Reproducibility of tract segmentation between sessions using an unsupervised modelling-based approach

Jonathan D. Clayden[1†], Amos J. Storkey[2], Susana Muñoz Maniega[3] & Mark E. Bastin[4]

[1]Institute of Child Health, University College London; [2]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh; [3]Division of Clinical Neurosciences, University of Edinburgh; [4]Medical and Radiological Sciences (Medical Physics), University of Edinburgh. [†]*Corresponding author: j.clayden@ucl.ac.uk*

This work describes a reproducibility analysis of scalar water diffusion parameters, measured within white matter tracts segmented using a probabilistic shape modelling method. In common with previously reported neighbourhood tractography (NT) work, the technique optimises seed point placement for fibre tracking by matching the tracts generated using a number of candidate points against a reference tract, which is derived from a white matter atlas in the present study. No direct constraints are applied to the fibre tracking results. An Expectation–Maximisation algorithm is used to fully automate the procedure, and make dramatically more efficient use of data than earlier NT methods. Within-subject and between-subject variances for fractional anisotropy and mean diffusivity within the tracts are then separated using a random effects model. We find test–retest coefficients of variation (CVs) similar to those reported in another study using landmark-guided single seed points; and subject to subject CVs similar to a constraint-based multiple ROI method. We conclude that our approach is at least as effective as other methods for tract segmentation using tractography, whilst also having some additional benefits, such as its provision of a goodness-of-match measure for each segmentation.

Introduction

A rapidly accumulating clinical literature based on the technique of diffusion magnetic resonance imaging (dMRI; see Le Bihan, 2003) is lending weight to the proposition that the brain's white matter fasciculi may be be detrimentally affected in a broad spectrum of pathological scenarios. The development of diffusion tensor imaging (Basser *et al.*, 1994), and of derived measures such as fractional anisotropy (Basser & Pierpaoli, 1996), has provided tools for gaining insight into the microstructural properties of white matter *in vivo*. Clinical applications for these tools include a range of white matter diseases such as multiple sclerosis and Alzheimer's disease (Horsfield & Jones, 2002), as well as psychiatric disorders like schizophrenia and depression (Lim & Helpern, 2002). Some of these are thought to be caused, at least partly, by a breakdown in the connective efficacy of white matter. Such pathologies are known as *disconnection syndromes*, a denomination due to Geschwind (1965a,b).

Group contrast analysis is extremely important in clinical studies of white matter integrity, with a patient group of interest typically being compared against a matched control group. Whilst analysis techniques such as the recently developed tract-based spatial statistics method (Smith *et al.*, 2006) are useful for examining the white matter of the whole brain in the absence of spatially localised hypotheses, it is often desirable to focus on specific tracts of interest and thereby improve statistical power.

There is usually substantial variability in dMRI-visible white matter characteristics even between normal individuals, due to imaging noise and genuine biological disparity; but uncontrolled sources of variability within and between groups need to be minimised if any contrast is not to be masked or exaggerated. Tract segmentation is itself a potentially major source of variability, and the reproducibility of measures calculated under a particular segmentation method therefore need to be assessed.

Perhaps the simplest method for searching for tract-specific differences between populations involves manually superimposing regions of interest (ROIs) with fixed dimensions onto an MRI image with high grey matter–white matter contrast. Fractional anisotropy (FA), mean diffusivity (MD), and potentially other measures can then be averaged within these regions and compared statistically. This approach was employed in many of the first clinical comparative studies that used diffusion tensor imaging (e.g. Ellis *et al.*, 1999; Jones *et al.*, 1999a), and it is not yet obsolete.

The manual placement of ROIs is a simple but incomplete, time consuming and subjective segmentation technique; and owing to the complex shapes of most tracts, complete segmentation by hand is extremely difficult. The advent of tractography techniques (Basser *et al.*, 2000; Conturo *et al.*, 1999; Jones *et al.*, 1999b; Mori *et al.*, 1999), which reconstruct the trajectories of white matter structures algorithmically, provides an alternative approach to segmentation. Although development of more sophisticated tractography methods continues apace, the issues of initialisation and constraint are likely to remain crucial if consistent segmentation across subjects is to be achieved. Tractography algorithms are typically initialised using a seed point or region, from which the tract reconstruction begins, but placing seed points by hand reintroduces subjectivity into the results. Ciccarelli *et al.* (2003) found test–retest coefficients of variation (CVs) of 5.0–7.1% for FA across three tracts, based on landmark-guided seed points. More recently, Heiervang *et al.* (2006) have shown that a "multiple ROI" approach, in which a number of constraints are imposed on a tractography algorithm so that only pathways which follow an expected trajectory are retained, can produce less downstream variability. Over a set of five tracts, the authors obtained test–retest CVs of 1.3–4.3% for FA and 1.0–2.1% for MD.

The multiple ROI method can be thought of as an intermediate stage between manual ROI placement and unconstrained tractography. The form of tract trajectory expected by the observer is used to form constraints for an otherwise automatic segmentation process, rather than being themselves used for segmentation directly. Heiervang *et al.* (2006) use "termination" ROIs as well as "waypoint" regions, meaning that they define rules for truncating included streamlines whilst excluding others entirely from the segmentation. An alternative, and less direct, form of constraint is provided by "neighbourhood tractography" (NT), in which a single seed point is selected in each individual brain volume based on the similarity of the tract it generates to a predefined reference tract (Clayden *et al.*, 2006). The evaluation of similarity is performed algorithmically, based on the topological characteristics of the candidate tract relative to the reference. A probabilistic model-based method based on the NT principle has recently been described, in which an explicit model of the topological variability of equivalent tracts between subjects is used to establish candidate tract plausibility (Clayden *et al.*, 2007). This approach has the benefit of giving a clear probabilistic significance to the outcome of the tract similarity algorithm, and of giving an explicit indication of the goodness-of-match between the best matching candidate tract and the reference—something which is not provided by multiple ROI methods.

Although the tract shape modelling approach has been previously shown to be successful for cross-subject tract segmentation, the algorithmic framework used was a "supervised learning" one, in which the model must first be "trained", increasing the data requirements of the procedure. In this work we introduce a refinement of the model-based NT method which makes considerably more efficient use of data by removing the need for separate training; and then investigate the reproducibility of the method amongst a set of major fasciculi.

Methods

The probabilistic NT approach to tractography aims to maximise consistency of segmentation by optimising the initialisation of a fibre tracking algorithm. A reference tract, independent of the main data sets, first needs to be created for use as a guide to the expected topology of each tract of interest. Then, for each individual brain volume, a two phase process is applied. During the "matching phase", a set of candidate tracts are created by seeding the tractography algorithm within a small region of the brain volume, and the results are algorithmically assessed as plausible counterparts to the reference tract. In the "segmentation phase", the tractography algorithm is seeded again at the point which generated the most likely match to the reference. We outline these steps in more detail below.

