NeuroImage 256 (2022) 119249

Contents lists available at ScienceDirect

NeuroImage

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

Test-retest reliability of multi-parametric maps (MPM) of brain microstructure

Norman Aye[a,*], Nico Lehmann[a,h], Jörn Kaufmann[b], Hans-Jochen Heinze[b,c,d,e], Emrah Düzel[c,d,f,g], Marco Taubert[a,d], Gabriel Ziegler[c,f]

[a] Faculty of Human Sciences, Institute III, Department of Sport Science, Otto von Guericke University, Zschokkestraße 32, 39104 Magdeburg, Germany
[b] Department of Neurology, Otto von Guericke University, Leipziger Straße 44, 39120 Magdeburg, Germany
[c] German Center for Neurodegenerative Diseases (DZNE), Leipziger Straße 44, 39120 Magdeburg, Germany
[d] Center for Behavioral and Brain Science (CBBS), Otto von Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany
[e] Leibniz-Institute for Neurobiology (LIN), Brenneckestraße 6, 39118 Magdeburg, Germany
[f] Institute of Cognitive Neurology and Dementia Research, Otto von Guericke University, Leipziger Str. 44, 39120 Magdeburg, Germany
[g] Institute of Cognitive Neuroscience, University College London, Alexandra House, 17-19 Queen Square, Bloomsbury, London, WC1N 3AZ United Kingdom
[h] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103 Leipzig, Germany

ARTICLE INFO

Keywords:
Multiparameter mapping (MPM)
Quantitative MRI
HMRI
Scan-rescan reliability
Reproducibility

ABSTRACT

Multiparameter mapping (MPM) is a quantitative MRI protocol that is promising for studying microstructural brain changes in vivo with high specificity. Reliability values are an important prior knowledge for efficient study design and facilitating replicable findings in development, aging and neuroplasticity research. To explore longitudinal reliability of MPM we acquired the protocol in 31 healthy young subjects twice over a rescan interval of 4 weeks. We assessed the within-subject coefficient of variation (WCV), the between-subject coefficient of variation (BCV), and the intraclass correlation coefficient (ICC). Using these metrics, we investigated the reliability of (semi-) quantitative magnetization transfer saturation (MT_{sat}), proton density (PD), transversal relaxation ($R2^*$) and longitudinal relaxation (R1). To increase relevance for explorative studies in development and training-induced plasticity, we assess reliability both on local voxel- as well as ROI-level. Finally, we disentangle contributions and interplay of within- and between-subject variability to ICC and assess the optimal degree of spatial smoothing applied to the data. We reveal evidence that voxelwise ICC reliability of MPMs is moderate to good with median values in cortex (subcortical GM): MT: 0.789 (0.447) PD: 0.553 (0.264) R1: 0.555 (0.369) $R2^*$: 0.624 (0.477). The Gaussian smoothing kernel of 2 to 4 mm FWHM resulted in optimal reproducibility. We discuss these findings in the context of longitudinal intervention studies and the application to research designs in neuroimaging field.

1. Introduction

Magnetic resonance imaging (MRI) has become an indispensable tool to investigate structural and functional aspects of brain organization and its pathologies. Using the noninvasive technique of MRI, we can quantify changes in gray matter (GM) induced by training, aging and disease and their response to potential treatments (Cercignani et al., 2018). This endeavor clearly goes beyond mapping of macroscopic and mesoscopic morphometric aspects of neuroanatomy such as local brain volumes and cortical thickness (Lerch et al., 2017).

There is a growing field of quantitative MRI (qMRI), that targets new imaging and analysis methods to improve our understanding of the underlying physical brain parameters and image contrast mechanisms driving brain changes in aging and disease. For example, associating certain neurological disorders with specific multi-parametric contrast patterns using quantitative MRI might facilitate conclusions about microstructural processes such as (de-) myelination and iron-accumulation that can characterize disease pathology (Weiskopf et al., 2015).

In commonly used MRI protocols T1 weighted (T1w) images are acquired very often. However, the T1w intensities are influenced by multiple factors, such as sequence type (MPRAGE, (Mugler and Brookeman, 1990) or MDEFT (Deichmann et al., 2004)), sequence parameters for instance repetition time, TR, echo time, TE, or flip angle, and hardware effects (e.g. transmit and receive profiles and any scaling factors). Importantly, local intensity values also depend on multiple physical tissue properties, such as longitudinal and transverse relaxation times, or proton density (Helms et al., 2010, 2009). qMRI aims to account for several of these effects, which should lead to improved speci-

* Corresponding author.
E-mail address: norman.aye@ovgu.de (N. Aye).

https://doi.org/10.1016/j.neuroimage.2022.119249.
Received 8 September 2021; Received in revised form 22 April 2022; Accepted 25 April 2022
Available online 27 April 2022.

Abbreviations

MPM	multiparameter mapping
MT_{sat}	magnetization transfer saturation
PD	proton density
R1	longitudinal relaxation rate
R2*	effective transversal relaxation rate
RF	radio frequency
MS	mean sum of squares
MS_R	mean sum of squares for rows
MS_C	mean sum of squares for columns
MS_E	mean sum of squares for error/residuals
MBSS	mean between subject sum of squares
MWSS	mean within subject sum of squares
ICC	intra-class-correlation coefficient
T1w	T1 weighted images
WCV	within subject coefficient of variation
BCV	between subject coefficient of variation
GM	gray matter
WM	white matter
k	number of scans/raters
n	number of subjects

ficity with respect to (micro-) structural tissue properties of the brain (Cercignani et al., 2018; Lutti et al., 2010; Weiskopf et al., 2013).

Multiparameter mapping (MPM) has been introduced as a protocol that deduces (semi-) quantitative and reproducible maps of multiple physical imaging parameters such as magnetization transfer saturation (MT_{sat}), proton density (PD), longitudinal relaxation rate (R1), effective transversal relaxation rate (R2*) (Helms et al., 2008a; Weiskopf et al., 2013). MPM has potential for assessing abnormalities in brain tissue microstructure of diseases like multiple sclerosis (MS) (Lommers et al., 2019) or neurodegeneration, (Ziegler et al., 2018) age-related changes (Callaghan et al., 2014; Draganski et al., 2011; Lambert et al., 2013; Lorio et al., 2014; Steiger et al., 2016; Whitaker et al., 2016) and training-induced plasticity (Carey et al., 2017; Dick et al., 2017; Helbling et al., 2015; Lutti et al., 2014). Especially gray matter is an often investigated tissue in studies observing structural changes throughout interventions, aging or diseases (Gallardo-Ruiz et al., 2019; Kodama et al., 2018; Krajcovicova et al., 2019; Terribilli et al., 2011). However, it is currently not known whether the MPM acquisition protocol yields microstructural maps with high (long-term) reproducibility. Quantification is also expected to increase comparability across measurement timepoints in longitudinal studies and imaging sites in multicentric studies (Deoni et al., 2008; Weiskopf et al., 2013), which might improve the sensitivity and reduce biases in longitudinal studies of development, plasticity and disease progression in rare conditions.

Long-term reproducibility of imaging biomarkers is highly relevant for sensitivity during early stage detection of neurodegenerative diseases such as MS, Parkinson, and Alzheimer's disease. Moreover, good reproducibility of qMRI maps is especially important in intervention and training studies, when expected effect sizes of brain changes are comparably small (Tardif et al., 2016), or when focusing on individual differences analysis in longitudinal studies (Ziegler et al., 2020, 2019). Reliability-based choices of power analysis and standardized workflows for design of future studies ultimately lead to less noisy observations that could protect us from weak evidence, biased reports and false overestimations (Loken and Gelman, 2017; Poldrack et al., 2016).

In psychology and neuroimaging, reliability of a measure is related to two aspects of variability: (A) the amount of variation over repeated observations and (B) the ability to distinguish among individuals (Cercignani et al., 2018; Hopkins, 2000). More specifically, within-subject variation (WCV) quantifies the inconsistency (or standard deviation) of repeated measures of the same subjects, thus also called random error or noise in the measurements. Its value is often expressed as the standard deviation over repetitions in percentage units of the measure's mean value. The smaller the WCV the better the reproducibility. Between-subject variation (BCV) is expressed as the sample standard deviation (after averaging over available repeated observations) in units of the sample mean. BCV is an indicator of sample heterogeneity or amount of individual differences. Moreover, the intra-class coefficient (ICC) is a popular indicator for reliability that integrates both above aspects (Bartko, 1966; Shrout and Fleiss, 1979).

Although reliability of traditional macrostructural measures has been studied repeatedly (Goto et al., 2015; Schnack et al., 2010; Shuter et al., 2008), only a few studies have examined long-term reproducibility of novel qMRI parameters. Leutritz et al. (2020) observed the reliability of the MPM protocol for six different scanners (including the vendors Philips and Siemens) using the hMRI toolbox for preprocessing (Tabelow et al., 2019). They found for MT_{sat}, PD and R1 an intra-site CoV between 4 and 10% and for R2* 16% over time points and sites (Leutritz et al., 2020).

This and other previous studies are limited in that (A) the sample size was comparably small ($n \leq 11$) (Leutritz et al., 2020; Péran et al., 2007; Schwartz et al., 2019; Zou et al., 2010) which might limit the generalizability due to potential small sample biases; (B) studies only explored a single quantitative parameter (Levesque et al., 2010; Péran et al., 2007; Schwartz et al., 2019; Zou et al., 2010); (C) studies reported reliability only on regional level (Schwartz et al., 2019; Zou et al., 2010) although prior knowledge about location of effects can be scarce; (D) covered only one metric of reliability (Péran et al., 2007; Schwartz et al., 2019); and (E) focused on a very short scan-rescan interval (Leutritz et al., 2020; Péran et al., 2007). The influence of smoothing on MRI data is also not fully understood in processing MRI data and have a huge impact on the results of statistical tests (Ashburner and Friston, 2000; Jones et al., 2005).

Here we aim to overcome some limitations and study the longitudinal reliability of quantitative MRI in cortical and subcortical gray matter brain areas using the MPM protocol. We examine reliability in all 4 MPM's with an emphasize on MT_{sat} and R2* maps since these are increasingly used in quantitative MRI studies. The key contribution is a comparison of multiple reliability metrics (WCV, BCV, ICC) on the voxel- and region-based level. We finally identify the optimal amount of smoothing to facilitate more reproducible results in future studies.

2. Material & methods

2.1. Participants and experimental design

In order to assess the test-retest reliability of MPM, two MRI measurements separated by four weeks were acquired. A scan-rescan interval of several weeks is commonly used in neuroimaging plasticity studies (Valkanova et al., 2014). The $n = 31$ participants (6 ♀, 25 ♂; age: $M = 22.9$, SD = 3.92, range 19–35; BMI: $M = 21.3$, SD = 3, range 15.9–28.2) had no history of systemic, psychiatric or neurological diseases. In terms of cognitive functions, the age range from 19 to 35 years reflects a comparably homogenous phase of human ontogeny (Li et al., 2004) and structural brain development (Mills et al., 2016). The study was carried out in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Otto von Guericke University Magdeburg (approval number 106/98). Written informed consent was obtained from all participants.

2.2. MR image acquisition

The data were acquired by a 3T MAGNETOM Prisma system (Siemens Healthcare, Erlangen, Germany) using a 64-channel head coil. We used the same measurement protocol for each volunteer and session. The second scan was 27 to 44 (mean: 34 SD: 4,04) days after the first scan at nearly the same time of day to minimize the influence

of time-of-day variability (Trefler et al., 2016). We instructed participants to not consume coffee or energy drinks directly before and alcohol 24 h before the measurement and documented what and how much they drank. We observed no difference of fluid intake between the two measurement points. To minimize the impact of head orientation variations on different measurements, all subjects were carefully positioned in every examination by a trained medical technical assistant (MTA). The receive brain coil was fixed on the table at the same position using a hold-down groove. All subjects lied down with head-first supine posture and isocenter landmarks were positioned between the eyebrows. The body of the subjects adjusted parallel to main magnetic field before examination. Subjects were introduced to relax and keep their mind free of any thoughts while move as little as possible.

