Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2010, Article ID 618747, 11 pages
doi:10.1155/2010/618747

Research Article

Statistical Evaluations of the Reproducibility and Reliability of 3-Tesla High Resolution Magnetization Transfer Brain Images: A Pilot Study on Healthy Subjects

Kelly H. Zou,[1] Hongyan Du,[2] Shawn Sidharthan,[2] Lisa M. DeTora,[3] Yunmei Chen,[4] Ann B. Ragin,[5] Robert R. Edelman,[2,6] and Ying Wu[2,6]

[1] *Pfizer Inc., New York, NY, USA*
[2] *NorthShore University HealthSystem, Evanston, IL, USA*
[3] *Albany Medical College, Albany, NY, USA*
[4] *University of Florida, Florida, FL, USA*
[5] *Northwestern University, Chicago, IL, USA*
[6] *University of Chicago, Chicago, IL, USA*

Correspondence should be addressed to Ying Wu, ywu@northshore.org

Received 29 September 2009; Accepted 4 December 2009

Academic Editor: Shan Zhao

Magnetization transfer imaging (MT) may have considerable promise for early detection and monitoring of subtle brain changes before they are apparent on conventional magnetic resonance images. At 3 Tesla (T), MT affords higher resolution and increased tissue contrast associated with macromolecules. The reliability and reproducibility of a new high-resolution MT strategy were assessed in brain images acquired from 9 healthy subjects. Repeated measures were taken for 12 brain regions of interest (ROIs): genu, splenium, and the left and right hemispheres of the hippocampus, caudate, putamen, thalamus, and cerebral white matter. Spearman's correlation coefficient, coefficient of variation, and intraclass correlation coefficient (ICC) were computed. Multivariate mixed-effects regression models were used to fit the mean ROI values and to test the significance of the effects due to region, subject, observer, time, and manual repetition. A sensitivity analysis of various model specifications and the corresponding ICCs was conducted. Our statistical methods may be generalized to many similar evaluative studies of the reliability and reproducibility of various imaging modalities.

1. Introduction

Magnetization transfer (MT) imaging is a quantitative approach for detecting subtle or occult abnormalities in brain tissue. In previous studies, the Magnetization Transfer Ratio (MTR), an index of MT imaging, was sensitive to brain changes in patients with mild cognitive impairment, an Alzheimer's disease prodrome [1, 2], to new lesions in patients with multiple sclerosis, [3] and to changes associated with progression in chronic neurological disorders [4]. The higher magnetic field strength afforded by 3T allows MT image resolution to be augmented compared with conventional MT acquisition at 1.5T [5–7]. We developed a high resolution MT technique to detect subtle changes in anatomically small, functionally eloquent brain structures. The increased field strength affords whole-brain coverage with considerably thinner slices, potentially reducing partial volume artifacts. However, even among healthy subjects, numerous factors may introduce variability in measures derived from magnetic resonance (MR) data, such as static field B_0 signal dropout and RF nonuniformity. Measurement variation may be introduced by scan repetitions, repositioning at different time points, and image post-processing. Moreover, 3T may be susceptible to variation associated with increased field strength [8]. Such variability may pose limitations when conducting clinical comparisons to differentiate normal and diseased brains or in developing statistically predictive algorithms.

To validate high resolution MT for detecting early disease or for monitoring progression in chronic neurological

disease, it is necessary to collect information on normative values and to evaluate the reliability and reproducibility of the measurements when measured across time in healthy controls. This investigation evaluated observer-agreement of high-resolution MT measurements determined from repeated brain scans of 9 healthy volunteers. We postulated that MT values would remain stable during the one month study interval. We evaluated the reliability and reproducibility of the high resolution MT measurements in 12 brain regions of interest (ROIs), applied statistical measures to the data and used complex multivariate mixed-effects models to test the statistical significance of several effects due to region, subject, observer, time, and manual repetition.

2. Materials and Methods

2.1. Study Subjects. The study was approved by the IRB at the North Shore University Health System, and conducted following the ethical principles outlined in the Declaration of Helsinki. Eleven healthy adult volunteers were randomly selected from a database maintained at the Center for Advanced Imaging, Radiology Department, NorthShore University Health System provided written informed consent and evaluated for eligibility criteria. To protect the subjects' confidentiality, all data were de-identified and handled according to the guidelines specified by the Health Insurance Portability and Accountability Act (HIPAA) in the USA.

2.2. Image Acquisition. Brain images were acquired using a 3T General Electric (GE) HDx system (Waukesha, WI, USA). Each volunteer was scanned twice in a randomly-selected time interval between 1 to 4 weeks. Methods for reducing random errors in image acquisition included the use of a body-coil for excitation to control B1 non-uniformities and an 8-channel quadrature receive-only coil [9]. MT pulses with (M_s) and without saturation (M_0) were applied at an offset frequency from water resonance. To accelerate the scan for whole-brain coverage, while maintaining thin slices, the image protocol was optimized based on 3T using 3D SPGR [5]. The Gaussian Sinc MT pulse was applied in 8 ms at a 1200 HZ offset. The stability of the scanner and set-up procedure were addressed with a fixed set of parameters per subject. MT pulse was based on a three-dimensional spoiled gradient recalled (3D SPGR) acquisition. The image protocol included the following parameters: TR 34 to 35 ms, TE 4 to 8 ms, imaging FA 5°, bandwidth 15.6 kHz, 0.75 NEX, phase FOV 0.75, voxel dimensions $0.9 \times 0.9 \times 0.9 \sim 1.3\,\text{mm}^3$. The whole brain was covered in 90 to 140 slices with acquisition time ranging from 7 minutes 40 seconds to 10 minutes 20 seconds using a partial k-space acquisition.