In this work, all tracts are represented as uniform cubic B-spline curves during the matching phase. The comparison itself is performed automatically using a probabilistic model. In the segmentation phase the tract is represented as a thresholded and binarised visitation map, within which FA and MD are averaged. We then assess the variability of these measures, which are commonly used in group contrast work, using a statistical random effects model. The tractography algorithm used for both phases was the probabilistic BEDPOST/Probtrack algorithm (Behrens *et al.*, 2003).

Finally, with a view to reducing run times for the technique, we investigate the effect of using fewer probabilistic streamlines to characterise the tract shape and establish its plausibility as a match.

Data acquisition and preprocessing

Eight healthy, right-handed volunteer subjects (four male and four female, mean age 31.9 ± 5.3 years) underwent a dMRI protocol on three separate occasions over a period of no more than two months. Scans were performed on a GE Signa LX 1.5 T clinical system (GE Healthcare, Milwaukee, Wis., USA), using a self-shielding gradient set with maximum gradient strength of 33 mT m^{-1}, and standard "birdcage" quadrature head coil. Echo-planar diffusion weighted images were acquired for an isotropic set of 64 noncollinear directions (Jones *et al.*, 2002), using a weighting factor of b = 1000 s mm^{-2}; along with seven T_2-weighted (b = 0) volumes. 57 contiguous slices of width 2 mm were imaged, using a field of view of 256 × 256 mm and 128 × 128 voxel acquisition matrix, for a final image resolution of 2 × 2 × 2 mm. Echo time was 78 ms and repetition time was 17 s per volume, producing a total scan time of approximately 20 min. This dMRI protocol is very similar to that applied by Heiervang *et al.* (2006). The local ethics committee approved the study and informed consent was obtained from each subject.

Each data set was skull-stripped (Smith, 2002) and corrected for eddy-current induced distortions using FSL tools (FMRIB, Oxford, UK). The FSL software library was also used to fit a diffusion tensor at each brain voxel, and calculate values of FA and MD.

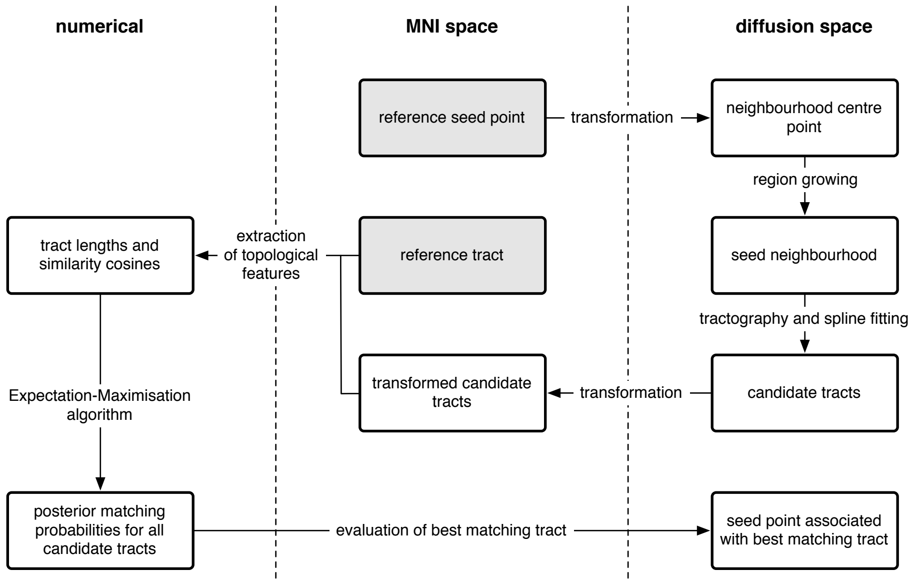

numerical
MNI space
diffusion space
reference seed point
transformation
neighbourhood centre point
region growing
tract lengths and similarity cosines
extraction of topological features
reference tract
seed neighbourhood
tractography and spline fitting
transformed candidate tracts
transformation
candidate tracts
Expectation-Maximisation algorithm
posterior matching probabilities for all candidate tracts
evaluation of best matching tract
seed point associated with best matching tract

Figure 1: Graphical illustration of the matching process for each tract. The grey boxes indicate the initial material required for the process: a reference tract and associated seed point.

Reference tracts

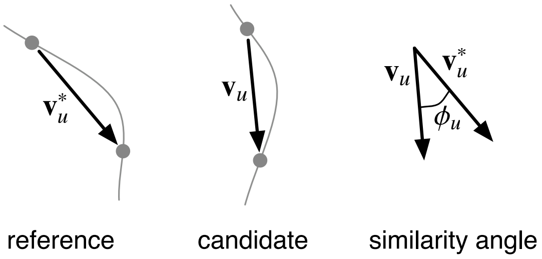
The method for creating reference tracts for use in this work is based on that described by Muñoz Maniega *et al.* (2008). Our aim is to construct standardised B-spline curves representing the typical trajectory of each fasciculus, in a manner that is independent of the data used to check reproducibility.

Regions of the brain representing the left and right cingulum bundles (CBs), left and right corticospinal or pyramidal tracts (PTs) and the genu of the corpus callosum were extracted directly from a digital human white matter atlas, which was created using tractography in a population of 28 young healthy adults (Hua *et al.*, 2008). Each region, in MNI standard space (Evans *et al.*, 1993), was then thinned to form a core pathway of single voxel thickness; and these voxel locations were used to fit the spline curve, after applying a small random perturbation to move them off the regular grid. Care was taken to ensure the MNI space seed points associated with each curve avoided regions of the brain where crossing fibre pathways might be expected.

Candidate tracts

The matching phase of our approach involves the creation of a set of candidate tracts, whose similarity to the reference tract is evaluated using the modelling methods described below. Each subject's $b = 0$ image was registered to the MNI standard brain template using the FLIRT linear registration algorithm (Jenkinson & Smith, 2001), in order to establish a transformation between MNI space and the subject's diffusion space. The reference seed point was transformed into diffusion space using this transformation, and used as the centre of a $7 \times 7 \times 7$ voxel neighbourhood, which supplied the seed points for the candidate tracts. Each seed voxel with an FA of at least 0.2 was

Figure 2: Illustration of the similarity angles derived from interknot vectors in the reference and candidate tracts.



passed to the tractography algorithm, which generates 5000 probabilistic streamlines for each seed. A B-spline curve is subsequently fitted to the spatial median of each of these sets of streamlines, and the knot points of the splines are transformed back into MNI space for subsequent comparison with the reference. The entire process for each brain volume is laid out graphically in Fig. 1.

