NeuroImage: Clinical 26 (2020) 102168

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Reproducibility, reliability and variability of FA and MD in the older healthy population: A test-retest multiparametric analysis

Pedro A. Luque Laguna[a,b,c,1,*], Anna J.E. Combes[a], Johannes Streffer[f,g], Steven Einstein[e,f], Maarten Timmers[d,g], Steve C.R. Williams[a], Flavio Dell'Acqua[b,c]

[a] Department 5 of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK
[b] Natbrainlab, Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK
[c] Sackler Institute for Translational Neurodevelopment, Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK
[d] Janssen Research and Development, a division of Janssen Pharmaceutica NV, Beerse, Belgium
[e] Janssen Research and Development LLC, Titusville, NJ, US
[f] UCB Biopharma SPRL, Chemin du Foriest B-1420 Braine-l'Alleud, Belgium
[g] Reference Center for Biological Markers of Dementia (BIODEM), Institute Born-Bunge, University of Antwerp, Antwerp, Belgium

1. Introduction

Diffusion tensor imaging (DTI) is a diffusion-weighted magnetic resonance imaging (DW-MRI) technique used to probe in vivo the properties of the grey-matter and white-matter tissue. By measuring the random displacement of water molecules inside the brain, DTI can provide quantitative metrics sensitive to many properties of tissue microstructure such as cell density, membrane permeability, axonal complexity, degree of myelination, etc. (Pierpaoli et al., 2001; Beaulieu 2002; Vos et al., 2012). Its remarkable ability to detect a wide range of biological changes and its clinical availability have made of DTI a method of choice for the investigation of many neurodegenerative disorders and white-matter pathologies (for a review on the application of DW-MRI and DTI in neurology see Goveas et al. (2015) and Raja et al. (2019)). In particular, DTI metrics such as fractional anisotropy (FA) and mean diffusivity (MD) (Basser and Pierpaoli 1996) are widely used today to investigate the onset and the progression of many neurological conditions and they have been proposed as surrogate biomarkers in studies and clinical trials developing new treatments for Alzheimer disease (AD) (Galluzzi et al., 2016; the PharmaCog Consortium et al., 2019), amyotrophic lateral sclerosis (ALS) (Duning et al., 2011), multiple sclerosis (MS) (Kapoor 2017; Vavasour et al., 2019; Zivadinov et al., 2018; Zhou et al., 2018), cerebral small vessel disease (SVD) (Croall et al., 2017), brain tumours (Ellingson et al., 2010; Moffat et al., 2005) and other neurological conditions (Boespflug et al., 2011; Egger et al., 2013; Huhn et al., 2018; Nath et al., 2010; Paldino et al., 2012). In these studies, differences in FA and MD are interpreted as brain changes that reflect the progression of a neurological condition or the biological response to treatment. To be useful, however, study results have to be not only statistically significant but also reproducible and reliable.

Reproducibility and reliability quantify two different aspects of the consistency of the measurements. Reproducibility is the ability to obtain similar values from different acquisitions of the same subject and it is measured by indices of within-subject variability across different scanning sessions. High reproducibility is defined by a low within-subject variability and implies a low measurement error (repeatability). Having a high reproducibility is particularly important in neuroimaging studies that rely on robust quantitative measurements. On the other hand, reliability refers to the overall consistency of the measurements across subjects and it is measured by indices that compare the within-subject variability with the 'true' between-subject variability (i.e. the 'error-free' variability between subjects). If the within-subject variability is small compared to the between-subject variability, then there is high reliability because it is relatively easy to distinguish between different subjects based on the quantitative values of the measurements. Reciprocally, if the within-subject variability is large compared to the between-subject variability, the reliability is low. Under low reliability, two similar subjects could yield very different measurements purely because of the errors related to the within-subject variability rather than for the existence of a genuine difference between their true values. Therefore, high reliability is required when interpreting changes in the metrics' values as genuine changes across subjects.

Reproducibility and reliability are separate properties of the metrics and do not imply each other. Ignoring this fact can lead to a dubious interpretation of (unreliable) results or inappropriate use of (inexact) measurements. Metrics that show good reproducibility may or may not demonstrate also good reliability (and vice-versa) depending on the scale of the between-subject variability. The between-subject variability is also related to the statistical power associated with a metric, for

* Corresponding author.
E-mail addresses: pedro.luque.laguna@gmail.com (P.A. Luque Laguna), steve.williams@kcl.ac.uk (S.C.R. Williams), flavio.dellacqua@kcl.ac.uk (F. Dell'Acqua).
[1] The new permanent address for the corresponding author Pedro A. Luque Laguna is: Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Manidy Road, Cardiff CF24 4HQ, UK

https://doi.org/10.1016/j.nicl.2020.102168
Received 7 July 2019; Received in revised form 10 January 2020; Accepted 10 January 2020
Available online 25 January 2020
2213-1582/ © 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

example in terms of minimum sample sizes required to detect changes of a given magnitude in the metrics' values. A metric will be most useful in those regions where it is reproducible, reliable and has sufficient statistical power. The design of neuroimaging studies (and the interpretation of their results) benefits from having a good characterisation of the employed metrics in terms of their reproducibility, reliability and between-subject variability across the brain.

In the case of FA and MD, there have been several studies investigating different aspects of the test-retest variability of diffusion MRI measurements in healthy populations. While some of these studies have provided results for reproducibility, reliability and between-subject variability [Grech-Sollars et al. (2015); Marenco et al. (2006); Vollmar et al. (2010); Wang et al. (2012);], most of them have focused on one or two properties only, such as reproducibility and reliability (Albi et al., 2017; Bisdas et al., 2008; Boekel et al., 2017; Jansen et al., 2007; Lemkaddem et al., 2012; Liu et al., 2014), reproducibility and between-subject variability (Heiervang et al., 2006; Pannek et al., 2011; Veenith et al., 2013; Willats et al., 2014), only reproducibility (Pfefferbaum et al., 2003), only reliability (Brandstack et al., 2016; Duan et al., 2015) or between-subject variability only (Sadeghi et al., 2015). The reference values provided by these studies show how the reproducibility, the reliability and the between-subject variability of FA and MD change from region to region across the entire brain anatomy. It remains unclear, however, the extent to which these results, based on data acquired mostly on young healthy adults, can also be applied to an older healthy population. Compared to young adults, older healthy adults show substantially different patterns of age-related DTI changes across the brain (Madden et al., 2012) and different age trajectories at each anatomical region (Barrick et al., 2010; Bender and Raz 2015; Cox et al., 2016; Lebel et al., 2012; Lövdén et al., 2014; Sexton et al., 2014; Sullivan et al., 2010). It would be expected that the different trajectories observed in older healthy adults would also translate into differences in terms of the statistical properties of the DTI metrics. In consequence, it would be more effective for studies and trials investigating diseases and disorders in the older population to consider as reference the statistical values derived from data acquired on older individuals. This is the case with most neurodegenerative conditions, which show a much higher prevalence amongst older adults and are characterised by peak ages of onset typically around 50 years or later (Andren et al., 2017).

To evaluate the reproducibility, the reliability and the between-subject variability of FA and MD across the brain in the older population, we analysed DTI data from a longitudinal study aimed at the test-retest reliability of structural and functional neuroimaging in healthy adults using conventional MRI acquisition protocols. We used the within-subject coefficient of variation (CV_{ws}), the intra-class correlation coefficient (ICC) and the between-subject coefficient of variation (CV_{bs}) to characterise the test-retest reproducibility and reliability of the metrics as well as their between-subject variability. We computed values for these coefficients voxel-wise in the brain differentiating between grey-matter and white-matter voxels. In addition, we compared the coefficients on six predefined white-matter regions of interest (ROIs) where we also carried out statistical power calculations to compare the ability of the metrics to detect changes in these regions. We then produced whole-brain statistic maps of reproducibility, reliability and between-subject variability from the CV_{ws}, ICC and CV_{bs} values obtained at each voxel in the brain. Finally, we analysed the statistical properties of the metrics and their interaction across the brain anatomy using a multiparametric segmentation approach that allowed us to integrate and visualise straightforwardly the information from the three statistic maps combined.

2. Materials and methods

2.1. Participants

Sixteen healthy adults (8 females) with an age range of 53–65 years were recruited for the study. All subjects were scanned on three separate occasions over a period of two months, having 1-week and 4-week follow-ups after the first scanning session under the assumption that only minimal within-subject differences are to be expected between the different time intervals. Each participant gave written informed consent before taking part in the study, which was approved by the King's College London Psychiatry, Nursing and Midwifery Research Ethics subcommittee. The MRI data were acquired using a 3T *GE MR750* system equipped with a C-*GE* HNS 12-channel head coil.

2.2. Image acquisition

For each scanning session, a total of 60 diffusion-weighted MRI volumes (b-value 1500s/mm^2) and 9 $b = 0$ s/mm^2 diffusion-weighted MRI volumes (3 with reverse phase-encoding direction) were acquired with isotropic resolution of $2 \times 2 \times 2$ mm using a spin-echo single shot EPI pulse sequence with echo time TE = 75.4 milliseconds and repetition time TR = 12 s (peripherally pulse gated).

2.3. Preprocessing of DTI data

All diffusion data were corrected for artefacts due to subject motion, Eddy currents and magnetic field susceptibilities distortions using FSL "eddy" (Andersson and Sotiropoulos 2015) and FSL "top-up" preprocessing tools (Andersson et al., 2003; Smith et al., 2004).

2.4. Computation of DTI metrics

FA and MD maps were computed for each diffusion dataset following the fitting of the diffusion tensor at every voxel using the implementation in ExploreDTI (Leemans et al., 2009) of the 'Robust Estimation of Tensors by Outlier Rejection' (RESTORE) algorithm (Chang et al., 2005). In total, three different pairs of FA and MD maps were produced for every participant (one pair for each time-point).

2.5. Image normalisation to MNI space

The FA and MD maps for each participant were all normalised to the same anatomical space. For this purpose, each FA map from every subject and time-point was spatially normalised to the FMRIB58 FA $1 \times 1 \times 1$ mm isotropic template with a non-linear elastic registration (Smith et al., 2006). Consecutively, each mathematical transformation obtained during the normalisation of the FA volumes was applied to the corresponding MD pair. The normalisation process allows the statistical analysis of the images in an anatomically congruent manner across all subjects. In particular, the statistical properties of the metrics can be evaluated at the level of single voxels, on specific anatomical regions or any other regions obtained through specific segmentation of the anatomical space.

2.6. White-matter and grey-matter segmentation

A preliminary objective was to investigate differences in the properties of the metrics that are to be expected in the white matter compared to the grey matter. To discriminate between cortico-spinal fluid (CSF), white-matter voxels and grey-matter voxels, we used a template of anisotropic power (AP) values (Dell'Acqua et al., 2014) obtained from whole-brain AP maps obtained from data acquired on 200 healthy subjects from the Human Connectome Project (Van Essen et al. 2013). In the AP template, values near zero are associated with isotropic diffusion and indicate the presence of water or CSF. The highest AP values,

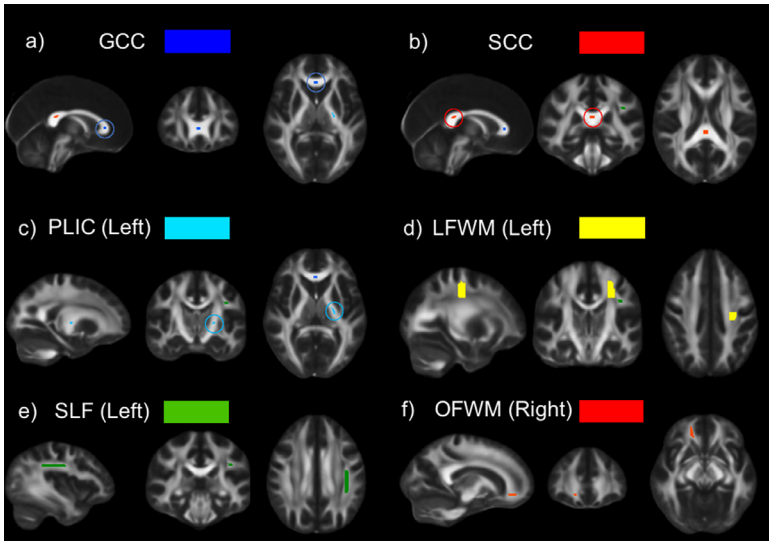

Fig. 1. Anatomical Regions of Interest. Six regions of interest manually delineated in the FMRIB58 FA 1 × 1 × 1 mm isotropic template to define anatomical areas commonly used for detecting white-matter changes in clinical populations: (a) genu of the corpus callosum (GCC) - 45 voxels, (b) the splenium of the corpus callosum (SCC) - 84 voxels, (c) the posterior limb of the left internal capsule (PLIC) - 52 voxels, (d) the centrum semiovale on the left frontal white matter (LFWM) - 1180 voxels, (e) the third component of the left superior longitudinal fasciculus (SLF) - 435 voxels, and (f) the right orbitofrontal white matter region (OFWM) - 122 voxels.

on the other hand, are associated with the presence of underlying microstructure typical of white-matter regions. We used an arbitrary threshold around the 20th percentile of the total distribution of AP values to distinguish between voxels in the white matter (above) or the grey matter (below). For each tissue type, the mean FA and MD values were calculated over all the voxels included in the corresponding region.

