Title
Multisite reliability and repeatability of an advanced brain MRI protocol

Journal
Journal of Magnetic Resonance Imaging, 50(3)

ISSN
1053-1807

Authors
Schwartz, Daniel L
Tagge, Ian
Powers, Katherine
et al.

Publication Date
2019-09-01

DOI
10.1002/jmri.26652

Peer reviewed

HHS Public Access

Author manuscript

J Magn Reson Imaging. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

J Magn Reson Imaging. 2019 September ; 50(3): 878–888. doi:10.1002/jmri.26652.

Multisite reliability and repeatability of an advanced brain MRI protocol

Daniel L. Schwartz, BA[#1,2], **Ian Tagge, PhD**[#1], **Katherine Powers, BS**[1], **Sinyeob Ahn, PhD**[6], **Rohit Bakshi, MD**[3], **Peter A. Calabresi, MD**[4], **R. Todd Constable, PhD**[5], **John Grinstead, PhD**[1,6], **Roland G. Henry, PhD**[7], **Govind Nair, PhD**[8], **Nico Papinutto, PhD**[7], **Daniel Pelletier, MD**[10], **Russell Shinohara, PhD**[11], **Jiwon Oh, MD, PhD**[4,9], **Daniel S. Reich, MD, PhD**[8], **Nancy L. Sicotte, MD**[12], **William D. Rooney, PhD**[1,2] **NAIMS Cooperative**

[1]Advanced Imaging Research Center, Oregon Health & Science University, Portland, OR

[2]Neurology, Oregon Health & Science University, Portland, OR

[3]Brigham & Women's Hospital, Harvard Medical School, Boston, MA

[4]Johns Hopkins University, Baltimore, MD

[5]Yale University, New Haven, CT

[6]Siemens Healthineers, Malvern, PA

[7]University of California San Francisco, San Francisco, CA

[8]National Institute of Neurological Disease and Stroke, National Institute of Health, Bethesda, MD

[9]University of Toronto, Ontario, Canada

[10]University of Southern California, Keck School of Medicine, Los Angeles, CA

[11]University of Pennsylvania, Philadelphia, PA

[12]Cedars-Sinai Medical Center, Los Angeles, CA

[#] These authors contributed equally to this work.

Abstract

BACKGROUND: MRI is the imaging modality of choice for diagnosis and intervention assessment in neurological disease. Its full potential has not been realized due in part to challenges in harmonizing advanced techniques across multiple sites.

PURPOSE: To develop a method for the assessment of reliability and repeatability of advanced multisite-multisession neuroimaging studies and specifically to assess the reliability of an advanced MRI protocol, including multiband fMRI and diffusion tensor MRI, in a multi-site setting.

STUDY TYPE: Prospective.

Corresponding Author: D.S. or W.R. : Advanced Imaging Research Center, Oregon Health & Science University, Mailcode:L452, 3181 SW Sam Jackson Pk. Rd., Portland, OR, 97239, United State: schwartd@ohsu.edu or rooneyw@ohsu.edu.

[13]a list of other North American Imaging in Multiple Sclerosis (NAIMS) Cooperative participants in this study is provided in the Acknowledgements

POPULATION: Twice repeated measurement of a single subject with stable relapsing-remitting multiple sclerosis (MS) at seven institutions.

FIELD STRENGTH/SEQUENCE: A 3T MRI protocol included higher spatial resolution anatomical scans, a variable flip-angle longitudinal relaxation rate constant ($R_1 \equiv 1/T_1$) measurement, quantitative magnetization transfer imaging, diffusion tensor imaging, and a resting-state fMRI (rsFMRI) series.

ASSESSMENT: Multiple methods of assessing intra-site repeatability and inter-site reliability were evaluated for imaging metrics derived from each sequence.

STATISTICAL TESTS: Student's t, Pearson's r, and ICC(2,1) were employed to assess repeatability and reliability. Two new statistical metrics are introduced which frame reliability and repeatability in the respective units of the measurements themselves.

RESULTS: Intra-site repeatability was excellent for quantitative R_1, magnetization transfer ratio (MTR), and diffusion-weighted imaging (DWI) based metrics ($r>0.95$). rsFMRI metrics were less repeatable ($r=0.8$). Inter-site reliability was excellent for R_1, MTR, and DWI (ICC>0.9), and moderate for rsFMRI metrics (ICC~0.4).

DATA CONCLUSION: From most reliable to least, using a new reliability metric introduced here, MTR>R_1>DWI>rsFMRI; for repeatability, MTR>DWI>R_1>rsFMRI. A graphical method for at-a-glance assessment of reliability and repeatability, effect sizes, and outlier identification in multisite-multisession neuroimaging studies is introduced.

Keywords

multisite; multiple sclerosis; MRI; reliability; repeatability

INTRODUCTION

Magnetic resonance imaging (MRI) is frequently used for the investigation of the pathophysiology and progression of neurological disease, but metrics derived from imaging data may be inadequate diagnostic and prognostic markers for clinical symptoms (e.g. in multiple sclerosis, (1)). Efforts to develop more reliable, sensitive and specific measurements in the clinic have been successful in improving their relationship to imaging, but variability in imaging metrics remains high, especially when comparing values across different sites. Demonstrating that similar outcome measures can be realized at multiple sites by independent investigators is an essential step in establishing a standardized imaging protocol. Quality assurance metrics of raw images (signal- and contrast-to-noise) and image analysis results must be rigorously compared across imaging sessions and sites. One of the first tasks undertaken by the North American Imaging in Multiple Sclerosis (NAIMS) Cooperative was harmonization of an advanced one-hour MRI protocol with whole brain coverage (2).

The protocol included: 1) R_1 measurement ($R_1 \equiv 1/T_1$) using a variable flip angle method, a putative proxy for macromolecular content and structural organization within tissue (3); 2) magnetization transfer ratio (MTR) imaging, shown to be sensitive for tissue damage in gray matter (GM) and white matter (WM)(4) that precedes the development of acute lesions (5–

8); 3) diffusion weighted imaging (DWI), a method that quantifies restricted movement of water in the brain parenchyma; and 4) resting state functional MRI (rsFMRI), a dynamic method that maps concordance among spontaneous fluctuations in blood-oxygen level dependent signal across brain regions. R_1 is sensitive to inflammatory and demyelinating processes (9) and is decreased in normal appearing white matter (NAWM) in MS (10,11). MTR may predict cognitive deterioration in MS (12,13), and has been shown to be relatively stable (14,15) though one study has needed to "normalize" MTR (6). The association of DWI metrics with MS has been variable, ranging from high (16) to nonexistent (17), though sensitivity for lesion detection is high (18). rsFMRI has been the most variable of the four with respect to studies of MS, reporting decreases (19) and increases (20,21) in MS patients relative to controls, though a recent report found strong correlations with well-established disability assessment tools (22).

Assessment of the reliability of a given imaging metric or method across different sites or sessions is complex. While the intraclass correlation coefficient (ICC, (23)) is a well-documented metric for assessment of agreement, it is inadequate for the assessment of all sources of variance within each imaging modality; additionally, the ICC(2,1) metric is a dimension-less quantity for which interpretation is loosely defined (24). A metric that is suited to providing effect sizes for multisite imaging studies would be valuable. We propose new metrics for reliability (RAJ, agreement across sites) and repeatability (RPT, agreement within site) that employ departure from the measured sample distribution and introduce a graphical approach for assessing both metrics at a single glance. Each metric uses a jackknife procedure to anchor it within the distribution of the study, which provides an assessment of reliability in situations for which ground truth is difficult to assess (e.g. *in vivo* MRI). The method is similar to Bland-Altman analysis; however, Bland-Altman is used to assess differences between two methods of measuring the same quantity. Multisite imaging studies measure the same quantity with the *same* method. The new method (RAJ plot) provides reliability assessment at a single glance, visual representations of effect sizes (in this study between lesional and control tissue measurements), and repeatability in the form of bar ranges.

The purpose of this study was to evaluate the reliability and repeatability of advanced imaging metrics derived from R_1 mapping, MTR imaging, DWI and rsFMRI in the context of known pathology and to present a novel method of assessing measurement variability.

