Test-Retest Reliability and Reproducibility of Short-Echo-Time Spectroscopic Imaging of Human Brain at 3T

Charles Gasparovic,[1,2*] Edward J. Bedrick,[3] Andrew R. Mayer,[1,4] Ronald A. Yeo,[2] HongJi Chen,[5] Eswar Damaraju,[1] Vince D. Calhoun,[1,6] and Rex E. Jung[1,7]

A ^{1}H magnetic resonance spectroscopic imaging study at 3T and short echo time was conducted to evaluate both the reproducibility, as measured by the interscan coefficient of variation (CV), and test-retest reliability, as measured by the intraclass correlation coefficient (ICC), of measurements of glutamate (Glu), combined glutamate and glutamine (Glx), myo-inositol (mI), *N*-acetylaspartate, creatine, and choline in 21 healthy subjects. The effect of partial volume correction on these measures and the relationship of reproducibility and reliability to data quality were also examined. A ^{1}H magnetic resonance spectroscopic imaging slice was prescribed above the lateral ventricles and single repeat scans were performed within 30 min to minimize physiologic variability. Interscan CVs based on all the voxels varied from 0.05 to 0.07 for *N*-acetylaspartate, creatine, and choline to 0.10–0.13 for mI, Glu, and Glx. Findings on the reproducibility of gray and white matter estimates of *N*-acetylaspartate, creatine, and choline are consistent with previous studies using longer echo times, with CVs in the range of 0.02–0.04 and ICC in the range of 0.65–0.90. CVs for Glu, Glx, and mI are much lower than reported in previous studies at 1.5T, while white matter mI (CV = 0.04, ICC = 0.93) and gray matter Glx (CV = 0.04, ICC = 0.68) demonstrated both high reproducibility and test-retest reliability. Magn Reson Med 66:324–332, 2011. © 2011 Wiley-Liss, Inc.

Key words: magnetic resonance spectroscopy; brain; reproducibility

INTRODUCTION

Proton magnetic spectroscopic imaging (^{1}H-MRSI) is being used increasingly at high field ($\geq$3T) and short echo times (TEs) to measure brain metabolites with multiplet signals that are generally more challenging to resolve at lower fields or longer TEs (1,2). The concentrations of glutamate (Glu), glutamine (Gln), combined Glu and Gln (Glx), or myo-inositol (mI), for example, have been measured with ^{1}H-MRSI at 3T in several recent studies on brain disorders, including multiple sclerosis (3,4), cancer (5,6), and traumatic brain injury (7). Higher field strength improves signal-to-noise and the separation of neighboring signals, even though this is somewhat offset by an increase in line broadening (8). Short TEs, on the other hand, reduce signal loss due to T_2 relaxation and cancelation of J-coupled signals by phase modulation, though the judicial use of the latter effect can be used to improve detectability of selected peaks (9–11). Even with these advantages, of course, the factors that impact the reliability of measuring metabolite signals at lower fields or with longer TEs are present at higher fields and shorter TEs. These include voxel size, magnetic field homogeneity, instrument noise, patient motion, and data processing. Furthermore, if longitudinal studies are undertaken, the accuracy of relocating the spectroscopic region of interest in subsequent scans becomes critical.

Although a number of past studies have examined the reproducibility or reliability of ^{1}H-MRSI (12–20), most have been based on few subjects, some were conducted with nonstandard pulse sequences, and none have been conducted at 3T using a short TE. Furthermore, the majority of these studies were not based on data corrected for partial volume effects, nor has a consistent interscan interval or measure of reproducibility or reliability been applied across studies. Only two studies (13,16) used a standard measure of test-retest measurement reliability, the intraclass correlation coefficient (ICC), while two others (12,14) used an analysis of variance approach to calculate coefficients of variation (CVs) as a measure of reproducibility, dividing the square root of separated variance components (for subject, scan, or voxel) by the mean across all voxels. The remainder of the studies reported more conventional CVs, based on the means and standard deviations of metabolite intensities over similar voxels across subjects or repeated scans. Only two studies calculated CVs based on metabolite intensities corrected for tissue content or brain region (16,20). The fact that the CVs determined from these various approaches vary, even for the most prominent ^{1}H-MRS signals, from a few percent (16,20) to over 10% (14,15,17,18) underscores the general lack of comparability between approaches, experimentally as well as

[1] The Mind Research Network, Albuquerque, New Mexico, USA.

[2] Department of Psychology, University of New Mexico, Albuquerque, New Mexico, USA.

[3] Department of Internal Medicine, University of New Mexico, Health Sciences Center, Albuquerque, New Mexico, USA.

[4] Department of Neurology, University of New Mexico, Health Sciences Center, Albuquerque, New Mexico, USA.

[5] Department of Psychiatry, University of New Mexico, Health Sciences Center, Albuquerque, New Mexico, USA.

[6] College of Electrical and Computer Engineering, University of New Mexico, Albuquerque, New Mexico, USA.

[7] Department of Neurosurgery, University of New Mexico, Health Sciences Center, Albuquerque, New Mexico, USA.

Grant sponsor: The Mind Research Network; Grant number: DOE DE-FG02-99ER62764; Grant sponsor: National Institutes of Health; Grant number: P20 RR021938-01.

*Correspondence to: Charles Gasparovic, Ph.D., The Mind Research Network, Pete & Nancy Domenici Hall, 1101 Yale Blvd. NE, Albuquerque, NM 87106. E-mail: chuck@mrn.org

Received 29 September 2010; revised 7 January 2011; accepted 10 January 2011.

DOI 10.1002/mrm.22858

Published online 28 February 2011 in Wiley Online Library (wileyonlinelibrary.com).

statistically; and, understandably, neuroimaging researchers may be either disheartened or encouraged to use ^{1}H-MRSI in their studies, depending on which reports they read.

In this report we present the results of a study on ^{1}H-MRSI reproducibility and test-retest reliability at 3T and short TE, involving 21 healthy subjects and a standard double spin echo pulse sequence. Our primary goals were to evaluate both the reproducibility and reliability of measurements of Glu, Glx, and mI and other major visible metabolites at 3T in healthy subjects and, more generally, to examine the effect of partial volume correction on these measures. A single ^{1}H-MRSI slice was prescribed above the lateral ventricles and repeat scans were performed within 30 min to minimize any physiologic variability. The data were corrected for tissue composition and relaxation factors as previously described (7,21) and estimates of pure gray and white matter concentrations were estimated by linear regression. Signals from *N*-acetylaspartate (NAA), combined NAA and *N*-acetyl-aspartylglutamate (tNAA), total creatine (Cr) and total choline (Cho), Glu, Glx, and mI, were examined. To compare our findings to previous studies, reproducibility was assessed with CVs calculated from subject, scan, voxel and error variances separated using an analysis of variance method and absolute test-retest reliability was assessed with ICCs. All measures were calculated with and without partial volume correction. Finally, the relationship of reproducibility and reliability to data quality was investigated.