Tract shape modelling

Our approach to modelling the variability in shape and length between a set of tracts drawn from a range of subjects here is closely related to that described in Clayden *et al.* (2007). Here, as there, we are interested in characterising the probability that any given tract, labelled i, represents the best match amongst a number of candidates to the reference tract. However, it was previously necessary to choose by hand a number of examples of good matches to the reference—in effect, additional reference tracts—in order to provide training data for the model. By contrast, in the present work we apply an Expectation–Maximisation (EM) algorithm (Dempster *et al.*, 1977) to fit the model whilst concurrently using it to distinguish between matching and nonmatching tracts. We have previously proposed this approach in Clayden (2008).

As in our previous work, we represent candidate tract trajectories using B-spline curves. The knot points of these curves are arranged such that one coincides with the seed point, and the remainder are notionally split into “left” and “right” sides, indexed by distance from the seed knot. (The actual directions corresponding to “left” and “right” here will depend on the orientation of the reference tract near to the seed point.) The reference tract has a left length, L^*_1, representing the number of knots on the left side; and a right length, L^*_2. The equivalent lengths in candidate tract i are denoted L^i_1 and L^i_2. The shape similarity between the two tracts is based on the angles, ϕ^i_u, between interknot vectors in each (see Fig. 2). We then model

$$s^i_u = \cos \phi^i_u \tag{1}$$

as well as the lengths. By convention, $u < 0$ on the left side of the tract and $u > 0$ on the right side.

We introduce a variable z^i which indicates whether tract i represents the best matching tract ($z^i = 1$) or not ($z^i = 0$), subject to the restriction that only one tract can be the best match. Hence, if $z^1 = 1$, say, then $z^i = 0$ for all other values of i. We additionally allow the special value $i = 0$ to indicate the case where no candidate tract is a suitable match to the reference—the most likely reason for this being poor data quality.

The model that we use for tract shape and length is then dependent on the value of this variable. Given the data vector $\mathbf{d}^i$ describing tract i, and a set of modelling parameters which affect the shapes of the distributions, $\mathbf{A} = (\alpha_u)$, we use the likelihood

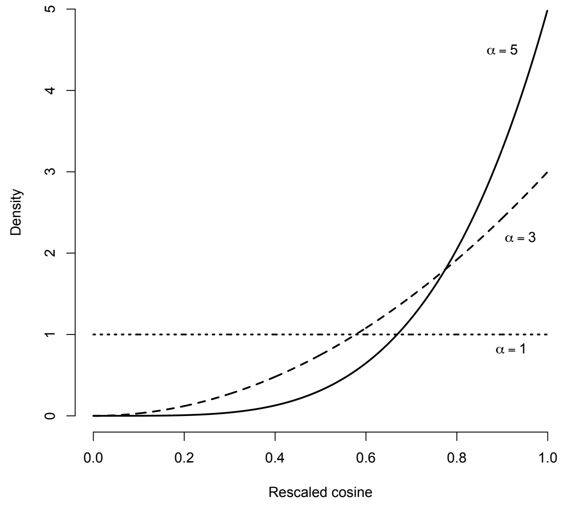


Figure 3: Probability density functions for the beta distribution with $\beta = 1$ and three different values of α. Note that the larger the value of the α, the more emphatically the probability mass is concentrated around large cosines—and therefore small deviations from the reference tract. The beta distribution is equivalent to the uniform distribution over $[0, 1]$ when $\alpha = \beta = 1$.

function

$$\begin{aligned} P(\mathbf{d}^i \,|\, \mathbf{A}, z^i = 1) = P(L_1^i \,|\, L_1^*, z^i = 1)\, P(L_2^i \,|\, L_2^*, z^i = 1) \\ \times \prod_{u=1}^{\check{L}_1^i} P(s_{-u}^i \,|\, \alpha_u, z^i = 1) \prod_{u=1}^{\check{L}_2^i} P(s_u^i \,|\, \alpha_u, z^i = 1)\ , \end{aligned} \tag{2}$$

for $z^i = 1$. Here, $\check{L}_1^i = \min\{L_1^i, L_1^*\}$, and equivalently for $\check{L}_2^i$; and

$$\frac{s_v^i + 1}{2} \sim \mathrm{Beta}(\alpha_u, 1) \quad \text{for } v \in \{u, -u\}\ . \tag{3}$$

The left hand side of Eq. (3) takes this form because the similarity cosine values need to be rescaled into the interval [0,1] over which the beta distribution is defined. When $z^i = 0$ we use the uninformative likelihood function

$$P(\mathbf{d}^i \,|\, z^i = 0) = P(L_1^i \,|\, z^i = 0)\, P(L_2^i \,|\, z^i = 0) \left(\frac{1}{2}\right)^{\check{L}_1^i + \check{L}_2^i}\ . \tag{4}$$

For both the matching and nonmatching likelihood functions, the length distributions are modelled as multinomial, subject to some maximum allowable length value. However, the shape distribution is uniform in Eq. (4), and therefore the direction of a nonmatching candidate tract is completely unconstrained by the reference. The uniform distribution is a special case of the beta distribution used in Eq. (3), which appears when $\alpha_u = 1$. For all $\alpha_u > 1$, the smallest angular deviations from the reference tracts are considered most likely; and larger values of α_u imply more “concentrated” distributions of orientations around the direction of the reference (see Fig. 3). Further details of these distributions are given in Appendix A.

To establish the best matching tract given a data set consisting of the shape and length information for candidate and reference tracts, D, we need to characterise the posterior distribution $P(z^i \mid D)$ for each tract, i. This, in turn, requires an estimate for the parameter vector, $\mathbf{A}$. The EM method provides a framework for performing these two tasks iteratively, using the whole data set to refine the model and find matches in turn (see Appendix B). Consequently, no separate training data are required for fitting the model parameters.

Completing the matching phase of the NT process is then simply a matter of extracting the seed point with the greatest posterior probability of matching the reference. This is used for the segmentation phase, which involves generating a visitation map using the tractography algorithm, thresholding the voxel data at the 1% level, and then binarising the resulting image to produce a mask.

This whole modelling framework has been implemented using the R language (R Development Core Team, 2008), as part of the TractoR package for fibre tracking and analysis (http://code.google.com/p/tractor/).

Variance components analysis

The value of some measure, such as FA, averaged within a segmented tract of interest, may be assumed to represent a sample from an unknown distribution over all comparable measurements. Multiple scans of a single individual, or of different individuals with similar ages and clinical statuses, would be expected to yield similar measurements.