2.3. Image Analysis. MTR maps were generated off-line on a General Electric AW Workstation (General Electric, Milwaukee, WI, USA) using the standard equation:

$$\text{MTR} = \frac{M_0 - M_S}{M_0} \times 100\%, \tag{1}$$

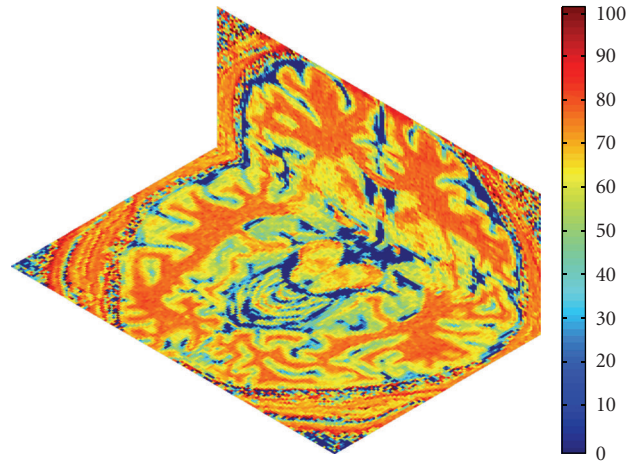


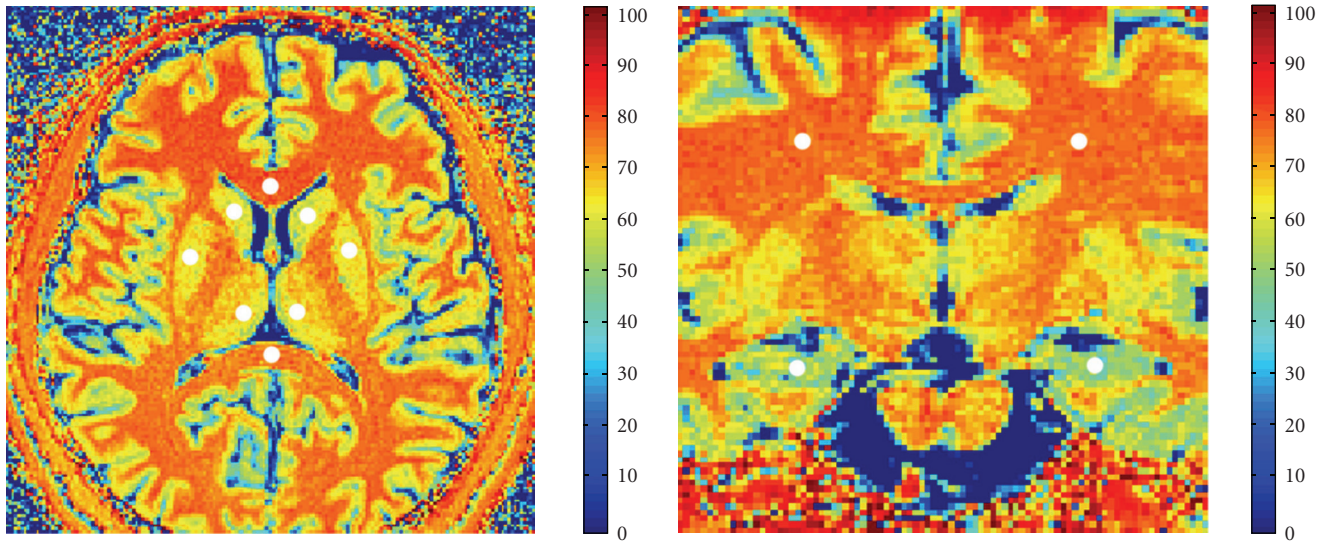
Figure 1: High resolution three-dimensional MTR map displayed both for the original view of the Axial plane and the reconstructed view of the Coronal plan. The MTR maps have excellent tissue conspicuity and high image resolution in all three dimensions.

where M_S and M_0 were the signal intensities in a given voxel obtained, with and without the MT saturation pulse, respectively. MTR maps generated based on the high resolution MT are demonstrated in Figure 1. The 12 ROIs were: genu, splenium, left and right hemispheres of the hippocampus, caudate, putamen, thalamus, and cerebral white matter. Figure 2 illustrated the 12 ROIs that were investigated. Each ROI was sized approximately 30 to 43 mm^2 and manually and independently placed by Observers 1 and 2 (Authors S.S. and Y.W.) following procedures in classical and standard agreement studies [10]. After an initial consensus decision was drawn regarding the sizes and locations of the 12 ROIs, the observers performed manual segmentations of the ROI independently on each set of images. This ROI placement procedure was repeated by each observer in the following week.

MTR values were extracted using the manually-defined ROIs with the combinations of observer, time point, and repetition (Table 1). The mean and SDs of the ROI values were calculated. Meta-data were stored in a SAS 9.1 (SAS, Cary, NC, USA) dataset, with individual volunteer identification numbers withheld and replaced by a sequence of 1 to 9 for each subject.

2.4. Statistical Methods. Statistical analyses were performed using SAS 9.1 (SAS Institute, Cary, NC, USA; http://www.sas.com). The SAS analytic procedures conducted included "Proc Univariate," "Proc Means," "Proc Corr," and "Proc Mixed." Bar diagrams were constructed using Microsoft Excel (http://www.microsoft.com). Age and gender were not controlled for in analyses.

2.4.1. Descriptive Statistics. Let $Y = Y_{ijklm}$ having the indices described in Table 1 be a random variable representing the mean ROI value. For the mth ROI, we first computed



(a) (b)

Figure 2: The Axial (a) and Coronal (b) views of high resolution MTR maps. Twelve brain ROIs are illustrated (white dots).

Table 1: The random or fixed effects in the data structure for the repeated measures MT study.