MATERIALS AND METHODS

Subjects

Twenty-one healthy subjects (males = 12, mean age = 24.7 ± 5.9) with no history of neurological or psychiatric disorders were recruited and scanned in accordance with a protocol approved by the Institutional Review Board for human research at the University of New Mexico.

MRI and MRS

MRI and ^{1}H-MRSI experiments were performed on a Siemens 3T Tim Trio scanner. Foam padding and paper tape was used to restrict motion within the scanner. High resolution sagittally prescribed T_1-weighted anatomic images were acquired with a 5-echo multiecho MPRAGE sequence [TE (echo time) = 1.64, 3.5, 5.36, 7.22, 9.08 ms, TR (repetition time) = 2.53 s, TI (inversion time) = 1.2 s, 7° flip angle, number of excitations = 1, slice thickness = 1 mm, field of view = 256 mm, resolution = 256 × 256]. Only the root-mean-square of the five images generated by this sequence was used in subsequent analyses. T_2-weighted images were collected with a fast spin echo sequence [TE = 77.0 ms, TR = 1.55 s, flip angle 152°, number of excitations = 1, slice thickness = 1.5 mm, field of view = 220 mm, matrix = 192 × 192, voxel size = 1.15 × 1.15 × 1.5 mm^3]. The T_2-weighted image was aligned axially, parallel to the anterior–posterior commissure axis as it appeared in the sagittal plane of the T_1-weighted image. The T_2-weighted image was used to prescribe the ^{1}H-MRSI slice.

Each subject was scanned with localizer, T_1-weighted, T_2-weighted, and ^{1}H-MRSI sequences, removed completely from the scanner, placed back in the scanner within 15 minutes, and rescanned once. Careful replication of the prescription of the T_2-weighted and ^{1}H-MRSI sequences was accomplished by visual comparison to the initial images. Only the T_1-weighted image was not repeated. This image was used for tissue segmentation for both sets of data, the results of which were registered to the each of the two T_2-weighted images to compute the tissue fractions in each spectroscopic voxel for either scan (see ^{1}H-MRSI data processing section later).

^{1}H-MRSI was performed with a phase-encoded version of a point-resolved spectroscopy sequence both with and without water presaturation (TE = 40 ms, TR = 1500 ms, slice thickness = 15 mm, field of view = 220 × 220 mm, circular *k*-space sampling (radius = 24), total scan time = 582 s). A TE of 40 ms was chosen to improve detection of the glutamate signal (11). The nominal voxel size was 6.9 × 6.9 × 15 mm^3 (0.71 cm^3) after zero-filling in *k*-space to 32 × 32 samples. Using the width at half maximum of the theoretical point spread function (with circular phase encoding and a Hamming filter with a 0.5 width) as the effective diameter of the voxel, the effective voxel volume is estimated to be 2.4 cm^3. The ^{1}H-MRSI volume of interest was selected with strong saturation bands to reduce chemical shift artifacts and was prescribed with the T_2-weighted image to lie immediately above the lateral ventricles and parallel to the anterior–posterior commissure axis (in-plane T_2-weighted image), and included portions of the cingulate gyrus and the frontal and parietal lobes. To further minimize the chemical shift artifact, the transmitter was set to the frequency of the NAA methyl peak during the acquisition of the metabolite spectra and to the frequency of the water peak during the acquisition of the unsuppressed water spectra. Additionally, the outermost rows and columns of the volume of interest were excluded from analysis. This resulted in a total voxel number of 48–80 analyzed voxels per subject, depending on the head size, for a grand total of 1496 voxels across all 21 subjects.

^{1}H-MRSI data processing: After zero-filling to 32 × 32 points in *k*-space, applying a Hamming filter with a 50% window width, and 2D spatial Fourier transformation, the time domain ^{1}H-MRSI data were analyzed using LCModel (22) from 4.2 to 1.8ppm. The basis set for LCModel was generated using spectrum simulation software, based on the theoretical chemical shifts and coupling constants of 15 metabolites, and provided by the developer of LCModel (S. Provencher). Parameterized macromolecule intensities were included over the fitted spectral region (the LCModel set MM20). The Cramer-Rao lower bounds of the fit to the peak of interest output by LCModel were used as a criterion to exclude poor quality data (>20% for a metabolite of interest) from the final analysis. This resulted in a total of 1496 voxels for NAA, tNAA, Cr, and Cho, 1491 voxels for mI, 1340 voxels for Glu, and 1337 voxels for Glx that were analyzed further. The glutamine signal met the CRLB criterion in less than 50% of all spectra and, therefore, was not