The simple two-level random effects model described here separates out the variances due to between-subject and within-subject effects. We model the measurement of a metric in the jth scan of the ith subject, m_{ij}, according to

$$m_{ij} = \mu + \delta_i + \varepsilon_{ij} \; , \tag{5}$$

where μ is the underlying population mean and

$$\delta_i \sim N(0, \sigma_b^2) \qquad\qquad \varepsilon_{ij} \sim N(0, \sigma_w^2) \; . \tag{6}$$

The between-subject variance, σ_b^2, is thus separated from the within-subject variance, σ_w^2; although both are assumed to be independent of the subject. Sources of within-subject variance might include differences in noise characteristics, magnetic field properties and subject placement; while between-subject variance captures genuine microstructural differences between individuals. The theory of this kind of random effects model has been well characterised in the statistics literature (see in particular Laird & Ware, 1982).

Given a full set of measurements for a particular tract, the model parameters, $\{\mu, \sigma_b, \sigma_w\}$, were fitted using the "nlme" package for R, version 3.1, using the restricted maximum likelihood method (http://stat.bell-labs.com/NLME/; see also Bates & Pinheiro, 1998).

Impact of using fewer streamlines

To test the effect of reducing the number of streamlines in the tract matching phase, we recalculated B-splines using 1000, 500, 100, 50, 10 and 5 streamlines for two example fasciculi (left CB and genu). Fixing the model parameters to those chosen using the original 5000 streamlines in each case, we recalculated log-likelihoods for each candidate tract using Eq. (2), and computed a correlation coefficient between each set of likelihood values and those generated using B-splines based on 5000 streamlines.

Metric	Tract	Mean (μ)	Interscan CV, % (σ_w/μ)	Intersubject CV, % (σ_b/μ)
FA	CB, left	0.376	7.00	4.92
	CB, right	0.374	6.79	6.63
	CB, both	0.375	7.04	5.55
	CC, genu	0.379	3.93	9.16
	PT, left	0.476	7.35	3.83
	PT, right	0.478	4.83	5.01
	PT, both	0.477	6.30	4.22
MD, $\text{mm}^2\ \text{s}^{-1}$ $(\times 10^{-3})$	CB, left	0.786	3.37	3.88
	CB, right	0.769	3.28	3.65
	CB, both	0.777	3.59	3.66
	CC, genu	0.910	6.95	4.49
	PT, left	0.809	6.72	6.11
	PT, right	0.785	5.20	4.44
	PT, both	0.797	6.68	4.62

Table 1: Means and CVs of FA and MD in each tract of interest, as estimated using our variance components analysis.

Since the log-likelihoods are not expected to be normally distributed, we used a rank-based correlation coefficient (Spearman's ρ). We then examined how the correlation decreases with the number of streamlines used.

Results

Fig. 4 demonstrates the variability in segmentation across the group of subjects, for each tract of interest. In each case, the segmented tract from the first scan of each individual was transformed into standard space, and overlaid to form a group map. It can be seen that the segmented tracts closely follow the trajectories of the reference tracts, which are superimposed and coloured green.

A simple scatter plot of FA against MD, across all scans, is shown in Fig. 5. It is immediately apparent that the CBs, genu and PTs form three distinctive clusters when plotted in this way. Relative to the other tracts, the CBs have low FA and low MD; genu has low FA and high MD; and the PTs have high FA and low MD. This figure therefore provides evidence that values of these quantitative parameters measured using our segmentation method are specific to each tract.

Table 1 shows the results of the variance components analysis for FA and MD in each tract. The estimated mean, μ, is given as an absolute value, whilst the variance parameters are expressed in terms of CVs, σ_w/μ and σ_b/μ, to facilitate comparison between our results and those of previous studies. In the cases of the two bilateral tracts, the left and right values are given separately, and then a third set of parameters was estimated by combining the left and right data together, treating them as repeated measurements. These estimated CVs, along with their 95% confidence intervals, are also shown in Fig. 6. It can be seen that the confidence intervals are invariably wider between subjects (represented with lighter bars), than within subjects (darker bars).

To confirm that our random effects model was appropriate in each case, we tested the fit residuals for normality using the Shapiro–Wilk test. In no case was the null hypothesis of a normal source distribution rejected ($p > 0.05$).

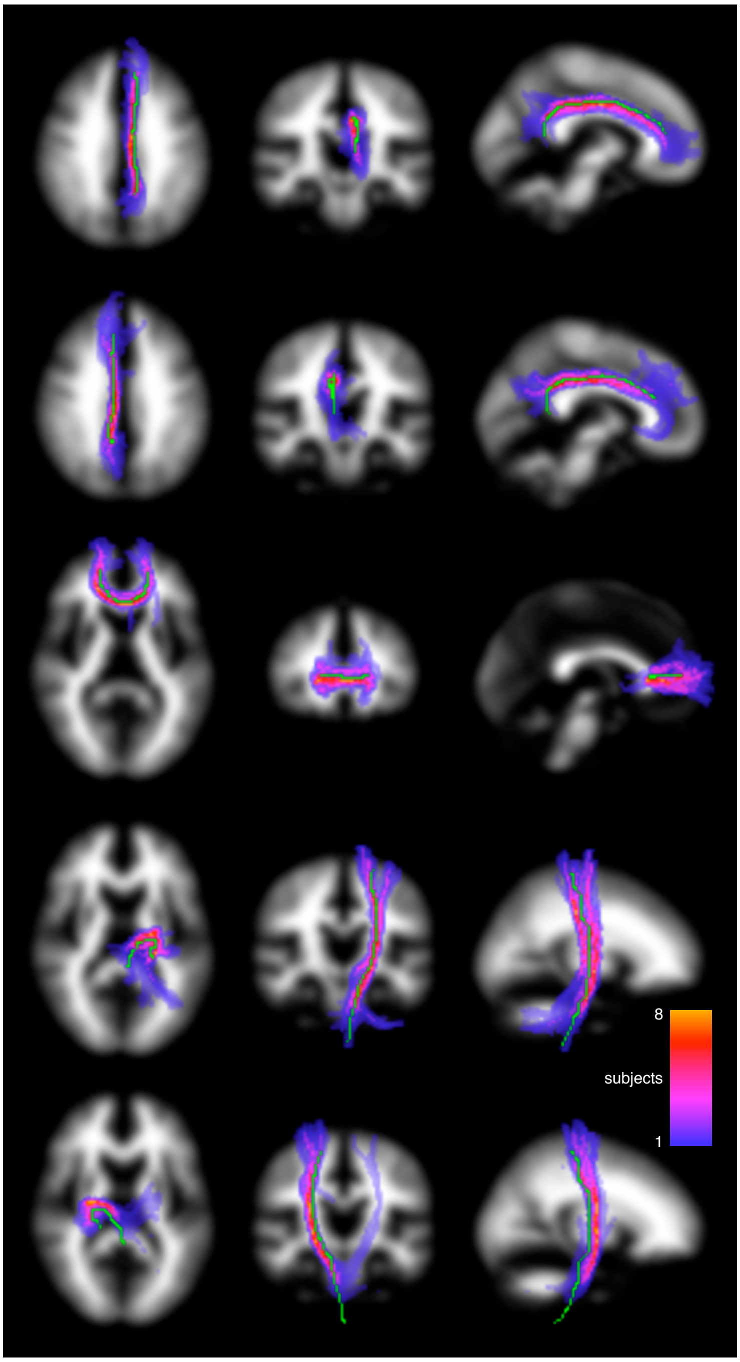


Figure 4: Standard space group maps of the segmented tracts of interest, overlaid on an MNI white matter map. From top to bottom, the tracts are: left CB, right CB, genu, left PT and right PT. The first, second and third columns show axial, coronal and sagittal maximum intensity projections respectively. The reference tract is superimposed in green in each case.