We acquired the MPM protocol (for details and references see e.g. Tabelow et al. (2019)) using three different predominant T1-, PD-, and MT-weighted images with multi-echo FLASH scans by appropriate choice of the repetition time (TR) and the flip angle α: TR/α = 23.0 ms/25° for T1w scan, 23.0 ms/5° for PDw scan, and 37.0 ms/7° for MTw scan. Multiple gradient echoes were acquired with alternating readout polarity at 8 equidistant echo times (TE) between 2.46 ms and 19.68 ms for T1w and PDw acquisitions and at 6 equidistant TE between 2.46 ms and 14.76 ms for MTw acquisition. Other acquisition parameters were: 0.8 mm isotropic resolution, 224 sagittal partitions, field of view (FOV) = 230 × 230 mm. The total acquisition time was 34.23 min. Transmit and receive field correction acquisition was done before every weighted image (56 sagittal partitions, field of view FOV = 230 × 230 mm, TR = 4,1 ms, TE = 1,98 ms for B1- and TR = 2000 ms, TE1 = TE2 = 14 ms, 24 sagittal slices, slice thickness = 5 mm, FA = 90, 120, 60, 135, 45° for the rf map which was used for the B1+ correction as a part of the hMRI toolbox (Lutti et al., 2012, 2010).

2.3. MPM generation

The generation of (semi-) quantitative maps were performed using the hMRI toolbox (version 0.2.0, www.hmri.info, Tabelow et al. (2019) hMRI-toolbox, RRID:SCR_017682; SPM, RRID:SCR_007037; MATLAB, RRID:SCR_001622) using default parameters. The signal from the multi-echo PDw, T1w and MTw acquisitions are modelled using the Ernst equation (Ernst and Anderson, 1966; Helms et al., 2008b, 2008a). We derived the effective transverse relaxation rate R2* from the TE dependence of the signal, combining all three contrasts using the ESTATICS model (Weiskopf et al., 2014). This provides a more robust estimation of R2* with a higher signal-to-noise ratio in comparison to separate estimations. The hMRI toolbox uses approximations of the signal equations for small repetition time TR and small flip angles α, and estimates the longitudinal relaxation rate R1, the apparent signal amplitude A* map (proportional to the proton density map PD) and the MT_{sat}. The maps might be biased by the B_1 transmit f_T and receive f_R field inhomogeneities. We used the B_1 (transmit) field and receive field sensitivity correction to eliminate the bias field errors for all maps (Tabelow et al., 2019). The local flip angle of all three maps (R1, PD, MT_{sat}) are influenced by f_T, the PD is additionally influenced by the radio frequency (RF) sensitivity bias field f_R (in absence of subject motion). The toolbox generates (semi-quantitative) magnetization saturation (MT_{sat}) maps. The MT_{sat} was adjusted for T_1 and B_1 contributions, which often leads to additional variability. Nevertheless, this problem is addressed with the MPM approach and should not influence the main findings of this study (Tabelow et al., 2019).

The maps were reoriented towards a standard pose using the auto-reorient module within the hMRI toolbox. The anterior commissure was automatically placed at the origin and both anterior and posterior commissure (AC/PC) in the same axial plane using rigid-body registration (Tabelow et al., 2019). This is a common step to increase the consistency in individual head positions prior to normalization and/or segmentation (Mazziotta et al., 2001b, 2001a, 1995). SPM's segmentation for example is sensitive to the initial orientation of the images (Ashburner and Friston, 2005) which is addressed by this step. The output resolution of every multi-parameter map was set to 1 mm isotropic.

2.4. Spatial processing

We used the standard processing pipeline of the hMRI toolbox from SPM12 with the default settings (for overview see Fig. 1).

In the first step of spatial processing we segmented the MT_{sat} images into different tissue classes (gray matter - GM, white matter - WM, etc.). This followed a probabilistic approach taking advantage of a prior tissue probability map (TPM) which was specifically developed for MPM's (denoted as eTPM) and is expected to provide favorable segmentation results (Lorio et al., 2014; Tabelow et al., 2019). Using MT_{sat}-based segmentations might lead to improved subcortical contrast because of the improved delineation of WM laminae embedded in GM structures (Helms et al., 2009). In a second step we normalized individual native space images to the MNI space by generating an average shaped template using DARTEL non-linear spatial registration (Ashburner, 2007). All maps and GM tissue segments were registered to the MNI space by applying the obtained deformation fields followed by an affine transform. In the last step of the spatial preprocessing we applied so called tissue weighted smoothing to account for potential registration inaccuracies. This specific technique developed for qMRI (Draganski et al., 2011) was used to improve the spatial realignment, preserving quantitative values within tissue classes (not smoothing across tissue boundaries) while also accounting for partial volume contribution of the tissue density in each voxel in subject space. All maps and GM tissue segments were registered to the MNI space by applying the obtained deformation fields followed by an affine transform. In the last step of the spatial preprocessing we applied so called tissue weighted smoothing to account for potential registration inaccuracies. This specific technique developed for qMRI (Draganski et al., 2011) was used to improve the spatial realignment, preserving quantitative values within tissue classes (not smoothing across tissue boundaries) while also accounting for partial volume contribution of the tissue density in each voxel in subject space. Tissue-weighted smoothing of MPM's per toolbox defaults uses individual 95% tissue probability masks (Draganski et al., 2011; Tabelow et al., 2019) All further group level voxel-based indices of reliability were calculated and presented using a 5% template GM probability. For further quantitative analysis of voxel-based indices in cortical regions we used a slightly more conservative explicit mask (GM thresholded at 20%) in MNI space. For the ROI-based analysis we used unsmoothed normalized MPMs and read out qMRI values inside ROIs of the neuromorphometrics atlas. We used the default smoothing kernel of 6 mm which was additionally varied in a subsequent analysis (see below).

3. Statistical analysis

3.1. Reliability metrics

Reproducibility (or reliability) of a measure in cognitive neurosciences refers to the agreement of multiple assessments of the same subject (Bartko, 1991). Here we focused on MPM reliability assessment via the Intraclass Correlation Coefficient (ICC). Moreover, we disentangled both contributions of (A) noise/precision (in terms of WCV) and (B) individual differences (in terms of BCV) to voxel-based and regional variations of the ICC. Reliability analysis according to classical test theory is based on the decomposition of observed scores into between-subject variability ("true score") and within subject variability (Bartko, 1991; Hopkins, 2000; Paul S. Tofts, 2018). The within-subject variability can be described as the inconsistency (or dispersion) of observations when measuring a single individual repeatedly. This is reflective of amount of random error (or irreducible noise) of the qMRI measurements. First, to calculate the within-subject variability we follow procedures proposed

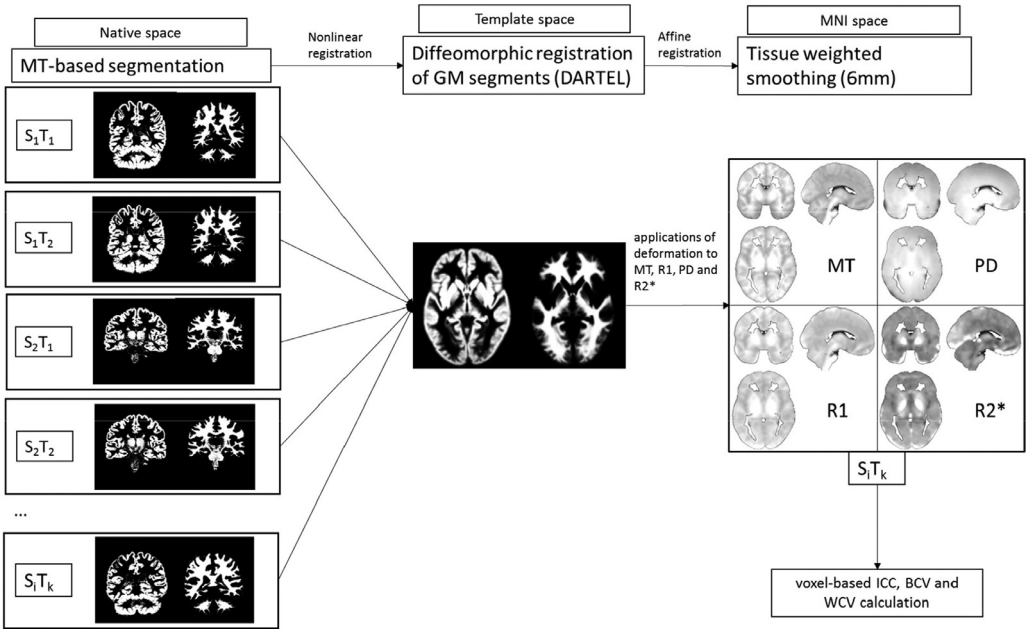


Fig. 1. Preprocessing of the MPM's from native space trough template in MNI space. All $i = 1,\ldots,31$ subjects' MT images from both timepoints $k = 1,2$ were first segmented into gray and white matter (left) and a study-wise template was generated using DARTEL (middle). All 4 qMRI maps (MT, R1, R2*, PD) were normalized using obtained nonlinear and affine registrations to MNI template space for all subjects and timepoints. Tissue weighted smoothing was applied to preserve tissue boundaries (right). The ICC, WCV and BCV was calculated in the smoothed images (right).

in Hopkins (2000) and log-transformed the measurements of each subject according to Lehmann et al. (2021)

$$\tilde{y}_{ik} = 100 * \log y_{ik} \tag{1}$$

for subject i and timepoint k. In the next step we calculated the difference (change or scores) from scan to rescan using

$$\Delta_i = \tilde{y}_{i1} - \tilde{y}_{i2} \tag{2}$$

Then the standard deviation of within-subject differences was obtained according to the

Formula

$$SD_\Delta = \frac{\sigma(\Delta)}{\sqrt{2}} \tag{3}$$

were σ refers to the standard deviation (Cercignani et al., 2018; Hopkins, 2000). Finally, standard deviation of within-subject difference was converted to the within-subject coefficient of variation (WCV) using

$$WCV = 100 \times \left(e^{SD_\Delta/100} - 1\right) \tag{4}$$

where e is the base of the (natural) exponential function (Hopkins, 2000). The WCV is equivalent to the SD of replicated measurements, expressed as a percentage of the mean value (Hopkins, 2000). This means a WCV of 10% indicates a variation about the mean value about 1/1.1 to 1.1 times the mean, or ≈ 0.91 to 1.1. The interpretation is straightforward: lower WCV values indicate better reproducibility of the measurement.

The BCV is another source of variability and representing the sample heterogeneity (or sometimes called true-score variability) (Bartko, 1991; Cercignani et al., 2018). The calculation started with averaging of the two log transformed measurements within-subject (i.e., pre and post), followed by calculation of the mean ($\bar{x}$) and the standard deviation (σ) of the resulting scores across subjects. Then, the BCV is obtained using the following ratio

$$BCV = \frac{\sigma}{\bar{x}} \times 100 \tag{5}$$

Many reliability studies in social sciences and psychology rather report the ICC, which implicitly depends on both above sources of variation (WCV and BCV). The definition of ICC is often based on ANOVA and random-effects models, and requires further model assumptions partially affected by the specific study design (Baumgartner et al., 2018; Koo and Li, 2016; Matheson, 2019; McGraw and Wong, 1996; Shrout and Fleiss, 1979). Here we follow recommendations of Koo and Li (2016) and focus on ICC (2.1) for scan-rescan reliability study where the "2" refers to the two-way random model and the "1" to the reliability of single repeated measurement instead of several measurements where agreement is defined as a term of consistency (Eq. (6)). It reflects the fraction of observed variation that is attributed to (reproducible) between-subject variation (Cercignani et al., 2018). A related interpretation often used in psychology is how much of the error-free true-score variability can be recovered by measurement. The data is arranged in a convenient matrix with rows representing subjects and columns for repeated measurements. The ICC (2.1) is calculated using this formula

$$ICC(2.1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}\left(MS_C - MS_E\right)} \tag{6}$$

formula where MS refers to the mean sum of squares subdivided into MSR for rows (i.e. between-subject), MSE for the error (i.e. reflecting random measurement noise) and MSC for the columns (containing the repeated measurements), and k referring to the number of raters (repeated measures) and n to the number of subjects. It is important to note that the ICC increases with decreasing WCV and increasing BCV. The interpretation of the ICC values lower than 0.4, between 0.4 and 0.59, between 0.6 and 0.75 and greater than 0.75 are indicative of poor, fair, good and excellent reliability, respectively (Cicchetti, 1994). For the voxel-based and region-based calculations of WCV, BCV and ICC we used custom-made MATLAB code implementing the formula presented above. All other statistical analyses (interference statistics) were performed using SPSS version 27 (IBM SPSS Statistics, RRID:SCR_019096).