2.7. Anatomical regions of interest

To evaluate the statistical properties of the metrics across different white-matter regions and to compare the results with the existing literature, we delineated (manually) six different regions of interest (ROIs) in the FMRIB58 FA template (Fig. 1) to match ROIs used by previous studies. We used Marenco et al. (2006) for the genu of the corpus callosum (GCC), the splenium of the corpus callosum (SCC), the posterior limb of the left internal capsule (PLIC) and the orbitofrontal white matter (OFWM). For the centrum semiovale of the left frontal white matter (LFWM), we followed Vollmar et al. (2010). Finally, the third component of the left superior longitudinal fasciculus (SLF) was manually drawn in the FMRIB58 FA template using the Johns Hopkins University (JHU–ICBM) white matter atlas as a reference (Mori et al., 2008). For each ROI, we computed the mean values of the metrics over all the voxels occupied by each region for further statistical analysis.

2.8. Statistical model for test-retest variability

For the analysis of variance, we used a two-way mixed-effects model by considering subjects as random effects and sessions as fixed effects. This allowed us to separate the total variability of the data into two components: the within-subject variability and the between-subject variability. The within-subject variability represents the differences in the values of the metrics observed when scanning the same subjects on multiple occasions. Such differences can be caused by a combination of factors, such as changes experienced by the subjects between different sessions, differences in the scanning conditions and the experimental noise intrinsic to the acquisition. The between-subject variability is the variability not explained by factors related to the within-subject variability. Under test-retest conditions akin to our study, and assuming no significant bias introduced by the age differences between the participants, the between-subject variability can be interpreted as the inter-subject biological variability of the metrics, i.e. the variability explained by true inter-individual biological differences that are not explained by factors related to within-subject variability (Sadeghi et al., 2015).

2.9. Statistical coefficients

The following statistical coefficients were computed to characterise the reproducibility, reliability and between-subject variability of the metrics: CV_{ws}, ICC and CV_{bs}. By separating the total variability into within-subject variability and between-subject variability, we could estimate the corresponding values for the within-subject coefficient of variation (CV_{ws}) and the between-subject coefficient of variation (CV_{bs}). For the ICC, we used the third form of the intraclass correlation coefficient ($ICC_{3,\ 1}$) as defined by Shrout and Fleiss (1979). Refer to Appendix A for the mathematical definitions and calculation formulas for each index. The CV_{ws} summarises the within-subject variability of the metrics. Lower CV_{ws} values identify voxels or regions where the metrics' values were more reproducible across the three sessions. The ICC estimates the correlation between metrics values corresponding to different sessions in terms of their consistency across subjects. Statistical group differences detected in voxels or regions with higher ICC are considered to be more reliable in the sense that changes in the metric values are expected to be consistent across all subjects. The CV_{bs} summarises the between-subject variability of the metrics values. High CV_{bs} values are associated with increased capacity of the metrics to discriminate between individuals within the same population. However, higher CV_{bs} values also require larger sample sizes to detect predefined percentage changes in the metric values at the group level (they are associated with lower statistical power).

2.10. Computation of CV_{ws} ICC and CV_{bs} across the brain anatomy

We used the statistical toolbox designed for ICC analysis of neuroimaging data by Caceres et al. (2009) to calculate CV_{ws}, ICC and CV_{bs} values at each voxel, anatomical ROIs and also for the whole-brain white mater and grey matter regions. For each combination of metric and statistical coefficient, the ICC toolbox produced a whole-brain statistic parametric map with the values of the corresponding coefficient at every voxel. We also evaluated the distribution of each statistical coefficient across all-brain voxels, white-matter voxels and grey-matter voxels by computing the corresponding histogram of frequency densities. For each anatomical region, the ICC toolbox yielded the value of the coefficients calculated from the regional averages of each metric.

Table 1
Percentage rates previously reported for longitudinal changes in FA and MD. Each row shows the detected longitudinal percentage change in a metric's average over a specific brain region in subjects from the same clinical population. Clinical conditions: amyotrophic lateral sclerosis (ALS), mild cognitive impairment (MCI), frontotemporal dementia (FTD), Alzheimer's disease (AD) and primary progressive multiple sclerosis (ppMS). Brain regions: corticospinal tract (CST), the body of the corpus callosum (CC body), left and right uncinate fascicle (UF), right parahippocampal cingulum (PH-CING) and cerebellar, frontal and occipital white matter (WM)
.

Study	Condition	Brain Region	Metric	% change	Time
Study	Condition	Brain Region	Metric	% change	Time
Blain et al. 2007	ALS	CST	FA	−2.70%	6 months
Teipelet al. 2010	MCI	CC body	FA	−2.50%	12 months
Mahoney et al. 2015	FTD	CC body	FA	−3.30%	12 months
		UFLeft	FA	−7.90%	12 months
Kitamura et al., al. 2013	AD	UF Right	FA	−3.20%	18 months
		UF Left	FA	−4.50%	18 months
Blain et al. 2007	ALS	CST	MD	0.58%	6 months
Mahoney et al. 2015	FTD	PH-CING Right	MD	2.70%	12 months
		UF Left	MD	5.20%	12 months
Schmiereret al. 2004	ppMS	Cerebellar WM	MD	2.30%	12 months
		Frontal WM	MD	2.70%	12 months
		Occipital WM	MD	3.60%	12 months

2.11. Statistical power calculations

We calculated the minimum sample size required to detect longitudinal changes of predetermined magnitudes in the average values of FA and MD over each ROI assuming that ROI average values are normally distributed in the population. More specifically, we calculated the number of subjects required to detect 2%, 4% and 6% group-level longitudinal changes from baseline values (either increase or decrease) with statistical power of at least $1-\beta = 0.8$ and statistical significance level $\alpha = 0.05$ (see Appendix B. Statistical power calculations). The chosen percentage values are in line with previously reported rates of longitudinal changes in FA and MD detected on various neurodegenerative conditions across different brain regions (Schmierer et al., 2004; Blain et al., 2007; Teipel et al., 2010; Kitamura et al., 2013; Mahoney et al., 2015) (Table 1).

2.12. Multiparametric segmentation of statistical maps

In addition to the analysis based on anatomical regions, we also wanted to investigate the relationship between the three statistical properties of the metrics (reproducibility, reliability and between-subject variability). We were particularly interested in identifying brain regions where the statistical properties of each metric were comparatively homogeneous to provide whole-brain statistical atlases for each metric based on the information from the three statistical coefficients combined. For that purpose, we segmented each CV_{bs} map into separate regions according to different (low or high) levels of reproducibility and reliability using the following threshold criteria: $CV_{ws} = 10\%$ to differentiate between regions of low ($CV_{ws} > 10\%$) and high ($CV_{ws} < 10\%$) reproducibility, as in (Marenco et al., 2006), ICC = 70% to differentiate between regions of low (ICC < 70%) and high (ICC > 70%) reliability, as levels above 70% are usually associated with reasonably good to very good reliability (Boekel et al., 2017; Duan et al., 2015; Marenco et al., 2006; Vollmar et al., 2010).

3. Results

3.1. Whole-brain, white matter and grey matter

When considering average values of FA and MD across the whole brain, the white matter and the grey matter, the reproducibility and reliability of both metrics are good with CV_{ws} values below 5% and ICC values above 80% (Table 2). The between-subject variability summarised by the CV_{bs} is similar between all averaged measures with CV_{bs} values around 7.5% except for the case of grey-matter MD averages that yield $CV_{bs} = 17.74\%$ (Table 2).

Table 2
Whole-brain CV_{ws}, ICC and CV_{bs} of FA and MD. Statistical indices are calculated from the mean values of the metrics across all voxels in the brain, across all white-matter (WM) and grey-matter (GM) voxels.

Metric	Region	Mean (SD)	CV_{ws}	ICC	CV_{bs}
FA	Brain	0.30 (0.022)	0.69%	97.78%	7.43%
	WM	0.382 (0.028)	0.70%	97.82%	7.46%
	GM	0.228 (0.017)	0.76%	97.35%	7.65%
Metric	Region	Mean (SD) mm^2/s	CV_{ws}	ICC	CV_{bs}
MD	Brain	0.86e-03(0.10e-03)	2.72%	87.03%	12.76%
	WM	0.76e-03(0.05e-03)	1.45%	89.75%	7.46%
	GM	0.95e-03(0.16e-03)	4.42%	82.58%	17.74%

Across different voxels, the variability of both metrics becomes more heterogeneous and shows clear differences between the white matter and the grey matter (Fig. 2). In the white matter, both metrics demonstrate good reproducibility with CV_{ws} values well below 10% levels in most voxels (FA: $CV_{ws} = 10\% \pm 6\%$), MD: $CV_{ws} = 4\% \pm 2\%$). In the grey matter, MD has good reproducibility (MD: $CV_{ws} = 4\% \pm 6\%$) while FA is found reproducible in less than 25% of grey-matter voxels (FA: $CV_{ws} = 14\% \pm 5\%$). Both metrics show good reliability levels in the white matter with ICC values well above 70% in most voxels (FA: ICC = 77% ± 13%, MD: ICC = 74% ± 19%). In the grey matter, the reliability is still good in approximately half of the voxels, with a larger proportion of high ICC values in the case of FA (ICC = 67% ± 20%) compared to MD (ICC = 62% ± 47%). Finally, the between-subject variability of the metrics in white-matter voxels is twice as high for FA ($CV_{bs} = 40\% \pm 26\%$) compared to MD ($CV_{bs} = 21\% \pm 18\%$). The same effect is also found across the grey matter, with an increase in the between-subject variability (FA: $CV_{bs} = 49\% \pm 26\%$ and MD: $CV_{bs} = 26\% \pm 35\%$).