MATERIALS AND METHODS

Experimental Design

A 45 year old male with clinically definite, stable relapsing-remitting multiple sclerosis and mild-to-moderate physical disability (13 years disease duration, no new lesions or significant clinical progression for at least 400 days prior to this study) was selected from a cohort at the National Institutes of Health (NIH). Further disease characteristics of this subject are published separately (25). The subject travelled to seven North American sites and underwent two distinct 3T MRI sessions at each site. All MRI instruments were manufactured by Siemens, though instrument models, software versions, and hardware (including RF coils) varied by site. Informed consent was obtained at each imaging center.

The timeline, hardware and software details for each site can be found in Figure 1. The subject was removed from the magnet between intra-site scanning sessions and was re-registered prior to the second scan.

MRI Acquisition Parameters

All sites collected data using Siemens 3T MRI instruments with a body RF coil transmitter and a head only or head/neck RF coil receiver (see Figure 1). Employees of the instrument vendor were instrumental in the development and application of many of the sequences used in the protocol, but did not contribute to study design, data acquisition, or data analysis. Each site supplied their own MRI operator and imaging slabs were positioned manually. The variable flip angle (VFA) method was employed to create whole-brain quantitative T_1 (qT_1), reported here as R_1 ($\equiv 1/T_1$) maps. VFA images were acquired using whole-brain 3D gradient recalled echo (GRE) sequences with 2.3 ms echo time (TE), 20 ms repetition time (TR), four read-RF pulses of nominal flip angle (FA): 3°, 6°, 10°, 20°; field of view (FOV) 19.2cm x 25.6cm x 14.4 cm [LR x AP x HF], 192 × 256 × 48 matrix, 1 min 20 sec acquisition time per FA. B_1 field maps were acquired using a Siemens calibration pulse sequence for 2D B_1 mapping by measuring the ratio of spin echo and stimulated echo with TR 1000 ms, TE 14 ms, FA 90° (FA 80° at site 6), FOV $(25.6\ cm)^2$, 64 × 64 matrix, and 24 5 mm thick slices, 1 min 9 sec acquisition time. Site 2 also acquired a B_1 field map using the Siemens turbo-flash (TFL) based product protocol with TR 10,200 ms, TE 2.02 ms, FA 8°, FOV $(25.6\ cm)^2$, 64 × 64 matrix, 40 slices, and 3 mm thick slices, 0 min 21 sec acquisition time. Magnetization transfer imaging (MTI) data were acquired with a prototype 3D GRE pulse sequence providing a flexible MT frequency offset, with two saturation pulse offset frequencies (4 kHz for MT effect on, 100 kHz for MT effect off) at the peak B_1 of 8 μT amplitude and 15 ms Gaussian saturation pulse, sufficient for MTR calculation. Other MTI acquisition parameters include: TR 43 ms, TE 2.3 ms, FA 10°, FOV 19.2 cm x 25.6 cm x 14.4 cm, 192 × 256 × 48 matrix, 2 min 45 sec for each MT contrast. Total acquisition time for all MTR and VFA images was 11 min 59 sec. rsFMRI data were collected with whole-brain 2D echo planar imaging (EPI) with a multiband factor of 4 (MB=4, C2P provided by Center for Magnetic Resonance Research, University of Minnesota, USA) (26), TR 1 sec, TE 30 ms, FA 55°, FOV $(22\ cm)^2$, 110 × 110 matrix, 60 2 mm thick contiguous slices, iPAT 4, number of volumes acquired was between 280 and 340, ~5 min acquisition time. DWI data were collected with whole-brain 2D EPI and a bipolar diffusion scheme with a multiband factor of 2 (MB=2, C2P provided by Center for Magnetic Resonance Research, University of Minnesota, USA) (27), 64–65 isotropic directions at b=2000 sec/mm^2, with 4(site 3) or 8 (all other sites) unweighted (b=0) volumes, TR 4.3 sec, TE 96 ms, FA 90°, FOV $(19.8\ cm)^2$, 86 × 86 imaging matrix, iPAT 2, total acquisition time ~5 min 36 sec. Site two acquired the DWI at b=1000 sec/mm^2, and site five acquired DWI at both b=2000 sec/mm^2 and b=~750 sec/mm^2 (a “multishell” experiment). These acquisitions were processed the same as all other DWI acquisitions. High spatial resolution T_1-w 3D MPRAGE (TE 2.52 ms, TR 1900 ms, inversion time (TI) 900 ms, FA 9°, 256 × 246 × 176 matrix, FOV $(25\ cm)^2$ × 17.6 cm, sagittal orientation, and T_2-w 3D FLAIR (TE 355 ms, TR 4800 ms, TI 1800 ms, FA 120°, 256 × 256 × 176 matrix, FOV $(25.6\ cm)^2$ × 17.6 cm sagittal orientation) were used for segmentation.

Preprocessing

All data were processed at a single site (OHSU).

Anatomical ProcessingAnd Lesion Segmentation—MPRAGE and FLAIR acquisitions were used for lesion segmentation. FLAIR was rigid body registered (6 DOF, OAR/FLIRT) to the MPRAGE and the MPRAGE was skull stripped (SPECTRE/TOADS-CRUISE (28)). The resultant brain mask was applied to the registered FLAIR, and both acquisitions were submitted to LesionTOADS (28) with inhomogeneity correction. Output binary lesion masks were visually inspected for accuracy against the FLAIR acquisition.

R_1 And MTR Processing—Images were converted to NIFTI format, coregistered to an average of all MT and VFA images acquired at session 1 at site 1 (inclusive, FLIRT/FSL), and skull stripped (BET/FSL). All images were averaged together to create a new population mean reference space and then subsequently re-registered to the population mean to ensure consistent manipulation of all images. MTR maps were calculated voxelwise for each session by [$MTR_i = (S_{0,i} - S_{sat,i}) / S_{0,i}$]; where $S_{0,i}$ is the signal intensity obtained from the i^{th} voxel with 100 kHz saturation pulse offset and $S_{sat,i}$ is the signal intensity obtained from the i^{th} voxel with 4 kHz saturation pulse offset. R_1 maps were calculated by fitting voxel signal intensity (S_i)

$$S_i(TR, \alpha) = M_{0i} \frac{\sin(B_{1i} \cdot \alpha) \cdot \left(1 - e^{-TR \cdot R_{1i}}\right)}{1 - e^{-TR \cdot R_{1i}} \cdot \cos(B_{1i} \cdot \alpha)} \quad (1)$$

using nonlinear least squares regression with equal weighting, where B_{1i} is the flip angle correction factor for the i^{th} voxel calculated from the acquired B_1 map and α is the nominal flip angle, M_{0i} is the total magnetization for the i^{th} voxel. Limited spatial SNR (SNR; defined here as [mean signal divided by the standard deviation over voxels]) and extensive B_1 inhomogeneities in the most superior and inferior slices in VFA imaging prohibited accurate R_1 estimation in these areas, so twelve mid-axial slices demonstrating good homogeneity were selected for R_1 analysis. Interestingly, although R_1 values were generally reasonable, the Siemens service protocol for B_1 mapping exacerbated B_1–related inhomogeneity in the site 2 R_1 maps. The TFL-based B_1 map for site 2 produced flat maps, but underestimated R_1 relative to other sites due to global underestimation of actual flip angles. Thus, the TFL-based B_1 map was histogram matched to the service sequence B_1 before being used for B_1 correction for site 2 R_1 mapping.

Tissue-specific quantification of R_1 and MTR metrics was performed in site 1 session 1 space (i.e. ROIs were applied in the within subject template space).