3.2. Reliability analysis using Voxel- and Region-based approaches

We analyzed ICC reliability, WCV and BCV of the MPM's on a voxel-based level focussing on gray matter regions of normalized MPMs (MT_{sat}, R2*, R1, and PD). To characterize these voxel-based ICCs we

additionally calculated their mean (and SD) within a cortical and a subcortical GM ROI (including amygdala, caudate, hippocampus, putamen, pallidum, accumbens area and thalamus). The latter values were used for various statistical comparisons involving t- and F-tests to assess the differences between cortex and subcortex ROIs. The local intercorrelation of reliability metrics was calculated. Based on the concept of modulation in the context of studies of voxel-based Morphometry (VBM) (Ashburner and Friston, 2000; Kurth et al., 2015) we additionally assessed the contribution of anatomical variability to ICC metrics using the voxel-based standard deviation of the Jacobian determinant of deformation fields obtained from DARTEL normalization. The determinant of the Jacobian tensor describes the (individual) voxel-wise volume changes induced by the nonlinear mapping during normalization to template space (Ashburner, 2007).

The main voxel-based results are reported for 6 mm smoothing kernel used for tissue weighted smoothing of MPMs (hMRI toolbox default). However, we additionally studied the effect of systematically varying smoothing kernels from 1 to 8 mm. Notably, this procedure used ROIs for aggregates of local voxel-level ICC and not actual ROI-level ICC (where averaging happens before ICC calculation, see subsequent section on ROI analysis). Quantitative effects of smoothing were assessed using paired t-tests of MPM data and comparing selected pairs of kernels with values 1–8 mm. Based on the specific t-values we calculated the effect size of each comparison. Rules of thumb for interpreting the effect size of the differences are $|d| \langle 0.30$ "small", $0.30 < |d| < 0.50$ "medium", and $|d| \rangle 0.50$ "large" effects, respectively (Cohen, 1988).

In addition to the above voxel-based ICC analysis a conventional region-based analysis of ICC was performed. For this purpose we used the neuromorphometrics atlas (neuromorphometrics.com) to first aggregate ROI-level mean values of unsmoothed normalized MPMs inside 57 gray matter regions in cortical and subcortical locations in template space. After extraction of these mean ROI values of all MPMs we calculated the reliability metrics for each of the 57 ROI's using SPSS. We report averaged values across hemispheres.

4. Results

4.1. Mapping local longitudinal reproducibility of MPMs

The results of the whole-brain voxelwise analysis of multiple reliability indices for WCV, BCV and ICC are presented in Fig. 2. The reliabilities are summarized for cortical and subcortical GM regions in Fig. 3. The analysis revealed that MT_{sat} and PD parameters had favorably low WCV (median cortex (subcortical GM): MT_{sat}: 2.3% (2.6%) PD: 1.9% (1.8%) R1: 2.9% (3.2%) $R2^*$: 4.1% (5.0%)) in widespread GM areas. For MT_{sat} and PD the pattern of local WCV was relatively homogenous across the cortical mantle. The R1 and especially the $R2^*$ parameter map were found to have higher WCV values compared to MT_{sat} and PD ($F(1,1,053,559) = 5,859,339.79$, $p = <10^{-16}$ partial $\eta^2 = 0.957$, $n = 1,053,560$). A particularly high WCV indicating more substantial noise or artefact presence was observed in orbitofrontal cortex areas for MT_{sat}, R1 and especially for $R2^*$. WCV in subcortical GM regions as opposed to cortical GM was found to be higher in MT_{sat}, R1 and $R2^*$ maps ($F(1,1,053,558) = 566,311.747$, $p < 10^{-16}$ partial $\eta^2 = 0.683$, $n = 1,053,560$). The voxelwise analysis of BCV revealed the highest values in $R2^*$ and MT_{sat} (median cortex (subcortical GM): MT_{sat}: 7.1% (4.3%) PD: 3.6% (2.4%) R1: 5.6% (4.7%) $R2^*$: 8.0% (8.8%)). The PD maps showed the lowest BCV especially in subcortical GM. All maps except $R2^*$ showed higher BCV in cortical areas when compared to subcortical regions ($F(1,1,053,558) = 460,459.044$, $p < 10^{-16}$ partial $\eta^2 = 0.636$, $n = 1,053,560$). Visual inspection of the BCV maps suggested larger individual differences in fronto-temporo-parietal brain regions and less differences in midline structures, basal ganglia and subcortical gray matter. When averaging over cortical or subcortical voxels, BCV was found to exceed WCV in all MPM's and most of the ROI's. In contrast, there are also a few brain regions in which noise level locally exceeds the amount of individual differences, resulting in low ICC e.g. in medial frontal cortex or cerebellum. Stronger posterior gradients in WCV and BCV were observed close to the boundary between GM and CSF and due to smoothing appeared partially in CSF or close to the left and right occipital pole. However, most of these artificial values are excluded when applying explicit masking during analysis (Figure S1).

The ICC was found to be higher in the MT_{sat} and $R2^*$ compared to PD and R1 maps, especially in the cortex (median cortex (subcortical GM): MT_{sat}: 0.789 (0.447) PD: 0.553 (0.264) R1: 0.555 (0.369) $R2^*$: 0.624 (0.477)). Generally, all four maps had higher ICC in the cortex compared to the subcortex ($F(1,1,053,558) = 830,696.519$, $p < 10^{-16}$ partial $\eta^2 = 0.759$, $n = 1,053,560$). We observed regionally high ICC of MT_{sat} as a result of combined contribution of low WCV and high BCV values. A similar pattern of contributions to ICC was observed for the $R2^*$.

To further explore the contribution of noise and individual differences to ICC in MPMs we conducted a voxelwise regression analyses with predictors WCV (Fig. 4A) and BCV (Fig. 4B) for MT_{sat} in cortical gray matter regions. We observed that ICC decreases with more noise (WCV, $r = -0.606$, $p < 10^{-16}$) and increases with individual differences (BCV, $r = 0.509$ $p < 10^{-16}$). This suggests that having less noise in a voxel aligns with a more favorable ICC but that the amount of true individual variability is another important factor for reliability in terms of ICC. It is worth noting that both contributors to ICC can jointly vary across brain areas as will be demonstrated in subsequent analyses.

4.2. Exploring the interplay of reliability metrics and anatomical variability

Next, we asked the question whether local differences of irreducible error variability are associated to the amount of individual differences of a qMRI parameter? The inter-correlation (across voxels) between reliability parameters for MT_{sat} and $R2^*$ is presented in Fig. 5C & 5D respectively. Indeed we observed generally positive inter-correlations of WCV and BCV in cortical and subcortical regions ranging from 0.085 to 0.771 suggesting that measurement error partly aligns with higher true score variance of mapped parameters.

Then, we explored the potential contribution of local anatomical variability to the substantial reliability differences of MPM parameters across the cortical mantle and gray matter nuclei. Although complex non-linear registration of MPMs is applied to enable optimal alignment and subsequent group-level qMRI analysis, individual morphometric variability of local gray matter volume, folding and shape might influence reliability differences across brain regions (Alexander-Bloch et al., 2013; Pizzagalli et al., 2020; Wonderlick et al., 2009). As a measure of anatomical variability, we focused each voxel's standard deviation (across subjects) of the Jacobian determinant of the deformations used for normalization to template space (Fig. 5A). For instance, a higher standard deviation of the Jacobian determinant is expected in brain areas with more individual variability of local volumes or thickness (across subjects SD of the Jacobian determinant has the median in the cortex (subcortex) 0.151 (0.093), Fig. 5B).

Using this measure, ICC reliability of MT_{sat} increased in voxels with higher anatomical variability in both cortical ($r = 0.35$) and subcortical ($r = 0.208$) gray matter suggesting that the underlying increase of BCV might outweigh increases in measurement error. However, for $R2^*$ the pattern was had less straightforward dependencies. The contributions to ICC might be further elucidated when exploring the differential effects that anatomical variability has on ICC's key components WCV and BCV. Our analysis suggested that the interplay with anatomical variability is more complex since WCV decreased with larger anatomical differences in cortical areas of both maps unexpectedly. Moreover, anatomical variability did not consistently align with larger true score variations in terms of BCV, suggesting morphometric aspects and qMRI parameters reflecting partially independent aspects.

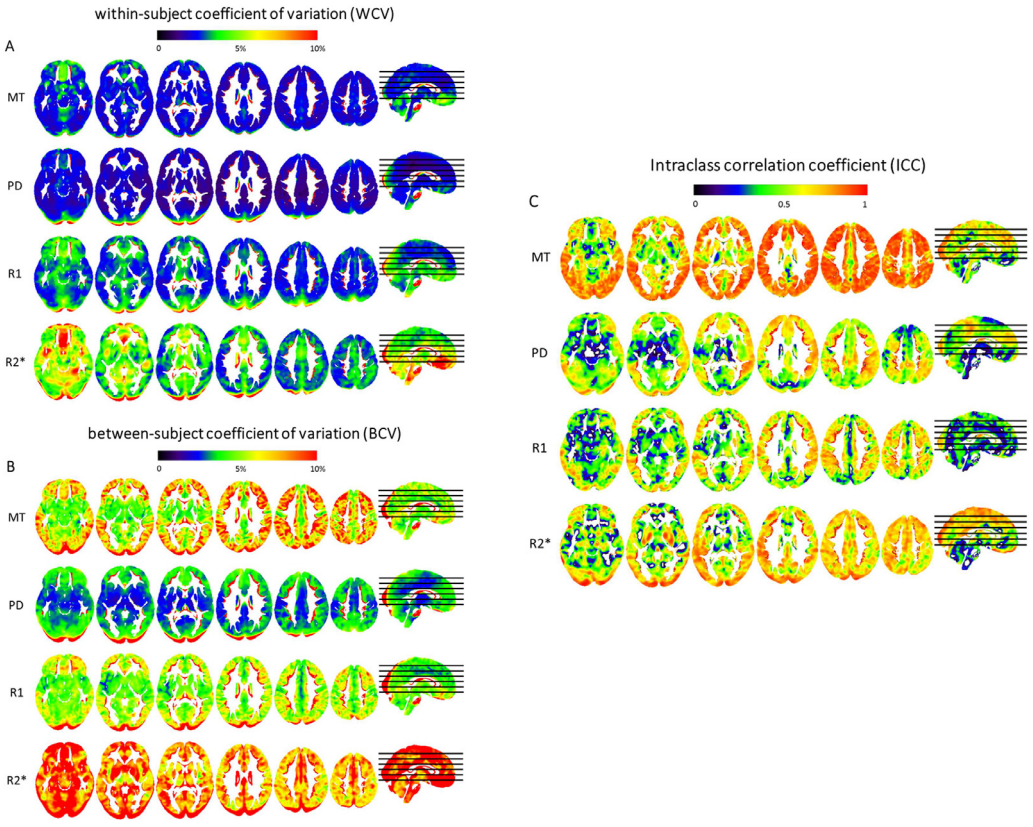


Fig. 2. Local voxel-based reliability analysis of MPM parameters. We show MT_{sat}, PD, R1, and $R2^*$ reliability metrics using scan-rescan data of $n = 31$ participants and an interscan interval of 1 month. (A) Within-subject coefficient of variation (WCV) (B) Between subject coefficient of variation (BCV); and (C) Intraclass correlation coefficient (ICC) with the smoothing kernel for tissue weighted smoothing of 6 mm (current default of hMRI toolbox, for effects of alternative kernels see also supplementary Figure S5).