3.2. ROI analysis: reproducibility and reliability

Overall, the reproducibility and reliability of FA and MD improved with the use of anatomical ROIs compared to the voxel-level results (Table 3). Both FA and MD show good reproducibility in all ROIs with CV_{ws} values that ranged between 1% and 1.8% for FA and between 1.1% and 2.8% for MD. Reproducibility is higher for FA than for MD in the SCC, the GCC and the PLIC. The reliability of both metrics is also generally high, with ICC values above 70% in all ROIs except for the MD in the GCC (ICC = 54%). Reliability is always higher for FA than for MD, with ICC values that ranged from 76% to 97% for FA, compared

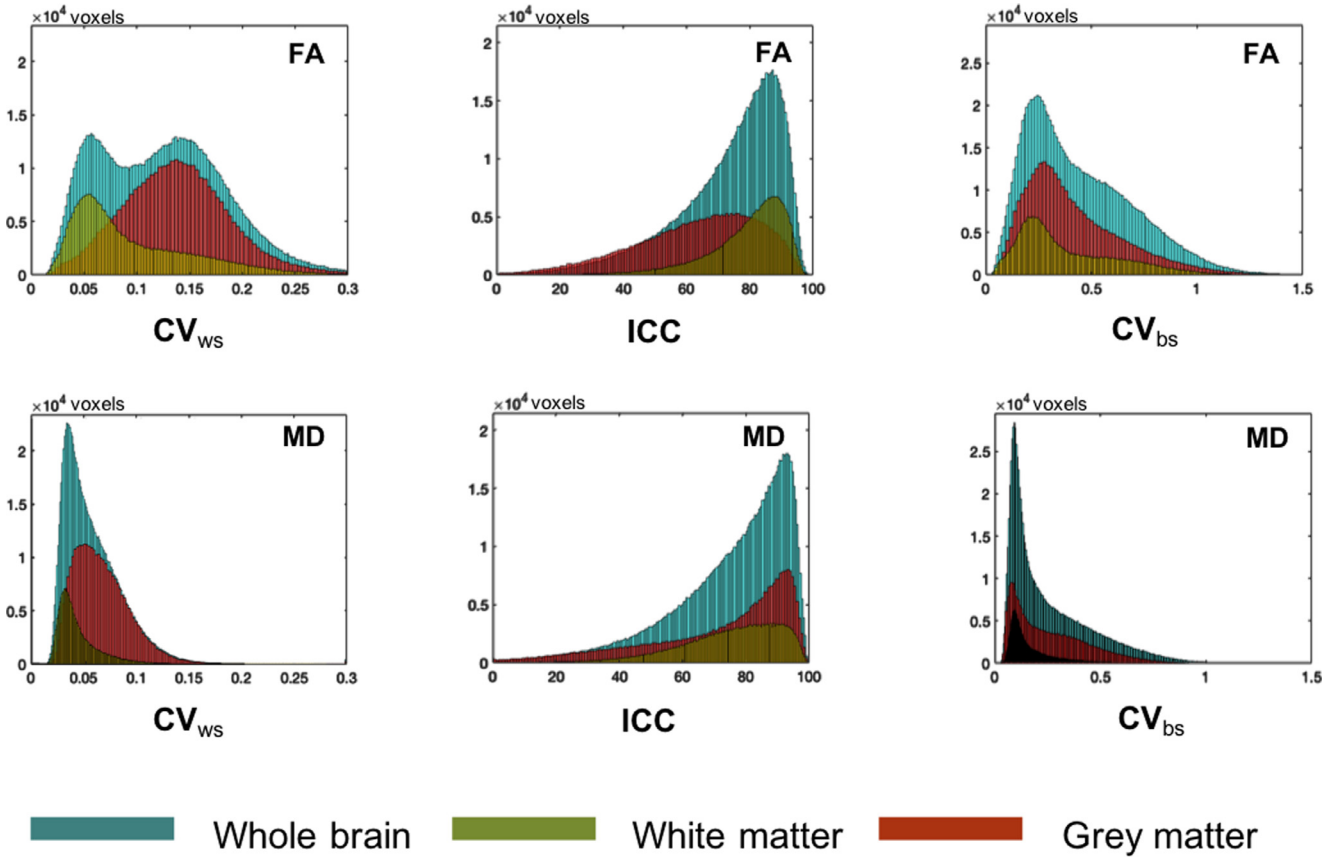

Fig. 2. Reproducibility, Reliability and Variability of FA and MD metrics across the brain. Histograms describing the distribution of the different coefficients for FA (top) and MD (bottom) for the whole brain, white-matter and grey-matter voxels. Left: within-subject coefficients of variation. Centre: intraclass correlation coefficient. Right: between-subject coefficient of variation.

to values between 71% and 88% for MD. In the case of MD, there is a substantial agreement between the reproducibility and the reliability of the metric. The ROIs where MD is the most reproducible are also the ROIs where MD is most reliable (SLF CV_{ws} = 1.2% ICC = 88%, LFWM CV_{ws} = 1.1% ICC = 92%). Analogously, the ROIs with lowest MD reproducibility are also the ROIs where MD is less reliable (SCC CV_{ws} = 2.5% ICC = 73%, GCC CV_{ws} = 2.8% ICC = 54%). In contrast, there appeared to be less correlation between the reproducibility and the reliability of FA. For example, the ROIs where FA is most reproducible (SCC and PLIC) are not the ROIs where the FA is most reliable (OFWM and LFWM) (see Table 3).

3.3. ROI analysis: statistical power

The bar charts in Fig. 3 shows the statistical power of FA (left) and MD (right) in terms of the minimum number of subjects needed by each metric to detect at the group level a longitudinal increase (or decrease) from the baseline values at each ROI. There are clear differences between the statistical power of the two metrics and also between the power of the same metric across different ROIs. Differences in statistical power become larger as the percentage change to be detected decreases. FA is most sensitive in the SCC (3–16 subjects) and the GCC (4–18 subjects) followed by the PLIC (5–28 subjects) and SLF (8–62 subjects). The ROIs where FA is the less sensitive are the OFWM (21–176 subjects) and the LFWM (25–213 subjects). Compared to FA, the MD shows roughly the opposite trend: the ROIs where the MD is less sensitive are the SCC (5–30 subjects) and the GCC (4–24 subjects) followed by the OFWM (4–23 subjects) and the SLF (4–22 subjects). The ROIs where MD is the most sensitive are the LFWM (3–15 subjects) and the PLIC (3–17 subjects).

3.4. Reproducibility CV_{ws} maps

In the FA CV_{ws} map (Fig. 4 top-left slide), it is possible to appreciate that voxels where the reproducibility is high (CV_{ws} < 10%) are found mostly in the deep white matter (corticospinal tract and the corpus callosum). Within these regions, the spatial distribution of the CV_{ws} values is smoother and relatively homogeneous. A gradient in CV_{ws}

Table.3
CV_{ws} and ICC of FA and MD across ROIs. The reproducibility of FA and MD is high across all ROIs with a CV_{ws} between 1% and 3% for all ROIs. The reliability of FA is also high across all ROIs, with ICC values above 70% in all ROIs except the GCC.

ROI	Mean FA	FA CV_{ws}	FA ICC	Mean MD mm^2/s	MD CV_{ws}	MD ICC
Splenium Corpus Callosum	0.84	1.00%	86%	7.30E-04	2.50%	73%
Genu Corpus Callosum	0.81	1.80%	76%	7.50E-04	2.80%	54%
Left Internal Capsule	0.73	1.10%	92%	6.40E-04	1.80%	71%
Superior Longitudinal Fasciculus	0.59	1.40%	94%	6.30E-04	1.20%	88%
Orbito Frontal WM	0.55	1.70%	97%	7.50E-04	1.50%	84%
Centrum Semiovale (LFWM)	0.44	1.80%	97%	6.30E-04	1.10%	85%

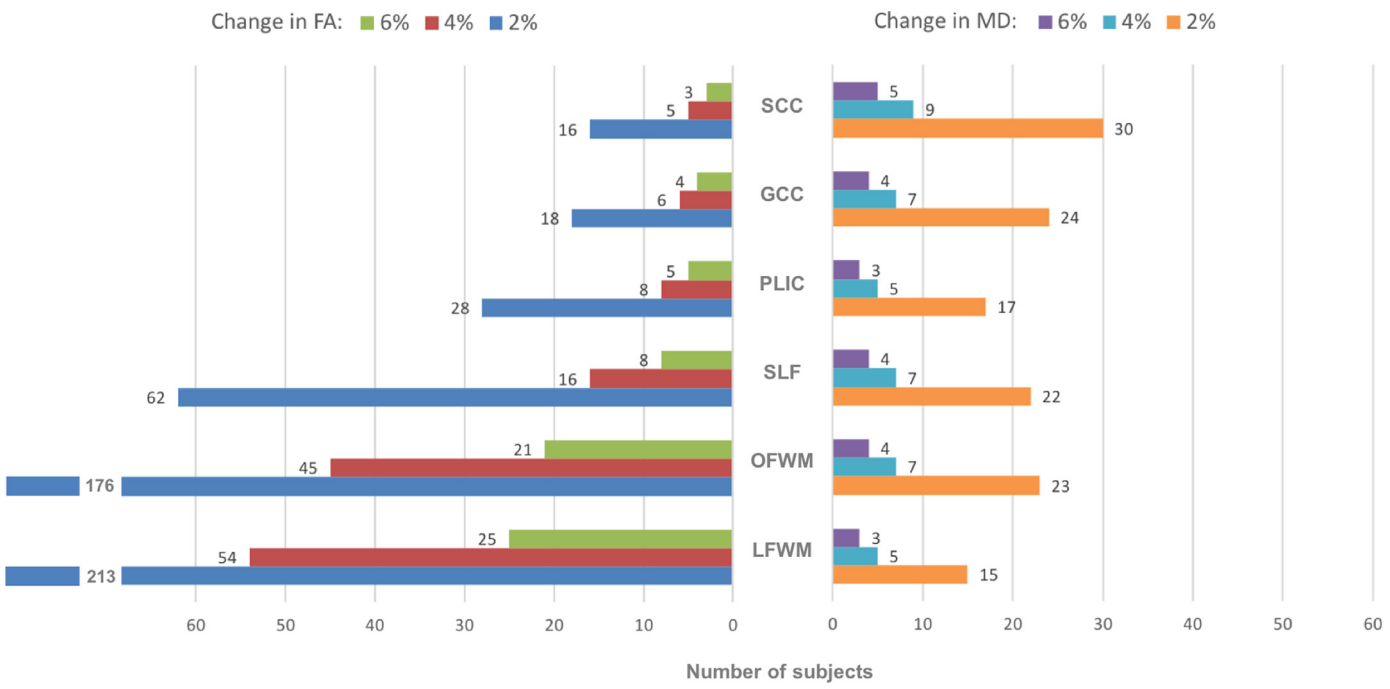

Fig. 3. Statistical power of FA and MD across ROIs. Number of subjects needed to detect longitudinal changes in FA (left) and MD (right) from nominal values at $p < 0.05$ significance level and $1 - \beta$ statistical power $= 0.80$ (1-tailed). FA Nominal values: SCC 0.84, GCC 0.81, PLIC 0.73, SLF 0.59, OFWM 0.55, LFWM 0.44. MD Nominal values: SCC .73 $\times 10^{-3}$ s/mm^2, GCC 0.75 $\times 10^{-3}$ s/mm^2, PLIC 0.64 $\times 10^{-3}$ s/mm^2, SLF 0.63 $\times 10^{-3}$ s/mm^2, OFWM 0.75 $\times 10^{-3}$ s/mm^2, LFWM 0.63 $\times 10^{-3}$ s/mm^2.