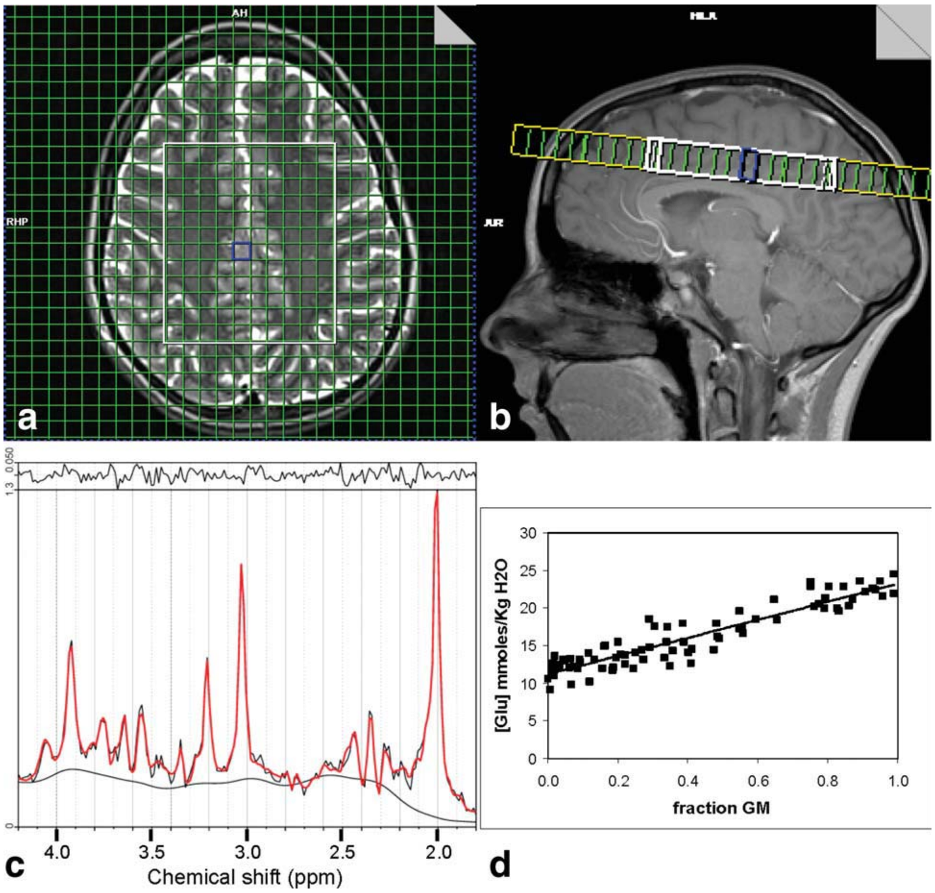

FIG. 1. **a:** Location of ^{1}H-MRSI excitation volume (white rectangle) and measured voxels (green grid within white rectangle) overlaid on T_2-weighted image. **b:** Sagittal view of ^{1}H-MRSI excitation volume. **c:** LCModel fit of representative spectrum from gray matter voxel outlined in blue in images of a and b. **d:** Regression analysis of Glu data from all analyzed voxels. The horizontal axis is the GM fraction of the total tissue (GM + WM) fraction.

examined further in this study, other than as a fraction of the Glx signal. Subsequent processing of the derived metabolite values has been described previously (21). Briefly, concentration values were corrected for partial volume and relaxation effects using gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) maps generated by segmenting the T_1-weighted images with SPM5 (23) and taking into account the variable water densities and relaxation times in each tissue or CSF compartment. In the present study, we used a CSF T_1 value of 3.55 s and a CSF T_2 estimate of 2.47 s based on previous measurements at our site (24). Otherwise, the previously reported T_1, T_2, and water density (WD) values used were as follows: GM: $T_1 = 1.304$ s, $T_2 = 0.093$ s (25), WD = 0.78 (26); WM: $T_1 = 0.660$ s, $T_2 = 0.073$ s (25); WD = 0.65 (26); CSF: WD = 0.97 (26). Estimates of metabolite T_1 and T_2 values at 3T were drawn from Mlynarik et al. (27). The Gln T_1 and T_2 values were assumed to be equal to the Glu values.

Estimates of metabolite concentrations in both GM and WM were generated by linear regression of the metabolite concentration in each against the normalized GM fraction of the voxel (GM fraction divided by the sum of the GM and WM fractions) and extrapolating to a GM fraction of one (pure GM) or zero (pure WM) (see example in Fig. 1).

Statistics

Measurement reproducibility for most ^{1}H-MRS studies has most often been evaluated in terms of measurement variance (28), usually cast in the form of a CV. However, different approaches have been taken to calculate CVs related to reproducibility, and CVs per se do not reflect the capability of a device to obtain the same value for a measurement repeatedly. This latter capability is evaluated with a measure of test-retest reliability, such as some form of the ICC. In the present study, CVs to evaluate ^{1}H-MRSI reproducibility and ICCs to measure absolute test-retest reliability were calculated from subject, voxel, scan, and error variances separated using a straightforward analysis of variance (ANOVA) approach with various multifactorial random effects models, as introduced by others in seminal early reports on ^{1}H-MRSI reproducibility (12,14). This approach utilizes the restricted maximum likelihood (REML) method to ensure non-negative variance terms. The ANOVA model for the data sets that included each voxel's metabolite estimate across all subjects was

$$\mathrm{Y_{ijk}} = \mu + \mathrm{Sub_i} + \mathrm{Vox_{ij}} + \mathrm{Scan_k} + \mathrm{e_{ijk}} \qquad [1]$$

where μ is the mean across all voxels; $\mathrm{Sub_i}$, $\mathrm{Vox_{ij}}$, and $\mathrm{Scan_k}$ are the subject, voxel, and scan random effects, respectively; and $\mathrm{e_{ijk}}$ is the residual or error term. For the calculation of the ICC for individual subject data, this model was reduced to

$$\mathrm{Y_{jk}} = \mu + \mathrm{Vox_j} + \mathrm{Scan_k} + \mathrm{e_{jk}}. \qquad [2]$$

For the data sets with only gray or white matter estimates of metabolite concentrations across all subjects, the model was