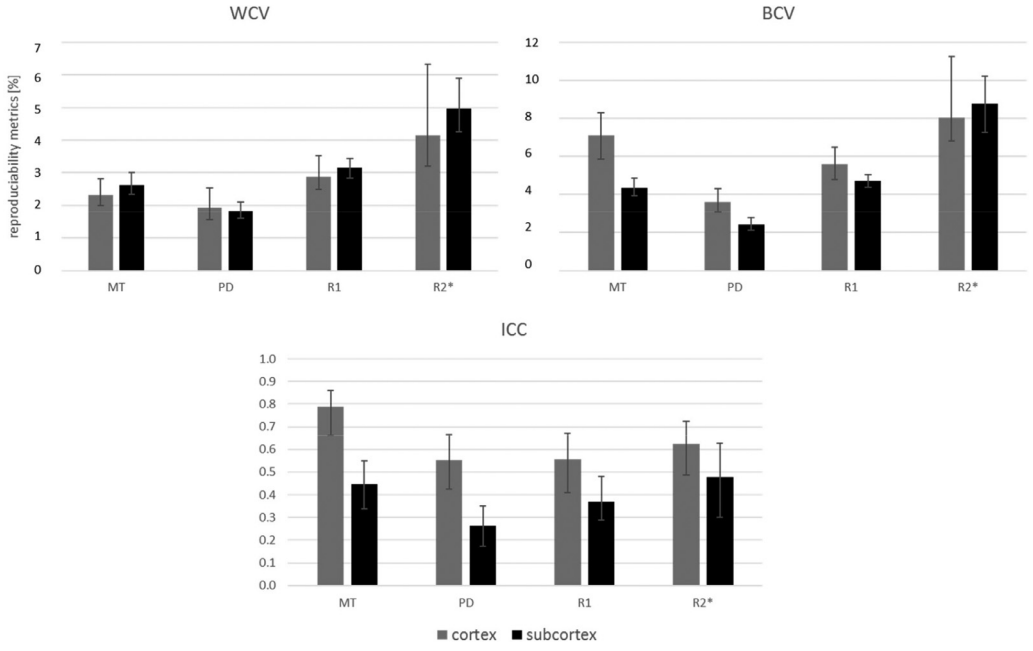


Fig. 3. Summarized reliability metrics in cortical and subcortical gray matter. WCV, BCV and ICC for a voxelwise analysis approach, separated for cortical (gray) and subcortical (black) gray matter. Mean and standard deviation (error bars) of reliability indices over all voxels in cortical and subcortical ROIs are shown.

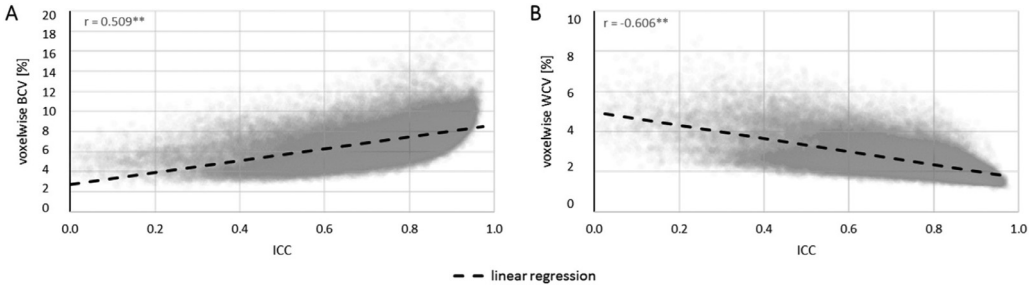
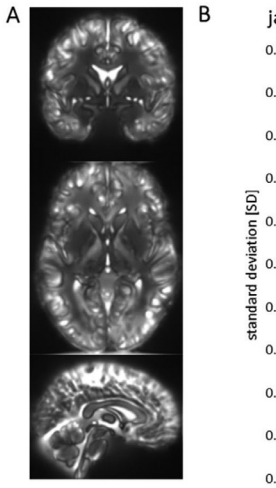
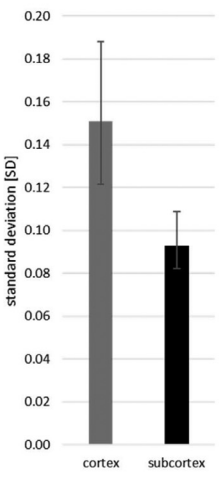


Fig. 4. Exploring the voxelwise contribution of WCV and BCV to ICC for cortical $MT_{sat.}$ We show the linear regression of WCV (A) and (BCV) (B) on ICC over all voxel inside the specific cortical GM mask. We additionally provide the Pearson correlation coefficient ® between the ICC and the WCV and the BCV. To improve illustration we reduce the number of presented voxels to 10% of the voxel inside the region, the correlation and regression where calculated with all voxels inside the cortical GM mask.





C

Pearson correlation cortex				
MTsat	SD Jacobian det.	WCV	BCV	ICC
SD Jacobian det.	1			
WCV	-.343**	1		
BCV	.099**	.181**	1	
ICC	.350**	-.664**	.549**	1
R2*	SD Jacobian det.	WCV	BCV	ICC
SD Jacobian det.	1			
WCV	-.322**	1		
BCV	-.277**	.771**	1	
ICC	.069**	-.202**	.326**	1

D

Pearson correlation subcortex				
MTsat	SD Jacobian det.	WCV	BCV	ICC
SD Jacobian det.	1			
WCV	.023**	1		
BCV	.350**	.358**	1	
ICC	.208**	-.719**	.374**	1
R2*	SD Jacobian det.	WCV	BCV	ICC
SD Jacobian det.	1			
WCV	.162**	1		
BCV	-.015**	.085**	1	
ICC	-.118**	-.713**	.593**	1

Fig. 5. Association of anatomical variability and associations of WCV and BCV Pearson correlation (over voxels) of standard deviation (across participants) of the individual Jacobian determinant with ICC, BCV and WCV within cortical (C) and subcortical (D) ROIs. (A) Standard deviation (across participants) of Jacobian determinant of individual deformation fields for mapping to study-wise template space. The cortical and subcortical variability (Jacobian determinant [SD]) with 25 and 75 percentile error bars (B). Higher SD indicates stronger anatomical variability. * indicates significant correlation $p < 0.05$; ** indicates high significant correlation $p < 0.01$.

4.3. Mapping regional longitudinal reproducibility of MPMs

Next, we calculated the regional ICC using 57 gray matter ROI's illustrated for MT_{sat} and R2* in Fig. 6 (for R1 and PD see supplementary Figure S3 & S4). The obtained regional findings were coarsely consistent with above voxel-based findings. The values for MT_{sat} ranged from lowest regional ICC seen in the amygdala with 0.157 to highest ICC of 0.796 in the occipital pole. We observed a significantly increased ICC of MT_{sat} in cortical areas compared to subcortical gray matter regions ($z(56) = -2.093$; $p = 0.036$; Fig. 6). Interestingly, we found that the ICC of R2* was generally more favorable using the ROI approach resulting in the highest ICC median over the ROI's of all four maps (R2* 0.730, MT_{sat} 0.486, PD 0.514, R1 0.417). In contrast, ICC of MT_{sat} was found to be slightly decreased when using a ROI compared to voxel-based approach.

Our results suggested that the brain area with the best average regional reproducibility of all four MPMs is the occipital pole, followed by the parietal operculum and the middle and superior temporal gyrus (supplementary Table 1). The lowest ICC averaged over four maps has the amygdala with 0.253. 9 regions had poor (less than 0.40), 32 had fair (0.40 to 0.59), 15 had good (0.60 to 0.75) and 1 excellent (above 0.75) ICC's (Cicchetti, 1994).

4.4. Influence of smoothing kernel size on reproducibility metrics

Since the application of smoothing in local qMRI image analysis often requires *a priori* choices of filter size, we finally aimed to study the optimal amount of smoothing that is beneficial for scan-rescan reliability in terms of voxel-based ICC. Here we report averages of voxel-based ICC within the total cortical and subcortical gray matter. Additionally, we inspected voxel-based ICC in cingulate cortex, precentral gyrus and hippocampus ROIs since these are often focus of investigations in context of development, aging and training induced plasticity studies.

The obtained reliability curves of MT_{sat} parameter over varying filter size from 1 to 8 mm (FWHM) aggregated within five ROIs are presented in Fig. 7 (selected filtered maps are shown in supplementary Figure S5). For $MT_{sat,}$ we observed a consistent decrease of ICC, WCV and BCV with increasing smoothing kernel size in all regions except for the ICC in the subcortical ROI. In contrast reliability curves of R2* did exhibit rather minor positive dependency of ICC on filter size (with exception of the cingulate ROI, Fig. 8). Interestingly, although WCV and BCV of R2* also declined with larger kernels, the relative decrease of BCV was not as emphasized which might have resulted in the observed stability of ICC. We observed that R2* had for all filter sizes a higher BCV and WCV com-

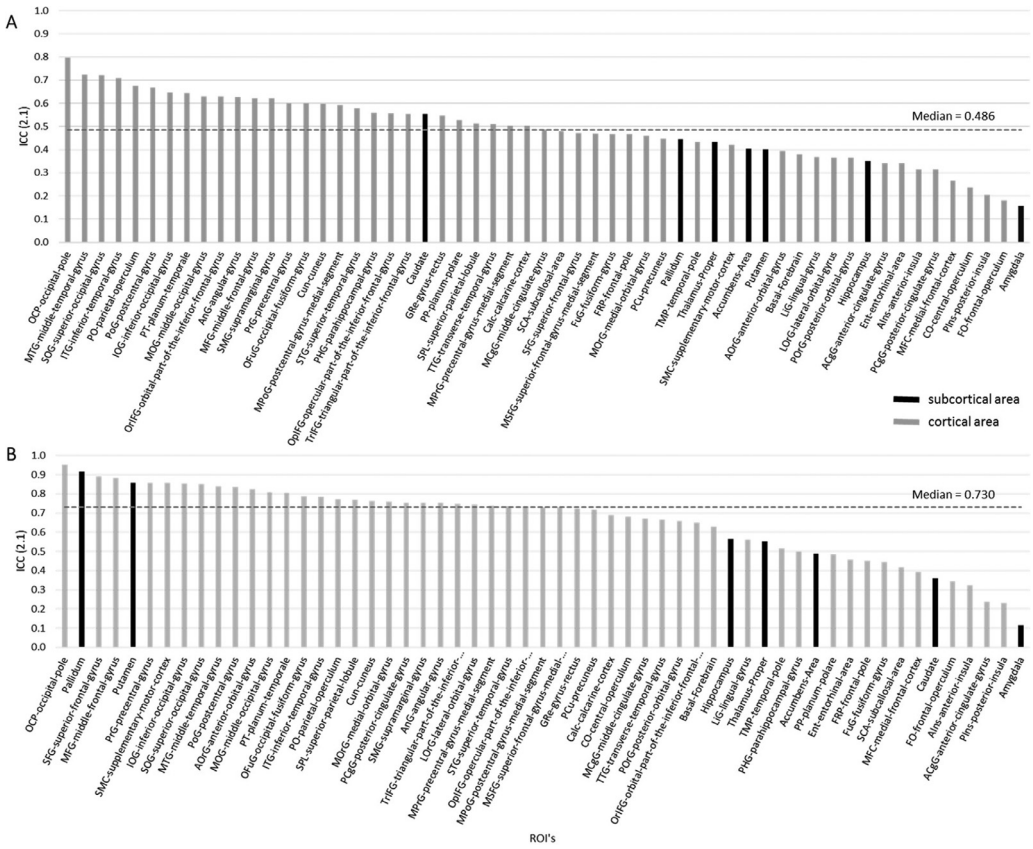


Fig. 6. ICC of MT_{sat} and R2* using neuromorphometric atlas ROIs. Regions are sorted from highest to lowest ICC and colored according to cortical (gray) and subcortical (black) brain areas for MT(A) and for R2* (B). We used the default smoothing kernel (6 mm). The ROI extraction was conducted in the MNI space and the ICC was calculated on the ROI-average intensities in each quantitative map. **The** ICC was averaged across both hemispheres for every ROI.

pared to MT_{sat} in subcortical GM regions (subcortex and hippocampus). Differences of ICC values which can be attributed to smoothing kernel size varied from 0.436 to 0.901 for MT_{sat} and from 0.382 to 0.661 for R2* over ROIs also suggested partially different underlying mechanisms. The optimal smoothing filter size for reproducible MT_{sat} was found to be between 2 and 4 mm, while 2 mm yielded the highest ICC value. The smoothing-related differences of ICC increased with larger kernels (see supplementary Table 2 for effect sizes). For R2* we found optimal reproducibility with larger smoothing kernel size above 6 mm. However, we found that the effect size of these smoothing-related differences were rather small (supplementary Table 3).