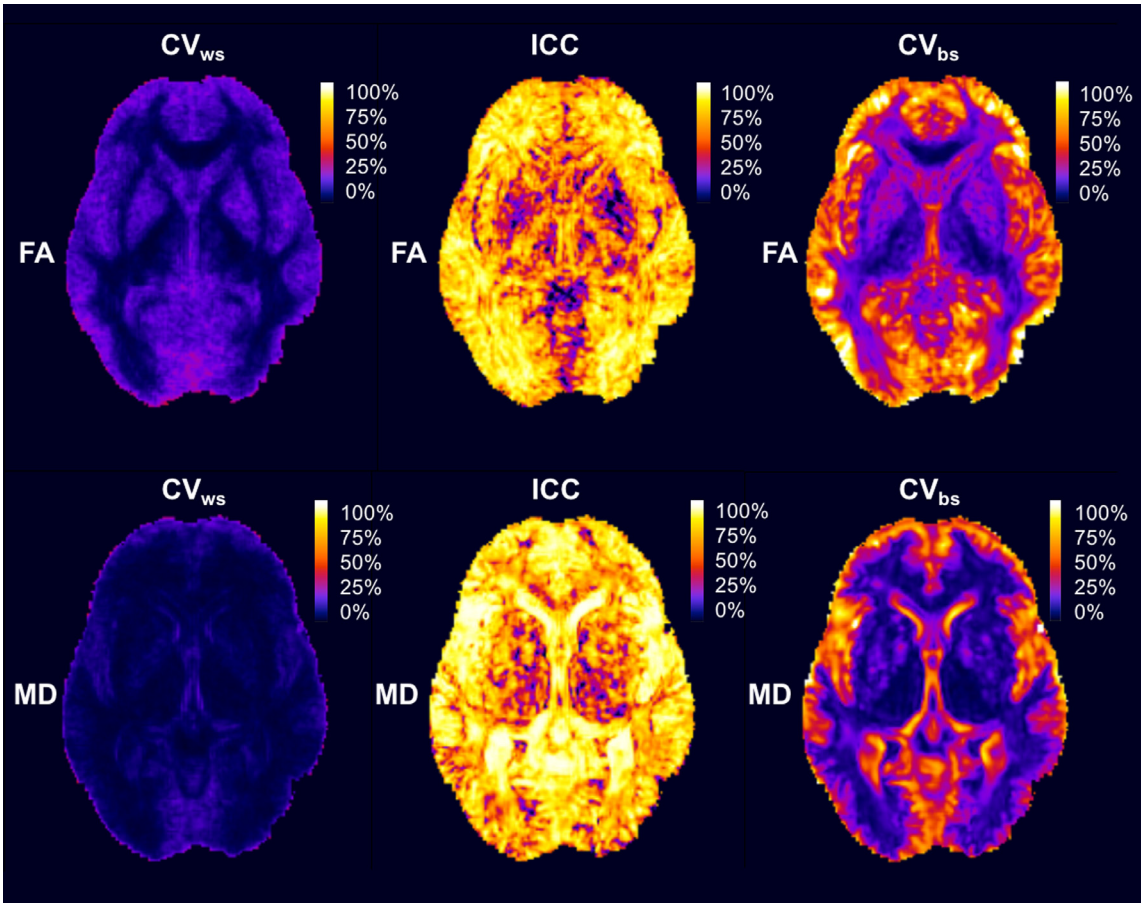

Fig. 4. Reproducibility, Reliability and Variability of FA and MD metrics across a representative axial slice of the brain. On the left, the within-subject coefficients of variation for FA and MD show patterns of high and homogeneous levels of reproducibility across different tissue types. In the centre, the intraclass correlation coefficient shows a more heterogeneous pattern of reliability of these measures while still high. On the right, the between-subject coefficient of variation describes also a high variation in the between-subject variability of these metrics across different regions.

values can be observed between the voxels in the deep white matter towards the subcortical white matter (CV_{ws} around 20%). Voxels with lowest FA reproducibility (CV_{ws} > 20%) are found in cortical regions, grey-matter/white-matter boundary (U-shape fibres), the striatum, the ventricles and voxels with CSF partial volume. In these regions, the spatial distribution of the CV_{ws} values is noisy and more heterogeneous.

In the case of the MD, the majority of voxels across all regions show good or very good reproducibility with CV_{ws} values well below 10% (Fig. 4 bottom-left slide). We can identify only small groups of voxels where the MD CV_{ws} slightly increase beyond 10% in cortical regions, in periventricular areas and regions with CSF partial volume contamination. Only in the periphery of the cortex, we find voxels where the CV_{ws} of MD falls below acceptable reproducibility values (CV_{ws} > 10%).

3.5. Reliability ICC maps

The voxels with the highest reliability for FA are found in cortical areas and most of the white matter (see Fig. 4 top-centre slide). Some other regions like the internal capsule, part of the corpus callosum and part of the thalamus contain voxels of medium to high FA reproducibility (50% < ICC < 70%). Then, voxels with low reproducibility (ICC < 50%) are found in deep grey-matter structures (striatum and thalamus) and regions affected by large CSF presence, such as outside the insular cortex and other inter-lobular regions. Compared to the FA CV_{ws} map, the spatial distribution of FA ICC values is overall quite heterogeneous and with less anatomical contrast.

In the case of the MD ICC map (Fig. 4 bottom-centre slide), the voxels with the highest reliability are located in the CSF within the ventricles, near the insular cortex and the space between the two hemispheres. In contrast, voxels close to ventricular regions, in the grey-matter/white-matter boundary and close to the deep grey matter structures (e.g. thalamus) are amongst those with the lowest ICC for MD with values between 40% and 70%. The rest of the white matter and the cortex consistently show voxels with a relative high ICC value (ICC > 70%). The spatial distribution of ICC values in the MD ICC map is irregular and heterogeneous, which makes it difficult to distinguish between different anatomical regions except for the ventricles.

3.6. Between-subject variability CV_{bs} maps

The voxels with the highest CV_{bs} values for FA (Fig. 4 top-right slide) are found in the cortex, in the outermost white-matter, in the posterior horns of the lateral ventricles and inter-lobar areas. Medium levels of FA CV_{bs} values are found in voxels in the striatum and intra-lobar white-matter regions. Deep white-matter regions such as the external capsule (EC) the anterior limb of the internal capsule (ALIC) contain large numbers of voxels with low FA CV_{bs}, while the lowest CV_{bs} values for FA are found in the GCC, the PLIC and the thalamus. The contrast in the FA CV_{bs} maps allows to see distinguish some anatomical features like some deep white matter structures such as the external capsule and the striatum.

For the MD, the voxels with the highest CV_{bs} values (Fig. 4 bottom-right slide) are found in the cortex, in the grey-matter/white-matter boundary, in the ventricles and periventricular areas. Medium levels of MD CV_{bs} values are found in voxels within the striatum and also in the subcortical white matter. The lowest MD CV_{bs} values correspond to voxels in the thalamus and most of the remaining white-matter regions. The MD CV_{bs} map shows the highest anatomical contrast amongst all statistic maps, clearly separating voxels in the cortex, in the grey-matter/white-matter boundary and the deep white matter.

3.7. Multiparametric segmentation of FA CV_{bs}

Fig. 5 shows the multiparametric segmentation of the FA CV_{bs} maps on three representative slices (mid-axial, mid-coronal and mid-left-sagittal) based on the corresponding values of FA CV_{ws} and FA ICC in each voxel. Overall, the cortical grey matter has the highest between-subject FA variability with low reproducibility and high reliability, whereas most deep grey-matter structures have much lower between-subject variability also with low reproducibility and reliability. In comparison, the FA in the white matter has much lower between-subject variability, high reproducibility and high reliability, except around the putamen and parts of the corpus callosum where the reliability is low.

Quadrant I: High-reproducibility/high-reliability regions are formed mostly by white-matter voxels characterised by low to medium levels of between-subject variability. In particular, the between-subject variability is relatively lower at voxels in the internal capsule and deepest white-matter regions, with higher levels of variability towards the grey-matter/white-matter boundary.

Quadrant II: Low-reproducibility/high-reliability regions are formed by voxels with the highest between-subject variability of FA in the brain. These regions include the cerebral cortex, the subcortical white matter and periventricular areas.

Quadrant III: Low-reproducibility/low-reliability regions contain two family of voxels where the between-subject variability of FA is either moderate (20% < CV_{bs} < 40%) corresponding to the striatum (including the putamen and the head of caudate nucleus) or high (40% < CV_{bs} < 60%) corresponding to inter-lobular regions within the insular, prefrontal and occipital cortices.

Quadrant IV: High-reproducibility/ low-reliability regions contain voxels where the between-subject variability of FA is relatively lower across the entire brain (CV_{bs} < 10%) We can identify a part of the genu of the corpus callosum (GCC) and sections of the posterior limb of the internal capsule (PLIC). With slightly higher between-subject variability but still low (10% < CV_{bs} < 20%), we can clearly identify the rest of the internal capsule, the external capsule and the thalamus.

3.8. Multiparametric segmentation of MD CV_{bs}

Fig. 6 shows the multiparametric segmentation of the MD CV_{bs} map on the same representative slices and using the same criteria for the definition of areas of low/high reproducibility and reliability as in Fig. 5. The between-subject variability of MD is higher in the cortical grey matter, where the reliability is always high and the MD is (mostly) reproducible. Deep grey-matter structures have the lowest between-subject variability, with low reliability and high reproducibility. In the white matter, the between-subject variability of MD is relatively low and with high reliability except in some voxels along the internal capsule, around the thalamus and some subcortical WM regions where the MD is not reliable.

Quadrant I: There are many high-reproducibility/ high-reliability voxels with a broad range of between-subject variability. At the lowest values (CV_{bs} < 20%) we found voxels in the putamen, along the external capsule and the deep white matter. With higher between-subject variability (20% < CV_{bs} < 40%) we found some of the subcortical white-matter voxels. Finally, the anatomical areas found with the highest between-subject variability in this quadrant (CV_{bs} > 50%) are the ventricles and the periventricular areas in the occipital lobe, the ventral areas of the prefrontal cortex and the inter-lobular regions surrounding the insula.

Quadrant II: like in the case of the FA, low-reproducibility/ high-reliability regions for the MD contain the voxel with the highest between-subject variability for MD in the brain. In this case, these regions include only the most external cortical areas where the between-subject variability is consistently high (CV_{bs} > 50%).

Quadrant III: In the case of the MD, the low-reproducibility/ low-reliability region is formed mostly by voxels interfacing with the CSF along the periphery of the brain.

Quadrant IV: High-reproducibility/ low-reliability regions are formed by voxels where the between-subject variability of the MD is lower than in any of the other three regions, just as in the case of FA. In the case of the MD, these regions cover a larger range of well-defined

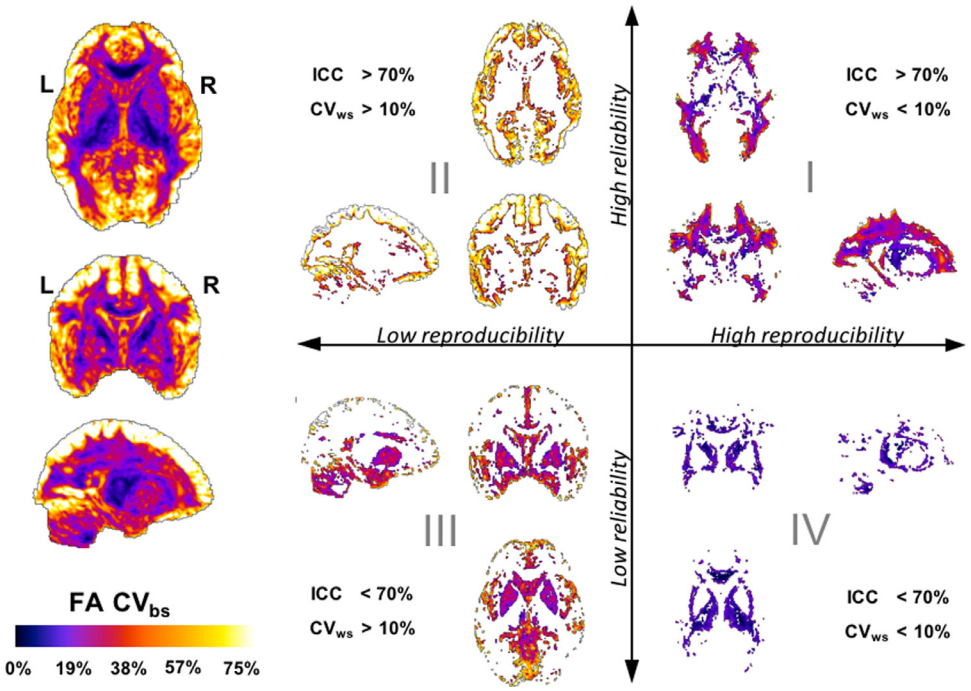

Fig. 5. Multiparametric segmentation of whole-brain FA CV_{bs} by reproducibility and reliability. The FA CV_{bs} whole-brain map (left) is segmented into four groups of anatomical regions showing similar levels of reproducibility and reliability (right). Quadrant I: regions formed by voxels with high FA reproducibility (CV_{ws} <10%) and high FA reliability (ICC >70%) - Quadrant II: regions formed by voxels with low FA reproducibility (CV_{ws} >10%) and high FA reliability (ICC >70%) - Quadrant III: regions formed by voxels with low FA reproducibility (CV_{ws} >10%) and low FA reliability (ICC <70%) - Quadrant IV: regions formed by voxels with high FA reproducibility (CV_{ws} <10%) and low FA reliability (ICC <70%).

anatomical areas. Firstly, the thalamus can be identified as the structure with the lowest between-subject variability in MD (CV_{bs} < 10%), together with white-matter voxels along the internal capsules and the genu of the corpus callosum. With slightly higher between-subject variability (10% < CV_{bs} < 20%) we can find the striatum (putamen and head of caudate) and a few subcortical white-matter regions likely following the trajectory of intra-lobar fibres (Howells et al., 2018).