Outcome Variable Y_{ijkln}	Effect in the Variance-Component Analysis	Type of Effect	Mathematical Symbol	Index	Maximum of the Index
Mean ROI Value via Manual Segmentations	Subject	Random	S_i	$i = i,\ldots,I$	$I = 9$
	Observer	Fixed or Random	O_j	$j = 1,\ldots,J$	$J = 2$
	Time Point	Fixed or Random	T_k	$k = 1,\ldots,K$	$K = 2$
	Repetition	Fixed or Random	R_l	$l = 1,\ldots,L$	$L = 2$
	Region of Interest	Fixed	K_m	$m = 1,\ldots,M$	$M = 12$
	Interaction Terms	Generally Mixed	$\{S_i; O_j; T_k; R_l; K_m\}$	$\{i; j; k; l; n\}$	Based on the Appropriate Model Specification

the sample mean and standard deviation of all mean ROI values:

$$\widehat{\text{Mean}}(Y_m) = \overline{Y_{\bullet\bullet\bullet\bullet m}} = \frac{1}{N_m}\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{j=1}^{2}\sum_{i=1}^{9} Y_{ijklm},$$

$$\widehat{\text{SD}}(Y_m) = \left\{\widehat{\text{Var}}(Y_m)\right\}^{1/2} = \left\{\frac{1}{N_m - 1}\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{j=1}^{2}\sum_{i=1}^{9}\left(Y_{ijklm} - \overline{Y_{\bullet\bullet\bullet\bullet m}}\right)^2\right\}^{1/2}, \quad (2)$$

where $N_m = I \times J \times K \times L = 9 \times 2^3 = 72$ measurements and the operator "$\bullet$" means the marginal sum over the particular index.

The 95-percentile normality range was approximately within the following interval, with the following lower and upper bounds:

$$\left(\widehat{\text{Mean}}(Y_m) - 2 \times \widehat{\text{SD}}(Y_m),\ \widehat{\text{Mean}}(Y_m) + 2 \times \widehat{\text{SD}}(Y_m)\right). \quad (3)$$

The term "normality range" as used in Europe, could be arbitrarily-defined according to the number of standard deviations away from the mean [11]. Thus, it should not be viewed as the range of the entire dataset, but rather an interval useful for estimating the population value by one or several standard deviations away from the mean. Here the critical value of 2 was chosen as recommended by Bland and Altman [12].

Additionally, we justified using a Student's t-distribution with $N_m - 1 = 71$ degrees of freedom. For any tail probability of $\alpha/2$ (e.g., 0.025 for a 95-percent normality range), we

used the quantile of the corresponding to particular t-distribution, such that

$$t_{N_m-1}^{-1}\left(1-\frac{\alpha}{2}\right)=t_{71}^{-1}(0.975)=1.994, \tag{4}$$

This value happened to be close to the recommended multiplier of 2. Therefore, we rounded it to 2 in (3) for convenience.

2.4.2. Concordance Using Spearman's Rank Coefficient Coefficients. We first explored and measured the concordance between the various measurements fully nonparametrically via Spearman's rank correlation coefficient. Suppose that we correlated the ROI values by Observers $j = 1$ and $j' = 2$, then denoted the marginal ranks, $R_{ijklm} = \text{rank}_i(Y_{ijklm})$ and $R_{ij'klm} = \text{rank}_{i'}(Y_{i'jklm})$, respectively, for all $j \neq j'$ with $j = 1$ and $j' = 2$. The sample version of Pearson's product-moment correlation coefficient between the ranks of the data was equivalent to Spearman's rank correlation coefficient [13]:

$$\begin{aligned}\widehat{\text{Cor}}\left(r_{ijklm}, r_{ij'klm}\right) &= \frac{(N_m/2)\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}(R_{i1klm}R_{i2klm})-\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}R_{i1klm}\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}R_{i2klm}}{\left\{(N_m/2)\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}R_{i1klm}^2-\mathcal{A}\right\}^{1/2}\left\{(N_m/2)\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}R_{i2klm}^2-\mathcal{B}\right\}^{1/2}},\\ &= \frac{\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}(R_{i1klm}R_{i2klm})-(N_m/2)\overline{R_{i1klm}R_{i2klm}}}{(N_m/2-1)\text{SD}(R_{i1klm})\text{SD}(R_{i2klm})}.\end{aligned} \tag{5}$$

where $\mathcal{A}$ denotes $\left(\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}R_{i1klm}\right)^2$ and $\mathcal{B}$ denotes $\left(\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}R_{i2klm}\right)^2$.

Assuming that there was no presence of any ties since the ROI values were of continuous random variables, the Spearman's rank correlation coefficient between Observers j and j' was

$$\text{Corr}\left(r_{ijklm}, r_{ij'klm}\right)=1-\frac{6\sum_{l=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{9}D_{i\bullet klm}^2}{(N_m/2)(N_m^2/4-1)}, \tag{6}$$

where the difference of an arbitrary pair of marginal ranks for Observer j and j' was denoted by $D_{i\bullet klm} = R_{ijklm} - R_{ij'klm}$, for all $j \neq j'$. Consequently, all of the raw mean ROI values were converted to their marginal ranks and the differences between the ranks of each observation on the two variables were computed. Spearman's rank correlation coefficient was also computed for the ROI values between any two different time points $k = 1$ and $k' = 2$.

The strength of the concordance and the benchmark values have been discussed [14]. Bar diagrams were made to display the Spearman's rank correlation coefficients between observers or time points for each ROI.

2.4.3. Reproducibility Using Coefficients of Variations. We used the normalized measure of dispersion of a distribution to evaluate the reproducibility of the measurement [15]. The measure was the coefficient of variation (CV), defined as the ratio of the SD to the mean.

$$\widehat{\text{CV}}(Y_m)=\frac{\widehat{\text{SD}}(Y_m)}{\widehat{\text{Mean}}(Y_m)}, \tag{7}$$

where both the numerator (i.e., sample SD) and the denominators (i.e., sample mean) in the above expression for CV are provided in (2). Skewed data, such as those generated by an exponential distribution for which the underlying population mean and standard deviation would be equal, and thus the CV became 1. Hence, CV < 1 would generally represent low variability, and CV > 1 would represent high variability. As in (4) and (6), further stratified computations of CV for different observers, time point, or repetitions were achieved using formulae similar to (7).

2.4.4. Normality and Significance Tests for the Effects via a Multivariate Regression Analysis. As overall variability was likely a result of the effects illustrated in Table 1. We employed a multivariate mixed-effects regression analysis to direct model the ROI values.