5. Discussion

In this study we assessed the longitudinal four weeks scan-rescan reproducibility of the MPM protocol of four quantitative maps on single-voxel level in whole-brain gray matter using a sample healthy young adults. The longitudinal reproducibility was found to be good to high in most areas of the brain across all quantitative maps. The within-subject coefficient of variation (WCV) in all maps ranged between 2% and 5%, which confirms previous results of MPM reliability over vendors, time points and centers (Leutritz et al., 2020; Weiskopf et al., 2013). Overcoming limitations of previous studies on MPM reliability, we additionally provide new estimates for the ICC, a ratio measure of between-coefficient of variation (BCV) and WCV indicating the absence of noise relative to the distinguishability between subjects. Depending on the specific quantitative map, the MPM protocol might be successfully used to detect trait- or risk-factor-related individual differences of microstructure in observational studies (Carrasco et al., 2014; Zuo et al., 2019). We found that the ICC of R2* was generally more favorable using the ROI approach. In contrast, ICC of MT_{sat} was found to be slightly decreased when using a ROI compared to voxel-based approach. This is an important result, suggesting that apart from statistical differences due to applied multiple comparison corrections, a local voxel-based analysis without strong assumptions on atlas or ROIs can be recommended. We also found that spatial smoothing affected MPM reproducibility in a map- and brain region-dependent fashion, suggesting that specific analysis strategies are warranted depending on the specific research question at hand.

A previous study by *Leutritz et al. (2020)* also used the hMRI toolbox for reliability analysis of MPM's (acquired at six 3 T MRI systems) in five subjects using a rescan interval of 2 h (compared to 4 weeks and 31 participants in our study). The authors reported an intra-site WCV of 16% for R2* and 4–10% for MT_{sat}, R1 and PD. Our study includes thirty participants and a longer scan-rescan interval of 4 weeks. Results revealed slightly lower WCV values (median cortex (subcortical GM): MT_{sat}: 2.3% (2.6%) PD: 1.9% (1.8%) R1: 2.9% (3.2%) R2*: 4.1% (5.0%)). Some differences might be attributed to differences in the length of the measurement protocol (34.23 min vs 20 min) or the differences in resolution (0.8 mm vs 1 mm) compared to Leutritz et al. (2020). The larger sample recruited in the present study did enable solid estimates of multiple reliability indices and to focus for the first time on the ICC accounting for individual variability (BCV), which was not reported by Leutritz et al. (2020). Notably there are other methodological differences in both studies such as (A) focus on voxel-level vs. ROI-level (ROI-results being more comparable) (B) single vs. multi-site; and (C) individual level vs. group level statistics. The novel voxel-based group-level results of ICC, BCV and WCV in GM do suggest comparably good

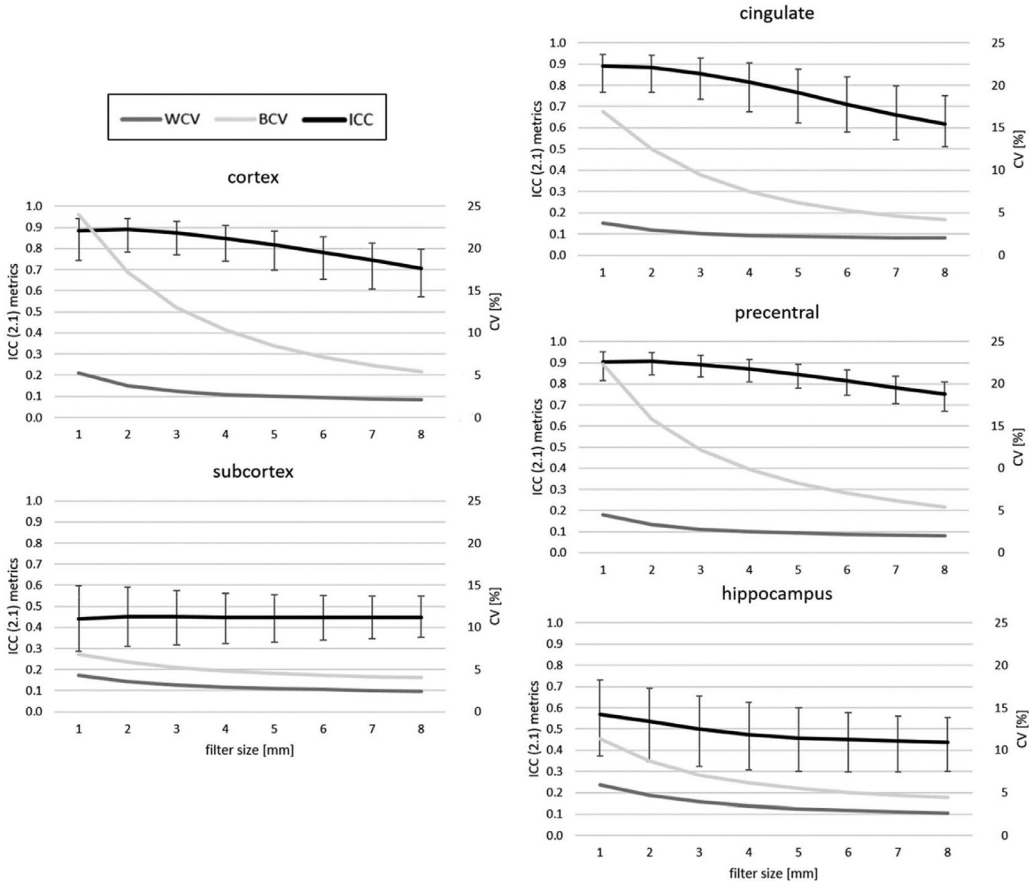


Fig. 7. Exploring the effect of smoothing kernels on reproducibility of MT_{sat} map. Within (WCV; dark gray) and between (BCV; light gray) coefficient of variation (CV [in%]), ICC (black) over different smoothing kernels separated by cortex, subcortex (including amygdala, caudate, hippocampus, putamen, pallidum, accumbens area and thalamus), mid cingulate, precentral cortex, hippocampus. The presented ICC curves were obtained using the median of the voxelwise ICC of the spatially normalized and smoothed MT_{sat} inside the particular ROI, the error bars show the 25 and 75 percentile. Unit on the secondary y-axes is coefficient of variation in percent.

reliabilities. With combined results of BCV, WCV and ICC we expand the knowledge about a broader spectrum of reliability and variability metrics for voxel- and ROI-based qMRI.

We observed a particularly high WCV for MT_{sat} and especially for $R2^*$ in orbitofrontal cortex areas, which indicates more noise or artefact presence. This problem in certain inferior and brain regions close to tissue boundaries was noted before in qMRI and might be slightly reduced by increasing the GM probability threshold applied during analysis (see supplemental Figure S1). Other sources of undesired variability might influence reproducibility of MPM parameters. In addition to resolution differences, the number and timing of echoes in the multi-echo FLASH is expected to affect quality of the parameter estimates such as $R2^*$ (Weiskopf et al., 2014). The latter maps are particularly prone to motion artifacts, since their estimation requires images acquired at long echo times that are more affected by motion. The other quantitative maps (R1, MT, PD) generated from the MPM protocol are significantly less affected by motion, since they are estimated from averages across multiple echo times (including short echo times, Weiskopf et al., 2013). However, the ESTATICS model uses multiple (available) contrast-weightings simultaneously which has been show to provide more robust estimation of $R2^*$ with a higher signal-to-noise ratio compared to separate estimations in the presence of motion artifacts (Tabelow et al., 2019; Weiskopf et al., 2014). The ESTATICS model assumes mono-exponential free induction decay in line with previous correction approaches (Nöth et al., 2014) which will be violated in brain areas suffering from significant susceptibility artifacts (Neeb et al., 2006) or partial volume effects. The applied procedures did not account for field inhomogeneities due to susceptibility variations which might have affected $R2^*$ reliability in some areas of the brain (e.g. air-bone-soft tissue transitions). Future studies might focus on exploring these effects on MPMs and potential correction methods (e.g. using fieldmaps). As pointed out by Weiskopf et al., 2014, the comparably high resolution of 0.8 mm (isotropic) used in this study is expected to reduce the impact of susceptibility artifacts and partial volume effects on the $R2^*$ maps but imaging protocols with lower resolution or studies using 7 Tesla scanners might be affected more. The reported reliability estimates might also depend on accuracy of the implemented methods that correct for effects of transmit (Lutti et al., 2012, 2010) and receive fields on parameter maps such as PD and R1 (Tabelow et al., 2019). The applied RF sensitivity correction uses an additional receive sensitivity field that is acquired before each of the PDw, T1w and MTw contrasts and combines a correction for motion-related relative receive sensitivity variations with rigid-body realignment (Papp et al., 2016). The MT_{sat} parameter map is relatively robust against differences in relaxation times and RF transmit and receive field inhomogeneities – unlike the conventional MT ratio, which is affected by variations of the R1 and RF transmit field (Helms et al., 2010, 2008b; Weiskopf et al., 2013). It is important to mention that typical motion trajectories may differ between volunteers and patients but also between different (patient) populations. Thus the reliability metrics reported in this study might not necessarily generalize to clinical studies. Interestingly, the hMRI-toolbox provides summary measures of head motion within- and between the acquisitions of each image volume (intra- and inter-scan motion) (Castella et al., 2018) that could potentially be used in future studies to exclude or downweight poor-quality data of individuals in