4. Discussion

In this study, we have assessed the test–retest reproducibility, reliability, between-subject variability and the statistical power associated with FA and MD measurements in an older healthy population. The study is based on an acquisition setup of 60 diffusion directions, 9 non-diffusion weighted volumes, b-value = 1500s/mm^2 and isotropic resolution of 2 mm. With these settings, the results show that the reproducibility of MD is good (CV_{ws} < 10%) across the entire brain and very good for most white-matter regions (CV_{ws} < 5%). In contrast, FA showed good reproducibility only in white-matter voxels, but this is expected because of the very low FA values and increased sensitivity to noise in CSF and grey-matter regions. Overall, the voxel-wise reliability of both FA and MD is good, with voxel values well above ICC > 70% in the cortex and the subcortical white matter for both metrics. In the deep white-matter near grey-matter structures, however, only FA showed consistently good reliability while only MD showed good reliability in some deep grey-matter structures. In respect to the between-subject

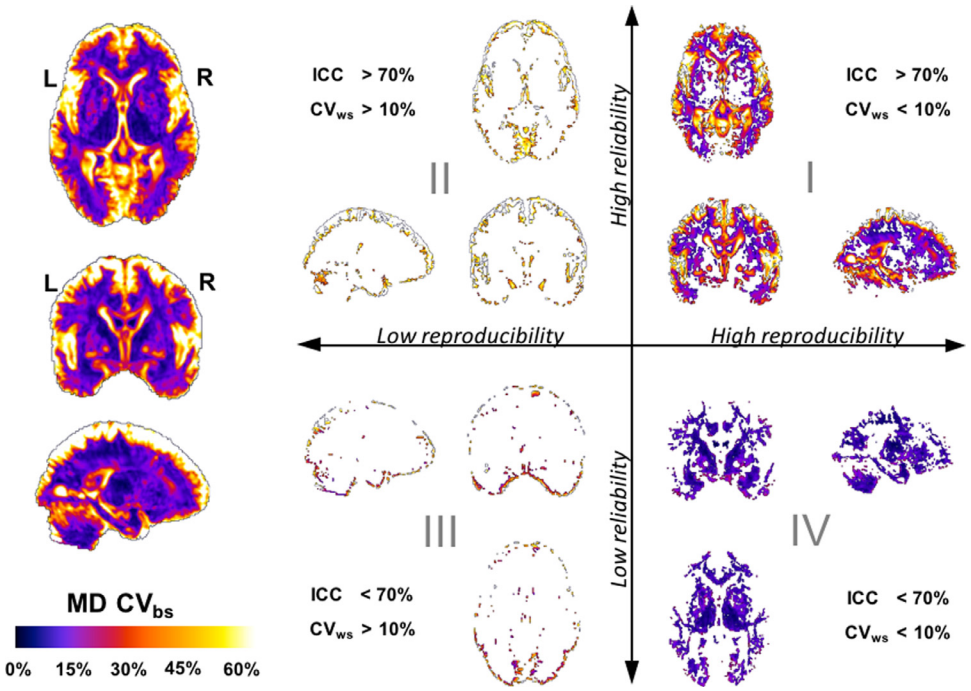

Fig. 6. Multiparametric segmentation of whole-brain MD CV_{bs} by reproducibility and reliability. The MD CV_{bs} whole-brain map (left) is segmented into four groups of anatomical regions showing similar levels of reproducibility and reliability (right). Quadrant I: regions formed by voxels with high MD reproducibility (CV_{ws} <10%) and high MD reliability (ICC >70%) - Quadrant II: regions formed by voxels with low MD reproducibility (CV_{ws} >10%) and high MD reliability (ICC >70%) - Quadrant III: regions formed by voxels with low MD reproducibility (CV_{ws} >10%) and low MD reliability (ICC <70%) - Quadrant IV: regions formed by voxels with high MD reproducibility (CV_{ws} <10%) and low MD reliability (ICC <70%).

variability, it changed greatly from region to region for both metrics. The regions with the lowest between-subject variability were the regions with low reliability and high reproducibility and the other way around.

4.1. Reproducibility

Compared to previous studies, the reproducibility of overall FA and MD values in white-matter and grey-matter regions is high. In particular, in the case of FA, our results have shown better CV_{ws} values compared to previous reproducibility studies reporting values between 1% and 5% (Pfefferbaum et al., 2003; Vollmar et al., 2010; Willats et al., 2014; Veenith et al., 2013; Grech-Sollars et al., 2015). Interestingly, only the CV_{ws} for MD in overall grey matter is slightly higher (CV_{ws} = 4.42%) than previously reported values for this index. When looking at selected regions of interest, the reproducibility values for both FA and MD are consistently better (between 1% and 2.8%) than those reported by a previous study (Marenco et al., 2006) for FA (2.5–10.16%) and MD (2.49–6.20%) with data collected with 8 averages of 6 diffusion directions at 1100 s/mm^2 b-values on a 1.5 Tesla system. In our study, the GCC appeared the least reproducible of all the regions for both FA and MD (FA CV_{ws} = 1.8%, MD CV_{ws} = 2.8%), showing a much higher CV_{ws} than the SCC (FA CV_{ws} = 1.0%, MD CV_{ws} = 2.5%) – the same effect observed by Veenith et al., al. (2013). Other studies showed GCC values closer to those of the SCC, as in Marenco et al. (2006) and Willats et al. (2014), but in these studies, the GCC ROI had a substantially larger size. As expected, the reproducibility of both DTI metrics decreased when evaluated at the voxel level. This is in line with the voxel-wise values reported in previous studies (Jansen et al., 2007; Pannek et al., 2011; Willats et al., 2014). In the white matter, it remained within acceptable levels (below 10%) in a majority of voxels for both metrics. In grey-matter voxels, only MD maintained good or very good reproducibility levels for most voxels. On the contrary, FA showed good reproducibility in less than 25% of grey-matter voxels. The observed decrease in reproducibility for overall measures and ROIs replicates the effect reported by Vollmar et al. (2010), where the CV_{ws} was found to be more than twice as large when using voxel-wise values compared to ROI averages. The CV_{ws} maps show consistently low FA reproducibility in grey-matter voxels, but good reproducibility for MD as in Marenco et al. (2006) and Willats et al. (2014). Notwithstanding, many voxels in grey-matter structures such as the thalamus share similar CV_{ws} values with white-matter voxels. Overall, these results suggest that even in an older populations DTI metrics provide a good reproducibility across all white matter and most of grey matter regions.

4.2. Reliability

The reliability of overall whole-brain measures is high (above 82% in MD) or very high (over 97% for FA) and always better in the white matter compared to the grey matter. The ICC for overall FA values is also very similar to those reported by Vollmar et al. (2010) and higher than by Grech-Sollars et al. (2015) (88% in grey matter, 53% white matter). The ICC for overall MD values is also higher than in Grech-Sollars et al. (2015) (48% in grey matter, 41% white matter) although in that study all reported values seem particularly low. Looking at regions of interest, the SCC, the OFWM and the LFWM, our ICC results show higher values for MD and similar values for FA compared to Marenco et al. (2006). Compared to another study from Duan et al. (2015), the ICC values from our study are higher in most of the selected regions except the corpus callosum regions (76% vs 87% in GCC and 86% vs 94% in SCC) again probably because of the different sizes of these regions). At the voxel-level, our results are either equivalent to or better than previous studies. E.g. in this work, the mean voxel-wise values of 77% (FA) and 74% (MD) are comparable to the 80% (FA) and 73% (MD) mean values reported by Jansen et al. (2007). Also similar to our results, in Duan et al. (2015) the number of voxels showing more than moderate reliability (ICC > 40%) was around 90% of voxels for FA and 75% of voxels for MD.

4.3. Statistical power

Considering the older population recruited in this study, we can report that the statistical power of our diffusion metrics is more than adequate to detect, from a relatively small number of subjects ($N < 55$), longitudinal changes greater than 4% in most white-matter regions and changes greater than 2% in the deeper white-matter structures. In all regions except for the corpus callosum, MD requires substantially fewer subjects than FA to detect the same percentage change in the metrics values.

4.4. Between-subject variability

Except for deep CSF regions (i.e. ventricles, intra-lobar and intra-hemispheric CSF), a gradient showing high to low between-subject variability is evident when going from more external to internal regions. Here, deep white- and grey-matter structures show the lowest between-subject variability. The multiparametric segmentation of the CV_{bs} maps reveals in anatomical terms the relationship between the reproducibility and the reliability of the metrics with the between-subject variability. Both FA and MD metrics show the highest between-subject variability in the regions where they are also most reliable but less reproducible (cortical regions) while they show the lowest between-subject variability where they are less reliable but more reproducible (subcortical grey matter and deep white matter). In the case of FA, these are the subcortical and the intra-lobar white-matter regions where FA shows medium levels of between-subject variability (CV_{bs} between 20% and 40%). In comparison, the between-subject variability of MD ranges more widely from CV_{bs} = 10% to CV_{bs} = 40% in the corresponding high-reproducibility/high-reliability regions (the cortex and some subcortical white-mater and grey-matter subregions).

4.5. Data quality and age of participants

The improved reproducibility of our results respect to previous studies could be in part explained by the increased quality of the MRI data collected for this study (number of diffusion-weighted directions, number of b0s images directions, the use of peripheral pulse gating, the quality of the hardware). Also, the low within-subject variability of the metrics could have benefited from the fact that each participant was re-scanned at approximately the same time of the day of the first session (Thomas et al., 2018). Altogether, the results suggest that investing in good acquisition protocols may be very beneficial also for the estimation of traditional DTI metrics and not only for more advanced High Angular Resolution Diffusion Imaging (HARDI) methods (Dell'Acqua and Tournier 2018) and tractography applications.

The characteristics of our older cohort could be another factor explaining some of the results in our study. For example, the lower CV_{ws} for the overall FA and MD could be a consequence of older brains being more stable compared to those from younger adults, and therefore less likely to experience significant changes between sessions. Another example could be the slightly higher CV_{ws} for the overall MD in grey-matter regions, which could be explained by the fact that cortical grey matter reduces in the older population, increasing the probability of partial volume with other structures and CSF.

4.6. Effect of the anatomy

In agreement with previous studies (Marenco et al., 2006; Venkatraman et al., 2015), the results from the multiparametric segmentation make evident that the statistical properties of both metrics across the brain are determined not only by the tissue type (grey matter

versus white matter) but also by the anatomy. For example, for both FA and MD, there are clear differences between the high reliability and low reproducibility of voxels within the cortical grey-matter compared to the high reproducibility and mostly low reliability in deep grey-matter. Within the deep grey-matter, the properties of the metrics are further determined by each specific grey-matter structure, for example, the difference in reproducibility values between the thalamus and the basal ganglia. This effect of the anatomy in the properties of the metrics is also visible in the white matter where the reliability goes from being high in intra-lobar white-matter regions to be below 70% near deep grey-matter structures. The anatomy has also a clear effect on the statistical power estimated for each metric across different brain regions. For example, both metrics require far fewer subjects to detect changes in deep white-matter regions compared to more superficial white-matter regions. These findings closely reflect the regional differences depicted by the CV_{bs} maps, because changes in statistical power across regions are related largely to changes in the between-subject variability of the metrics.