A variance-component approach has advantages over many stratified analyses, especially studying studies with a limited sample size. Here, because of the novel imaging modality using MT and 3T acquisitions with labor-intensive manual segmentation procedures, large number of subjects would not have been feasible. To conduct an analysis of variance (ANOVA) based on the various effects, a distributional assumption of normality was necessary and convenient. Therefore, we conducted marginal normality tests using the Shapiro-Wilk test [16]. We would demonstrate (see Section 3.4) that the normality assumption was generally satisfactory.

Thus, we could then consider adopting a linear random-effects model with all pair-wise interactions, in addition to a third-order interaction term:

$$\begin{aligned}Y_{ijklm} &= \mu_m + S_i + O_j + T_k + R_l\\ &\quad + S_i\times O_j + S_i\times T_k + S_i\times R_j + O_j\times T_k\\ &\quad + O_j\times R_l + T_k\times R_j + O_j\times T_k\times R_l + \varepsilon_{ijklm},\\ &\quad \forall i = 1,\ldots,9,\ j = 1,2,\ k = 1,2,\ l = 1,2.\end{aligned} \tag{8}$$

The effects represented the following: μ_m as intercept, S_i as subjects, O_i as observers, T_i as time points, R_i as repetitions,

and ε_{ijklm} as the error team. A random-effects model assumed that each of the effects would have independent normal distributions with mean and variance.

If normality had failed and because the data were mean ROI values that were positively-valued, we would recommend a Box-Cox transformation, $h(Y_{ijklm}, \lambda)$, of the outcome variable with an optimal power coefficient λ [17–19]. Note that the log-normal becomes a special case when the power coefficient $\lambda = 0$. This normality transformation is given by:

$$Y'_{ijklm} = h\left(Y_{ijklm}, \lambda\right) = \begin{cases} \dfrac{{Y_{ijklm}}^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log\left({Y_{ijklm}}^{\lambda}\right), & \lambda = 0 \end{cases} \quad (9)$$
$$\forall i = 1, \ldots, 9,\ j = 1, 2,\ k = 1, 2,\ l = 1, 2.$$

A profile log-likelihood, llik of λ given the observations y_{ijklm}, would be maximized to estimate an optimal Box-Cox transformation via a nonlinear minimization routine, where the log-likelihood was

$$\begin{aligned} &\text{llik}\left(\lambda \mid y_{ijklm}\right) \\ &= -N_m \log\left\{SD\left(y'_{ijklm}\right)\right\} + (\lambda - 1)\left\{\sum_{i=1}^{N_m} \log\left(y'_{ijklm}\right)\right\} + c, \end{aligned} \quad (10)$$

where c was a constant free of the power coefficient to be optimized.

Due to the limited number of subjects, however, even with an optimal normality transformation, over-fitting and non-convergence might be issues. Alternatively, we could regard all of the observers, time points, and repetitions as fixed and specify a mixed-effects model. The significances of the sources of variability were tested via a restricted maximum likelihood (REML) approach. For our multivariate analysis, the significance threshold for two-tailed P-values was set if $P \leq .05$.

2.4.5. Interobserver Reliability Using the ICCs. Stratified by the time points within each ROI, a two-way ANOVA was performed by regarding all of the observers, time points, and repetitions as fixed. We specified a mixed-effects model for simplicity. Due to the complexity of the variance components, we instead adopted a hybrid approach by considering two effects at once. For example, all subjects were segmented by the same observers who were from an entire population of observers. In other words, the subject effect was always assumed to be random, while the remaining effect (e.g., here the observer) was assumed to be fixed. We computed the Case-3 ICCs, accordingly [20].

We simplified our notations by only keeping the indices for the subject and observer effects of interest. We decomposed the data as follows:

Table 2: Various strengths of correlation coefficients as a measure of concordance.

Absolute Value of the Correlation Coefficient	Strength of the Concordance Between Samples
0.0	No
0.2	Weak
0.5	Moderate
0.8	Strong
1.0	Perfect

Table 3: Two-way ANOVA table for the mixed-effects model.

Source of Variation	Degrees of Freedom	Mean Squares	
(A) Between Subjects	$I - 1$	BSMS	$J\sigma_S^2 + \sigma_E^2$
(B) Within Subjects	$I(J-1)$	WSMS	$\theta_o^2 + J\sigma_{S\times o}^2/(J-1) + \sigma_E^2$
(B.1) Between Observers	$J - 1$	OMS	$I\theta_o^2 + J\sigma_{S\times o}^2/(J-1) + \sigma_E^2$
(B.2) Error	$(I-1)(J-1)$	EMS	$J\sigma_{S\times o}^2/(J-1) + \sigma_E^2$

Note: BSMS: Between Subjects Mean Squares; WSMS: Within Subject Mean Squares; OMS: Observer Mean Squares; EMS: Error Mean Squares.

$$Y_{ij} = \mu + S_i + o_j + S_i \times o_j + \varepsilon_{ij}, \quad \forall i = 1, \ldots, 9,\ j = 1, 2, \quad (11)$$

where the subject effect S_i was assumed to be random in an upper-case letter, which had a normal distribution with mean 0 and variance σ_S^2, for all $i = 1, \ldots, I$ (here $I = 9$); the observer effect o_j was considered to be a fixed effect in a lower-case letter, with the constraint $\sum_{j=1}^{J} o_j = 0$, with the corresponding parameter to the variance being $\theta_o^2 = (1/(J-1))\sum_{j=1}^{J} o_j^2$, for all $j = 1, \ldots, J$ (here $J = 2$); the interaction term between the subject and the observer $S_i \times o_j$ was the degree to which the jth observer departed from his or her usual rating tendencies for the ith subject, which had a normal distribution with a mean of 0 and variance $\sigma_{S\times o}^2$; the errors terms ε_{ij} were assumed to have an independent and identical distribution (iid) normal distribution with a mean of 0 and variance σ_E^2. For the same ith subject, the effects are further assumed to be subjected to the constraint $\sum_{j=1}^{J} (S \times o)_{ij} = 0$ over all of the observers. The corresponding two-way ANOVA table was listed (Table 3).