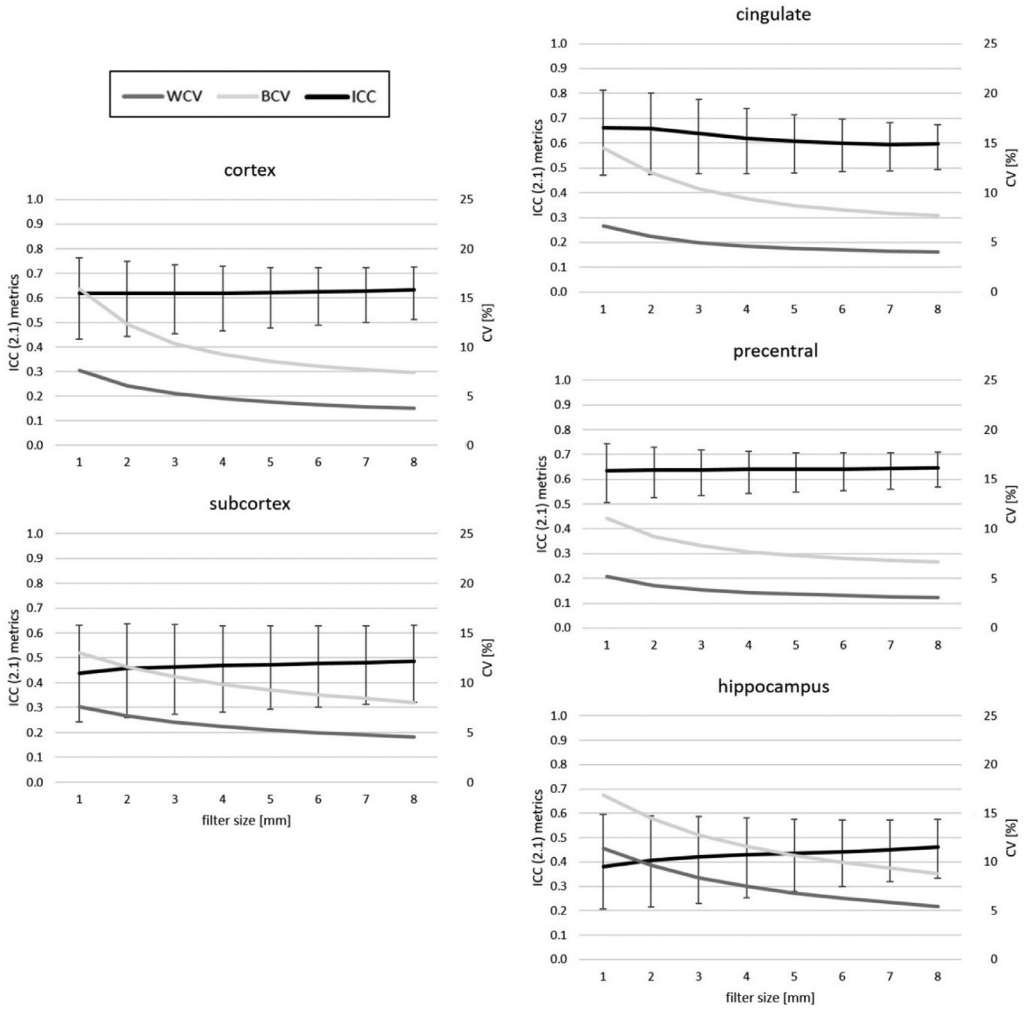


Fig. 8. Exploring the effect of smoothing kernels on reproducibility of R2* map. Within (WCV; dark gray) and between (BCV; light gray) coefficient of variation (CV [in%]), ICC (black) over different smoothing kernels separated by cortex, subcortex (including amygdala, caudate, hippocampus, putamen, pallidum, accumbens area and thalamus), mid cingulate, precentral cortex, hippocampus. The presented ICC curves were obtained using the median of the voxelwise ICC of the spatially normalized and smoothed R2* inside the particular ROI, the error bars show the 25 and 75 percentile. Unit on the secondary y-axes is coefficient of variation in percent.

a statistical group analysis (Lutti et al., 2022). Based on the slightly better ICC using a region-based approach, we would carefully suggest using this method for analyzing R2* map. The WCV results reported in our study may be further used as benchmarks for monitoring individuals (e.g., patients) over time in longitudinal studies (Hopkins, 2000; Lehmann et al., 2021).

The often-reported WCV (or CV) only covers one important aspect of reliability and therefore reproducibility. A very common neuroimaging research question is to detect potential relationships between qMRI brain measures with other (behavioral or psychological) constructs. From a statistical perspective, the ability to detect such correlations crucially depends on the reliability index ICC (Zuo et al., 2019), a ratio measure that essentially places the amount of random measurement error (WCV) in the context of real biological (or psychological) variation between subjects (BCV) (Bartko, 1991). Our results revealed ICC values ranging from fair to excellent (according to (Cicchetti, 1994)), depending on the specific MPM map, thus suggesting a favorable balance between WCV and BCV. On the one hand, this indicates that MPM's are generally suited for use in correlational or explorative observational studies. On the other hand, by multiplying the expected population effect size by the square root of the ICC (our results can be used for sample size planning in future studies using the MPM protocol (Kanyongo et al., 2007; Zuo et al., 2019).

It is a common assumption in many neuroimaging studies that smoothing improves signal to noise ratio and therefore might benefit sensitivity of the analysis (Ashburner and Friston, 2000). It additionally compensates imperfect registration and lowers within- and between-subject variability in terms of neural organization, folding etc. (Maisog and Chmielowska, 1998; Mikl et al., 2008). In theory, there is an *a priori* unknown optimal amount of smoothing that optimizes sensitivity for a given spatial size of true effects. However, typically the optimal smoothing is hard to determine empirically (Penny et al., 2003) and coarse heuristics do exist only for established methods such as VBM and fMRI. In principle, it is therefore important to study the extent of spatial smoothing resulting in the highest voxel-wise reliability for novel quantitative MRI features. Using informed processing procedures are expected to have beneficial implications for sensitivity, specificity, and general replicability of studies.

In this study we reveal evidence that increased spatial smoothing filter size leads to a decrease of both "bad" (WCV) and "good" (BCV) variability. In the best case, there is a reasonable balance between WCV and BCV, thus resulting in a high ICC. The current default of smoothing in most of the toolboxes like hMRI or SPM12 is 6 mm for T1w and quantitative MRI. In contrast our findings suggest that 2–4 mm smoothing kernels might result in the highest reproducibility for voxelwise analysis of MT_{sat} (WCV mostly $\leq$ 4%, subcortical ICC = 0.453 and cortical

ICC = 0.890). This also enables preservation of more spatial structural details. Since the ICC of R2* was found to slightly increase with larger smoothing kernels, we would recommend kernel sizes of 6 mm or higher for voxel-based analysis of this parameter map. The higher optimal R2* smoothing kernel might be attributed to the different noise structure between the maps, which is alleviated in R2* with higher smoothing. However, one might lose substantial spatial information about structural differences and longitudinal changes when using a larger smoothing kernels. Given also that the gradient of ICC with respect to filter sizes was found to be rather small for R2* we would recommend not to apply kernels larger than 8 mm for tissue-weighted smoothing. It appears that optimal smoothing kernels in terms of reliability must be chosen dependent on the respective map and research question. If *a priori* information exists, the ROI approach for certain areas with higher ICCs might provide a reasonable alternative.

We observed a higher ICC of MPM parameters in the cortex compared to subcortex. According to our separate analysis this is likely to be driven by a comparably high inter-individual true score variability of MPMs in the cortex in terms of BCV, while WCV values were rather similar between cortex and subcortex for MT_{sat} and R2*. This aligns with conclusions about morphometric variability inside the cortex in previous studies (Alexander-Bloch et al., 2013; Pizzagalli et al., 2020; Wonderlick et al., 2009). However, macro- and microstructural differences should be differentiated if possible (Ziegler et al., 2019). One might speculate that the macro-anatomical variability drives some of the observed differences in subcortical and cortical MPM reliability. To address this hypothesis, we analyzed whether the standard deviation (SD) of the Jacobian determinant, a proxy for macro-structural variability, would predict cortical and subcortical MPM reliability such as ICC. Our findings that ICC of MT_{sat} was positively linked to anatomical variability supports this hypothesis, whereas more complex patterns for R2* were found (Fig. 5C & D). We further showed that anatomical variability explains a small but significant amount of the BCV (true score variance), an effect that was more pronounced in the cortex than in subcortical areas. It is important to note that local qMRI parameters that are partially affected by increased macro-structural differences (via BCV) might indeed have a high reliability but do not necessarily represent a neuroimaging biomarker with high (internal or external) validity. It might still be the brain areas with less contributions of macro-structural variability (and lower ICC) where qMRI parameters might show their relevance for cognitive neuroscience. We would argue for a careful consideration of the complex interplay reflected in these different metrics and factors when conducting a new study. The observed *negative* correlation of cortical anatomical variability with WCV was an unexpected finding of our study. We can only speculate that this might be a statistical effect where different scaling and varying heterogeneity of voxel-wise MPM parameters affects noise variance estimates (based on only 2 scans). Maybe this correlation might also point towards potential correction methods, or generative models that incorporate both macro- and microstructural aspects using MPMs or other multimodal data.

We are aware of a number of limitations of the present study. First, as with all reliability studies, the tacit assumption that the brain is not changing during the scan-rescan might be not tenable in an experimental setting (Cercignani et al., 2018). To this end, it cannot be ruled out that biological variation added unwanted variability to our data that might have influenced our results (Trefler et al., 2016). Unlike in repeated phantom measurements, we do not know the “true” biological variation of young adulty brains over a month. Such changes might influence WCV and BCV in our ICC model although they are meaningful additional sources of variance and not just noise. Therefore, more potential influences to the WCV (e.g., positioning, head-movement, multi-run variability, day-to-day variability) needs to be analyzed in further studies using more sophisticated designs (Brandmaier et al., 2018). The second limitation is that ICC and BCV are potentially sample-specific. We focused on a homogeneous group of young healthy adults and against this background it must be noted that the ICC crucially depends on the amount of between-subject variation present in the analyzed sample, which is unknown *a priori* (Hopkins, 2000). Therefore, the presented/reported ICC values might not generalize to samples with different characteristics. We would expect a higher ICC in a more heterogeneous population (e.g., inclusion of diseased subject, wider age range etc.) and lower values in a more homogeneous population (e.g., only one sex, narrower age range, similar expertise levels etc.). However, the WCV can be estimated from a sample of individuals which are not particularly representative for the whole population (Hopkins, 2000). Therefore, the WCV results presented in this study provide a potentially useful orientation which can be used for sample size planning in the future (Button et al., 2013). Third, the application of bi-polar read-out gradients during multi-echo acquisition of MPMs might have resulted in distortion due to susceptibility artifacts that caused slightly lower reliability estimates. Future studies might incorporate accurate and time-efficient correction methods. Fourth, there are considerable differences between the processing tools for example SPM, FSL, freesurfer etc. It was beyond the scope of this study to explore all the specifications of the different processing pipelines for qMRI data. Instead, we focused on the qMRI preprocessing defaults from the hMRI toolbox running in SPM12 (Tabelow et al., 2019). Novel generative modeling approaches incorporating spatial, longitudinal or multimodal priors might enable more robust MPM parameter estimation in future (Balbastre et al., 2020).

6. Conclusion

The present study revealed good reproducibility estimates of the multi-parameter mapping (MPM) qMRI protocol for voxel- and ROI-level analyses over a comparably long-time interval of four weeks. The voxelwise between- and within-subject variation (BCV and WCV) were found to be between 2 and 9% and 2–5% for all 4 MPMs (MT_{sat}, R2*, PD, R1). Our results pave the way to scientific application studies assessing microstructural changes induced by e.g. training intervention. Furthermore, especially the ICC values as reported here can be used for sample size planning in future studies. Specifically, for the most common statistical models (independent and dependent samples *t*-test and nonparametric equivalents, one-way ANOVA), it has been suggested to multiply the expected population effect size by the square root of retest reliability/ICC (Kanyongo et al., 2007; Zuo et al., 2019). Findings support the important role of between-subject- and anatomical variability for regional reliability differences. Our results pave the way to scientific application studies assessing microstructural changes induced by e.g. training intervention.

Data and code availability statement

Data availability: The datasets generated during and/or analyzed during the current study will be available on request from the corresponding author [NA] without undue reservation

Code availability: All previously unpublished computer code used to generate results that are reported in the paper will be available on request from the corresponding author [NA] without undue reservation.

Funding

Norman Aye was founded by the innovation fund of the Otto von Guericke University Magdeburg (2039236003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Material

supplementary_files_NI.docx

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationship that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Norman Aye: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Nico Lehmann:** Conceptualization, Methodology, Writing – review & editing. **Jörn Kaufmann:** Methodology, Investigation, Data curation, Writing – review & editing. **Hans-Jochen Heinze:** Resources. **Emrah Düzel:** Conceptualization, Resources. **Marco Taubert:** Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition. **Gabriel Ziegler:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision.