DWI Processing—Mosaic DICOM images, b-vectors, and b-values were compiled into NIFTI format (MRIConvert), eddy current/motion corrected (3dAllineate --EPI –mutualinfo/ AFNI), unweighted volumes were affine registered to session 1 at site 1 (inclusive, flirt/ FSL), unweighted volumes were averaged together and nonlinear tensor estimation

performed (3dDWItoDT –eigs/AFNI). Tractography was performed with MedINRIA. *A priori* ROI masks (right and left middle frontal gyrus [R/L MFG], right and left inferior parietal lobule [R/L IPL], and bilateral PCC [BiPCC]) were chosen using TT_Daemon ROIs (@auto_tlrc, 3dWarp, convert_xfm, whereami, 3dAutomask, 3dcalc/AFNI+FSL) and limited to juxtacortical WM (FA > 0.1, 3dclust, 3dcalc, 3dmaskave/AFNI). A single normal appearing white matter (NAWM) ROI contralateral to the lesion was constructed using the lesion mask generated as above and flipped from right to left in Talairach space (3dLRflip). Lesion, GM, and WM ROIs were originally generated on the T_1 and FLAIR images acquired at site 1 session 1, in an effort to avoid segmentation or spatially-specific ROI differences to bias site-specific DWI results. All ROIs were then back transformed to native space (convert_xfm, flirt/FSL).

RsFMRI Processing—Mosaic DICOM images were compiled to NIFTI with .json dump (dcmstack), slice-time corrected (3dTshift [custom schedule]/AFNI), motion corrected (3dvolreg/AFNI), affine registered to session 1 at site 1 (inclusive, flirt/FSL), bandpass filtered at 0.01–0.1 Hz and blurred at $(4mm)^3$ FWHM (3dBandpass, 3dmerge/AFNI), and volume truncated to be the same number of samples at every site for every session (256 volumes). Time courses were averaged over the voxels in each *a priori* ROI constructed as above (without juxtacortical WM limitation, 3dmaskave/AFNI) and correlated with one another (1ddot/AFNI). Each region pair was r-to-Fisher's Z transformed (1deval/AFNI, ten total region pairs). Temporal and spatial SNR measurements (mean/standard deviation) were taken over the L MFG ROI.

Statistical Analysis

Pearson's r was employed to assess within-site agreement. Multivariate 2-way ANOVA ([site] and [measurement type]) was employed to assess site effects across measurements. In the case of rsFMRI, ROI-based diffusion analyses, and lesion to non-lesional tissue in MTR and R_1 analyses, a multivariate 3-way ANOVA was employed ([site] and [measurement type] and [region/interregional correlation]). In each case measurements were assessed as grand means across sessions. ICC(2,1) was calculated by

$$ICC(2,1) = \frac{MSV_M - MSV_E}{MSV_M + (n-1) * MSV_E + n * \left(MSV_B - MSV_E\right)/k} \quad (2)$$

where MSV_M is the mean squared variance (MSV) in measurement, MSV_E is the MSV in the residual, n is the number of sites, MSV_B is the MSV in site, and k is the number of measurements, which produced a metric for which "1" is perfect agreement. Excellent agreement was loosely defined as values greater than 0.8 (24).

Two new methods were employed to assess reliability and repeatability for most imaging metrics within each tissue class in an attempt to provide guidance for expected effect sizes across multisite imaging studies and to test an application to real data. The first is reliability assessment with jackknifing (RAJ), for which the jackknife difference of each sample from

the sample mean without that sample (RAJ_i "site bias pseudovalue") is calculated in normalized units as:

$$RAJ_i = \frac{\left| \left(\frac{1}{n-1} \sum_{i \neq 1}^{n} a_i \right) - a_i \right|}{\bar{a}} \quad (3)$$

where n is the number of samples, a is the measurement value at each sample averaged over repeated measurements (in this case, sessions at each site), and i is site $\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$ (the mean of all measurements); RAJ_i is the ordinate axis of graphs in Figure 5. RAJ is the mean of all RAJ_i,

$$RAJ = \frac{1}{n} \sum_{i=1}^{n} RAJ_i \quad (4)$$

which yields a study-wide metric of jackknife unit deviation from the sample distribution for which high values denote less reliable measurements and low values denote more reliable measurements, an inverse relationship to that of ICC (Figure 5). Note that this metric is normalized to the mean magnitude of the measurement itself, and thus is easily comparable across methods and measurements.

Repeatability (RPT), defined as the variability of the same measurement on the same sample with the same equipment, is the average of the distance to the midpoint of repeated measurements in samples normalized to the mean of the measurement over all samples:

$$RPT = \frac{\frac{1}{n} \sum_{i=1}^{n} \left| \frac{a_i^x - a_i^y}{2} \right|}{\bar{a}} \quad (5)$$

where a^x and a^y are the first and second sessions at site i, and for which higher values indicate a measurement is less repeatable (Figure 5). This metric is easily modifiable for designs in which more than two repeated measurements were taken by replacing the numerator by the average of a variance metric over sites.

For the purpose of comparing RAJ metrics to ICC metrics, ICC(2,1) was computed over mean squared values from single factor ANOVAs on each tissue type in each modality (or inter/intrahemispheric connections in the case of rsFMRI). These values are displayed on the abscissa in Figure 5.

A variable flip angle method was used to collect R_1 measurements, which is faster than an inversion recovery experiment but suffers from B_1 inhomogeneity, and was corrected with a separately acquired B_1 map. MTR and R_1 images were acquired with relatively thick slices (3mm) in order to cover the whole brain in a reasonable timeframe. Spatial and temporal SNR measurements, as well as common post-processing metrics (e.g. for DWI, fractional anisotropy [FA], mean diffusivity [MD], and longitudinal diffusivity [LD]) were taken at each site for each measurement, and agreement assessed within site ("intrasite", N=2) and across sites ("intersite", N=7), both in normal appearing brain and in lesional and perilesional tissue.

RESULTS

The subject completed studies at all sites within a 4-month window. The subject successfully completed two MRI sessions at each site with the average time between the start of each session was 83 minutes with a range to 64 – 122 minutes (Figure 1).

All statistical testing is presented as: [test type]$_{(\text{degrees of freedom})}$=[test result],[significance value]. Summary statistics for each technique can be found in Table 1.

R_1 And MTR Mapping

Spatial SNR in VFA measurements over voxels in a right hemisphere WM volume were not stable within- and across-site at flip angles 3°, 10°, and 20° (between session $r_{(5)}$=.16, .29, .39, respectively, p>.05, Supplementary Figure 1, top left bar graph) but reasonably stable within and across-site at 6° ($r_{(5)}$=.85, p<.02); overall within- and across site spatial SNR was extremely stable ($r_{(26)}$=.86, p<.0001). Spatial SNR was stable in MTR both within-site and across-site for 4 kHz measurements (between session $r_{(5)}$=0.87,p=0.01, between site mean/stdev=42.9/4.2, top right bar graph in Supplementary Figure 1); 100 kHz measurements were stable between sites but not between sessions (between session $r_{(5)}$=0.22,p>0.5 between site mean/stdev=52.0/4.0). Calculated MTR and R_1 were stable across sessions in NAWM, cortex, and lesion masks ($r_{(19)}$=.99 and .96, p<.0001, respectively, Figure 2). MTR was relatively stable across sites as were R_1 measurements (ICC=.98 and .97, respectively); R_1 across-voxel averages remained stable relative to between session variance (Figure 2). Representative axial slices of MTR and R_1 maps can be found in Supplementary Figure 1. Reliability and repeatability for WM R_1 as measured by RAJ and RPT were RAJ/RPT_{WM}=2.7/2.4%, RAJ/RPT_{GM}=4.3/1.8%, RAJ/RPT_{lesion}=4.4/2.3%. Reliability and repeatability for MTR as measured by RAJ and RPT were RAJ/RPT_{WM}=1.7/.8%, RAJ/RPT_{GM}=1.4/.8%, RAJ/RPT_{lesion}=2.6%/1.5% (Figure 5).