Table 1
Reproducibility and Reliability Results

		Variance										
Metabolite	Mean	Subject	%	Scan	%	Voxel	%	Error	%	Total	CV	ICC
NAA												
LCmodel	17.51	0.612	25.3	0.021	0.9	0.842	34.9	0.941	38.9	2.416	0.06	0.60
Corrected	16.45	0.446	10.5	0.019	0.4	2.874	67.8	0.899	21.2	4.238	0.06	0.78
GM	19.54	0.886	76.3	0.008	0.7			0.267	23.0	1.161	0.03	0.76
WM	14.23	0.627	83.4	0.006	0.8			0.119	15.8	0.752	0.02	0.83
tNAA												
LCmodel	19.63	0.640	28.7	0.018	0.8	0.739	33.1	0.836	37.4	2.233	0.05	0.62
Corrected	18.37	0.476	21.7	0.016	0.7	0.936	42.7	0.765	34.9	2.193	0.05	0.64
GM	19.50	0.521	65.1	0.026	3.3			0.253	31.6	0.800	0.03	0.65
WM	17.48	0.727	89.5	0.001	0.1			0.084	10.3	0.812	0.02	0.90
Cr												
LCmodel	12.42	0.316	9.6	0.007	0.2	2.446	74.3	0.523	15.9	3.292	0.06	0.84
corrected	12.88	0.609	8.8	0.008	0.1	5.668	81.4	0.675	9.7	6.960	0.06	0.90
GM	17.52	1.373	82.0	0.062	3.7			0.239	14.3	1.674	0.03	0.82
WM	9.67	0.230	82.1	0.000	0.0			0.050	17.9	0.280	0.02	0.82
Cho												
LCmodel	3.58	0.040	14.5	0.002	0.7	0.179	64.9	0.055	19.9	0.276	0.07	0.79
Corrected	3.21	0.038	17.7	0.002	0.9	0.131	60.9	0.044	20.5	0.215	0.07	0.79
GM	3.27	0.061	74.4	0.004	4.9			0.017	20.7	0.082	0.04	0.75
WM	3.17	0.058	90.6	0.000	0.0			0.006	9.4	0.064	0.02	0.90
mI												
LCmodel	12.08	0.750	13.2	0.033	0.6	3.296	58.1	1.594	28.1	5.673	0.11	0.71
Corrected	10.64	0.820	12.5	0.028	0.4	4.395	67.1	1.304	19.9	6.547	0.11	0.80
GM	14.48	0.685	49.7	0.159	11.5			0.533	38.7	1.377	0.06	0.50
WM	7.96	1.084	93.1	0.000	0.0			0.080	6.9	1.164	0.04	0.93
Glu												
LCmodel	16.77	0.324	2.9	0.004	0.0	7.918	70.4	2.995	26.6	11.241	0.10	0.73
Corrected	15.49	0.431	2.8	0.001	0.0	12.071	78.7	2.836	18.5	15.339	0.11	0.82
GM	22.06	1.286	54.1	0.000	0.0			1.093	45.9	2.379	0.05	0.54
WM	10.49	0.097	21.7	0.000	0.0			0.349	78.3	0.446	0.06	0.22
Glx												
LCmodel	20.67	0.765	3.0	0.126	0.5	17.824	69.2	7.052	27.4	25.767	0.13	0.72
Corrected	19.17	1.305	4.0	0.095	0.3	25.157	76.2	6.476	19.6	33.033	0.13	0.80
GM	28.47	3.380	67.9	0.000	0.0			1.595	32.1	4.975	0.04	0.68
WM	11.93	0.768	45.0	0.222	13.0			0.717	42.0	1.707	0.08	0.45

Results of test-retest statistical analyses for LCModel data without partial volume correction "LCModel," LCModel data after partial volume correction "corrected," and gray matter "GM" and white matter "WM" estimates of concentrations. Percent of total variance "%" for each variance component appears in the adjacent column to the right.

$$Y_{\mathrm{ik(grayorwhite)}} = \mu + \mathrm{Sub_i} + \mathrm{Scan_k} + e_{\mathrm{ik}} \quad [3]$$

where μ is the gray or white matter mean concentration across subjects.

ICCs to assess test-retest measurement reliability were based on these random-effects analysis of variance models and were computed as follows (29):

$$\mathrm{ICC} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2 + \sigma_e^2} \quad [4]$$

where σ_{B}^2 is a generalized between-subject variance of metabolite concentrations, σ_{W}^2 is the within-subject variance between scans 1 and 2 (i.e., the interscan variance), and σ_{e}^2 is the variance due to random noise. For the data sets involving all voxels across all subjects, σ_{B}^2 is the sum of the subject and voxel within-subject variances, as estimated using REML and model [1] above. For ICCs based on individual subject data, σ_{B}^2 is simply the voxel variance as estimated using model [2], and for the data sets involving gray and white matter estimates, σ_{B}^2 is the subject variance as estimated using model [3].

CVs for interscan (test-retest) reproducibility were calculated based on the variances separated as earlier, using just the interscan (σ_{W}^2) and error (σ_{e}^2) variances along with the concentration mean μ.

$$\mathrm{CV} = \frac{\sqrt{\sigma_W^2 + \sigma_e^2}}{\mu}.$$

RESULTS

Figure 1 shows the ^{1}H-MRSI slice location, a representative fit by LCModel to a spectrum from a voxel with primarily GM, and a plot of the regression analysis used to estimate the Glu concentration in either pure gray or white matter. Table 1 summarizes the various measures of reproducibility or reliability calculated for all data. As expected, the interscan (test-retest) CVs, based on the interscan and error variances alone, did not differ substantially between the partial volume corrected and uncorrected LCModel output, since the substantial variance due to different fractions of GM, WM, and CSF in