Acknowledgements

The authors thank Arturo Cardenas-Blance for the kind support in preparing the imaging protocols, Claus Tempelmann for resources and helpful comments on the MR physics and correction methods.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119249.

References

Alexander-Bloch, A., Giedd, J.N., Bullmore, E., 2013. Imaging structural co-variance between human brain regions. Nat. Rev. Neurosci. 14, 322–336. doi:10.1038/nrn3465.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95–113. doi:10.1016/j.neuroimage.2007.07.007.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry–the methods. Neuroimage 11, 805–821. doi:10.1006/nimg.2000.0582.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851. doi:10.1016/j.neuroimage.2005.02.018.

Balbastre, Y., Brudfors, M., Azzarito, M., Lambert, C., Callaghan, M.F., Ashburner, J., 2020. Joint Total Variation ESTATICS for Robust Multi-Parameter Mapping, 11 pp. http://arxiv.org/pdf/2005.14247v1.

Bartko, J.J., 1966. The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. 19, 3–11. doi:10.2466/pr0.1966.19.1.3.

Bartko, J.J., 1991. Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17, 483–489. doi:10.1093/schbul/17.3.483.

Baumgartner, R., Joshi, A., Feng, D., Zanderigo, F., Ogden, R.T., 2018. Statistical evaluation of test-retest studies in PET brain imaging. EJNMMI Res. 8, 13. doi:10.1186/s13550-018-0366-8.

Brandmaier, A.M., Wenger, E., Bodammer, N.C., Kühn, S., Raz, N., Lindenberger, U., 2018. Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). ELife 7, e35718. https://doi.org/10.7554/eLife.35718.

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

Callaghan, M.F., Freund, P., Draganski, B., Anderson, E., Cappelletti, M., Chowdhury, R., Diedrichsen, J., Fitzgerald, T.H.B., Smittenaar, P., Helms, G., Lutti, A., Weiskopf, N., 2014. Widespread age-related differences in the human brain microstructure revealed by quantitative magnetic resonance imaging. Neurobiol. Aging 35, 1862–1872. doi:10.1016/j.neurobiolaging.2014.02.008.

Carey, D., Krishnan, S., Callaghan, M.F., Sereno, M.I., Dick, F., 2017. Functional and quantitative MRI mapping of somatomotor representations of human supralaryngeal vocal tract. Cereb. Cortex. doi:10.1093/cercor/bhw393.

Carrasco, J.L., Caceres, A., Escaramis, G., Jover, L., 2014. Distinguishability and agreement with continuous data. Stat. Med. 33, 117–128. doi:10.1002/sim.5896.

Castella, R., Arn, L., Dupuis, E., Callaghan, M.F., Draganski, B., Lutti, A., 2018. Controlling motion artefact levels in MR images by suspending data acquisition during periods of head motion. Magn. Reson. Med. 80, 2415–2426. doi:10.1002/mrm.27214.

Cercignani, M., Dowell, N.G., Tofts, P.S, 2018. Quantitative MRI of the Brain: Principles of Physical Measurement. CRC Press, Milton, p. 611 Second edition.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6, 284–290. doi:10.1037/1040-3590.6.4.284.

Cohen, J., 1988. Statistical Power Analysis For the Behavioral Sciences, 2nd ed. Erlbaum, Hillsdale, NJ [u.a.], XXI, 567 S.

Deichmann, R., Schwarzbauer, C., Turner, R., 2004. Optimisation of the 3D MDEFT sequence for anatomical brain imaging: technical implications at 1.5 and 3 T. Neuroimage 21, 757–767. doi:10.1016/j.neuroimage.2003.09.062.

Deoni, S.C.L., Williams, S.C.R., Jezzard, P., Suckling, J., Murphy, D.G.M., Jones, D.K., 2008. Standardized structural magnetic resonance imaging in multicentre studies using quantitative T1 and T2 imaging at 1.5 T. Neuroimage 40, 662–671. doi:10.1016/j.neuroimage.2007.11.052.

Dick, F.K., Lehet, M.I., Callaghan, M.F., Keller, T.A., Sereno, M.I., Holt, L.L., 2017. Extensive Tonotopic Mapping across auditory cortex is recapitulated by spectrally directed attention and systematically related to cortical myeloarchitecture. J. Neurosci. 37, 12187–12201. doi:10.1523/JNEUROSCI.1436-17.2017.

Draganski, B., Ashburner, J., Hutton, C., Kherif, F., Frackowiak, R.S.J., Helms, G., Weiskopf, N., 2011. Regional specificity of MRI contrast parameter changes in normal ageing revealed by voxel-based quantification (VBQ). Neuroimage 55, 1423–1434. doi:10.1016/j.neuroimage.2011.01.052.

Ernst, R.R., Anderson, W.A., 1966. Application of fourier transform spectroscopy to magnetic resonance. Rev. Sci. Instruments 37, 93–102. doi:10.1063/1.1719961.

Gallardo-Ruiz, R., Crespo-Facorro, B., Setién-Suero, E., Tordesillas-Gutierrez, D., 2019. Long-term grey matter changes in first episode psychosis: a systematic review. Psychiatry Investig. 16, 336–345. doi:10.30773/pi.2019.02.10.1.

Goto, M., Abe, O., Aoki, S., Hayashi, N., Miyati, T., Takao, H., Matsuda, H., Yamashita, F., Iwatsubo, T., Mori, H., Kunimatsu, A., Ino, K., Yano, K., Ohtomo, K., 2015. Influence of parameter settings in voxel-based morphometry 8. Using DARTEL and region-of-interest on reproducibility in gray matter volumetry. Methods Inf. Med. 54, 171–178. doi:10.3414/ME14-01-0049.

Helbling, S., Teki, S., Callaghan, M.F., Sedley, W., Mohammadi, S., Griffiths, T.D., Weiskopf, N., Barnes, G.R., 2015. Structure predicts function: combining non-invasive electrophysiology with in-vivo histology. Neuroimage 108, 377–385. doi:10.1016/j.neuroimage.2014.12.030.

Helms, G., Dathe, H., Dechent, P., 2008a. Quantitative FLASH MRI at 3T using a rational approximation of the Ernst equation. Magn. Reson. Med. 59, 667–672. doi:10.1002/mrm.21542.

Helms, G., Dathe, H., Dechent, P., 2010. Modeling the influence of TR and excitation flip angle on the magnetization transfer ratio (MTR) in human brain obtained from 3D spoiled gradient echo MRI. Magn. Reson. Med. 64, 177–185. doi:10.1002/mrm.22379.

Helms, G., Dathe, H., Kallenberg, K., Dechent, P., 2008b. High-resolution maps of magnetization transfer with inherent correction for RF inhomogeneity and T1 relaxation obtained from 3D FLASH MRI. Magn. Reson. Med. 60, 1396–1407. doi:10.1002/mrm.21732.

Helms, G., Draganski, B., Frackowiak, R., Ashburner, J., Weiskopf, N., 2009. Improved segmentation of deep brain grey matter structures using magnetization transfer (MT) parameter maps. Neuroimage 47, 194–198. doi:10.1016/j.neuroimage.2009.03.053.

Hopkins, W.G., 2000. Measures of reliability in sports medicine and science. Sports Med. 30, 1–15. doi:10.2165/00007256-200030010-00001.

Jones, D.K., Symms, M.R., Cercignani, M., Howard, R.J., 2005. The effect of filter size on VBM analyses of DT-MRI data. Neuroimage 26, 546–554. doi:10.1016/j.neuroimage.2005.02.013.

Kanyongo, G.Y., Brook, G.P., Kyei-Blankson, L., Gocmen, G., 2007. Reliability and statistical power: how measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. J. Mod. App. Stat. Meth. 6, 81–90. doi:10.22237/jmasm/1177992480, jmasm.eP1126.

Kodama, M., Ono, T., Yamashita, F., Ebata, H., Liu, M., Kasuga, S., Ushiba, J., 2018. Structural gray matter changes in the hippocampus and the primary motor cortex on an-hour-to-one- day scale can predict arm-reaching performance improvement. Front. Hum. Neurosci. 12, 209. doi:10.3389/fnhum.2018.00209.

Koo, T.K., Li, M.Y., 2016. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15, 155–163. doi:10.1016/j.jcm.2016.02.012.

Krajcovicova, L., Klobusiakova, P., Rektorova, I., 2019. Gray matter changes in Parkinson's and Alzheimer's disease and relation to cognition. Curr. Neurol. Neurosci. Rep. 19, 85. doi:10.1007/s11910-019-1006-z.

Kurth, F., Gaser, C., Luders, E., 2015. A 12-step user guide for analyzing voxel-wise gray matter asymmetries in statistical parametric mapping (SPM). Nat. Protoc. 10, 293–304. doi:10.1038/nprot.2015.014.

Lambert, C., Chowdhury, R., Fitzgerald, T.H.B., Fleming, S.M., Lutti, A., Hutton, C., Draganski, B., Frackowiak, R., Ashburner, J., 2013. Characterizing aging in the human brainstem using quantitative multimodal MRI analysis. Front. Hum. Neurosci. 7, 462. doi:10.3389/fnhum.2013.00462.

Lehmann, N., Aye, N., Kaufmann, J., Heinze, H.-J., Düzel, E., Ziegler, G., Taubert, M., 2021. Longitudinal reproducibility of neurite orientation dispersion and density imaging (NODDI) derived metrics in the white matter. Neuroscience 457, 165–185. doi:10.1016/j.neuroscience.2021.01.005.

Lerch, J.P., van der Kouwe, A.J.W., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K.L., Smith, S.M., Fischl, B., Sotiropoulos, S.N., 2017. Studying neuroanatomy using MRI. Nat. Neurosci. 20, 314–326. doi:10.1038/nn.4501.

Leutritz, T., Seif, M., Helms, G., Samson, R.S., Curt, A., Freund, P., Weiskopf, N., 2020. Multiparameter mapping of relaxation (R1, R2*), proton density and magnetization transfer saturation at 3 T: a multicenter dual-vendor reproducibility and repeatability study. Hum. Brain Mapp. 41, 4232–4247. doi:10.1002/hbm.25122.

Levesque, I.R., Chia, C.L.L., Pike, G.B., 2010. Reproducibility of in vivo magnetic resonance imaging-based measurement of myelin water. J. Magn. Reson. Imaging 32, 60–68. doi:10.1002/jmri.22170.