DWI

Spatial SNR (calculated as the mean over voxels in an ROI in NAWM in an unweighted volume contralateral to the lesion divided by the standard deviation of voxels in the same ROI) was stable both within-site and across-site (intrasite SNR mean of the difference between session one and session two was .17, with a standard deviation of .16; intersite mean of the SNR average over sessions within site was 3.55, with a standard deviation of .19, Supplementary Figure 2). For all seven sites, FA, MD, and LD in NAWM (averaged over

voxels in five juxtacortical *a priori* NAWM ROIs, Figure 3C) were in excellent agreement within-site ($r_{(33)}$=.94,p<.0001, $r_{(33)}$=.96,p<.0001, $r_{(33)}$=.97,p<.0001, respectively), and extremely high agreement across-site (ICC=.904, Supplementary Figure 2). The effect of site ($F_{(6,12)}$=3.4,p=.03) and measurement ($F_{(2,12)}$=120.5,p<.0001) were significant. Measurements of FA (top), MD (middle), and LD (bottom) in the lesion and contralateral NAWM ROI (Figure 3) were extremely well correlated within-site ($r_{(6)}$=.91-.99,p<.0001 for all seven sites) and across-site (ICC=.914, Figure 3). The effect of site was significant ($F_{(6,42)}$=210.9,p<.0001) as was measurement ($F_{(3,42)}$=12.6,p<.05) and the effect of lesion ($F_{(1,42)}$=184.1,p<.0001, Figure 3). Reliability and repeatability for diffusion metrics in NAWM as measured by RAJ and RPT were RAJ/RPT_{FA}=6.5/.8%, RAJ/RPT_{MD}=16.7/1.4%, RAJ/RPT_{LD}=16.4/1.4%. Reliability and repeatability in the lesion as measured by RAJ and RPT were RAJ/RPT_{FA}=5.8/.7%, RAJ/RPT_{MD}=13.2/.7%, RAJ/RPT_{LD} =13.5/.7% (Figure 5). Tractography seeded by the lesion and a contralateral ROI were qualitatively similar across all sites and sessions and clearly displayed limited cross-callosal, anterior-posterior and cortical projections in the lesion seeded tractographic map (Supplementary Figure 2). Diffusion metrics within tracts were not compared quantitatively across sites or sessions.

RsFMRI

Temporal SNR (tSNR; calculated as the mean time course of all voxels in LMFG divided by the standard deviation of those voxels over time) was extremely reliable within-site (intrasite mean of the difference in tSNR between session 1 and session 2 was 51.1, with a standard deviation of 46.4), but there was considerable variability across-site (intersite mean of the average tSNR over sessions within site was 278.9, with a standard deviation of 117.7, Supplementary Figure 3); spatial SNR (calculated as the mean over voxels in LMFG at a single timepoint divided by the standard deviation of those same voxels) was stable both within-site (intrasite mean of the difference in SNR between session one and session two was .44, with a standard deviation of .30) and across-site (intersite mean of the average SNR over sessions within site was 5.40, with a standard deviation of 1.12, Supplementary Figure 3), though site 2 had approximately half the spatial SNR of the other sites (3.0 compared to 5.8 [mean over sites 1, 3–7]). Qualitatively, motion estimates were smaller at session 2 at all sites other than site 2 (Supplementary Figure 3). For all seven sites, interregional correlations (Fisher's z as calculated in rsFMRI Methods) were reliable between sessions (Figure 4, total $r_{(68)}$=.80, p<0.0001, intrasite session correlations (over Fisher's z in each of ten region pairs) ranged from $r_{(8)}$=.65-.96). Intersite measurements of agreement for the ten ROI pairs (ROIs shown in Supplementary Figure 3) were not reliable (ICC=.408); the effect of site ($F_{(6,54)}$=123.5, p<.0001) and connection ($F_{(9,54)}$=63.2, p<.0001) were significant. Reliability and repeatability for intrahemispheric (LIPL-LMFG and RIPL-RMFG) and interhemispheric (LIPL-RIPL and LMFG-RMFG) correlations as measured by RAJ and RPT were RAJ/RPT=20.6/10.7% and RAJ/RPT=32.1/11.5%, respectively (Figure 5).

RAJ And ICC

A log fit to the relationship between ICC(2,1) and RAJ for all measurements is displayed in Figure 5, for which the fit equation was $RAJ = .32 + .079e^{(1 - ICC)}$. The RAJ metric is significantly inversely correlated with ICC(2,1), $r_{(12)}$=.83, p=.0002.

DISCUSSION

This study investigated the reliability and reproducibility of several advanced 3T MRI techniques on a single relapsing-remitting MS subject across seven institutions with two sessions at each site. Consistency for all tested MRI modalities, including MTR, R_1 mapping, DWI, and rsFMRI, was excellent between sessions within a single site. The most quantitative measurement, R_1, is in good agreement with literature values obtained from VFA experiments (29), which can underestimate R_1 compared to the gold standard inversion recovery method (3,29). With few exceptions, qualitative (tractography) and semi-quantitative (MTR, diffusion tensor metrics) measurements were reliable across all sites. rsFMRI metrics, while relatively reliable within site, were not reliable across site, likely due to between site variability in physiological motion characteristics or differences in functional status of the subject across time. In addition, one site showed substantial differences with other sites, perhaps related to greatly reduced spatial SNR in FMRI acquisitions. All image post-processing was largely automated and therefore these methods are well-suited for continued development with the ultimate goal of using these metrics in much larger multisite studies. We used ICC(2,1) (23) to test the reliability of each metric across site. ICC is a well-recognized statistic for the assessment of test-retest reliability of MRI data and has been used in both functional and anatomical multisite studies (30,31). MTR measurements in NAWM were in excellent agreement across sites, as were diffusion measurements and R_1 measurements . However, rsFMRI correlations were not reliable.

This protocol was specifically designed to use advanced acquisition techniques in the shortest possible scan time in order to be easily portable to large scale multisite studies and to facilitate adding them *en masse* or piecemeal to existing imaging protocols. The use of advanced techniques in this study, such as MTI and multiband (or “simultaneous multislice”) echo planar imaging further increases their utility in future studies. Although comparing multiband to conventional multislice interleaving was not an explicit goal of the present study, the reliability of multiband DTI measures were excellent, similar to reports in the literature that used conventional multislice acquisitions. Site 2 demonstrated markedly lower (~50%) spatial SNR in MB=4 fMRI multi-band acquisitions compared to other sites, though had spatial SNR very similar to overall group in conventional and MB=2 acquisitions, perhaps indicating a head-coil RF receive issue. Unfortunately, there were no longitudinal data (phantom or otherwise) collected to confirm a coil sensitivity issue at this site. Sensitivity to pathology in this RRMS patient relative to healthy controls was not assessed by this study, as a comparison was not explicitly made and the primary goal of the study was to assess the application of these general methods in a multisite study. However, comparisons to normal appearing tissue are sufficient in this context to illustrate ostensible effect sizes for each modality. Though a MS patient was used for this study, the power of these findings lies in the study design and analysis and comparison of computed metrics across sites and sessions, as this method is widely applicable to all radiographically defined neuropathology and normal neurophysiology.

Intraclass correlation coefficients are qualitatively helpful for the overall assessment of the reliability of a given metric, but do not provide an immediately accessible quantitative measure of reliability or repeatability. For example, ICC may not be useful in comparing two

multisite studies which seek to assess the same biological process using the same method. In contrast, RAJ and RPT provide quantitative measurements of departure from reliability and repeatability in units of percent jackknife difference from sample mean, which is easily comparable across multisite studies. It is important to note that ICC(2,1) is inversely related to RAJ and RPT; that is, a *decrease* in RAJ or RPT indicates a more reliable and repeatable measurement, and an *increase* in ICC denotes a commensurate interpretation. The introduction of a graphical method for the facile scrutiny of multisite-multisession measurements may well serve the neuroimaging community, as multisite studies have become more common in recent years owing to the ready sharing of data and methods across platforms, and the identification of problem sites or measurements can be difficult using agglomerative metrics such as ICC. Reliability and repeatability in a multisite study may also be achieved by the inclusion of a phantom for site calibration. A recent review of the importance of the use of a phantom to calibrate data acquisition and analysis in longitudinal and/or in cross sectional multisite studies noted that phantoms that can deliver a quantitative reference or standard are of particular importance in multisite studies which make quantitative measurements, such as relaxation rates or lesion volumes(32). The use of such a phantom has been widely adopted by the Alzheimer's Disease Neuroimaging Initiative. Some studies have taken different approaches towards reproducibility measurements, such as performing measurements on two identical imaging systems(33), repeated within subject, or quantifying reproducibility using test-retest on human subjects while varying sequence parameters on a single imaging system(34). These studies provide some guidance for expectations of effect sizes and coefficients of variability when a multisite study must combine or compare data across different imaging systems, but also illustrate the difficulty of this task: reproducibility is governed by a host of factors for which the magnitude of variability may be study-, site-, or population-specific.