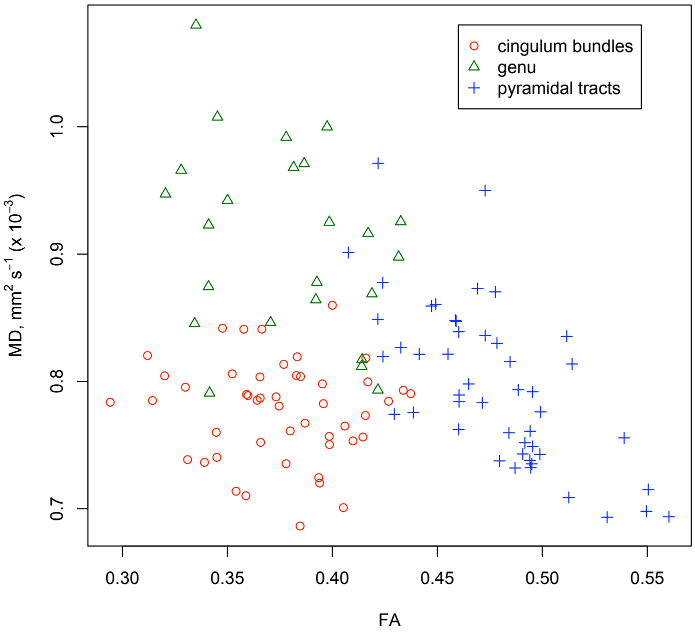


Figure 5: Scatter plot of FA against MD within all segmented tracts. The three structures appear to form three nearly separate clusters.

Metric	**Tract**	**Mean** (μ)	**Interscan CV, %** (σ_w/μ)	**Intersubject CV, %** (σ_b/μ)
FA	CB, left	0.378	7.80	7.40
	CB, right	0.364	10.00	10.80
	CB, both	0.371	9.51	8.66
	CC, genu*	0.385	6.99	9.58
	PT, left	0.478	6.63	4.16
	PT, right	0.451	6.38	2.77
	PT, both	0.465	6.78	4.18
MD, mm^2 s^{-1} ($\times 10^{-3}$)	CB, left	0.784	4.54	2.79
	CB, right	0.776	5.60	2.98
	CB, both	0.780	4.96	3.08
	CC, genu*	0.894	5.39	2.71
	PT, left	0.804	7.51	6.53
	PT, right	0.809	6.11	4.89
	PT, both	0.806	6.65	5.92

* *These figures are based on data from 7 subjects*

Table 2: Means and CVs of FA and MD in each tract of interest. In this case the tracts were segmented using the heuristic neighbourhood tractography method described in Clayden *et al.* (2006).

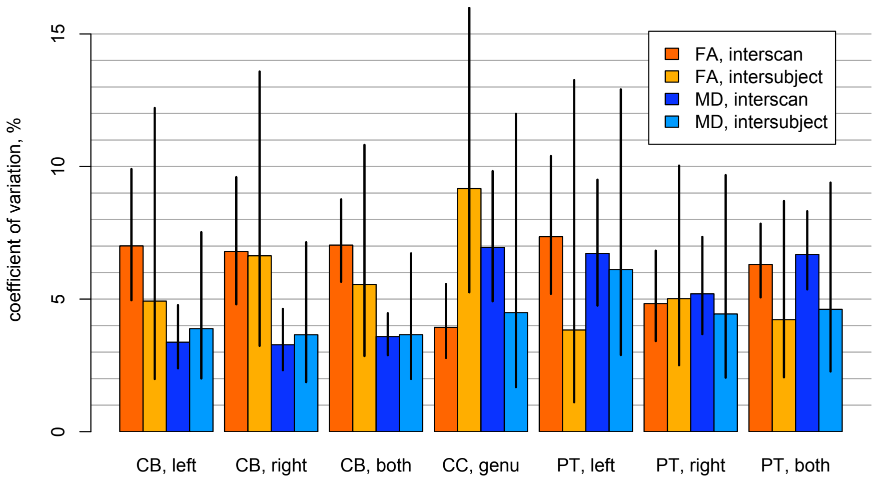


Figure 6: Bar plot of the estimated coefficients of variation, along with 95% confidence intervals, for FA and MD in each tract of interest. The orange bars—the first and second in each group of four—relate to FA, and the blue bars to MD. The darker bars—the first and third in each group—indicate within-subject CVs, while the lighter bars show between-subject CVs.

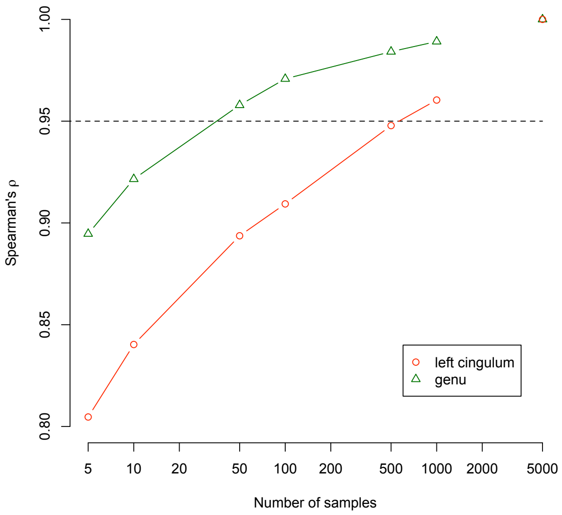


Figure 7: Graphical depiction of the effect of changing the number of streamlines used to characterise the shape of each tract, quantified by the value of Spearman's ρ statistic calculated between the log-ratios for each sample size and those with sample size 5000.

In order to provide a direct comparison with our earlier, “heuristic” NT method (Clayden *et al.*, 2006), we selected tracts from within the same neighbourhoods using that method and reran the variance components analysis. Results are shown in Table 2. Across the five tracts, the average interscan CV was 6.70%, against 5.54% for the probabilistic method. The average intersubject CV was 5.46% rather than 5.21%. The differences in estimated mean values were small, averaging 0.0054 for FA and -0.0016×10^{-3} for MD, relative to the values in Table 1. Due to an extraneous segmentation, one subject’s data for the genu tract had to be removed to avoid computational problems.