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., Baltes, P.B., 2004. Transformations in the couplings among intellectual abilities and con-

stituent cognitive processes across the life span. Psychol. Sci. 15, 155–163. doi:10.1111/j.0956-7976.2004.01503003.x.
Loken, E., Gelman, A., 2017. Measurement error and the replication crisis. Science 355, 584–585. doi:10.1126/science.aal3618.
Lommers, E., Simon, J., Reuter, G., Delrue, G., Dive, D., Degueldre, C., Balteau, E., Phillips, C., Maquet, P., 2019. Multiparameter MRI quantification of microstructural tissue alterations in multiple sclerosis. Neuroimage Clin. 23, 101879. doi:10.1016/j.nicl.2019.101879.
Lorio, S., Lutti, A., Kherif, F., Ruef, A., Dukart, J., Chowdhury, R., Frackowiak, R.S., Ashburner, J., Helms, G., Weiskopf, N., Draganski, B., 2014. Disentangling in vivo the effects of iron content and atrophy on the ageing human brain. Neuroimage 103, 280–289. doi:10.1016/j.neuroimage.2014.09.044.
Lutti, A., Corbin, N., Ashburner, J., Ziegler, G., Draganski, B., Phillips, C., Kherif, F., Callaghan, M.F., Di Domenicantonio, G., 2022. Restoring statistical validity in group analyses of motion-corrupted MRI data. Hum. Brain Mapp. doi:10.1002/hbm.25767.
Lutti, A., Dick, F., Sereno, M.I., Weiskopf, N., 2014. Using high-resolution quantitative mapping of R1 as an index of cortical myelination. Neuroimage 93 (Pt 2), 176–188. doi:10.1016/j.neuroimage.2013.06.005.
Lutti, A., Hutton, C., Finsterbusch, J., Helms, G., Weiskopf, N., 2010. Optimization and validation of methods for mapping of the radiofrequency transmit field at 3T. Magn. Reson. Med. 64, 229–238. doi:10.1002/mrm.22421.
Lutti, A., Stadler, J., Josephs, O., Windischberger, C., Speck, O., Bernarding, J., Hutton, C., Weiskopf, N., 2012. Robust and fast whole brain mapping of the RF transmit field B1 at 7T. PloS one 7, e32379. doi:10.1371/journal.pone.0032379.
Maisog, J.M., Chmielowska, J., 1998. An efficient method for correcting the edge artifact due to smoothing. Hum. Brain Mapp. 6, 128–136 https://doi.org/10.1002/(SICI)1097-0193(1998)6:3<128:AID-HBM2>3.0.CO;2-5.
Matheson, G.J., 2019. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. Peer J. 7, e6918. doi:10.7717/peerj.6918.
Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Boomsma, D., Cannon, T., Kawashima, R., Mazoyer, B., 2001a. A probabilistic atlas and reference system for the human brain: international Consortium for Brain Mapping (ICBM). Philosophical transactions of the Royal Society of London. Series B, Biol. Sci. 356, 1293–1322. doi:10.1098/rstb.2001.0915.
Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Feidler, J., Smith, K., Boomsma, D., Hulshoff Pol, H., Cannon, T., Kawashima, R., Mazoyer, B., 2001b. A four-dimensional probabilistic atlas of the human brain. J. Am. Med. Inf. Assoc. 8, 401–430. doi:10.1136/jamia.2001.0080401.
Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). Neuroimage 2, 89–101. doi:10.1006/nimg.1995.1012.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46. doi:10.1037/1082-989X.1.1.30.
Mikl, M., Marecek, R., Hlustík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., Krupa, P., 2008. Effects of spatial smoothing on fMRI group inferences. Magn. Reson. Imaging 26, 490–503. doi:10.1016/j.mri.2007.08.006.
Mills, K.L., Goddings, A.-L., Herting, M.M., Meuwese, R., Blakemore, S.-J., Crone, E.A., Dahl, R.E., Güroğlu, B., Raznahan, A., Sowell, E.R., Tamnes, C.K., 2016. Structural brain development between childhood and adulthood: convergence across four longitudinal samples. Neuroimage 141, 273–281. doi:10.1016/j.neuroimage.2016.07.044.
Mugler, J.P., Brookeman, J.R., 1990. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magn. Reson. Med. 15, 152–157. doi:10.1002/mrm.1910150117.
Neeb, H., Zilles, K., Shah, N.J., 2006. A new method for fast quantitative mapping of absolute water content in vivo. Neuroimage 31, 1156–1168. doi:10.1016/j.neuroimage.2005.12.063.
Nöth, U., Volz, S., Hattingen, E., Deichmann, R., 2014. An improved method for retrospective motion correction in quantitative T2* mapping. Neuroimage 92, 106–119. doi:10.1016/j.neuroimage.2014.01.050.
Papp, D., Callaghan, M.F., Meyer, H., Buckley, C., Weiskopf, N., 2016. Correction of inter-scan motion artifacts in quantitative R1 mapping by accounting for receive coil sensitivity effects. Magn. Reson. Med. 76, 1478–1485. doi:10.1002/mrm.26058.
Tofts, Paul S., 2018. Quality Assurance: accuracy, precision, controls and phantoms 1. In: Cercignani, M., Dowell, N.G., Tofts, P.S. (Eds.), *Quantitative MRI of the Brain: Principles of Physical Measurement.* CRC Press, Milton, pp. 33–53 Second edition.
Penny, W., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. Neuroimage 19, 727–741. doi:10.1016/S1053-8119(03)00071-5.
Péran, P., Hagberg, G., Luccichenti, G., Cherubini, A., Brainovich, V., Celsis, P., Caltagirone, C., Sabatini, U., 2007. Voxel-based analysis of R2* maps in the healthy human brain. J. Magn. Reson. Imaging 26, 1413–1420. doi:10.1002/jmri.21204.
Pizzagalli, F., Auzias, G., Yang, Q., Mathias, S.R., Faskowitz, J., Boyd, J.D., Amini, A., Rivière, D., McMahon, K.L., Zubicaray, G.I.de, Martin, N.G., Mangin, J.-.F., Glahn, D.C., Blangero, J., Wright, M.J., Thompson, P.M., Kochunov, P., Jahanshad, N., 2020. The reliability and heritability of cortical folds and their genetic correlations across hemispheres. Commun. Biol. 3, 510. doi:10.1038/s42003-020-01163-1.
Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M., Nichols, T.E., Poline, J.-B., Vul, E., Yarkoni, T., 2016. Scanning the Horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18 (2), 115–126. https://doi.org/10.1038/nrn.2016.167.
Schnack, H.G., van Haren, N.E.M., Brouwer, R.M., van Baal, G.C.M., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T.D., Huttunen, M., Lepage, C., Collins, D.L., Evans, A., Murray, R.M., Kahn, R.S., Hulshoff Pol, H.E., 2010. Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. Hum. Brain Mapp. 31, 1967–1982. doi:10.1002/hbm.20991.
Schwartz, D.L., Tagge, I., Powers, K., Ahn, S., Bakshi, R., Calabresi, P.A., Todd Constable, R., Grinstead, J., Henry, R.G., Nair, G., Papinutto, N., Pelletier, D., Shinohara, R., Oh, J., Reich, D.S., Sicotte, N.L., Rooney, W.D., 2019. Multisite reliability and repeatability of an advanced brain MRI protocol. J. Magn. Reson. Imaging 50, 878–888. doi:10.1002/jmri.26652.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.
Shuter, B., Yeh, I.B., Graham, S., Au, C., Wang, S.-.C., 2008. Reproducibility of brain tissue volumes in longitudinal studies: effects of changes in signal-to-noise ratio and scanner software. Neuroimage 41, 371–379. doi:10.1016/j.neuroimage.2008.02.003.
Steiger, T.K., Weiskopf, N., Bunzeck, N., 2016. Iron level and myelin content in the ventral striatum predict memory performance in the aging brain. J. Neurosci. 36, 3552–3558. doi:10.1523/JNEUROSCI.3617-15.2016.
Tabelow, K., Balteau, E., Ashburner, J., Callaghan, M.F., Draganski, B., Helms, G., Kherif, F., Leutritz, T., Lutti, A., Phillips, C., Reimer, E., Ruthotto, L., Seif, M., Weiskopf, N., Ziegler, G., Mohammadi, S., 2019. hMRI - A toolbox for quantitative MRI in neuroscience and clinical research. Neuroimage 194, 191–210. doi:10.1016/j.neuroimage.2019.01.029.
Tardif, C.L., Gauthier, C.J., Steele, C.J., Bazin, P.-L., Schäfer, A., Schaefer, A., Turner, R., Villringer, A., 2016. Advanced MRI techniques to improve our understanding of experience-induced neuroplasticity. Neuroimage 131, 55–72. doi:10.1016/j.neuroimage.2015.08.047.
Terribilli, D., Schaufelberger, M.S., Duran, F.L.S., Zanetti, M.V., Curiati, P.K., Menezes, P.R., Scazufca, M., Amaro, E., Leite, C.C., Busatto, G.F., 2011. Age-related gray matter volume changes in the brain during non-elderly adulthood. Neurobiol. Aging 32, 354–368. doi:10.1016/j.neurobiolaging.2009.02.008.
Trefler, A., Sadeghi, N., Thomas, A.G., Pierpaoli, C., Baker, C.I., Thomas, C., 2016. Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. Neuroimage 133, 41–52. doi:10.1016/j.neuroimage.2016.02.034.
Valkanova, V., Eguia Rodriguez, R., Ebmeier, K.P., 2014. Mind over matter–what do we know about neuroplasticity in adults? Int. Psych. 26, 891–909. doi:10.1017/S1041610213002482.
Weiskopf, N., Callaghan, M.F., Josephs, O., Lutti, A., Mohammadi, S., 2014. Estimating the apparent transverse relaxation time (R2(*)) from images with different contrasts (ESTATICS) reduces motion artifacts. Front. Neurosci. 8, 278. doi:10.3389/fnins.2014.00278.
Weiskopf, N., Mohammadi, S., Lutti, A., Callaghan, M.F., 2015. Advances in MRI-based computational neuroanatomy: from morphometry to in-vivo histology. Curr. Opin. Neurol. 28, 313–322. doi:10.1097/WCO.0000000000000222.
Weiskopf, N., Suckling, J., Williams, G., Correia, M.M., Inkster, B., Tait, R., Ooi, C., Bullmore, E.T., Lutti, A., 2013. Quantitative multi-parameter mapping of R1, PD(*), MT, and R2(*) at 3T: a multi-center validation. Front. Neurosci. 7, 95. doi:10.3389/fnins.2013.00095.
Whitaker, K.J., Vértes, P.E., Romero-Garcia, R., Váša, F., Moutoussis, M., Prabhu, G., Weiskopf, N., Callaghan, M.F., Wagstyl, K., Rittman, T., Tait, R., Ooi, C., Suckling, J., Inkster, B., Fonagy, P., Dolan, R.J., Jones, P.B., Goodyer, I.M., Bullmore, E.T., 2016. Adolescence is associated with genomically patterned consolidation of the hubs of the human brain connectome. Proc. Natl. Acad. Sci. U.S.A. 113, 9105–9110. doi:10.1073/pnas.1601745113.
Wonderlick, J.S., Ziegler, D.A., Hosseini-Varnamkhasti, P., Locascio, J.J., Bakkour, A., van der Kouwe, A., Triantafyllou, C., Corkin, S., Dickerson, B.C., 2009. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. Neuroimage 44, 1324–1333. doi:10.1016/j.neuroimage.2008.10.037.
Ziegler, G., Grabher, P., Thompson, A., Altmann, D., Hupp, M., Ashburner, J., Friston, K., Weiskopf, N., Curt, A., Freund, P., 2018. Progressive neurodegeneration following spinal cord injury: implications for clinical trials. Neurology 90, e1257–e1266. doi:10.1212/WNL.0000000000005258.
Ziegler, G., Hauser, T.U., Moutoussis, M., Bullmore, E.T., Goodyer, I.M., Fonagy, P., Jones, P.B., Lindenberger, U., Dolan, R.J., 2019. Compulsivity and impulsivity traits linked to attenuated developmental frontostriatal myelination trajectories. Nat. Neurosci. 22, 992–999. doi:10.1038/s41593-019-0394-3.
Ziegler, G., Moutoussis, M., Hauser, T.U., Fearon, P., Bullmore, E.T., Goodyer, I.M., Fonagy, P., Jones, P.B., Lindenberger, U., Dolan, R.J., 2020. Childhood socio-economic disadvantage predicts reduced myelin growth across adolescence and young adulthood. Hum. Brain Mapp. 41, 3392–3402. doi:10.1002/hbm.25024.
Zou, K.H., Du, H., Sidharthan, S., Detora, L.M., Chen, Y., Ragin, A.B., Edelman, R.R., Wu, Y., 2010. Statistical evaluations of the reproducibility and reliability of 3-tesla high resolution magnetization transfer brain images: a pilot study on healthy subjects. Int. J. Biomed. Imaging 2010, 618747. doi:10.1155/2010/618747.
Zuo, X.-.N., Xu, T., Milham, M.P., 2019. Harnessing reliability for neuroscience research. Nature Hum. Behav. 3, 768–771. doi:10.1038/s41562-019-0655-x.