4.7. Limitations of the study

While the preprocessing pipeline used in this study follows the recommendations of well-established guidelines for the pre-processing and post-processing of diffusion MRI data (Jones and Cercignani, 2010; Jones et al., 2013), it does not represent the current state of the art. Additional pre-processing steps recently developed specifically for DW-MRI such as data de-noising (Veraart et al., 2016) or the correction of Gibbs ringing artefacts (Veraart et al., 2015), could have improved the reproducibility and reliability of the DTI metrics. The cost-benefit of using extra pre-processing steps will also depend on the quality of the original data. For example, we observed decreased reproducibility in voxels located towards the most frontal and the temporal regions of the brain, suggesting that susceptibility geometrical distortions from EPI may have a significant effect even after pre-processing. These EPI distortions could have been attenuated by acquiring a full DTI dataset with reversed-encoding (Irfanoglu et al., 2015), at the cost of increasing the acquisition time.

The post-processing of the DTI metrics can also make a great difference in the final results. For example, methods based on free water elimination (FWE) (Pasternak et al., 2009) developed to ameliorate the limitations of the single diffusion compartment in DTI. The use of FWE can improve the reproducibility of FA and MD in elderly subjects and without detriment to their sensitivity (Albi et al., 2017). That said, for many studies, the most determinant step after the computation of the DTI maps is probably the normalisation of the images. The method used in this study is the same one provided by the TBSS pipeline (Smith et al., 2006). This approach was chosen over other alternatives because TBSS is a widely used method for the voxel-based analysis of DTI data. By using this normalisation approach, we keep our pipeline in line with the kind of methods used by the cited literature. However, the registration of the images in the TBSS pipeline, based on FA maps, is known to retain some extents of misalignment that might introduce uncertainty in grey matter and at the interface between white and grey matter, influencing later statistical results. This misalignment can negatively impact the reproducibility and reliability of the metrics in cortical regions and near deep grey-matter structures. The results from the multiparametric segmentation showed in Fig. 5 and 6 strongly suggest this possibility. This seems to be the case especially for MD, as normalising each MD map using the same transformations calculated from the corresponding FA map may be not optimal since FA is featureless in grey-matter regions. With an older population, this registration may be more problematic. Therefore, we recommend for the normalisation of the images the use of more advanced methods of image registration such as ANTs (Avants et al., 2011) or methods specific for the registration of DTI data to improve the alignment of brain structures in the FA and MD images (Irfanoglu et al., 2016).

4.8. Relevance of the results and potential applications

The results on the reproducibility and reliability of FA and MD confirm the validity of using these two metrics for the study of brain changes in older adults. The two metrics appear to complement each other in terms of their statistical properties across the brain: FA is more reliable in the white matter while MD is more reproducible and it is associated with higher statistical power in the grey matter.

The CV_{ws} and ICC values in the six anatomical regions and the provided power calculations can guide future studies to decide which metric to use in each region and the number of participants to be recruited. Adequate sample sizes provide sufficient statistical power to the study and help to contain scanning costs, computational needs and the overall running time of the studies (Ioannidis et al., 2014). If sample sizes are relatively small, results tend to be less reliable (Button et al., 2013) and to appear only in regions where the statistical power is comparatively high (usually regions with low between-subject variability). In this context, the reliability (ICC) maps can identify those regions (voxels) where statistical results are most likely to reflect genuine effects of interest in the population. Likewise, the reproducibility (CV_{ws}) maps can identify those regions where the metrics are also robust, a requirement for the satisfactory formulation of any useful biomarker (Strimbu and Tavel, 2010).

The multiparametric segmentation of the CV_{bs} maps reveals those regions where FA and MD are reproducible, reliable, and show adequate levels of between-subject variability and statistical power. Identifying such optimal regions in advance can simplify the statistical analysis and facilitate the control for false positives due to multiple statistical comparisons (Lindquist and Mejia, 2015). Finally, the multiparametric segmentation of the CV_{bs} maps can help researchers and clinicians to understand intuitively the relationship between reproducibility, reliability and between-subject variability, and the effect that the brain anatomy has on these properties of the metrics. Understanding this relationship becomes especially important when evaluating and comparing results across multiple brain regions where the performance of these two metrics may change from region to region (De Santis et al., 2014).

5. Conclusion

In this study, we have characterised the test-retest reproducibility, reliability, between-subject variability and statistical power associated with FA and MD in older healthy subjects. The values of the CV_{ws}, CV_{bs} and ICC across the brain confirm the validity and support the use of FA and MD metrics to study brain changes in this particular age group. Our results also show that FA and MD are very different in terms of their statistical properties, which change in a dissimilar manner across the entire brain anatomy. The statistical power calculations combined with the multiparametric segmentation of the between-subject variability can guide researchers to identify those brain regions where FA and MD would be most effective and reliable in finding statistical differences in diffusion MRI data. This information will be particularly useful for the design, sample-size calculations and interpretation of the results of future studies using DTI as a neuroimaging biomarker in older populations.

Funding

This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, by the King's College London and Imperial College London EPSRC Centre for Doctoral Training in Medical Imaging (grant EP/L015226/1) and by the Sackler Institute for Translational Neurodevelopment, Institute of Psychiatry, Psychology and Neuroscience, King's College London. The study is also part funded by

Janssen Research and Development, a division of Janssen Pharmaceutica N.V. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

CRediT authorship contribution statement

Pedro A. Luque Laguna: Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Anna J.E. Combes:** Resources, Writing - review & editing. **Johannes Streffer:** Writing - review & editing. **Steven Einstein:** Writing - review & editing. **Maarten Timmers:** Writing - review & editing. **Steve C.R. Williams:** Conceptualization, Project administration, Funding acquisition, Writing - review & editing. **Flavio Dell'Acqua:** Methodology, Supervision, Writing - review & editing.

Acknowledgements

We thank Ahmad Beyh and Cedric Ginestet for the support and helpful discussions, Pauline Harding for proofreading the final manuscript, all the participants in the study, and the Wellcome Trust for their ongoing support of our neuroimaging endeavours.

Appendix A. Indexes of reproducibility, reliability and between-subject variability

When data have been acquired for each subject on multiple occasions, the statistical properties of the metrics such as reproducibility, reliability and the between-subject variability can be summarised by statistical indices based on the different sources of variability. The within-subject variability summarises the variability observed when acquiring data from the same subjects on multiple occasions. The between-subject variability is the variability of the metric across different subjects (after excluding the within-subject variability). The mathematical definitions of each index based on the corresponding variabilities are defined next.

Within-subject coefficient of variation

The within-subject coefficient of variation (CV_{ws}) summarises the within-subject variability of the metric across multiple sessions,

$$CV_{ws} = \frac{\sigma_{ws}}{\mu} \times 100\%$$

where μ is the grand mean and σ_{ws} is the within-subject standard deviation which is estimated by

$$\hat{\sigma}_{ws} = \sqrt{\frac{n}{k-1} \times WMS}$$

where k is the total number of sessions, n is the total number of subjects and WMS is the within-subject mean square. A lower $CV_{ws} > 0$ values are associated with higher reproducibility.

Intraclass correlation coefficient

As an index of reliability, we use the third form of the intraclass correlation coefficient $ICC_{3,1}$. The $ICC_{3,1}$ consists of between-subject variability divided by the total variance of the metric according to a two-way mixed-effects model (fixed session effects random subject effects). Under this model, the $ICC_{3,1}$ can be consistently estimated (with bias) by

$$ICC_{3,1} = \frac{BMS - EMS}{BMS - (k-1)EMS}$$

where BMS is the between-subject mean square, EMS is the error mean square and k is the total number of sessions. A higher $0 < ICC_{3,1} < 1$ value is associated with a higher level of reliability (consistency) in the metric's values across the different scanning sessions.

Between-subject coefficient of variation

The between-subject coefficient of variation (CV_{bs}) summarises the between-subject variability of the metric across the population.

$$CV_{bs} = \frac{\sigma_{bs}}{\mu} \times 100\%$$

where μ is the grand mean and σ_{bs} is the between-subject standard deviation which is estimated by

$$\hat{\sigma}_{bs} = \sqrt{\frac{k}{n-1} \times BMS}$$

where k is the total number of sessions and n is the total number of subjects and BMS is the between-subject mean square.

Within our test-retest scenario, the BMS is computed using the mean values of the metric across three sessions. Therefore, the σ_{bs} can be considered to be a better estimation than the total variance σ to the between-subject variability of the metric. Lower $CV_{bs} > 0$ values are associated with higher statistical power when using the metric to detect group differences.

Appendix B. Statistical power calculations

We perform statistical power calculations based on the between-subject variability of the metrics to estimate the minimum sample size required by each metric to detect a statistically significant change in the metrics from their nominal values. More specifically, we calculated the number of subjects required per group to detect longitudinal changes in FA and MD of 2%, 4% and 6% (increase or decrease) in each ROI with statistical power $1\text{-}\beta = 0.8$ and statistical significance level $\alpha = 0.05$.

Statistical hypothesis test

For the power calculations, we used the statistical test most commonly employed to detect longitudinal changes in data, a *t*-test on the difference between two dependant means (matched pairs). The null hypothesis of this *t*-test states that the population means μ_x and μ_y of two matched observations *X* and *Y* are equal. In this case, we can reformulate this test in terms of the difference $Z = X - Y$ and the following null and alternative hypothesis:

H_0: $\mu_z = 0$

H_1: $\mu_z \neq 0$.

The number of subjects required to detect a certain value of $|\mu_z| \neq 0$ at significance level α and statistical power 1 - β is given by the minimum number of degrees of freedom n required for a central *t*-distribution to have a critical value *T* such as:

$$|T| = \frac{|\bar{Z}|}{\hat{\sigma_z}\sqrt{n}} > t_{n,1-\alpha/2}$$

and

$$1 - \beta < F_{n-1,\lambda}(t_{n,1-\alpha/2}) - F_{n-1,\lambda}(-t_{n,1-\alpha/2})$$

where $\bar{Z}$ is the difference between sample means, $\hat{\sigma}^2{}_z$ is the unbiased sample variance, $t_{n,1-\alpha/2}$ is the $(1 - \alpha/2)$th percentile of a central *t*-distribution with n degrees of freedom, and $F_{n-1,\lambda}$ is the cumulative function of a non-central *t*-distribution with $n - 1$ degrees of freedom and noncentrality parameter $\lambda = \sqrt{n}\mu_z/\sigma_z$.

Power calculations

The previous computation can be carried out by the statistical software package G*power (Faul et al., 2007) that requires passing to it as inputs the desired significance level α, the statistical power $1 - \beta$, and the effect size index for the corresponding *t*-test (Cohen 1988):

$$d_z = \frac{|\mu_z|}{\sigma_z} = \frac{|\mu_x - \mu_y|}{\sqrt{{\sigma_x}^2 + {\sigma_y}^2 - 2\rho_{xy}\sigma_x\sigma_y}}$$

where ${\sigma_x}^2$ and ${\sigma_y}^2$ are the population variances corresponding to the first and the second observations, and ρ_{xy} is the correlation between the random variables corresponding to the two observations. To obtain the size effect as required by G*power, we used the *in-vivo* MRI data from the first session to estimate the value of the population mean μ_x and the population variance ${\sigma_x}^2$ corresponding to the first observation *X*. We also pooled the MRI data from the second and third sessions to estimate the population variance ${\sigma_y}^2$ for the second observation *Y* and the correlation ρ_{xy} between the first observation *X* and the second observation *Y*. By setting $\mu_x - \mu_y$ to the required value for each percentage change (*pc*), we compute d_z as

$$d_z = \frac{|\mu_x \cdot pc|}{\sqrt{{\sigma_x}^2 + {\sigma_y}^2 - 2\rho_{xy}\sigma_x\sigma_y}}$$

where $pc = 0.02, 0.04, 0.06$. Finally, we used G*power to compute the minimum sample size required to detect each of the computed effect sizes with a statistical significance $\alpha = 0.05$ and a statistical power $\beta = 0.8$.