Finally, Fig. 7 shows the effect of reducing the number of streamlines on the log-likelihoods assigned to each candidate tract. As expected, the correlation with the original log-likelihoods, calculated using 5000 streamlines, drops as the sample size decreases. The rate of fall-off differs between the two tracts, but neither is precipitous; and it can be seen that a fivefold reduction in the sample size, to 1000, produces a ρ value well above 0.95 (marked with a dashed line) in both cases. Hence, appropriate selection of matching tracts should be scarcely affected by such a reduction, while calculation times will be shortened substantially: in this case from approximately 4 hr to 1 hr per brain volume.

Discussion

In this work we have brought together previous ideas on tract modelling and reference tract generation; and described an extension to the probabilistic NT method which makes the process of seed point optimisation fully “unsupervised”, removing the need for separate training data. Maximum use can therefore be made of any given data set, without losing the potential advantages of training a tract matching model on the specific population of interest. In addition, we have shown that the EM algorithm described here, along with the set of standardised reference tracts we have created, can be used to obtain downstream reproducibility in FA and MD measurements that is similar to that obtained by Ciccarelli *et al.* (2003) using labour-intensive hand placement of seed points with guidance from anatomical landmarks. A direct comparison between techniques is of course not possible without using exactly the same data set in each case, but we have endeavoured to choose a subject group and dMRI protocol which matches earlier work closely.

The range of test–retest CVs for FA obtained across the tracts we studied, 3.9–7.4%, compares favourably with the 5.0–7.1% range reported by Ciccarelli *et al.* (2003). This suggests that the probabilistic NT approach applied here is at least as effective as a human observer at choosing appropriate seed points, whilst also being more objective. Our estimated within-subject CVs were, however, larger than those obtained by Heiervang *et al.* (2006), using a constrained multiple ROI approach. There could be a number of reasons for this, but a significant contributing factor may be the truncation that the authors’ constraints effected on each tract, thereby preventing them from entering cortex and brain stem regions. Since FA, in particular, is less reliable in crossing-fibre and grey matter regions, this constraint is likely to remove a significant source of variance. It is interesting to note, then, that our between-subject CVs are in general closely comparable to those reported in that study, with our figures being in the range 3.8–9.2% for FA, compared to 3.3–9.3%. Since the authors transform visitation and FA maps into standard space, differences at the registration stage may increase their intersubject variances.

In this work we separate within-subject and between-subject variances using a random effects model. We have used this in preference to the approach of averaging CVs calculated from various subsets of the measurements, as applied in some other work, because it is more robust and statistically well-founded. It also facilitates the calculation of confidence intervals on these variances. Nevertheless the CVs are expressed in the same terms either way, making the results comparable.

There are some specific characteristics of our results which are worth noting. Firstly, the estimated mean values of the metrics—particularly FA—were very similar in the left and right hemispheres for the two bilateral tracts that we studied. This suggests that the diffusion characteristics of the CBs and PTs in this healthy young subject group are generally symmetric; and it is therefore unsurprising to see that combining data from the two hemispheres produces tighter confidence intervals in each case (see Fig. 6). Secondly, the estimated within-subject CVs are not consistently higher or lower for FA, compared with MD; nor are they consistently lower than the between-subject CVs. For the genu, the lowest CV was for within-subject FA, whereas for the right PT it was for between-subject MD. There were, however, very similar patterns across the two bilateral pairs included in the study; and it is really not surprising that the variances should follow different patterns in different tracts. Although it may seem counterintuitive for some within-subject CVs to be higher than the corresponding between-subject CVs, it should be remembered that the within-subject variance incorporates variability from a number of sources, including image noise and subject placement; and it is quite credible for the combined effect of these sources of variance to be larger than the effect of moving between subjects in this young, healthy cohort. Moreover, since the confidence intervals are invariably wider in the between-subjects case, some of these relationships would be likely to change given a larger test population. Heiervang *et al.* (2006) did report consistently higher intersubject CVs, but this may be partly due to registration effects, as mentioned above. Since previous studies have not calculated confidence intervals, more extensive comparison is difficult. Thirdly, it is clear from Fig. 5 that the three fasciculi examined in this study cover largely distinct regions of FA–MD space. FA and MD are not mathematically orthogonal measures (Ennis & Kindlmann, 2006), they are certainly not interdependent, and this result suggests that future tract-specific work might usefully consider treating their data in these two-dimensional terms. It may also be constructive to examine the variability in these measures along the tracts, when comparing either segmentation methods or clinical populations.

The choice of fibre tracking algorithm will undoubtedly have some effect on reproducibility. It should be noted that Heiervang *et al.* (2006) and the present study used the BEDPOST/Probtrack algorithm (Behrens *et al.*, 2003), whereas Ciccarelli *et al.* (2003) used the fast marching algorithm described by Parker *et al.* (2002a,b). Significantly, neither of these algorithms is capable of resolving multiple fibre orientations within a voxel. As a result, and in common with Ciccarelli *et al.* (2003), a few aberrant branching pathways remain visible in the segmented tract regions seen in Fig. 4. In particular, the confluence of a number of fibre pathways in the pons occasionally produces a small branch along the transpontine fibres projecting into cerebellum, which is visible on the left PT group maps; and, in one case, a right PT branches into the contralateral hemisphere. Although we have not done so here—so as to facilitate comparison with previous reproducibility work—the use of a more advanced algorithm which can resolve multiple populations (e.g. Behrens *et al.*, 2007; Jansons & Alexander, 2003; Tournier *et al.*, 2004; Tuch, 2004) would be expected to improve on the results of the segmentation phase.

A related issue is thresholding. The nature of probabilistic tractography is such that some outlying streamlines are often generated, and so we have set a threshold of 1% of the total number of streamlines and removed from the segmentation all voxels through which fewer than this number of streamlines pass. This type of thresholding is standard practice in the literature, but it is problematic because of its arbitrariness. It is certain that changing the threshold will affect both the mean and the variance parameters of the distributions over FA and MD, both within and between subjects, because of the differing extents of the segmentations that would result. In the limit, at a threshold of 100%, only the seed point itself and perhaps a few other nearby voxels would be retained. It is hard to balance the trade-off in a principled way. An appealing alternative strategy might be to use the tract shape model itself to identify and remove outlying streamlines, thereby obviating the need for a voxelwise threshold. Some care would be needed to develop such an approach, however, and so this is left as future work.