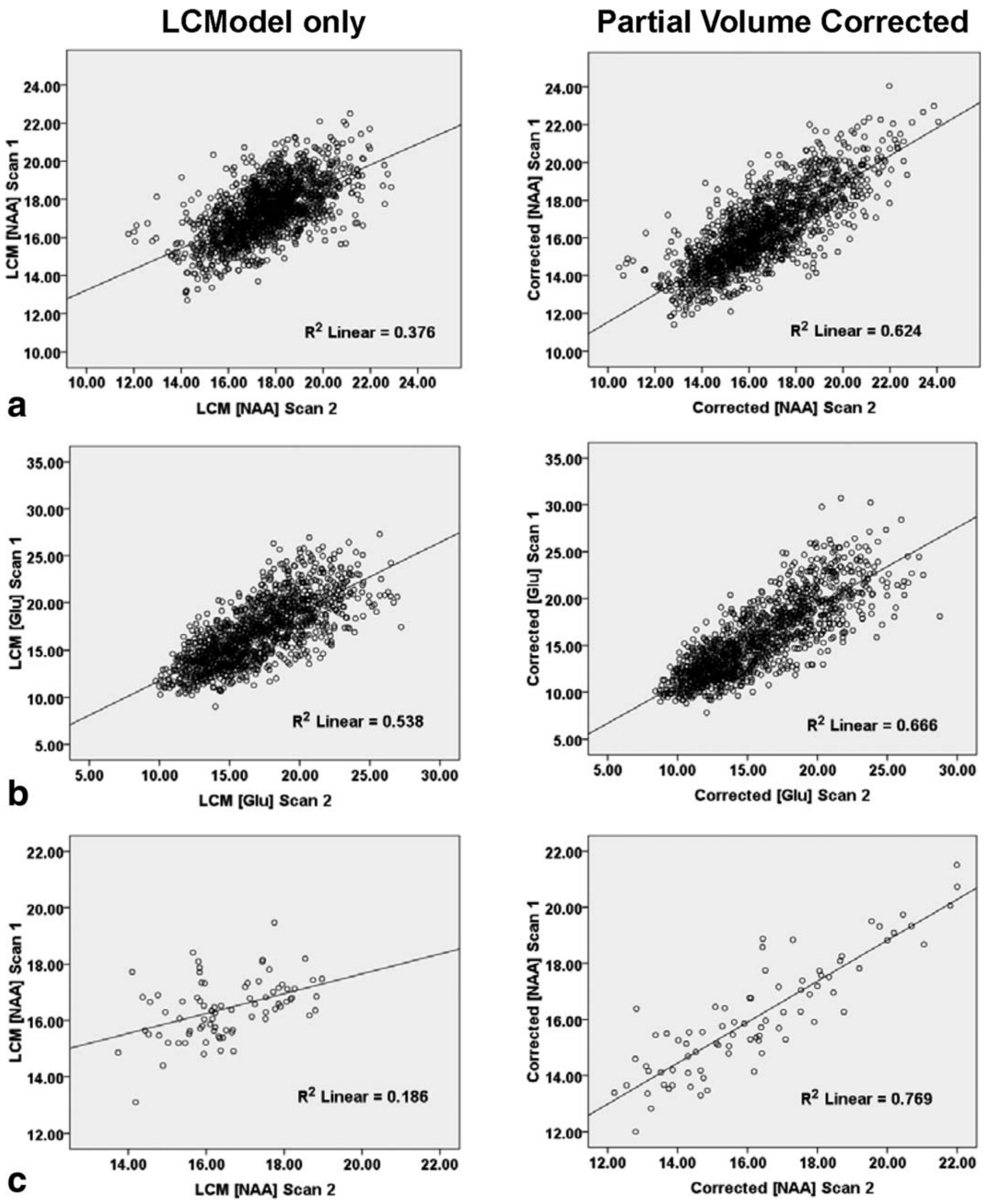

FIG. 2. **a,b:** Plots of first scan versus second scan data for NAA (a) and Glu (b) for all voxels from all subjects with (right panels) or without (left panels) partial volume and relaxation correction. **c:** Similar plots for NAA data from one subject.

different voxels is parceled into the voxel variance term in both cases and, hence, does not enter into the calculation of the interscan CV. Also consistent with expectations, the voxel variance is observed to be greater in the partial volume corrected data, due to adjusting the concentrations for the voxel CSF fraction and, hence, elevating the GM estimate of the metabolite from its uncorrected value and creating greater GM-WM metabolite differences across the brain. Interscan CVs based on all the voxels (ca. 1300–1500) across all 21 subjects in this study varied from lows in the range of 0.05–0.07 for tNAA, NAA, Cr, and Cho to highs of 0.10–0.13 for mI, Glu, and Glx signals defined entirely by their multiplet structures and routinely more difficult to measure.

The interscan CVs of the estimates of metabolite concentrations in either gray or white matter, on the other hand, are substantially less than those based on all voxels. This is also expected, due to reducing the voxel variability to single estimates of metabolite concentration in just gray or white matter, shifting the major source of variance to the subject-by-subject variability. These values ranged from 0.02–0.04 for tNAA, NAA, Cr, and Cho to 0.04–0.08 for mI, Glu, and Glx. Generally, GM CVs were slightly greater than WM CVs. However, this trend was reversed for Glu and Glx.

The effect of partial volume correction on improving reproducibility is also evident in the ICCs, which are the only true measures of absolute test-retest reliability in this study. As shown in Table 1 and suggested in the

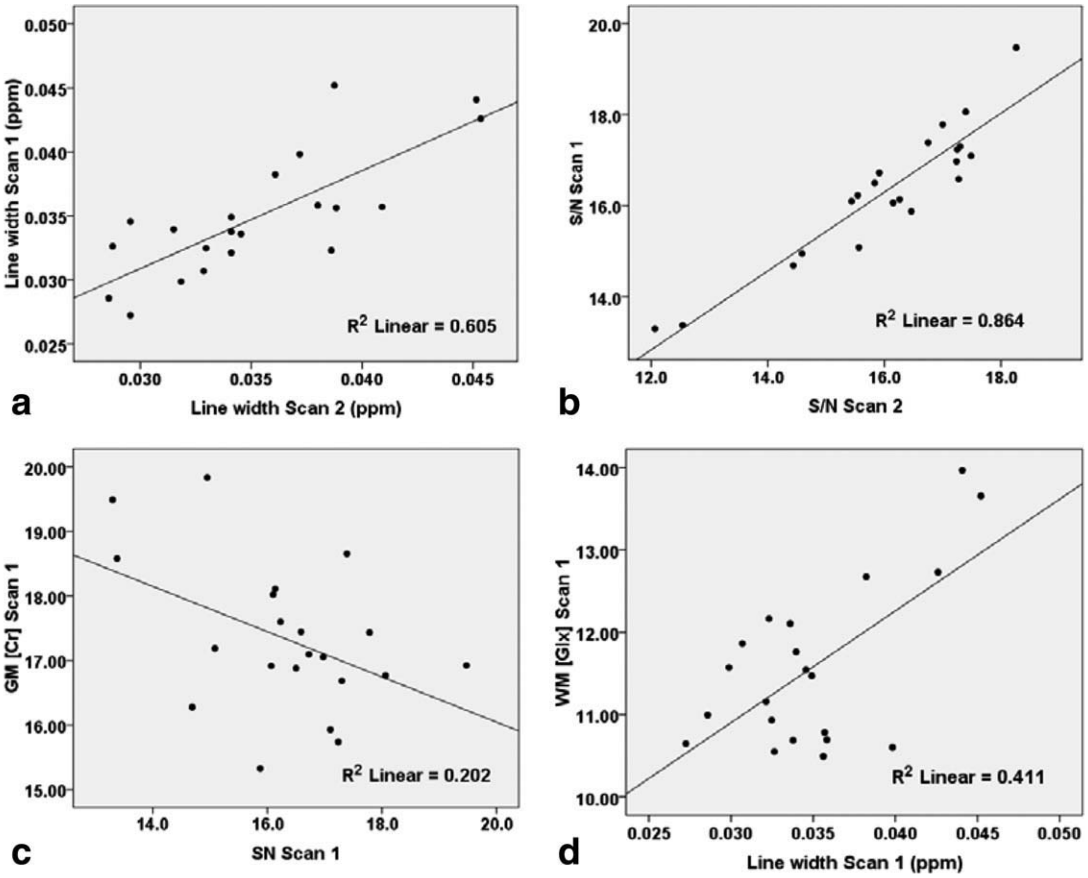

FIG. 3. **a:** Plot of mean ^{1}H-MRSI data set line width for first scan versus second scan. **b:** Plot of mean ^{1}H-MRSI data S/N for first scan versus second scan. **c:** Plot of estimate of GM Cr concentration versus mean S/N across subjects. **d:** Plot of estimate of WM Glu concentration versus mean S/N across subjects. The correlations were significant for data sets before removal of outliers (see Table 2).

representative scatter plots of Fig. 2, the ICCs based on all voxels were consistently higher for the data after partial volume correction. This was even more apparent in ICCs from individual subjects, as shown by a representative case in Fig. 2. Overall, the ICCs based on all the data from voxels across all subjects were high (0.6–0.9), including for Glu, Glx, and mI. However, ICCs based on GM and WM estimates, while still in this range for NAA, tNAA, Cr, and Cho, were sometimes substantially lower for Glu, Glx, and mI.