The data collection period occurred over a span of approximately 4 months; scheduling and travel to all the sites presented a logistical challenge. It is possible, though unlikely, that disease progression or regression occurred at a significant rate during this period and increased the amount of variance in the measurements across site; however, lesion load did not change when scanned at NIH pre- and post-study, suggesting stable disease (25). Additional post-processing techniques accounting for diffusion effects and imperfect spoiling in VFA R_1 mapping have been described. These corrections may improve R_1 fittings but were not explored in this work. Given the variability in fMRI measurements within site and the variability in motion estimates both between and within sites, differences across site reflected in poor reliability for this technique are more likely attributable to relatively noisy measurements as a result of motion than to disease state. A single manufacturer and B_0 field strength was used to make all MRI measurements reported in this study. Though this manufacturer and 3T magnets are well-represented in research centers, many sites use other vendors and/or 1.5T MRIs. It may be difficult to port or replicate this protocol on other manufacturer platforms or magnets with different B_0 strengths. The study design, specifically the use of a single patient, makes generalizability of the magnitude of agreement of these methods difficult. It is possible that subjects with a more severe disease burden, manifested as higher signal heterogeneity in a single tissue type, may decrease both repeatability and reliability. However, this report seeks in part to present a novel metric for

the assessment of intra- and intersite agreement which can be appropriately applied to data from other subjects, increasing the generalizability of this report. Finally, this study did not include a healthy control comparison, a reference standard, or the collection of phantom data, which might have served to elucidate a cause of a given measurement's departure from the rest of the sample. While one goal of the consortium is to quickly and effectively diagnose a data acquisition or processing issue arising from a given scan at a given site, this study and metric enables the consortium to frame the errant data point in the context of other scans across that site and across the consortium, which provides the information necessary to diagnose an issue without the necessity of a costly phantom purchase and/or distribution.

In conclusion, an advanced brain MRI protocol was able to be implemented and tested at seven sites across North America, comprised of differing software and hardware elements, though all manufactured by a single vendor, and the resulting derived metric of brain physiology were found to be largely reproducible. From most reliable to least, MTR>R_1>DWI>rsFMRI; for repeatability, MTR>DWI>R_1>rsFMRI. Finally, this work provides novel metrics for the assessment of the reliability and repeatability of multisite-multisession imaging studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

In addition to the authors listed, the following individuals contributed to this NAIMS study: Renxin Chu, Shahamat Tauhid, Subhash Tummala, Tobias Kober, Li Pan, Pascal Sati, Jack Simon, and William Stern. The authors declare no competing financial interests.

Grant support:

Major support for this study was provided by the Race to Erase MS foundation, and NIH S10 RR027694, S10 OD016356, and S10 OD018224.

REFERENCES

1. Barkhof F The clinico-radiological paradox in multiple sclerosis revisited. Curr Opin Neurol 2002;15(3):239–245. [PubMed: 12045719]
2. Oh J, Bakshi R, Calabresi PA, et al. The NAIMS cooperative pilot project: Design, implementation and future directions. Mult Scler 2017:1352458517739990.
3. Rooney WD, Johnson G, Li X, et al. Magnetic field and tissue dependencies of human brain longitudinal 1H2O relaxation in vivo. Magn Reson Med 2007;57(2):308–318. [PubMed: 17260370]
4. Tagge I, O'Connor A, Chaudhary P, et al. Spatio-Temporal Patterns of Demyelination and Remyelination in the Cuprizone Mouse Model. PLoS One 2016;11(4):e0152480. [PubMed: 27054832]
5. Chen JT, Easley K, Schneider C, et al. Clinically feasible MTR is sensitive to cortical demyelination in MS. Neurology 2013;80(3):246–252. [PubMed: 23269598]
6. Brown RA, Narayanan S, Arnold DL. Imaging of repeated episodes of demyelination and remyelination in multiple sclerosis. Neuroimage Clin 2014;6:20–25. [PubMed: 25610760]
7. Goodkin DE, Rooney WD, Sloan R, et al. A serial study of new MS lesions and the white matter from which they arise. Neurology 1998;51(6):1689–1697. [PubMed: 9855524]

8. Pike GB, de Stefano N, Narayanan S, Francis GS, Antel JP, Arnold DL. Combined magnetization transfer and proton spectroscopic imaging in the assessment of pathologic brain lesions in multiple sclerosis. AJNR Am J Neuroradiol 1999;20(5):829–837. [PubMed: 10369353]
9. Bonnier G, Roche A, Romascano D, et al. Advanced MRI unravels the nature of tissue alterations in early multiple sclerosis. Ann Clin Transl Neurol 2014;1(6):423–432. [PubMed: 25356412]
10. Vrenken H, Rombouts SA, Pouwels PJ, Barkhof F. Voxel-based analysis of quantitative T1 maps demonstrates that multiple sclerosis acts throughout the normal-appearing white matter. AJNR Am J Neuroradiol 2006;27(4):868–874. [PubMed: 16611780]
11. Berlow Y, Pollaro J, et al. Magnetic resonance imaging correlates of dichotic listening performance in multiple sclerosis. Seminars in Hearing 2012;33(3):283–294.
12. Filippi M, Preziosa P, Copetti M, et al. Gray matter damage predicts the accumulation of disability 13 years later in MS. Neurology 2013;81(20):1759–1767. [PubMed: 24122185]
13. Enzinger C, Barkhof F, Ciccarelli O, et al. Nonconventional MRI and microstructural cerebral changes in multiple sclerosis. Nat Rev Neurol 2015;11(12):676–686. [PubMed: 26526531]
14. Fernando KT, Tozer DJ, Miszkiel KA, et al. Magnetization transfer histograms in clinically isolated syndromes suggestive of multiple sclerosis. Brain 2005;128(Pt 12):2911–2925. [PubMed: 16219673]
15. Weiskopf N, Suckling J, Williams G, et al. Quantitative multi-parameter mapping of R1, PD(*), MT, and R2(*) at 3T: a multi-center validation. Front Neurosci 2013;7:95. [PubMed: 23772204]
16. Bester M, Jensen JH, Babb JS, et al. Non-Gaussian diffusion MRI of gray matter is associated with cognitive impairment in multiple sclerosis. Mult Scler 2015;21(7):935–944. [PubMed: 25392318]
17. Temel S, Keklikoglu HD, Vural G, Deniz O, Ercan K. Diffusion tensor magnetic resonance imaging in patients with multiple sclerosis and its relationship with disability. Neuroradiol J 2013;26(1):3–17. [PubMed: 23859160]
18. Inglese M, Bester M. Diffusion imaging in multiple sclerosis: research and clinical implications. NMR Biomed 2010;23(7):865–872. [PubMed: 20882528]
19. Lowe MJ, Phillips MD, Lurito JT, Mattson D, Dzemidzic M, Mathews VP. Multiple sclerosis: low-frequency temporal blood oxygen level-dependent fluctuations indicate reduced functional connectivity initial results. Radiology 2002;224(1):184–192. [PubMed: 12091681]
20. Dogonowski AM, Siebner HR, Sorensen PS, et al. Expanded functional coupling of subcortical nuclei with the motor resting-state network in multiple sclerosis. Mult Scler 2013;19(5):559–566. [PubMed: 23012251]
21. Hawellek DJ, Hipp JF, Lewis CM, Corbetta M, Engel AK. Increased functional connectivity indicates the severity of cognitive impairment in multiple sclerosis. Proc Natl Acad Sci U S A 2011;108(47):19066–19071. [PubMed: 22065778]
22. Wojtowicz M, Mazerolle EL, Bhan V, Fisk JD. Altered functional connectivity and performance variability in relapsing-remitting multiple sclerosis. Mult Scler 2014;20(11):1453–1463. [PubMed: 24619937]
23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–428. [PubMed: 18839484]
24. Cicchetti DV. Multiple comparison methods: establishing guidelines for their valid application in neuropsychological research. J Clin Exp Neuropsychol 1994;16(1):155–161. [PubMed: 8150886]
25. Shinohara RT, Oh J, Nair G, et al. Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject with Multiple Sclerosis. AJNR Am J Neuroradiol 2017.
26. Feinberg DA, Moeller S, Smith SM, et al. Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. PLoS One 2010;5(12):e15710. [PubMed: 21187930]
27. Setsompop K, Cohen-Adad J, Gagoski BA, et al. Improving diffusion MRI using simultaneous multi-slice echo planar imaging. Neuroimage 2012;63(1):569–580. [PubMed: 22732564]
28. Shiee N, Bazin PL, Ozturk A, Reich DS, Calabresi PA, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 2010;49(2):1524–1535. [PubMed: 19766196]
29. Stikov N, Boudreau M, Levesque IR, Tardif CL, Barral JK, Pike GB. On the accuracy of T1 mapping: searching for common ground. Magn Reson Med 2015;73(2):514–522. [PubMed: 24578189]