Shrout and Fleiss gave the true definition of ICC using the variance ratio of the subject variance over the total variance, with its estimated version using the quantities via ANOVA (Table 3) [19]:

$$\begin{aligned} \text{ICC} &= \frac{\sigma_S^2 - \sigma_{S\times o}^2/(J-1)}{\sigma_S^2 + \sigma_{S\times o}^2 + \sigma_E^2}, \\ \widehat{\text{ICC}}(3,1) &= \frac{\text{BSMS} - \text{EMS}}{\text{BSMS} + (J-1)\text{EMS}}. \end{aligned} \quad (12)$$

2.4.6. Intraobserver Reliability Using the ICCs. Similar to the analysis described above, we adopted a hybrid approach by considering two effects at once, with the subject effect always assumed to be random and the time point assumed to be fixed. The associate model was given by

$$Y_{ij} = \mu + S_i + t_k + S_i \times t_k + \varepsilon_{ik}, \quad \forall i = 1, \ldots, 9;\ k = 1, 2. \tag{13}$$

As in (12), the estimated intraobserver agreement and its estimate were provided by:

$$\begin{aligned} \text{ICC} &= \frac{\sigma_S^2 - \sigma_{S\times t}^2/(K-1)}{\sigma_S^2 + \sigma_{S\times t}^2 + \sigma_E^2}, \\ \widehat{\text{ICC}}(3,1) &= \frac{\text{BSMS} - \text{EMS}}{\text{BSMS} + (K-1)\text{EMS}}, \end{aligned} \tag{14}$$

where the interaction term the interaction term between the subject and the time $S_i \times t_k$ had a normal distribution with a mean of 0 and variance $\sigma_{S\times t}^2$.

2.4.7. Sensitivity Analyses of the ICCs under Various Models. We performed a sensitivity analysis by computing 6 different ICC values Shrout and Fleiss previously proposed assumptions for ICCs (Table 4) [18]. A SAS macro, written by Professor Robert Hamer, University of North Carolina School of Medicine, Chapel Hill, NC, USA (http://www.bios.unc.edu/~hamer), was run to perform the various ICC computations.

3. Results

3.1. Descriptive Statistics. Eleven healthy adults provided written informed consent to be evaluated and 9 underwent brain scans. Mean age of participants who received scans was 37.9 ± 14.2 years; 7 participants were men and 2 were women.

The mean ROI values varied across different region (Table 5). The left and right hemispheres tended to yield similar results when the average over these healthy subjects was considered.

3.2. Concordance Using Spearman's Rank Coefficient Coefficients. Spearman's rank correlation coefficients showed that a majority of correlations within each observer was above 0.5, suggesting a moderate to high concordance (Figure 3). Time point 2 tended to yield higher concordance between the observers, which suggested a possible learning effect over time (Figure 4). Due to limited sample sizes in this pilot study, in Figures 3 and 4, we demonstrated the effect of observers by averaging over repetitions by each observer. Similarly, we demonstrated the effect of time points by averaging over repetitions at each time point.

3.3. Reproducibility Using Coefficients of Variations. Overall, CVs ranged from 1.2% in the genu for Observer 2 to 7.0% in the right hippocampus for Observer 1 (Table 6). Since all of the CVs were within 7%, that is, all CVs were less than 10%, the reproducibility was reasonably high.

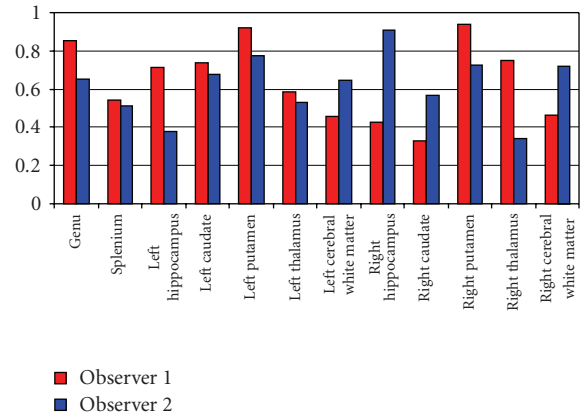


Figure 3: Spearman's rank correlation coefficients between the two different time points for the same observer (red = Observer 1; blue = Observer 2).

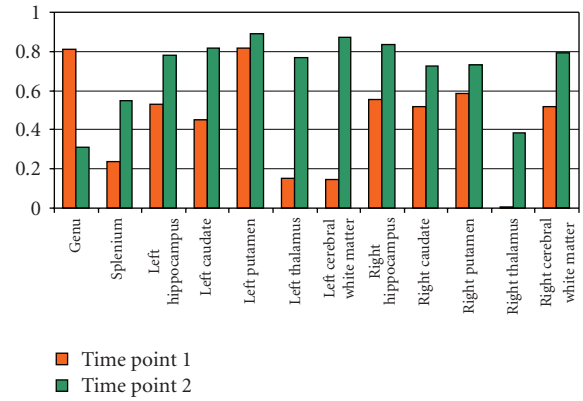


Figure 4: Spearman's rank correlation coefficients between the two different observers for the same time point (orange = Time Point 1; green = Time Point 2).

3.4. Normality and Significance Tests via a Multivariate Analysis. The tests of the normal distribution assumption marginally using the Shapiro-Wilk test indicated that only occasionally (e.g., for left caudate, left and right putamen, and right hippocampus), this assumption was not met (see Table 7). Therefore, it was reasonable to specify linear mixed-effects modeling and two-way ANOVA reported in Sections 3.5 and 3.6.