We also examined the relationship of data quality to reproducibility. Two measures of data quality reported by LCModel were examined: the NAA line width and signal-to-noise ratio (S/N), the latter approximated as the ratio of the peak height at 2.01 ppm to the root mean square of the residuals of the fit. The values of these measures for each spectrum were averaged across each subject's ^{1}H-MRSI data set for a particular scan to obtain one mean estimate of ^{1}H-MRSI line width and S/N for that scan. One unexpected finding of this analysis was that measures of data quality were highly reproducible, as illustrated in the scatter plots of Fig. 3. The ICC for line width and S/N on successive scans was 0.78 and 0.91, respectively. Furthermore, line width and S/N were predictive of concentration for several metabolites in linear regression models (Table 2). Inspection of scatter plots for these data, however, suggested that the significance of most of these correlations depended on a small number of cases with exceptionally low S/N or large line widths, as illustrated for representative cases in Fig. 3. Eliminating just three cases from the analysis substantially reduced the number of significant correlations (Table 2). These cases were selected on the basis of having a line width or S/N that was outside the 2.5 or 97.5 percentile points of a theoretical normal distribution of line width or S/N values. Hence, the three cases had either a line width that was 1.96 × SD greater than the mean line width across all subjects and scans or a S/N ratio that was 1.96 × SD less than the mean S/N across all subjects and scans. This is a relatively conservative but nonetheless arbitrary cut-off point to define statistical outliers, and applied here only to illustrate that a small number of observations with particular poor quality may be largely responsible for the high regression coefficients shown in Table 2.

DISCUSSION

This study examined both the reproducibility and absolute reliability of brain metabolite concentration estimates from ^{1}H-MRSI data collected with a short TE at 3T, with and without partial volume correction. The reproducibility and reliability of estimates of pure gray and white matter metabolite concentrations were also

Table 2
Significant Correlations of Tissue-Specific Metabolite Estimates with Line Width and S/N

	Normalized beta	
Metabolite (scan)	Line width	S/N
NAA GM (2)	0.563[a]	
NAA WM (2)		0.452[a]
tNAA GM (2)	0.475[a]	
Cr GM (1)		−0.449[a]
Cr WM (1)	0.639	
Cr GM (2)		−0.682[a]
Cr WM (2)		−0.669
mI GM (1)		−0.503[a]
mI WM (1)		−0.655
mI GM (2)		−0.646[a]
mI WM (2)		−0.718
Glu GM (2)	0.449	
Glx GM (1)		−0.531[a]
Glx WM (1)	0.641[a]	
Glx GM (2)		−0.637[a]
Glx WM (2)	0.590	

Normalized regression coefficients "beta" for linear regression analyses of GM or WM concentration estimates and either mean 1H-MRSI scan line width or S/N. Only significant correlations involving data from all 21 subjects are shown ($P \leq 0.05$).

[a]Significance vanished with outlier removal. Correlations no longer significant when outliers in line width or S/N removed.

measured. Our findings reveal a substantial improvement in both reproducibility, as assessed by interscan CVs, and reliability, as assessed by ICCs, with partial volume correction. CVs for corrected Glu, Glx, and mI levels in gray and white matter are substantially less than those reported in ^{1}H-MRSI studies at 1.5T and short TE (14,17), while the CVs obtained for gray and white matter NAA, tNAA, Cr, and Cho agree well with those reported for gray and white matter estimates of these metabolites at 3T but at longer TEs (16,20). ICCs for partial volume corrected metabolite estimates were also consistently higher than ICCs based on uncorrected data. These results demonstrate that differences in signal intensities between scans that arise from less-than-exact repositioning of the ^{1}H-MRSI slice, resulting in altered tissue fractions in each voxel, are partially compensated for by partial volume correction. Finally, the reproducibility (CVs) and reliability (ICCs) of NAA estimates in this study were comparable and sometimes superior to tNAA estimates (Table 1). This finding challenges the common assumption that NAAG cannot be reliably resolved from NAA at 3T and, therefore, that estimates of tNAA are more reliable than estimates of NAA alone.

The variety of field strengths, acquisition methods, and processing methods used in past studies has made arriving at a consensus on the reproducibility or reliability of ^{1}H-MRSI challenging. With respect to signals defined solely by their multiplet structures, Chard et al. used single-slice point-resolved spectroscopy sequence ^{1}H-MRSI at 1.5T with a 30-ms TE and relatively large nominal voxel size (2.3 cm^3) to obtain interscan CVs for Glu, Glx, and mI in the range of 0.16–0.19 (14). These values can be compared directly to the interscan CVs obtained for the all-subject, all-voxel data, both with and without partial volume correction, in the present study (0.11–0.13) and are substantially greater than the CVs for the estimates of pure gray and white matter metabolite concentrations reported here (0.04–0.08). Similarly, using a multislice point-resolved spectroscopy sequence sequence at 1.5T with a TE of 30 ms and a nominal voxel size of 1 cm^3, Langer et al. obtained median CVs for Glx and mI of 0.21 and 0.24, respectively (17). These CVs, however, were based on the square root of the total variance (the standard deviation) rather than solely on the scan and error variances and, therefore, are expectedly larger than those reported by either Chard et al. or in the present report. It is worth noting in this regard that the relatively low estimates of reproducibility reported by either Chard et al. or Langer et al. reflect different definitions of reproducibility as well as different acquisition and processing protocols. Along these lines, we note that these studies were conducted at lower field strength as well as in regions of brain that are generally characterized by greater field inhomogeneity than the region investigated in the present study. Nonetheless, judging from this study as well as whole-brain, multislice studies by others (16,20), the reproducibility of ^{1}H-MRSI measurements of tissue-specific metabolite levels, which are the values of interest to most researchers, substantially exceeds that suggested by studies that do not take regional variations of metabolite levels into account.