30. Caceres A, Hall DL, Zelaya FO, Williams SC, Mehta MA. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 2009;45(3):758–768. [PubMed: 19166942]
31. Barkhof F, Daams M, Scheltens P, et al. An MRI rating scale for amyloid-related imaging abnormalities with edema or effusion. AJNR Am J Neuroradiol 2013;34(8):1550–1555. [PubMed: 23436056]
32. Keenan KE, Ainslie M, Barker AJ, et al. Quantitative magnetic resonance imaging phantoms: A review and the need for a system phantom. Magn Reson Med 2018;79(1):48–61. [PubMed: 29083101]
33. Huang L, Wang X, Baliki MN, Wang L, Apkarian AV, Parrish TB. Reproducibility of structural, resting-state BOLD and DTI data between identical scanners. PLoS One 2012;7(10):e47684. [PubMed: 23133518]
34. Wonderlick JS, Ziegler DA, Hosseini-Varnamkhasti P, et al. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. Neuroimage 2009;44(4):1324–1333. [PubMed: 19038349]

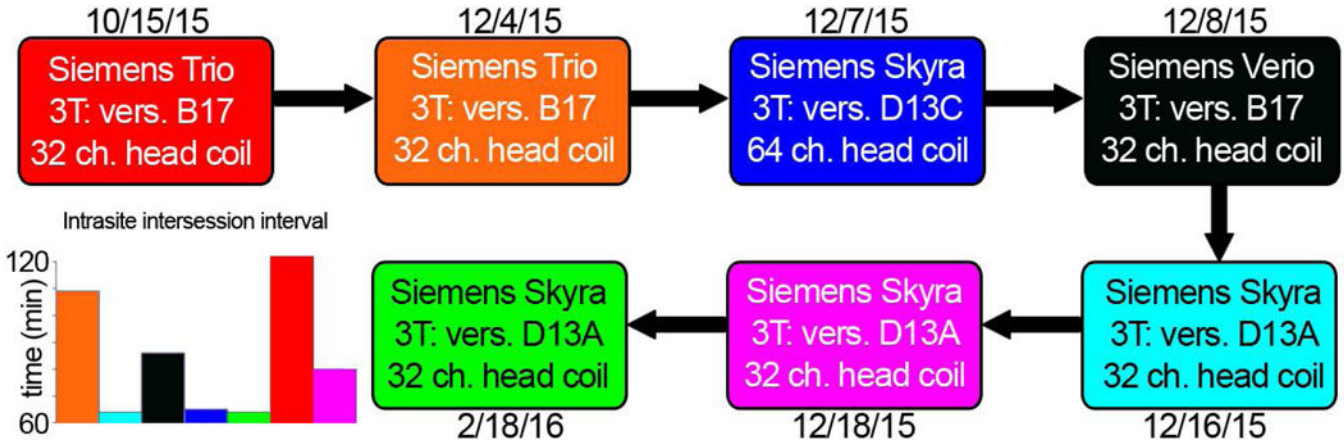

Figure 1.

Timeline, hardware and software details for data acquisition at each site. Sites are anonymized for reporting of all results, but are color coded the same throughout all figures. The bar plot illustrates the time between sessions at each site.

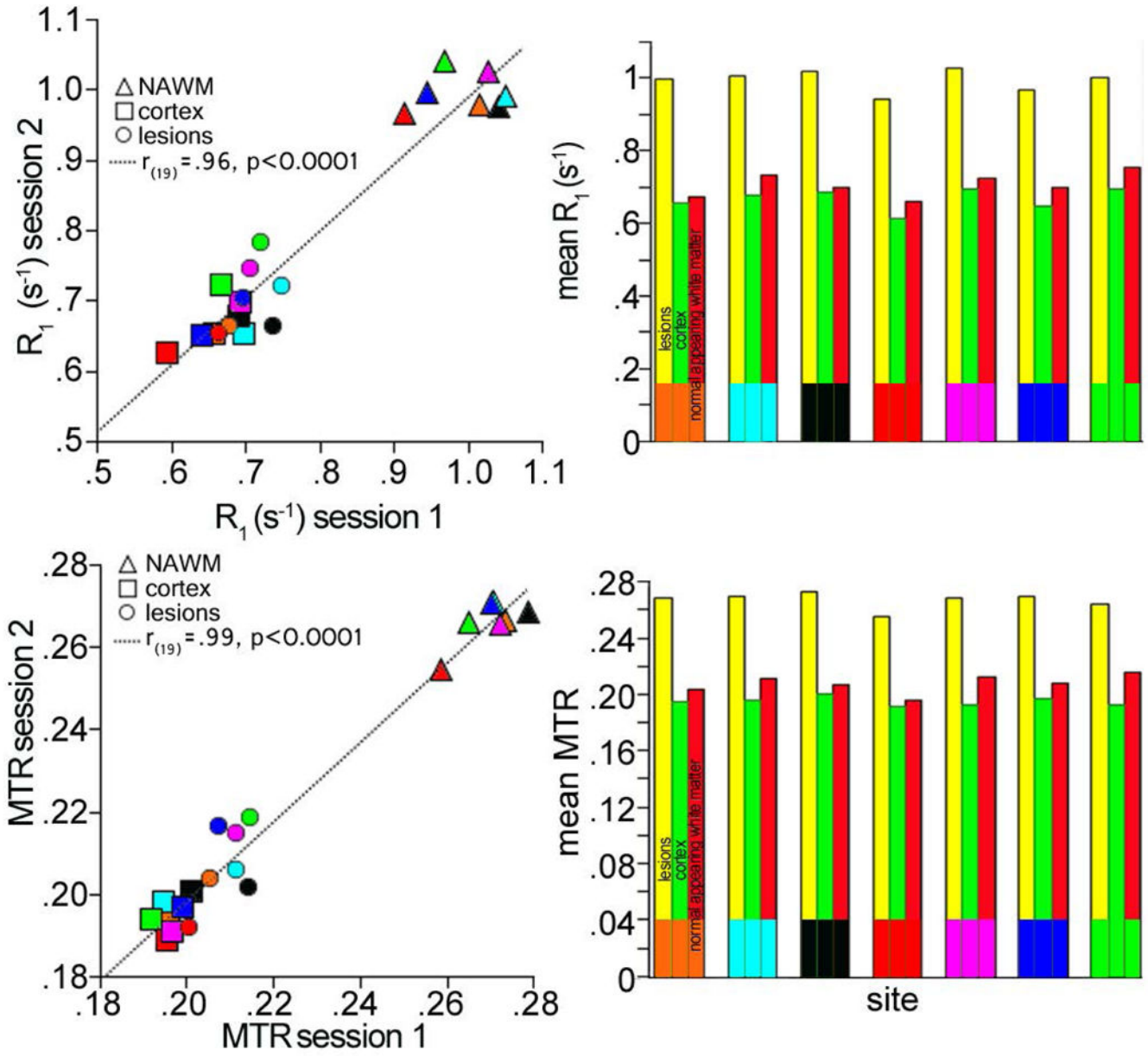

Figure 2.

MTR and R_1 image results. Left. Between session agreement for all sites for MTR (bottom) and R_1 (top) for each tissue class. Right. Between site agreement for all tissue classes for mean MTR (bottom) and R_1 (top).