References

Albi, A., Pasternak, O., Minati, L., Marizzoni, M., Bartrés-Faz, D., Bargalló, N., Bosch, B., et al., 2017. Free water elimination improves testretest reproducibility of diffusion tensor imaging indices in the brain: a longitudinal multisite study of healthy elderly subjects. Hum. Brain Mapp. 38 (1), 12–26. https://doi.org/10.1002/hbm.23350.

Andersson, J.L.R., Skare, S., Ashburner, J., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20 (2), 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7.

Andersson, J.L.R., Sotiropoulos, S.N., 2015. Non-parametric representation and prediction of single- and multi-shell diffusion-weighted MRI data using Gaussian processes. Neuroimage 122, 166–176. https://doi.org/10.1016/j.neuroimage.2015.07.067.

Avants, B.B, Nicholas, J.T., Gang, S., Philip, A.C., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54 (3), 2033–2044. https://doi.org/10.1016/J.NEUROIMAGE.2010.09.025.

Barrick, T.R., Charlton, R.A., Clark, C.A., Markus, H.S., 2010. White matter structural decline in normal ageing: a prospective longitudinal study using tract-based spatial statistics. Neuroimage 51 (2), 565–577. https://doi.org/10.1016/j.neuroimage.2010.02.033.

Basser, P.J., Pierpaoli, C., 1996. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. J. Magn. Reson. - Series B 111 (3), 209–219. https://doi.org/10.1006/jmrb.1996.0086.

Beaulieu, C., 2002. The basis of anisotropic water diffusion in the nervous system - A technical review. NMR Biomed. 15 (7–8), 435–455. https://doi.org/10.1002/nbm.782.

Bender, A.R., Raz, N., 2015. Normal-appearing cerebral white matter in healthy adults: mean change over 2 years and individual differences in change. Neurobiol. Aging 36 (5), 1834–1848. https://doi.org/10.1016/j.neurobiolaging.2015.02.001.

Bisdas, S., Bohning, D.E., Bešenski, N., Nicholas, J.S., Rumboldt, Z., 2008. Reproducibility, interrater agreement, and age-related changes of fractional anisotropy measures at 3T in healthy subjects: effect of the applied b-value. Am. J. Neuroradiol. 29 (6), 1128–1133. https://doi.org/10.3174/ajnr.A1044.

Blain, C.R.V., Williams, V.C., Johnston, C., Stanton, B.R., Ganesalingam, J., Jarosz, J.M., Jones, D.K., et al., 2007. A longitudinal study of diffusion tensor MRI. In: ALS." Amyotrophic Lateral Sclerosis. Taylor & Francis. 8. pp. 348–355. https://doi.org/10.1080/17482960701548139.

Boekel, W., Forstmann, B.U., Keuken, M.C., 2017. "A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control.". Kappenman, E.S., Keil, A. (Eds.), "A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control.". Psychophysiology 54 (1), 24–33. https://doi.org/10.1111/psyp.12769.

Boespflug, E.L., Storrs, J.M., Allendorfer, J.B., Lamy, M., Eliassen, J.C., Page, S., 2011. Mean diffusivity as a potential diffusion tensor biomarker of motor rehabilitation after electrical stimulation incorporating task specific exercise in stroke: a pilot study. Brain Imaging Behav. 8 (3), 359–369. https://doi.org/10.1007/s11682-011-9144-1.

Brandstack, N., Kurki, T., Laalo, J., Kauko, T., Tenovuo, O., 2016. Reproducibility of tract-based and region-of-interest DTI analysis of long association tracts. Clin. Neuroradiol. 26 (2), 199–208. https://doi.org/10.1007/s00062-014-0349-8.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., N.osek, B.A., Flint, J., Robinson, E.S.J., M.unafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14 (5), 365–376. https://doi.org/10.1038/nrn3475.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Chang, L.-C., Jones, D.K., Pierpaoli, C., 2005. RESTORE: robust estimation of tensors by outlier rejection. Magn. Reson. Med. 53 (5), 1088–1095. https://doi.org/10.1002/mrm.20426.

Cohen, J., 1988. Chapter 2. The t Test for Means." In Statistical Power Analysis for the Behavioral Sciences. Academic Press, CY, pp. 19–74.

Cox, S.R., Ritchie, S.J., Tucker-Drob, E.M., Liewald, D.C., Hagenaars, S.P., Davies, G., Wardlaw, J.M., Gale, C.R., Bastin, M.E., Deary, I.J., 2016. Ageing and brain white matter structure in 3,513 UK Biobank participants. Nat. Commun. 7 (1), 632. https://doi.org/10.1038/ncomms13629.

Croall, I.D., Lohner, V., Moynihan, B., Khan, U., Hassan, A., O'Brien, J.T., Morris, R.G., et al., 2017. Using DTI to assess white matter microstructure in cerebral small vessel disease (SVD) in multicentre studies. Clin. Sci. 131 (12), 1361–1373. https://doi.org/10.1042/CS20170146.

De Santis, S., Drakesmith, M., Bells, S., Assaf, Y., Jones, D.K., 2014. Why diffusion tensor MRI does well only some of the time: variance and covariance of white matter tissue microstructure attributes in the living human brain. Neuroimage 89, 35–44. https://doi.org/10.1016/j.neuroimage.2013.12.003.

Dell'Acqua, F., Donald Tournier, J., 2018. Modelling white matter with spherical deconvolution: how and why? NMR Biomed. 32 (4), e3945. https://doi.org/10.1002/nbm.3945.

Dell'Acqua, F., Lacerda, L., Catani, M., Simmons, A., 2014. Anisotropic power maps: a diffusion contrast to reveal low anisotropy tissues from Hardi data. In: ISMRM2014, Milan.

Duan, F., Zhao, T., He, Y., Shu, N., 2015. Test-retest reliability of diffusion measures in cerebral white matter: a multiband diffusion MRI study. J. Magn. Reson. Imaging 42 (4), 1106–1116. https://doi.org/10.1002/jmri.24859.

Duning, T., Schiffbauer, H., Warnecke, T., Mohammadi, S., Floel, A., Kolpatzik, K., Kugel, H., et al., 2011. G-CSF prevents the progression of structural disintegration of white matter tracts in amyotrophic lateral sclerosis: a pilot trial. PLoS ONE 6 (3), e17770. https://doi.org/10.1371/journal.pone.0017770.

Egger, K., Christian, C., Hohenberg, Michael, F., Schocke, C., G.uttmann, R.G., Demian Wassermann, M., Wigand, C., Wolfgang Nachbauer, et al., 2013. White matter changes in patients with Friedreich Ataxia after treatment with erythropoietin. J. Neuroimaging 24 (5), 504–508. https://doi.org/10.1111/jon.12050.

Ellingson, B.M., Malkin, M.G., Rand, S.D., Connelly, J.M., Quinsey, C., LaViolette, P.S., Bedekar, D.P., Schmainda, K.M., 2010. Validation of functional diffusion maps (fDMs) as a biomarker for human glioma cellularity. J. Magn. Reson. Imaging 31 (3), 538–548. https://doi.org/10.1002/jmri.22068.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 39. Springer-Verlag, pp. 175–191. https://doi.org/10.3758/BF03193146.

Galluzzi, S., Marizzoni, M., Babiloni, C., Albani, D., Antelmi, L., Bagnoli, C., Bartrés-Faz, D., et al., 2016. Clinical and biomarker profiling of prodromal Alzheimer's disease in workpackage 5 of the innovative medicines initiative pharmacog project: a European Adni study'. J. Intern. Med. 279 (6), 576–591. https://doi.org/10.1111/joim.12482.

Goveas, J., O'Dwyer, L., Mascalchi, M., Cosottini, M., Diciotti, S., Santis, SDe, Passamonti, L., Tessa, C., Toschi, N., Giannelli, M., 2015. Diffusion-MRI in neurodegenerative disorders. Magn. Reson. Imaging 33 (7), 853–876. https://doi.org/10.1016/j.mri.2015.04.006.

Grech-Sollars, M., Hales, PW., Miyazaki, K., Raschke, F., Rodriguez, D., Wilson, M., Gill, SK., et al., 2015. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. NMR Biomed. 28 (4), 468–485. https://doi.org/10.1002/nbm.3269.

Heiervang, E., Behrens, T.E.J., Mackay, C.E., Robson, M.D., Johansen-Berg, H., 2006. Between session reproducibility and between subject variability of diffusion MR and tractography measures. Neuroimage 33 (3), 867–877. https://doi.org/10.1016/j.neuroimage.2006.07.037.

Howells, H., Schotten, MTDe, Dell'Acqua, F., Beyh, A., Zappalà, G., Leslie, A., Simmons, A., Declan, G., Murphy, M.C., 2018. Frontoparietal tracts linked to lateralized hand preference and manual specialization. Cerebral Cortex 28 (7), 2482–2494. https://doi.org/10.1093/cercor/bhy040.

Huhn, S., Beyer, F., Zhang, R., Lampe, L., Grothe, J., Kratzsch, J., Willenberg, A., et al., 2018. Effects of resveratrol on memory performance, hippocampus connectivity and microstructure in older adults A randomized controlled trial. Neuroimage 174, 177–190. https://doi.org/10.1016/j.neuroimage.2018.03.023.

Ioannidis, J.P.A., Sander, G., Mark A, H., Muin J, K., Malcolm, R., Macleod, D.M., Kenneth, F.S., Robert Tibshirani, 2014. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383 (9912), 166–175. https://doi.org/10.1016/S0140-6736(13)62227-8.

Irfanoglu, M.O., Modi, P.., Nayak, A., H.utchinson, E.B., Sarlls, J., Pierpaoli, C., 2015. DR-BUDDI (Diffeomorphic registration for blip-Up blip-Down diffusion imaging) method for correcting echo planar imaging distortions. Neuroimage 106, 284–299. https://doi.org/10.1016/j.neuroimage.2014.11.042.

Irfanoglu, M.O., Nayak, A., Jenkins, J., Hutchinson, E.B., Sadeghi, N., Thomas, C.P., Pierpaoli, C., 2016. DR-TAMAS: diffeomorphic registration for tensor accurate alignment of anatomical structures. Neuroimage 132, 439–454. https://doi.org/10.1016/J.NEUROIMAGE.2016.02.066.

Jansen, JF.A., Eline Kooi, M., Kessels, AG.H., Nicolay, K., Backes, WH., 2007. Reproducibility of quantitative cerebral T2 relaxometry, diffusion tensor imaging, and 1H magnetic resonance spectroscopy at 3.0 tesla. Invest. Radiol. 42 (6), 327–337. https://doi.org/10.1097/01.rli.0000262757.10271.e5.

Jones, D.K., Cercignani, M., 2010. Twenty-five pitfalls in the analysis of diffusion MRI data. NMR Biomed. 23 (7), 803–820. https://doi.org/10.1002/nbm.1543.

Jones, D.K., Knösche, T..R., Turner, R., 2013. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. Neuroimage 73, 239–254. https://doi.org/10.1016/j.neuroimage.2012.06.081.

Kapoor, R., 2017. Advanced MRI measures like DTI or fMRI should be outcome measures in future clinical trials YES. Mult. Scler. 23 (11), 1454–1455. https://doi.org/10.71177173/13542458.

Kitamura, S., Kiuchi, K., Taoka, T., Hashimoto, K., Ueda, S., Yasuno, F., Morikawa, M., Kichikawa, K., Kishimoto, T., 2013. Longitudinal white matter changes in Alzheimer's disease: a tractography-based analysis study. Brain Res. 1515, 12–18. https://doi.org/10.1016/j.brainres.2013.03.052.

Andren, K., A., K., Gabel, NM., Stelmokas, J., Rich, AM., Bieliauskas, LA., 2017. Population base rates and disease course of common psychiatric and neurodegenerative disorders. Neuropsychol. Rev. 27 (3), 284–301. https://doi.org/10.1007/s11065-017-9357-1.

Lebel, C., Gee, M., Camicioli, R., Wieler, M., Martin, W., Beaulieu, C., 2012. Diffusion tensor imaging of white matter tract evolution over the lifespan. Neuroimage 60 (1), 340–352. https://doi.org/10.1016/j.neuroimage.2011.11.094.