3.5. Interobserver Reliability Using the ICCs. At time point 1, ICCs were greater than 0.7 in regions of genu, left and right putamen, whereas ICCs were from 0.5 to 0.7 in regions of splenium, left and right hippocampus, left caudate, and right cerebral white matter (Table 8). These results indicated moderate to strong interobserver reliability. In comparison, at time point 2, ICCs were greater than 0.7 in regions of genu, splenium, left and right caudate, putamen and cerebral white matter, and left hippocampus and thalamus, while ICCs were from 0.5 to 0.7 in right hippocampus and thalamus. These results suggested a learning effect over time. However, for some ROIs such as the left cerebral white matter, right

Table 4: Six different ICCs computed via a sensitivity analysis of the modeling choices.

Notation for the ICC Measure	Multivariate Modeling Assumptions
ICC(1,1)	Each subject is rated by multiple observers; the observers are assumed to be randomly assigned to the subjects; all subjects have the same number of observers.
ICC(2,1)	All subjects are rated by the same observers who are assumed to be a random subset of all possible observers.
ICC(3,1)	All subjects are rated by the same observers who are assumed to be the entire population of observers.
ICC(1,2)	Same assumptions as ICC(1,1) but reliability for the mean of 2 ratings.
ICC(2,2)	Same assumptions as ICC(2,1) but reliability for the mean of 2 ratings.
ICC(3,2)	Same assumptions as ICC(3,1) but reliability for the mean of 2 ratings. Assumes additionally there is no subject $\times$ observer interaction.

Table 5: Descriptive statistics and 95-percentile normality range of mean ROI values.

Region of Interest	Descriptive Statistics (Mean $\pm$ SD)	95% Normality Range (Mean $\pm$ 2 $\times$ SD)
Genu	77.0 $\pm$ 1.0	75.0–79.0
Splenium	72.8 $\pm$ 1.5	69.9–75.7
Left Hippocampus	51.5 $\pm$ 2.5	46.6–56.4
Left Caudate	59.5 $\pm$ 2.2	55.2–63.8
Left Putamen	62.0 $\pm$ 2.0	58.1–65.9
Left Thalamus	61.6 $\pm$ 2.3	57.1–66.1
Left Cerebral White Matter	73.2 $\pm$ 1.2	70.8–75.6
Right Hippocampus	52.0 $\pm$ 3.3	45.5–58.5
Right Caudate	61.3 $\pm$ 1.7	58.0–64.6
Right Putamen	62.8 $\pm$ 1.5	59.9–65.7
Right Thalamus	61.1 $\pm$ 2.5	56.2–66.0
Right Cerebral White Matter	73.0 $\pm$ 1.3	70.5–75.5

Note: Results were pooled among all 72 observations within each region of interest. SD: standard deviation.

caudate, right thalamus, ICCs increased from 0.2 (at time point 1) to 0.9 (at time point 2), making it difficult to determine whether this represents a learning effect.

3.6. Intraobserver Reliability Using the ICCs. At each time point, intraobserver agreement was at least 0.5 for a majority of the regions (Table 9).

3.7. Sensitivity Analyses of the ICCs under Various Models. Six different methods for generating ICCs exhibited similar patterns for high vs. low reliability results in different ROIs (Table 10). Thus, reliability appeared to be sensitive to ROI.

4. Conclusions and Discussion

We present mathematical methods for MT brain images using 3-T high resolution. Our image analysis may provide useful pilot information for future investigations. These mathematical and statistical methods may easily be generalized to practical studies with larger sample sizes or to studies of patients with active disease.

We acquired repeat brain measurements based on a high resolution MT imaging protocol at 3T in 9 healthy adults. Our results indicate moderate to high reproducibility, supporting the validity of this method for further studies. Overall, higher intraobserver reliability was observed at the second time point than that at the initial time point, suggesting a possible learning curve effect for both observers. Interobserver reliability was generally lower than intraobserver variability, suggesting a strong observer effect in this comparison, which may be a factor in future investigations using MT imaging.

Our analyses examined different aspects in a typical observer-agreement study, using measures for concordance, reproducibility, reliability, variance-component analysis, and multivariate analysis. In other studies, all or some of such methods may be considered. However, with a simpler study of either several observers, or one observer with several repetitions at different sessions or time points, then these scenarios may only require several of our methods. Only a small sample of healthy volunteers was evaluated in this initial pilot study. Therefore, the generalization of the 95-percentile normality range may be limited with respect to the wider spectrum of brain mechanisms represented in the broader population. For instance, demonstrating summary measures using all possible observer and time point combinations may not lead to meaningful interpretations in all cases. Nevertheless, since the technology is new, this

Table 6: Coefficient of Variation (CV) of the mean Region of Interest values for each observer.

Region of Interest	Observer 1		Observer 2	
	Mean ± SD ($N = 36$)	CV (%)	Mean ± SD ($N = 36$)	CV (%)
Genu	76.9 ± 1.0	1.3	77.1 ± 0.9	1.2
Splenium	73.1 ± 1.4	1.9	72.6 ± 1.5	2.1
Left Hippocampus	51.3 ± 2.4	4.7	51.6 ± 2.7	5.2
Left Caudate	59.7 ± 1.9	3.2	59.3 ± 2.5	4.2
Left Putamen	61.9 ± 2.2	3.6	62.1 ± 1.9	3.1
Left Thalamus	59.9 ± 1.5	2.5	63.3 ± 1.7	2.7
Left Cerebral White Matter	73.3 ± 1.3	1.8	73.1 ± 1.2	1.6
Right Hippocampus	52.5 ± 3.7	7.0	51.5 ± 2.7	5.2
Right Caudate	61.2 ± 1.9	3.1	61.5 ± 1.4	2.3
Right Putamen	62.7 ± 1.5	2.4	62.8 ± 1.5	2.4
Right Thalamus	59.7 ± 1.7	2.8	62.5 ± 2.5	4.0
Right Cerebral White Matter	73.2 ± 1.2	1.6	72.8 ± 1.4	1.9

Note. SD: standard deviation.

Table 7: *P*-value from the Shapiro-Wilk test of marginal normal distributions.