Test-retest reliability, as measured by the ICC, was uniformly high (0.60–0.90) for all metabolites in this study when the data from all voxels across subjects were used as input. However, ICCs based on pure gray and white matter metabolite estimates were roughly inversely related to the interscan CVs: high for NAA, tNAA, Cr, and Cho (0.65–0.9) but lower for particular mI, Glu, and Glx estimates (0.22–0.54). Inspection of Table 1 reveals that the low ICCs are primarily a consequence of an error variance term (σ_e^2) that was high relative to the between-subjects term (σ_B^2) in Eq. 4; whereas, in ICC analyses based on all voxels, the voxel variance accounted for much more, if not most, of the total variance, and thus led to a high σ_B^2. Regardless of the details of these differences, the conclusion that must be drawn from these results is that, under the acquisition and processing protocols of the present study, the test-retest reliability of the measurement of GM mI, WM Glu, or WM Glx is low (ICC $\leq$ 0.50) and the reliability of the GM Glu measurement is only slightly higher (ICC = 0.54). Hence, among the signals examined in this study that are entirely defined by J-coupled multiplets, only the measurement of WM mI (ICC = 0.93) and GM Glx (ICC = 0.68) appear to be highly reliable, in agreement with the relatively low interscan CVs obtained for these signals (0.04).

Given the much lower concentration of Glu and glutamine in white matter and the small nominal voxel size (0.71 cm^3) of this study, it is not unexpected that estimates of either Glu or Glx would be more reliable in gray matter than in white, nor that the more intense Glx signal could be detected more reliably than Glu alone. Hence, a larger voxel size and, consequently, greater S/N may improve the ICC for the detection of these molecules in both gray and white matter at 3T, i.e., by lowering the error variance term in Eq. 4. However, another factor in the calculation of the ICC is the between-subjects variance term, which can be seen to be much larger for the GM estimate of Glu or Glx relative to the WM estimate (Table 1) and thus elevates the ICC. Similarly, a greater between-subjects variance coupled with a lower error term in the estimates of mI in WM relative to GM underlies the higher ICC of mI in WM relative to GM. The between-subjects variance term is, in principle, related to real differences in metabolite concentrations among subjects. When interpreting the ICC, therefore, it is worth bearing in mind that, given a certain level of noise, the greater the real subject-to-subject differences in a measured quantity, the higher the apparent measurement reliability will be. In this sense, the ICC based on data from a sample of healthy control subjects may underestimate the reliability of measuring longitudinal differences in a patient group in a study, if the between-subject variability of the measured parameter is greater in the patient group while the within-subject variability is not. Nor does a low ICC based on test-retest data from healthy subjects indicate that differences between healthy subjects and patients cannot be measured reliably since, ultimately, the means and variance components of both groups need to be taken into account in group comparisons.

It is worth noting that the use of the water signal as a concentration reference, acquired in a separate 9.7-minute scan, undoubtedly introduces variance in the concentration estimates. This source of variance will be absent in metabolite ratio data, i.e., if the metabolite intensities are scaled to another metabolite intensity within the same spectrum, such as Cr. Furthermore, any variance introduced by the estimates of CSF needed for ‘absolute’ metabolite concentrations calculations will also be absent in metabolite ratio estimates. Though comparing concentration estimates to ratio data was not an aim of the present study, the reproducibility of metabolite ratios might be greater than the reproducibility exhibited for water-scaled absolute concentrations in this study, provided that the reference metabolite intensity does not vary independently from the metabolite of interest or that any independent variance is less than the combined variance introduced by the water signal and CSF estimation.

An unexpected finding of this study was the high reproducibility of line width and S/N in repeat scans. This could only derive from reproducible patterns of magnetic field inhomogeneity in each subject which, in turn, ultimately derive from the interaction between the scanner magnet, the shim routine, and the magnetic susceptibility of the subject. The latter factor is undoubtedly unique for each subject and primarily determined by factors such as head size and shape, proximity of orbit and nasal cavities to the ^{1}H-MRSI region of interest, and the amount of dental work. This reproducibility would be of little concern to researchers using ^{1}H-MRSI were spectral quality not related to the accuracy of spectral curve fitting. However, previous studies have shown that low S/N and broad line widths can indeed lead to under- or over-estimations of metabolite concentrations when using standard curving fitting routines (28,30,31). The present study supports these findings. Significant correlations between metabolite estimates and either mean ^{1}H-MRSI line width, S/N, or both accounted for a significant portion of the variance for some metabolites, which included Cr as well as Glu. The regression coefficients of this analysis are primarily negative, suggesting that low S/N leads to overestimates of concentrations by LCModel. The number of these correlations was reduced dramatically by eliminating just three cases (out of 21) that were outliers in terms of large line width or low S/N, underscoring the importance of screening data for spectral quality in ^{1}H-MRSI studies as well as avoiding any biases in spectral quality between the groups or time points that are to be compared.

In summary, the results of this study demonstrate high reproducibility and test-retest reliability of tissue-specific estimates of several metabolites using ^{1}H-MRSI at 3T and short TE in a group of healthy subjects. Our findings on the reproducibility of gray and white matter estimates of metabolites such as NAA, tNAA, Cr, and Cho are consistent with previous studies using longer TEs. Furthermore, these data show that, under the acquisition and processing protocols of this study, both WM mI and GM Glx in healthy subjects have relatively high reproducibility and test-retest reliability at 3T, and all other measurements of Glu, Glx, and mI demonstrate much lower CVs than reported in previously studies at 1.5T. Finally, the high reproducibility observed for spectral line widths and S/N ratios in ^{1}H-MRSI data from individual subjects, as well as the impact of these factors on spectral analysis, warrant further investigation.

ACKNOWLEDGMENTS

The authors thanks George Malloy for his assistance with data collection and Rachael Grazioplene for her assistance with data analysis.

