2001;20:45–57. [PubMed: 11293691]

Figure 1.

Overall schematics of permutation testing to calculate voxel-wise p-values and multiple-comparison-corrected cluster-wise p-values in a serial DTI study of a single subject. This corresponds to steps 4 and 5 in the PERVADE procedures (see human brain data analysis section on the methods). (a) DWIs (including b=0) are either kept as observed ($i=1$) or permuted ($i=2, \dots, N'$). For the simplicity of illustration, it is assumed that only seven images are acquired in each time point (A or B). Colored boundaries indicate the observed time point of each DWI. Permutation is done separately in each gradient direction (see figure 2). (b) Two FA maps are calculated from (original or reassigned) DWIs by diffusion tensor processing. (c) Two FA maps are subtracted to create ΔFA map. (d) All ΔFA maps (total N') represent the

map of voxel-wise permutation distribution of Δ F_A. **(e)** Voxel-wise p-value maps are estimated by comparing Δ F_A maps with the permutation distribution from (d). **(f)** Clusters are defined by thresholding p-value maps. **(g)** Sizes of the largest clusters in each cluster mask represent the permutation distribution of image-wise maximal cluster size. **(h)** All the clusters in the observed cluster mask are compared to the permutation distribution from (g) to estimate cluster-wise p-values, and only clusters $p < 0.05$ are declared significant.

Figure 2.

Detailed schematics of permutation testing to estimate voxel-wise p-values in a serial DTI study of a single subject. It is assumed that in each time point the DTI scan of seven DWIs (first DWI can be regarded as $b=0$) are repeated twice. In each DTI dataset (inside the black dotted box), diffusion-weighted signals $S(\vec{b})$ in the same row are acquired by applying the same diffusion gradients. $\vec{b}^A$ or $\vec{b}^B$ stresses that the effective gradients can be different in two time points due to the head positioning. Boxes around the signals are colored based on the diffusion gradients in the observed data while solid and dashed linestyle indicate the original time point A and B. On the right side is one possible permutation of DTI dataset. Permutation is done separately for each gradient (thus signals in the same gradient forms the exchangeability block), easily identified by the colors. For each $N'-1$ permuted and one observed DTI data, ΔFA from two time points are calculated and voxel-wise p-value of observed FA difference are estimated as the probability of more (or equal) extreme difference than the observed $\hat{\theta}$ from the null distribution of N' differences, i.e. proportion of counts outside the red dashed lines to all N' counts in the histogram of right lower corner.

Figure 3.

Quantile-quantile plots for expected p-values following the uniform distribution between 0 and 1 versus observed p-values by permutation testing under diverse simulation conditions. Blue and green symbols indicate uncorrected and corrected (permuting the rotated gradients paired with the diffusion-weighted signals and equalizing the gain factors) permutation scheme while red correspond to perfect match between the expected and the observed. (a) Directions of diffusion encoding gradients set rotated by 20 degrees each in x, y and z in one of two time points only. (b) Gradients rotated by 10 degrees. (c) MR signal gain factor larger by 10% in one of two time points. (d) Gain factor larger by 5%.

Figure 4.

Healthy volunteer. [Top row] p-value maps from the baseline (0 month) versus 0 month (same day, subject was moved in between), 1 month, and 2 month as well as FA map from the baseline. P-values are rescaled by log transformation, and the maximal intensity displayed is $p=0.03$. White matter regions (as well as the whole brain) used for masking before permutation procedures are shown in red contours. [Bottom left] Q-Q plot of estimated voxel-wise p-values in the whole brain white matter of baseline versus 2 month. [Bottom right] Histogram of baseline versus 2 month permutation distribution of image-wise maximal cluster size from $N=1000$ permutation trials. Critical value was 11 (i.e. probability of maximal cluster size ≥ 11 voxels was around 0.05). The largest cluster in the observed p-value map was 8 ($p=0.312$) thus no clusters were declared significant.

Figure 5.

Patient #1. [Top row] Registration of the baseline FA map to the reference FA map (10 months) by linear registration only and non-linear registration (after linear registration). Images are magnified to splenium of corpus callosum and nearby structures. Red contour denotes the border defined in the reference image. [Bottom row] P-value maps for the same anatomic structure either by linear or non-linear registration for baseline versus 10 months or 16 months. Significant clusters detected by PERVADE are displayed in the smaller embedded boxes.

Figure 6.

Patient #2. [Top row] $b=0$ images at different time points showing lesions around the left posterior horn of the lateral ventricle emerging and fading over time. Images are smoothed to the same degree as PERVADE and contrast / level were adjusted for the best visibility of lesions. Yellow contours demarcate the regions of the new lesions at the time of the appearance. [Middle row] Detected significant clusters for baseline versus subsequent time points. [Bottom row left] Permutation distribution of ΔFA as well as observed ΔFA for three voxels (NAWM and not within a significant cluster, NAWM and within a significant cluster, and core of a new lesion). Locations of these voxels are shown as well. [Bottom row right] 3D rendering of the tracks (light green), significant clusters (magenta) and lesions visible at $b=0$ images (cyan). Note that significant clusters are also present within the lesions though they are obscured by rendered lesions.

Figure 7. Patient #3. [Left column] $b=0$ images of baseline versus 3 months as well as detected clusters. Arrowheads and arrows indicate the lesion fading over time and detected NAWM cluster each. [Middle column] ΔFA , p-values and significant clusters in a sagittal slice (viewed from left) including these clusters. [Right column] 3D rendering of the tracks (light green), cluster within the lesion (orange) and cluster in the NAWM (blue). Cluster-wise p-values are shown as well.