Region of Interest	*P*-value Time Point 1		*P*-value Time Point 2	
	Observer 1	Observer 2	Observer 1	Observer 2
Genu	.29	.17	.70	.36
Splenium	.31	.06	.93	.61
Left Hippocampus	.14	.81	.45	>.99
Left Caudate	.97	<.0001[a]	.49	.92
Left Putamen	.20	.06	.01[a]	.01[a]
Left Thalamus	.86	.51	.63	.13
Left Cerebral White Matter	.82	.43	.21	.02
Right Hippocampus	.54	.86	.01[a]	.58
Right Caudate	.49	.80	.60	.89
Right Putamen	.07	.003[a]	.25	.03[a]
Right Thalamus	.50	.68	.82	.13
Right Cerebral White Matter	.79	.78	.16	.54

[a]Normal distribution was not met.

Table 8: Interobserver reliability between two observers for each time point.

Region of Interest	Inter-Reader ICC Time Point 1	Inter-Reader ICC Time Point 2
Genu	0.866	0.726
Splenium	0.537	0.758
Left Hippocampus	0.693	0.796
Left Caudate	0.580	0.902
Left Putamen	0.869	0.962
Left Thalamus	0.410	0.855
Left Cerebral White Matter	0.378	0.929
Right Hippocampus	0.653	0.656
Right Caudate	0.209	0.872
Right Putamen	0.725	0.882
Right Thalamus	0.264	0.572
Right Cerebral White Matter	0.637	0.896

Table 9: Intraobserver reliability within each observer between different repetitions.

Region of Interest	Intraobserver ICC Observer 1	Intraobserver ICC Observer 2
Genu	0.537	0.555
Splenium	0.598	0.756
Left Hippocampus	0.520	0.596
Left Caudate	0.709	0.362
Left Putamen	0.940	0.784
Left Thalamus	0.479	0.622
Left Cerebral White Matter	0.560	0.703
Right Hippocampus	0.411	0.826
Right Caudate	0.473	0.436
Right Putamen	0.659	0.657
Right Thalamus	0.687	0.308
Right Cerebral White Matter	0.570	0.770

Table 10: Sensitivity analysis of 6 different interobserver ICCs.

Region of Interest	ICC (1,1)	ICC (2,1)	ICC (3, 1)	ICC (1, 2)	ICC (2, 2)	ICC (3, 2)
Interobserver ICC at Time 1						
Genu	0.870	0.879	0.866	0.931	0.935	0.928
Splenium	0.497	0.463	0.537	0.664	0.633	0.699
Left Hippocampus	0.653	0.605	0.693	0.790	0.754	0.819
Left Caudate	0.562	0.542	0.580	0.719	0.703	0.734
Left Putamen	0.871	0.874	0.869	0.931	0.933	0.930
Left Thalamus	−0.015	0.114	0.410	−0.030	0.205	0.581
Left Cerebral White Matter	0.382	0.385	0.378	0.553	0.556	0.549
Right Hippocampus	0.660	0.669	0.653	0.795	0.802	0.790
Right Caudate	0.178	0.180	0.209	0.302	0.306	0.346
Right Putamen	0.725	0.732	0.720	0.840	0.845	0.837
Right Thalamus	−0.092	0.079	0.264	−0.202	0.146	0.417
Right Cerebral White Matter	0.630	0.621	0.637	0.773	0.766	0.779
Interobserver ICC at Time 2						
Genu	0.722	0.715	0.726	0.838	0.834	0.841
Splenium	0.758	0.757	0.758	0.862	0.862	0.863
Left Hippocampus	0.792	0.785	0.796	0.884	0.880	0.886
Left Caudate	0.905	0.909	0.902	0.950	0.952	0.949
Left Putamen	0.961	0.959	0.962	0.980	0.979	0.980
Left Thalamus	0.297	0.239	0.855	0.458	0.385	0.922
Left Cerebral White Matter	0.928	0.926	0.929	0.963	0.962	0.963
Right Hippocampus	0.640	0.620	0.656	0.781	0.765	0.793
Right Caudate	0.876	0.884	0.872	0.934	0.938	0.932
Right Putamen	0.884	0.887	0.882	0.938	0.940	0.937
Right Thalamus	0.419	0.347	0.572	0.591	0.516	0.728
Right Cerebral White Matter	0.889	0.876	0.896	0.941	0.934	0.945

research may provide useful pilot information for future investigations. Moreover, the statistical methods employed and illustrated here may easily be generalized to studies with larger sample sizes and diseased subjects.

Another limitation was that this study aimed to evaluate only the reproducibility and reliability, rather than the accuracy in a more comprehensive validation study. In the absence of a true gold standard, such as one based on digital phantoms where realistic variability may still not be simulated, or on histopathology, improved reliability may not be equated with improved accuracy [21]. Both sensitivity and specificity are of interest. Further research would benefit from a useful algorithm to perhaps statistically and optimally estimate the underlying spatial "ground truth" [22, 23].

Table 11: Sensitivity analysis of 6 different intraobserver ICCs.

Region of Interest	ICC (1,1)	ICC (2,1)	ICC (3, 1)	ICC (1, k)	ICC (2, k)	ICC (3, k)
Intraobserver for Observer 1						
Genu	0.537	0.537	0.537	0.699	0.699	0.699
Splenium	0.590	0.579	0.598	0.742	0.733	0.749
Left Hippocampus	0.531	0.544	0.520	0.694	0.705	0.684
Left Caudate	0.704	0.696	0.709	0.826	0.821	0.830
Left Putamen	0.942	0.946	0.940	0.970	0.972	0.969
Left Thalamus	0.481	0.484	0.479	0.650	0.653	0.647
Left Cerebral White Matter	0.550	0.539	0.560	0.710	0.701	0.718
Right Hippocampus	0.426	0.439	0.411	0.597	0.610	0.582
Right Caudate	0.470	0.467	0.473	0.640	0.637	0.643
Right Putamen	0.657	0.654	0.659	0.793	0.791	0.795
Right Thalamus	0.696	0.711	0.687	0.821	0.831	0.814
Right Cerebral White Matter	0.582	0.596	0.570	0.736	0.747	0.727
Intraobserver ICC for Observer 2						
Genu	0.563	0.572	0.555	0.720	0.728	0.714
Splenium	0.760	0.767	0.756	0.864	0.868	0.861
Left Hippocampus	0.607	0.623	0.596	0.756	0.767	0.747
Left Caudate	0.365	0.367	0.362	0.535	0.537	0.531
Left Putamen	0.790	0.800	0.784	0.883	0.889	0.879
Left Thalamus	0.632	0.645	0.622	0.774	0.784	0.767
Left Cerebral White Matter	0.712	0.726	0.703	0.832	0.841	0.826
Right Hippocampus	0.829	0.835	0.826	0.907	0.910	0.905
Right Caudate	0.432	0.429	0.436	0.603	0.601	0.607
Right Putamen	0.667	0.682	0.657	0.800	0.811	0.793
Right Thalamus	0.298	0.294	0.308	0.459	0.455	0.471
Right Cerebral White Matter	0.777	0.789	0.770	0.875	0.882	0.870