Leemans, A., Jeurissen, B., Sijbers, J., Jones, D.K., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. In: 17th Annual Meeting of Intl Soc Mag Reson Med. Hawaii, USA. pp. 3537. http://www.exploredti.com.

Lemkaddem, A., Daducci, A., Vulliemoz, S., O'Brien, K., Lazeyras, F., Hauf, M., Wiest, R., et al., 2012. A multi-center study: intra-scan and inter-scan variability of diffusion spectrum imaging. Neuroimage 62 (1), 87–94. https://doi.org/10.1016/j.neuroimage.2012.04.045.

Lindquist, M.A., Mejia, A., 2015. Zen and the art of multiple comparisons. Psychosom. Med. 77 (2), 114–125. https://doi.org/10.1097/PSY.0000000000000148.

Liu, X., Yang, Y., Sun, J., Yu, G., Xu, J., Niu, C., Tian, H., Lin, P., 2014. Reproducibility of diffusion tensor imaging in normal subjects: an evaluation of different gradient sampling schemes and registration algorithm. Neuroradiology 56 (6), 497–510. https://doi.org/10.1007/s00234-014-1342-2.

Lövdén, M., Köhncke, Y., Laukka, E.J., Grégoria, K., Alireza, S., Tie Qiang, Li, Fratiglioni, L., Bäckman, L., 2014. Changes in perceptual speed and white matter microstructure in the corticospinal tract are associated in very old age. Neuroimage 102 (P2), 520–530. https://doi.org/10.1016/j.neuroimage.2014.08.020.

Madden, D., Ilana, J., Bennett, J., Burzynska, A., Potter, G.G., Nan kuei, C., Song, A.W., 2012. Diffusion tensor imaging of cerebral white matter integrity in cognitive aging. Biochimica et Biophysica Acta - Mol. Basis Dis. 1822 (3), 386–400. https://doi.org/10.1016/j.bbadis.2011.08.003.

Mahoney, C.J., Simpson, I.J.A., Nicholas, J.M., Fletcher, P.D., Downey, L.E., Golden, H.L., Clark, C.N., et al., 2015. Longitudinal diffusion tensor imaging in rontotemporal dementia. Ann. Neurol. 77 (1), 33–46. https://doi.org/10.1002/ana.24296. John Wiley & Sons, Ltd.

Marenco, S., Rawlings, R., Rohde, G.K., Barnett, AS., Honea, RA., Pierpaoli, C., Weinberger, D.R., 2006. Regional distribution of measurement error in diffusion tensor imaging. Psychiatry Res. - Neuroimaging 147 (1), 69–78. https://doi.org/10.1016/j.pscychresns.2006.01.008.

Moffat, B.A., Chenevert, T..L., L.awrence, T.S., M.eyer, C.R., J.ohnson, T.D., Dong, Q., Tsien, C., et al., 2005. Functional diffusion map: a noninvasive MRI biomarker for early stratification of clinical brain tumor response. Proc. Natl. Acad. Sci. USA 102 (15), 5524–5529 10.1073 pnas.0501532102.

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., Hua, K., et al., 2008. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. Neuroimage 40 (2), 570–582. https://doi.org/10.1016/j.neuroimage.2007.12.035.

Nath, K., Ramola, M., Husain, M., Kumar, M., Prasad, K., Gupta, R., 2010. Assessment of therapeutic response in patients with brain abscess using diffusion tensor imaging. World Neurosurg. 73 (1), 63–68. https://doi.org/10.1016/j.surneu.2009.04.003.

Paldino, M.J., Desjardins, A., F.riedman, H.S., V.redenburgh, J.J., B.arboriak, D.P., 2012. A change in the apparent diffusion coefficient after treatment with bevacizumab is associated with decreased survival in patients with recurrent glioblastoma multiforme. Br. J. Radiol. 85 (1012), 382–389. https://doi.org/10.1259/bjr/24774491.

Pannek, K., Mathias, JL., Bigler, ED., Brown, G., Taylor, JD., Rose, S.E., 2011. The average pathlength map: a diffusion mri tractography-derived index for studying brain pathology. Neuroimage 55 (1), 133–141. https://doi.org/10.1016/j.neuroimage.2010.12.010.

Pasternak, O., Sochen, N., Gur, Y., Intrator, N., Assaf, Y., 2009. Free water elimination and mapping from diffusion MRI. Magn Reson Med 62 (3), 717–730. https://doi.org/10.1002/mrm.22055.

Pfefferbaum, A., Adalsteinsson, E., Sullivan, E.V., 2003. Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. J.Magn. Reson. Imaging 18 (4), 427–433. https://doi.org/10.1002/jmri.10377.

Pierpaoli, C., Barnett, A., Pajevic, S., Chen, R., Penix, LaR, Virta, A., Basser, P., 2001. Water diffusion changes in wallerian degeneration and their dependence on white matter architecture. Neuroimage 13 (6), 1174–1185. https://doi.org/10.1006/nimg.2001.0765.

Raja, R., Rosenberg, G., Caprihan, A., 2019. Review of diffusion mri studies in chronic white matter diseases. Neurosci. Lett. 694 (February), 198–207. https://doi.org/10.1016/j.neulet.2018.12.007.

Sadeghi, N., Nayak, A., Walker, L., Okan Irfanoglu, M., Albert, PS., Pierpaoli, C., Brain Development Cooperative Group, 2015. Analysis of the contribution of experimental bias, experimental noise, and inter-subject biological variability on the assessment of developmental trajectories in diffusion MRI studies of the brain. Neuroimage 109, 480–492. https://doi.org/10.1016/j.neuroimage.2014.12.084.

Schmierer, K., Altmann, D.R., Kassim, N., Kitzler, H., Kerskens, CM., Doege, CA., Aktas, O., et al., 2004. Progressive change in primary progressive multiple sclerosis normal-appearing white matter: a serial diffusion magnetic resonance imaging study. Mult. Scler. 10 (2), 182–187. https://doi.org/10.1191/1352458504ms996oa. SAGE Publications Ltd STM.

Sexton, C.E., Walhovd, K.B., Storsve, A.B., Tamnes, C.K., Westlye, L.T., Johansen-Berg, H., Fjell, A.M., 2014. Accelerated changes in white matter microstructure during aging: a longitudinal diffusion tensor imaging study. J. Neurosci. 34 (46), 15425–15436. https://doi.org/10.1523/JNEUROSCI.0203-14.2014.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 86 (2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31 (4), 1487–1505. https://doi.org/10.1016/j.neuroimage.2006.02.024.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 1 (SUPPL. 1), S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051. 23 Suppl.

Strimbu, K., Tavel, J.A., 2010. What are biomarkers? Curr. Opin. HIV AIDS 5 (6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177.

Sullivan, E.V., Rohlfing, T., Pfefferbaum, A., 2010. Longitudinal study of callosal microstructure in the normal adult aging brain using quantitative DTI fiber tracking. Dev. Neuropsychol. 35 (3), 233–256. https://doi.org/10.1080/87565641003689556.

Teipel, SJ., Meindl, T., Wagner, M., Stieltjes, B., Reuter, S., Hauenstein, K.H., Filippi, M., Ernemann, U., Reiser, MF., Hampel, H., 2010. Longitudinal changes in fiber tract integrity in healthy aging and mild cognitive impairment: a DTI follow-up study. J. Alzheimer's Dis. 22 (2), 507–522. https://doi.org/10.3233/JAD-2010-100234.

Consortium, P.C., Marizzoni, M., Ferrari, C., Jovicich, J., Albani, D., Babiloni, C., Cavaliere, L., et al., 2019. Predicting and tracking short term disease progression in amnestic mild cognitive impairment patients with prodromal Alzheimer's disease: structural brain biomarkers. J. Alzheimer's Dis. 69 (1), 3–14. https://doi.org/10.3233/JAD-180152.

Thomas, C., Sadeghi, N., Nayak, A., Trefler, A., Sarlls, J., Baker, C.I., Pierpaoli, C., 2018. Impact of time-of-day on diffusivity measures of brain tissue derived from diffusion tensor imaging. Neuroimage 173 (June), 25–34. https://doi.org/10.1016/j.neuroimage.2018.02.026.

Van Essen, DC., Smith, SM., Barch, DM., Behrens, TE.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.

Vavasour, I.M., Tam, R.., B. Li, D.K., Laule, C., Taylor, C., K.olind, S.H., M.ackay, A.L., Javed, A., Traboulsee, A., 2019. A 24-month advanced magnetic resonance imaging study of multiple sclerosis patients treated with alemtuzumab. Mult. Scler. J. 25 (6), 811–818. https://doi.org/10.811770/10358254.

Edited byVeenith, TV., Eleanor, C., Grossac, J., Newcombe, V.F.J., Outtrim, JG., Lupson, V., Williams, GB., Menon, DK., Coles, JP., 2013. “Inter subject variability and reproducibility of diffusion tensor imaging within and between different imaging sessions. Zuo, Xi-N (Ed.), “Inter subject variability and reproducibility of diffusion tensor imaging within and between different imaging sessions. PLoS ONE 8 (6), e65941. https://doi.org/10.1371/journal.pone.0065941.

Venkatraman, V.K., Gonzalez, C.E., Landman, B., Goh, J., R.eiter, D.A., An, Y., R.esnick, S.M., 2015. Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths. Neuroimage 119, 406–416. https://doi.org/10.1016/j.neuroimage.2015.06.078.

Veraart, J., Fieremans, E., Jelescu, I.O., Knoll, F., Novikov, D.S., 2015. Gibbs ringing in diffusion MRI. Magn. Reson. Med. 76 (1), 301–314. https://doi.org/10.1002/mrm.25866.

Veraart, J., Novikov, D.S., Christiaens, D., Ades-aron, B., Sijbers, J., Fieremans, E., 2016. Denoising of diffusion MRI using random matrix theory. Neuroimage 142 (November), 394–406. https://doi.org/10.1016/j.neuroimage.2016.08.016.

Vollmar, C., O'Muircheartaigh, J., Gareth, J., Mark, B., Symms, R., Thompson, P., Kumari, V., Duncan, JS., Richardson, MP., Koepp, MJ., 2010. Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. Neuroimage 51 (4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046.

Vos, S.B., Jones, D.K., Jeurissen, B., V.iergever, M.A., Leemans, A., 2012. The influence of complex white matter architecture on the mean diffusivity in diffusion tensor MRI of the human brain. Neuroimage 59 (3), 2208–2216. https://doi.org/10.1016/j.neuroimage.2011.09.086.

Wang, J.Y., Abdi, H., Bakhadirov, K., Diaz-Arrastia, R., Devous, M.D., 2012. A comprehensive reliability assessment of quantitative diffusion tensor tractography. Neuroimage 60 (2), 1127–1138. https://doi.org/10.1016/j.neuroimage.2011.12.062.

Willats, L., Raffelt, D., Robert, E., Smith, J., Tournier, D., Connelly, A., Calamante, F., 2014. Quantification of track-weighted imaging (TWI): characterisation of within-subject reproducibility and between-subject variability. Neuroimage 87, 18–31. https://doi.org/10.1016/j.neuroimage.2013.11.016.

Zhou, X., Sakaie, K.E., Debbins, J.P., Narayanan, S., Fox, R.J., Lowe, M.J., 2018. Scan-rescan repeatability and cross-scanner comparability of DTI metrics in healthy subjects in the Sprint-MS multicenter trial. Magn. Reson. Imaging 53, 105–111. https://doi.org/10.1016/j.mri.2018.07.011. November.

Zivadinov, R., Bergsland, N., Hagemeier, J., Tavazzi, E., Ramasamy, D.P., Durfee, J., Cherneva, M., et al., 2018. Effect of switching from glatiramer acetate 20 mg/daily to glatiramer acetate 40 mg three times a week on gray and white matter pathology in subjects with relapsing multiple sclerosis: a longitudinal DTI study. J. Neurol. Sci. 387, 152–156. https://doi.org/10.1016/j.jns.2018.02.023. April.