We have shown that our earlier, heuristic approach to NT produces generally higher variances over FA and MD than the probabilistic approach, particularly among the test–retest values which capture variability due to the technique. The difference is not vast, but since the heuristic method involves no transformation of candidate tracts into MNI space, we would expect the gap between the two techniques to widen in the presence of pathologies which substantially affect tract shape, or of patient head rotation. In the case of pathology, the present method could be easily adapted to use nonlinear registration to maximise performance in the matching phase if required—although it should be remembered that perfect registration is *not* a theoretical requirement of the method.

In patient studies, whole brains or specific tracts may be distorted or impeded by pathological effects. If such effects are substantial enough to prevent the tractography algorithm from generating any plausible tract trajectory, our method should reflect this fact in the calculated posterior probability of no match, $P(z^0 = 1 \,|\, D)$. Otherwise, a best matching tract should be established as usual, and the final model will reflect the tendency for greater deviation from the reference tract. The method looks only for a topological best match relative to other candidate tracts in the same brain volume, so homogeneity across the data set is not a requirement. Furthermore, since our matching process makes no use of FA or MD data—except for convenience, to prethreshold candidate seed points—there is no reason to expect a bias to arise in these measurements amongst a diseased cohort. Extensive exploration of the limits of the technique in disease is a major task, however, and must be left to future work.

Our finding of closely similar matching results based on substantially fewer streamlines is helpful, particularly for larger scale studies in which run time is an important consideration; but it should not be overgeneralised. We have seen that the exact magnitude of the difference is dependent on the tract of interest, and it will also depend on the tractography algorithm being used and the quality of the data. Algorithms capable of resolving crossing fibres may reduce the fall-off even further; while relatively noisy data, or data with low angular resolution, would be expected to increase it.

In conclusion, we have described an automated model-based tract matching and segmentation procedure; and demonstrated within-subject variance similar to that obtained using manual seed point placement and between-subject variance similar to that obtained using ROI constraints—although the latter measurements have significantly wider confidence intervals. Since between-subject variance within groups is of primary interest to group contrast studies, this result suggests that our procedure is at least equally viable as ROI-based methods for this kind of work. Moreover, it is not clear how portable a set of constraint ROIs will be between data sets, particularly in the relatively elaborate combination that they are used by Heiervang *et al.* (2006); whereas we have shown, in separate recent work, that reference tracts can be successfully reused across, for example, young and old populations (Bastin *et al.*, 2008). The reference tract approach is also more flexible than ROI methods, since it does not impose hard constraints on the routes of the tracts. We would therefore argue that our approach is a helpful one for tract-specific white matter characterisation and the investigation of disconnection syndromes.

Appendix A: Probability distributions

The beta distribution is a continuous probability distribution defined over the interval [0, 1]. It has parameters α and β, and the general probability density function (p.d.f.)

$$P(x\,|\,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \ , \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function. Any positive real numbers are valid for α and β, with their exact values affecting the shape of the distribution, but in our application we fix $\beta = 1$. In this case the p.d.f. becomes simply

$$P(x\,|\,\alpha) = \alpha x^{\alpha-1} \ . \tag{8}$$

We add this constraint because we always expect larger similarity cosines to be considered the most likely under the model—the added flexibility of fitting a value for β is not required. The value of α will directly affect the magnitude of this bias towards the larger cosine values, as shown in Fig. 3.

The left and right lengths of the candidate tracts are modelled using multinomial distributions. If $x_1, x_2, \ldots, x_l$ represent the observed number of tracts with lengths 1, 2, and so on—up to the cutoff value of l—the multinomial distribution amounts to the joint probability mass function

$$P(x_1, x_2, \ldots, x_l) = \frac{N!}{x_1!x_2!\ldots x_l!} p_1^{x_1} p_2^{x_2} \ldots p_l^{x_l} \ , \tag{9}$$

where N is the total number of tracts, and p_i is the probability of observing a tract with length i. It is necessarily true that $\sum_i x_i = N$ and $\sum_i p_i = 1$; and the maximum likelihood values of p_i are given by $p_i = x_i/N$.

Appendix B: EM algorithm

Given an initial estimate for the model parameter vector, $\hat{\mathbf{A}}$, and under the assumption of a single best matching tract in each brain volume, as described in the Methods, the posterior probability that tract i is the best match is given by

$$P(z^i = 1\,|\,D) = \frac{P(z^i = 1)\,P(\mathbf{d}^i\,|\,\hat{\mathbf{A}}, z^i = 1)\prod_{j\neq i} P(\mathbf{d}^j\,|\,z^j = 0)}{P(D)} \ . \tag{10}$$

The right hand side of Eq. (10) describes the fact that if tract i is the best match, it is drawn from the matching model, Eq. (2), while all other tracts, $j \neq i$, are drawn from the nonmatching model, Eq. (4). The probability of no match among the candidates is given by

$$P(z^0 = 1 \,|\, D) = \frac{P(z^0 = 1) \prod_j P(\mathbf{d}^j \,|\, z^j = 0)}{P(D)} \,, \tag{11}$$

and the evidence is

$$P(D) = \sum_i P(z^i = 1)\, P(\mathbf{d}^i \,|\, z^i = 1) \prod_{j \neq i} P(\mathbf{d}^j \,|\, z^j = 0) + P(z^0 = 1) \prod_j P(\mathbf{d}^j \,|\, z^j = 0) \,. \tag{12}$$

We assume that each candidate tract is *a priori* equiprobable, so the prior distribution $P(z^i = 1)$ is uniform over all $i \geq 0$. This includes the prior probability of no match. The E-step of our EM algorithm involves evaluating Eqs (10) and (11).

The M-step consists of updating our estimate of the parameters for the shape distributions. We use the maximum *a posteriori* estimate given by

$$\hat{\alpha} = \arg\max_{\alpha} \{ \ln P(\alpha \,|\, \mathbf{x}) \} = \arg\max_{\alpha} \left\{ \sum_i \ln P(x^i \,|\, \alpha) + \ln P(\alpha) \right\} \,, \tag{13}$$

for relevant data, $\mathbf{x} = (x^i)$. The prior for α is an exponential distribution with mean $1/\lambda$, defined by $P(\alpha) = \lambda e^{-\lambda\alpha}$. In this work we take $\lambda = 1$. This prior will favour smaller values of α, thereby counteracting the tendency for model overfitting when there is little data available. For each α value in turn, therefore, the estimator in Eq. (13) becomes

$$\hat{\alpha}_u = \frac{-2 \sum_{i>0} P(z^i = 1 \,|\, D)}{\sum_{i>0} P(z^i = 1 \,|\, D) \, \ln x_u^i - \lambda} \,, \tag{14}$$

where

$$x_u^i = \left(\frac{s_{-u}^i + 1}{2} \right) \left(\frac{s_u^i + 1}{2} \right) \,. \tag{15}$$

It should be noted that the M-step uses similarity cosine data from across all acquired brain volumes to inform the estimate of **A**.