Finally, future research may be directed to evaluating the diagnostic utility of high resolution MT for early detection of Alzheimer's disease, multiple sclerosis or other neurological disorders and for monitoring progression across the clinical course.

Acknowledgments

None of the authors on this study had any conflict of interest. This study was partially supported by research Grants 1R01MH080636-01A2, NorthShore University Health System Pilot Grant EH07-267 and Alzheimer's Drug Discovery Foundation (ISOA 271222). The authors are grateful for the assistance of Fiona Malone and Yuyuan Ouyang. In addition, they acknowledge with thanks for the SAS macro for computing various ICCs, developed by Dr. Robert M. Hamer, Professor of Psychiatry and Research Professor of Biostatistics, University of North Carolina School of Medicine, Chapel Hill, NC, USA. Dr. DeTora is a paid employee of Novartis Vaccines and Diagnostics, Cambridge MA, USA.

References

[1] N. J. Kabani, J. G. Sled, A. Shuper, and H. Chertkow, "Regional magnetization transfer ratio changes in mild cognitive impairment," *Magnetic Resonance in Medicine*, vol. 47, no. 1, pp. 143–148, 2002.

[2] W. M. van der Flier, D. M. J. van den Heuvel, A. W. E. Weverling-Rijnsburger, et al., "Magnetization transfer imaging in normal aging, mild cognitive impairment, and Alzheimer's disease," *Annals of Neurology*, vol. 52, no. 1, pp. 62–67, 2002.

[3] F. Agosta, M. Rovaris, E. Pagani, M. P. Sormani, G. Comi, and M. Filippi, "Magnetization transfer MRI metrics predict the accumulation of disability 8 years later in patients with multiple sclerosis," *Brain*, vol. 129, no. 10, pp. 2620–2627, 2006.

[4] J. T. Chen, D. L. Collins, H. L. Atkins, et al., "Magnetization transfer ratio evolution with demyelination and remyelination in multiple sclerosis lesions," *Annals of Neurology*, vol. 63, no. 2, pp. 254–262, 2008.

[5] M. Cercignani, M. R. Symms, M. Ron, and G. J. Barker, "3D MTR measurement: from 1.5 T to 3.0 T," *NeuroImage*, vol. 31, no. 1, pp. 181–186, 2006.

[6] G. Helms, B. Draganski, R. Frackowiak, J. Ashburner, and N. Weiskopf, "Improved segmentation of deep brain grey matter structures using magnetization transfer (MT) parameter maps," *NeuroImage*, vol. 47, no. 1, pp. 194–198, 2009.

[7] Y. Wu, P. Storey, A. Carrillo, et al., "Whole brain and localized magnetization transfer measurements are associated with cognitive impairment in patients infected with human immunodeficiency virus," *American Journal of Neuroradiology*, vol. 29, no. 1, pp. 140–145, 2008.

[8] R. R. Edelman, "MR imaging of the pancreas: 1.5T versus 3T," *Magnetic Resonance Imaging Clinics of North America*, vol. 15, no. 3, pp. 349–353, 2007.

[9] P. S. Tofts, S. C. A. Steens, M. Cercignani, et al., "Sources of variation in multi-centre brain MTR histogram studies: body-coil transmission eliminates inter-centre differences," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 19, no. 4, pp. 209–222, 2006.

[10] P. Graham, "Modelling covariate effects in observer agreement studies: the case of nominal scale agreement," *Statistics in Medicine*, vol. 14, no. 3, pp. 299–310, 1995.

[11] S. R. Filipović and V. S. Kostić, "Utility of auditory P300 in detection of presenile dementia," *Journal of the Neurological Sciences*, vol. 131, no. 2, pp. 150–155, 1995.

[12] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 1, no. 8476, pp. 307–310, 1986.

[13] T. S. Hettmansperger, *Statistical Inference Based on Ranks*, Krieger, Malabar, Fla, USA, 1991.

[14] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," *Radiology*, vol. 227, no. 3, pp. 617–622, 2003.

[15] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: method and validation," *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 716–724, 1994.

[16] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.

[17] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B*, vol. 26, pp. 211–252, 1964.

[18] K. H. Zou and A. J. O'Malley, "A Bayesian hierarchical nonlinear regression model in receiver operating characteristic analysis of clustered continuous diagnostic data," *Biometrical Journal*, vol. 47, no. 4, pp. 417–427, 2005.

[19] A. J. O'Malley and K. H. Zou, "Bayesian multivariate hierarchical transformation models for ROC analysis," *Statistics in Medicine*, vol. 25, no. 3, pp. 459–479, 2006.

[20] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[21] K. H. Zou, W. M. Wells III, R. Kikinis, and S. K. Warfield, "Three validation metrics for automated probabilistic image segmentation of brain tumours," *Statistics in Medicine*, vol. 23, no. 8, pp. 1259–1282, 2004.

[22] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[23] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation by estimating rater bias and variance," *Philosophical Transactions of the Royal Society A*, vol. 366, no. 1874, pp. 2361–2375, 2008.