REFERENCES

1. Srinivasan R, Vigneron D, Sailasuta N, Hurd R, Nelson S. A comparative study of myo-inositol quantification using LCmodel at 1.5 T and 3.0 T with 3 D 1H proton spectroscopic imaging of the human brain. Magn Reson Imaging 2004;22:523–528.
2. Di Costanzo A, Trojsi F, Tosetti M, Schirmer T, Lechner SM, Popolizio T, Scarabino T. Proton MR spectroscopy of the brain at 3 T: an update. Eur Radiol 2007;17:1651–1662.
3. Cianfoni A, Niku S, Imbesi SG. Metabolite findings in tumefactive demyelinating lesions utilizing short echo time proton magnetic resonance spectroscopy. AJNR Am J Neuroradiol 2007;28:272–277.
4. Baranzini SE, Srinivasan R, Khankhanian P, Okuda DT, Nelson SJ, Matthews PM, Hauser SL, Oksenberg JR, Pelletier D. Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis. Brain;133:2603–2611.
5. Li Y, Chen AP, Crane JC, Chang SM, Vigneron DB, Nelson SJ. Three-dimensional J-resolved H-1 magnetic resonance spectroscopic imaging of volunteers and patients with brain tumors at 3T. Magn Reson Med 2007;58:886–892.
6. Chawla S, Wang S, Wolf RL, Woo JH, Wang J, O’Rourke DM, Judy KD, Grady MS, Melhem ER, Poptani H. Arterial spin-labeling and MR spectroscopy in the differentiation of gliomas. AJNR Am J Neuroradiol 2007;28:1683–1689.
7. Gasparovic C, Yeo R, Mannell M, Ling J, Elgie R, Phillips J, Doezema D, Mayer AR. Neurometabolite concentrations in gray and white matter in mild traumatic brain injury: an 1H-magnetic resonance spectroscopy study. J Neurotrauma 2009;26:1635–1643.
8. Gonen O, Gruber S, Li BS, Mlynarik V, Moser E. Multivoxel 3D proton spectroscopy in the brain at 1.5 versus 3.0 T: signal-to-noise ratio and resolution comparison. AJNR Am J Neuroradiol 2001;22:1727–1731.

9. Schubert F, Gallinat J, Seifert F, Rinneberg H. Glutamate concentrations in human brain using single voxel proton magnetic resonance spectroscopy at 3 Tesla. Neuroimage 2004;21:1762–1771.
10. Mayer D, Spielman DM. Detection of glutamate in the human brain at 3 T using optimized constant time point resolved spectroscopy. Magn Reson Med 2005;54:439–442.
11. Mullins PG, Chen H, Xu J, Caprihan A, Gasparovic C. Comparative reliability of proton spectroscopy techniques designed to improve detection of J-coupled metabolites. Magn Reson Med 2008;60:964–969.
12. Tedeschi G, Bertolino A, Campbell G, Barnett AS, Duyn JH, Jacob PK, Moonen CT, Alger JR, Di Chiro G. Reproducibility of proton MR spectroscopic imaging findings. AJNR Am J Neuroradiol 1996;17:1871–1879.
13. Charles HC, Lazeyras F, Tupler LA, Krishnan KR. Reproducibility of high spatial resolution proton magnetic resonance spectroscopic imaging in the human brain. Magn Reson Med 1996;35:606–610.
14. Chard DT, McLean MA, Parker GJ, MacManus DG, Miller DH. Reproducibility of in vivo metabolite quantification with proton magnetic resonance spectroscopic imaging. J Magn Reson Imaging 2002;15:219–225.
15. Li BS, Babb JS, Soher BJ, Maudsley AA, Gonen O. Reproducibility of 3D proton spectroscopy in the human brain. Magn Reson Med 2002; 47:439–446.
16. Zhu XP, Young K, Ebel A, Soher BJ, Kaiser L, Matson G, Weiner WM, Schuff N. Robust analysis of short echo time (1)H MRSI of human brain. Magn Reson Med 2006;55:706–711.
17. Langer DL, Rakaric P, Kirilova A, Jaffray DA, Damyanovich AZ. Assessment of metabolite quantitation reproducibility in serial 3D-(1)H-MR spectroscopic imaging of human brain using stereotactic repositioning. Magn Reson Med 2007;58:666–673.
18. Ratai EM, Hancu I, Blezek DJ, Turk KW, Halpern E, Gonzalez RG. Automatic repositioning of MRSI voxels in longitudinal studies: impact on reproducibility of metabolite concentration measurements. J Magn Reson Imaging 2008;27:1188–1193.
19. Gu M, Kim DH, Mayer D, Sullivan EV, Pfefferbaum A, Spielman DM. Reproducibility study of whole-brain 1H spectroscopic imaging with automated quantification. Magn Reson Med 2008;60:542–547.
20. Maudsley AA, Domenig C, Sheriff S. Reproducibility of serial whole-brain MR spectroscopic imaging. NMR Biomed;23:251–256.
21. Gasparovic C, Song T, Devier D, Bockholt HJ, Caprihan A, Mullins PG, Posse S, Jung RE, Morrison LA. Use of tissue water as a concentration reference for proton spectroscopic imaging. Magn Reson Med 2006;55:1219–1226.
22. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. Magn Reson Med 1993;30:672–679.
23. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26: 839–851.
24. Gasparovic C, Neeb H, Feis DL, Damaraju E, Chen H, Doty MJ, South DM, Mullins PG, Bockholt HJ, Shah NJ. Quantitative spectroscopic imaging with in situ measurements of tissue water T1, T2, and density. Magn Reson Med 2009;62:583–590.
25. Vymazal J, Righini A, Brooks RA, Canesi M, Mariani C, Leonardi M, Pezzoli G. T1 and T2 in the brain of healthy subjects, patients with Parkinson disease, and patients with multiple system atrophy: relation to iron content. Radiology 1999;211:489–495.
26. Kreis R, Ernst T, Ross BD. Development of the human brain: in vivo quantification of metabolite and water content with proton magnetic resonance spectroscopy. Magn Reson Med 1993;30:424–437.
27. Mlynarik V, Gruber S, Moser E. Proton T (1) and T (2) relaxation times of human brain metabolites at 3 Tesla. NMR Biomed 2001;14: 325–331.
28. Kreis R. Issues of spectral quality in clinical 1H-magnetic resonance spectroscopy and a gallery of artifacts. NMR Biomed 2004;17: 361–381.
29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–428.
30. Schirmer T, Auer DP. On the reliability of quantitative clinical magnetic resonance spectroscopy of the human brain. NMR Biomed 2000;13:28–36.
31. Kanowski M, Kaufmann J, Braun J, Bernarding J, Tempelmann C. Quantitation of simulated short echo time 1H human brain spectra by LCModel and AMARES. Magn Reson Med 2004;51:904–912.