The E-step and M-step are repeated alternately until the algorithm converges, thereby producing a stable model and posterior distribution. We considered the algorithm converged when the log-evidence changed by less than 0.1, or when the mean change to the α_u parameters was less than 0.1, between successive iterations.

Acknowledgments

All scanning was performed at the SFC Brain Imaging Centre, University of Edinburgh. The authors are grateful to Dr Martin King for useful discussions in relation to this work.

References

Basser P., Mattiello J. & Le Bihan D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B* **103**(3):247–254.

Basser P., Pajevic S., Pierpaoli C., Duda J. & Aldroubi A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine* **44**(4):625–632.

Basser P. & Pierpaoli C. (1996). Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *Journal of Magnetic Resonance, Series B* **111**(3):209–219.

Bastin M., Piatkowski J., Storkey A., Brown L., MacLullich A. & Clayden J. (2008). Tract shape modelling provides evidence of topological change in corpus callosum genu during normal ageing. *NeuroImage* **43**(1):20–28.

Bates D. & Pinheiro J. (1998). Computational methods for multilevel modelling. *Tech. rep.*, Bell Laboratories.

Behrens T., Johansen-Berg H., Jbabdi S., Rushworth M. & Woolrich M. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage* **34**(1):144–155.

Behrens T., Woolrich M., Jenkinson M., Johansen-Berg H., Nunes R., Clare S., Matthews P., Brady J. & Smith S. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine* **50**(5):1077–1088.

Ciccarelli O., Parker G., Toosy A., Wheeler-Kingshott C., Barker G., Boulby P., Miller D. & Thompson A. (2003). From diffusion tractography to quantitative white matter tract measures: a reproducibility study. *NeuroImage* **18**(2):348–359.

Clayden J. (2008). *Comparative analysis of connection and disconnection in the human brain using diffusion MRI: New methods and applications*. Ph.D. thesis, University of Edinburgh.

Clayden J., Bastin M. & Storkey A. (2006). Improved segmentation reproducibility in group tractography using a quantitative tract similarity measure. *NeuroImage* **33**(2):482–492.

Clayden J., Storkey A. & Bastin M. (2007). A probabilistic model-based approach to consistent white matter tract segmentation. *IEEE Transactions on Medical Imaging* **26**(11):1555–1561.

Conturo T., Lori N., Cull T., Akbudak E., Snyder A., Shimony J., McKinstry R., Burton H. & Raichle M. (1999). Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences of the United States of America* **96**(18):10,422–10,427.

Dempster A., Laird N. & Rubin D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**(1):1–38.

Ellis C., Simmons A., Jones D., Bland J., Dawson J., Horsfield M., Williams S. & Leigh P. (1999). Diffusion tensor MRI assesses corticospinal tract damage in ALS. *Neurology* **53**(5):1051–1058.

Ennis D. & Kindlmann G. (2006). Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magnetic Resonance in Medicine* **55**:136–146.

Evans A., Collins D., Mills S., Brown E., Kelly R. & Peters T. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference*, vol. 3, pp. 1813–1817. IEEE.

Geschwind N. (1965a). Disconnexion syndromes in animals and man: Part I. *Brain* **88**(2):237–294.

Geschwind N. (1965b). Disconnexion syndromes in animals and man: Part II. *Brain* **88**(3):585–644.

Heiervang E., Behrens T., Mackay C., Robson M. & Johansen-Berg H. (2006). Between session reproducibility and between subject variability of diffusion MR and tractography measures. *NeuroImage* **33**(3):867–877.

Horsfield M. & Jones D. (2002). Applications of diffusion-weighted and diffusion tensor MRI to white matter diseases - a review. *NMR in Biomedicine* **15**(7-8):570–577.

Hua K., Zhang J., Wakana S., Jiang H., Li X., Reich D., Calabresi P., Pekar J., van Zijl P. & Mori S. (2008). Tract probability maps in stereotaxic spaces: Analyses of white matter anatomy and tract-specific quantification. *NeuroImage* **39**(1):336–347.

Jansons K. & Alexander D. (2003). Persistent angular structure: new insights from diffusion magnetic resonance imaging data. *Inverse Problems* **19**(5):1031–1046.

Jenkinson M. & Smith S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* **5**(2):143–156.

Jones D., Lythgoe D., Horsfield M., Simmons A., Williams S. & Markus H. (1999a). Characterization of white matter damage in ischemic leukoaraiosis with diffusion tensor MRI. *Stroke* **30**(2):393–397.

Jones D., Simmons A., Williams S. & Horsfield M. (1999b). Non-invasive assessment of axonal fiber connectivity in the human brain via diffusion tensor MRI. *Magnetic Resonance in Medicine* **42**(1):37–41.

Jones D., Williams S., Gasston D., Horsfield M., Simmons A. & Howard R. (2002). Isotropic resolution diffusion tensor imaging with whole brain acquisition in a clinically acceptable time. *Human Brain Mapping* **15**(4):216–230.

Laird N. & Ware J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**(4):963–974.

Le Bihan D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience* **4**(6):469–480.

Lim K. & Helpern J. (2002). Neuropsychiatric applications of DTI - a review. *NMR in Biomedicine* **15**(7-8):587–593.

Mori S., Crain B., Chacko V. & van Zijl P. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology* **45**(2):265–269.

Muñoz Maniega S., Bastin M., McIntosh A., Lawrie S. & Clayden J. (2008). Atlas-based reference tracts improve automatic white matter segmentation with neighbourhood tractography. In *Proceedings of the ISMRM 16th Scientific Meeting & Exhibition*, p. 3318. International Society for Magnetic Resonance in Medicine.

Parker G., Stephan K., Barker G., Rowe J., MacManus D., Wheeler-Kingshott C., Ciccarelli O., Passingham R., Spinks R., Lemon R. & Turner R. (2002a). Initial demonstration of in vivo tracing of axonal projections in the macaque brain and comparison with the human brain using diffusion tensor imaging and fast marching tractography. *NeuroImage* **15**(4):797–809.

Parker G., Wheeler-Kingshott C. & Barker G. (2002b). Estimating distributed anatomical connectivity using fast marching methods and diffusion tensor imaging. *IEEE Transactions on Medical Imaging* **21**(5):505–512.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Smith S. (2002). Fast robust automated brain extraction. *Human Brain Mapping* **17**(3):143–155.

Smith S., Jenkinson M., Johansen-Berg H., Rueckert D., Nichols T., Mackay C., Watkins K., Ciccarelli O., Cader M., Matthews P. & Behrens T. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage* **31**(4):1487–1505.

Tournier J.D., Calamante F., Gadian D. & Connelly A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *NeuroImage* **23**(3):1176–1185.

Tuch D. (2004). Q-ball imaging. *Magnetic Resonance in Medicine* **52**(6):1358–1372.