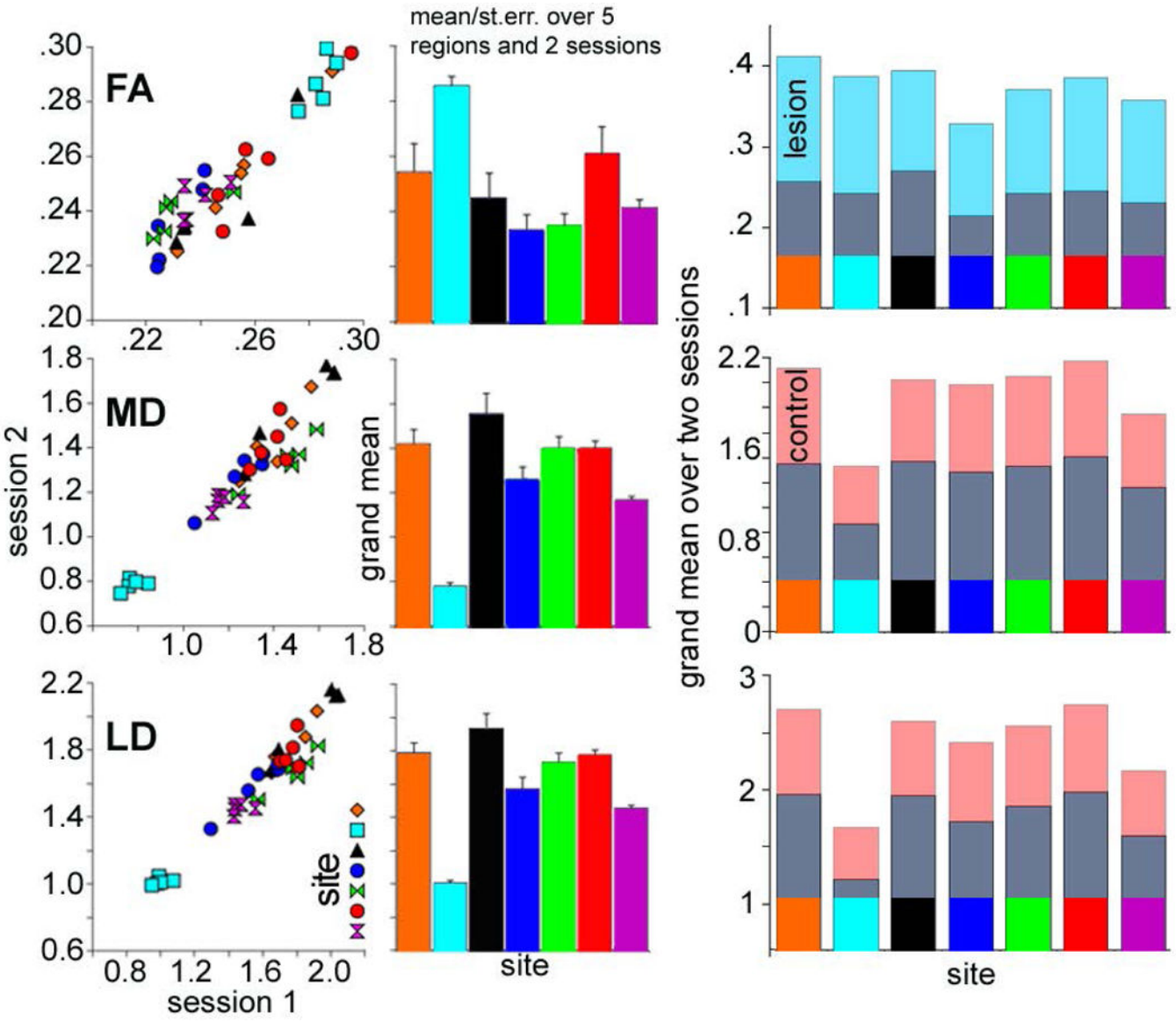

Figure 3.

Diffusion-weighted image results. Left and middle. Measurements of FA, MD, and the first eigenvalue (LD) over NAWM ROIs, grand site mean shown in bar graphs. Right. Mean over session within site for tensor measurements in lesion and NAWM.

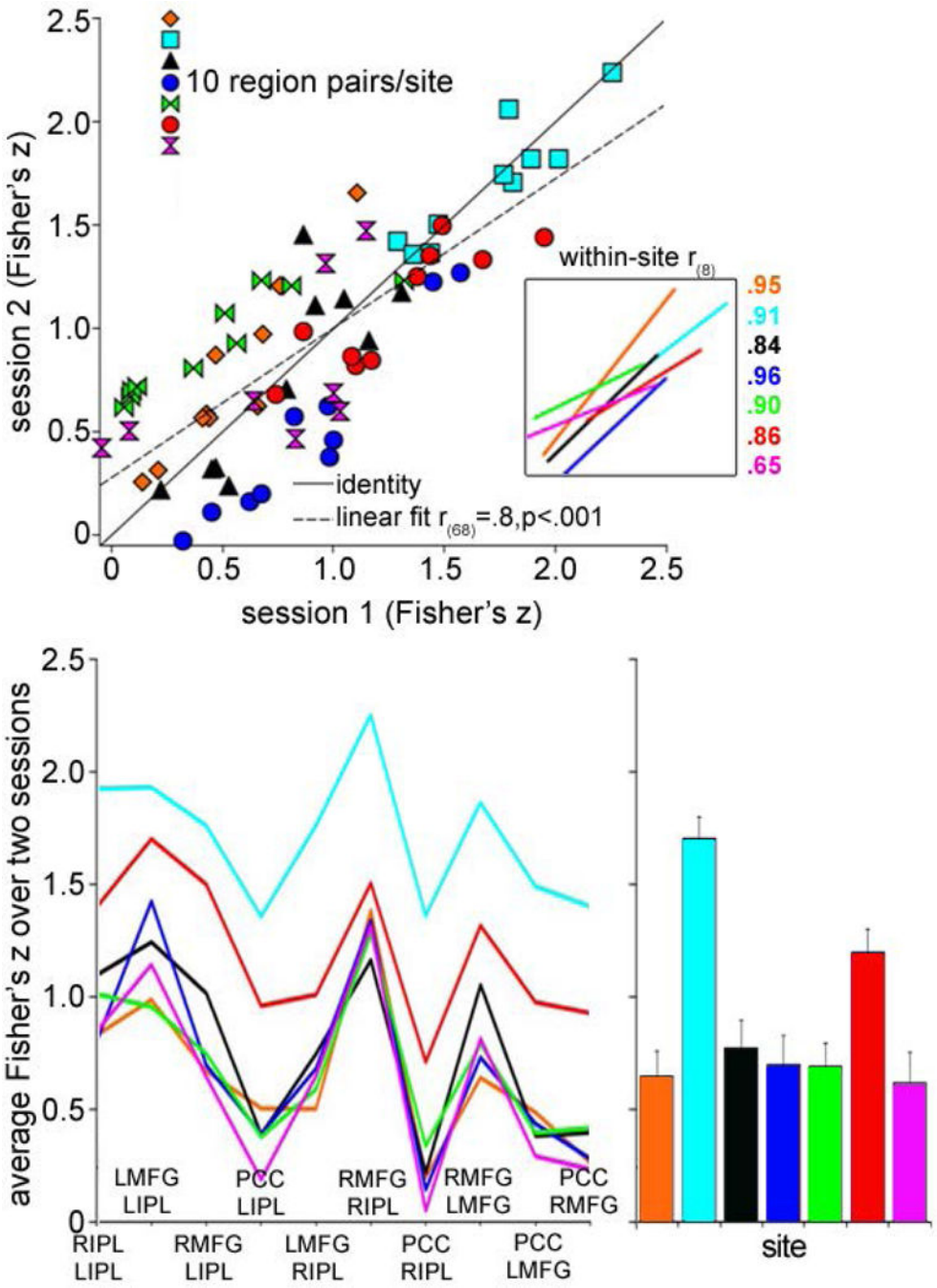

Figure 4.

rsFMRI image results. Top. Scatter plot depicting session-wise agreement for all sites (colors) and for each region pair (10 pairs), identity is depicted as a solid line, linear regression shown as a dotted line, and regression lines are drawn for each site (inset). Bottom. Lines drawn for each site across 10 region pairs (left), grand mean values between sites depicted in bar graph (right).

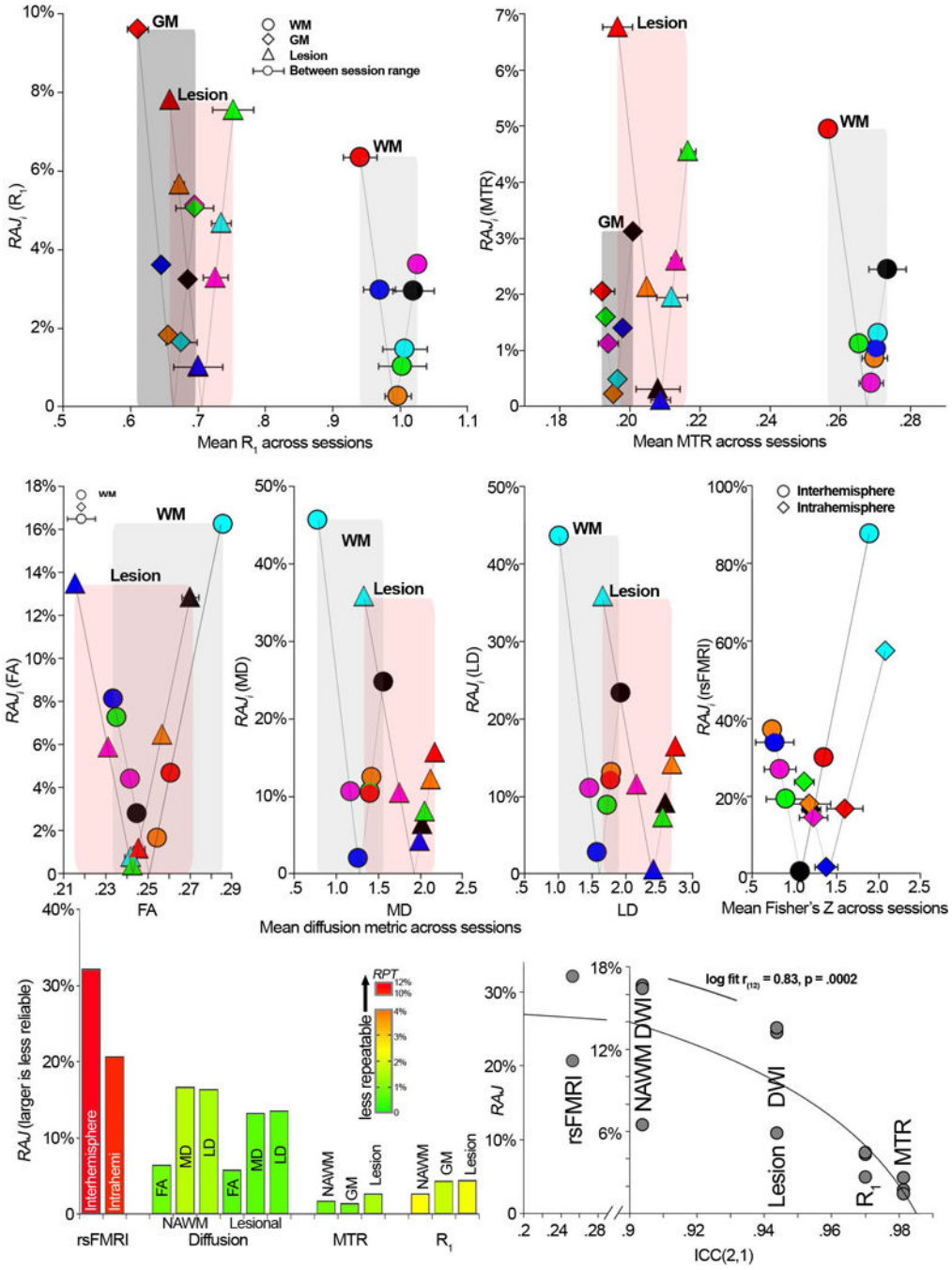

Figure 5.

Repeatability and reliability of each MRI metric. Top two rows. RAJ plots of R_1, MTR, DWI, and rsFMRI measurements, respectively. Interpretation of various elements of RAJ plots: a) taller tissue boxes indicate more variable measurement across site, b) proximity of a marker to the abscissa is proportional to that site's reliability, c) the magnitude of the horizontal range bars is inversely proportional to that site's repeatability, d) the horizontal overlap of tissue boxes is inversely proportional to the differentiability of tissue types using the method, e) the point on the abscissa of the traced "V" is the grand mean of that

measurement. Bottom left. RAJ over all methods; bars are colored by RPT. Bottom right. The inverse relationship of ICC(2,1) and RAJ.

J Magn Reson Imaging. Author manuscript; available in PMC 2020 September 01.

Table 1.

Repeatability (within site) and reliability (across sites) of measurements across all techniques

	Image quality metrics[a]	*Intrasite repeatability*[b]			*Intersite reliability*[c]				
Technique	**SNR(SD)/tSNR(SD)**[c]	**Δmean(SD)**	**Δrange^**	**r[p]**	**RPT**	**mean(SD)**	**range^**	**ICC$_{(2,1)}$**	**RAJ**
MTR	43.8(13.67)	-	-		-	-	-	.981	-
NAWM	-	.004(.004)	.011	.99 [<.001]	0.8%	.268(.006)	.017		1.7%
cortex	-	.003(.002)	.006		0.8%	.196(.003)	.009	-	1.4%
lesion	-	.006(.004)	.011		1.5%	.209(.007)	.020		2.6%
R_1	42.9(1.97)*	-	-		-	-	-	.970	-
NAWM	-	.048(.024)	.067	.96 [<.001]	2.4%	.995(.030)	.085		2.7%
cortex	-	.023(.021)	.052		1.8%	.666(.031)	.084	-	4.3%
lesion	-	.033(.026)	.067		2.3%	.706(.034)	.093		4.4%
DWI	3.55(.22)	-	-		-	-	-	.914	-
FA_{NAWM} MD_{NAWM}	-	.002(.001)	.003		0.8%	.251(.018)	.052		6.5%
		.021(.020)	.053		1.6%	1.29(.254)	.776	.904	17%
LD_{NAWM} FA_{lesion}	-	.022(.020)	.052	.99 [<.001]	1.4%	1.61(.307)	.924		16%
		.002(.001)	.004		0.7%	.243(.018)	.055		5.8%
MD_{lesion} LD_{lesion}	-	.013(.014)	.038		0.7%	1.92(.293)	.846	.944	13%
		.018(.016)	.047		0.7%	2.41(.378)	1.08		14%
rsFMRI[‡]	5.40(1.12) / 279(118)	-	-		-	-	-	.402	-
Interhemisphere	-	.267(.173)	.514	.80 [<.001]	11.5%	1.08(.417)	1.15	.253	32%
Intrahemisphere	-	.336(.177)	.560		10.7%	1.40(.343)	.969		21%

[a] Image quality metrics are collapsed across all measurements, sessions, and sites. Some metrics are not available for a given technique (i.e. tSNR for R_1 measurements); sub-metrics were not calculated and are displayed as "-". Detailed descriptions of each calculation can be found in "Methods" and "Results".

[b] Measurements (mean, range, standard deviation [SD]) are the absolute value of between session *differences* (Δ). Correlation values and their significance (r[p]) are across all measurements and sites between two sessions (i.e. Figure 5D).

[c] Measurements (mean, range, standard deviation [SD]) are collapsed within site *across seven sites*.

^ Range is calculated as [maximum − minimum].

* Note that SNR is variable over flip angle in this experiment (Supplementary Figure 1). Reported is the grand mean and standard deviation over all flip angles.

‡ All measurements are given in units of Fisher's z between ROIs except for image quality metrics. Interhemisphere connectivity refers to average correlations between left and right MFG, and left and right IPL; intrahemisphere connectivity refers to correlations between left IPL and left MFG, and right IPL and right MFG. r[p] is over all correlational measurements as is the first ICC; the second ICC is over correlational measurements averaged in inter- and intrahemispheric metrics.