Clinical Investigative Study

Comparative Reliability of Total Intracranial Volume Estimation Methods and the Influence of Atrophy in a Longitudinal Semantic Dementia Cohort

George Pengas, MRCP (UK), João M. S. Pereira, BEng, Guy B. Williams, MA, PhD, Peter J. Nestor, PhD, FRACP

From the Neurology Unit, Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom (GP, PJN); and Wolfson Brain Imaging Centre, Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom (JMSP, GBW).

ABSTRACT

BACKGROUND AND PURPOSE

Total intracranial volume (TIV) as a measure of premorbid brain size is often used to correct volumes of interest for interindividual differences in magnetic resonance imaging (MRI) studies. We directly compared the reliability of different TIV estimation methods to address whether such methods are influenced by brain atrophy in the neurodegenerative disease, semantic dementia.

METHODS

We contrasted several manual approaches using T1-weighted, T2-weighted, and proton density (PD) acquisitions with 2 automated methods (statistical parametric mapping 5 [SPM5] and FreeSurfer [FS]) in a cohort of semantic dementia subjects (n = 11) that had been imaged longitudinally.

RESULTS

Novel mid-cranial sampling of either PD or T2-weighted images were least susceptible to atrophy: of these, the PD method was both more precise and more user-friendly. SPM5 also produced good results, providing automation for only a small loss in precision compared to the best manual methods. The T1 method that underestimated TIV as atrophy progressed was the least reproducible and the most labor-intensive. Fully automated FS overestimated TIV with progressive atrophy, and the results were even worse after optimizing the transformation.

CONCLUSION

The mid-cranial sampling of PD images achieved the best combination of precision, reliability, and user-friendliness. SPM5 is an attractive alternative if the highest level of precision is not required.

Keywords: Total intracranial volume, Dementia, MRI, SPM, FreeSurfer, Atrophy.

Acceptance: Received January 21, 2008, and in revised form January 21, 2008. Accepted for publication January 28, 2008.

Correspondence: Address correspondence to Peter J. Nestor, PhD, FRACP, Ward R3, Box 83, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom. E-mail: pjn23@hermes.cam.ac.uk.

Conflict of Interest: The authors have reported no conflicts of interest.

J Neuroimaging 2008;XX:1-10.
DOI: 10.1111/j.1552-6569.2008.00246.x

Introduction

Volumetric magnetic resonance imaging (MRI) techniques have been extensively used to measure volumes of cerebral regions of interest in normal brain development,[1] ageing,[2] epilepsy,[3] multiple sclerosis,[4] and neurodegenerative diseases.[5] However, as cerebral volumes vary in the normal population according to gender and body and head size,[2,6] it is usual to correct such volumes for interindividual variability in cranial (and brain) size.

Various correction strategies have been proposed, including measures of cerebral area,[5,7] total brain volume (TBV),[8] and total intracranial volume (TIV). Free et al.[9] compared correction of hippocampal volumes in controls and patients with temporal lobe epilepsy using corpus callosum area, cranial area, parenchymal area, brain stem area, TIV, and TBV. They found the expected gender difference between male and female hippocampal volumes and that the standard deviation was most consistently reduced by correcting with TIV.

Estimation of TIV itself has been attempted with different MR acquisition techniques including proton density (PD)[10,11] and T1-[12] and T2-[13] weighted sequences. Even within each MR acquisition method, there is a potentially limitless number of permutations. For instance, PD images using the whole brain,[11] or every sixth slice rostral to the most caudal slice containing cerebellum,[10] have been used. T1-weighted images have been used to model the intracranial size as a sphere;[1] others have used sagittal[14] or axial[12] slices. Measuring every fifth[8] or every tenth slice[15] of T1-weighted sequences has been proposed. For T2-weighted images, every slice,[2] every slice rostral to the opening of the medulla,[16] or every slice rostral to the most caudal slice containing cerebellum[13] has been used to estimate TIV. Furthermore, once a TIV estimate has been obtained, different correction calculations have been described, including

dividing the observed regional volume by the TIV[8] or covariance methods.[14] In addition to manual approaches, various automated methods have also been described for TIV estimation such as statistical parametric mapping (SPM) with T1-[17] or T2-[4] weighted images, or using automated atlas normalization with T1-weighted images.[18] Another technique involves the use of T2 and PD sequences together.[19]

Although most authors quote intra- and/or interrater reliability statistics for their methods, there is lack of meaningful direct comparisons among methods. New automated methods are validated against existing techniques, but there is no consensus on whether there is a "gold standard" method against which novel methods are to be judged. Even the statistical analyses used to assess agreement among methods are inconsistent.[20]

As TIV is thought to closely reflect premorbid TBV,[1] a critical assumption is that TIV does not change with time (in adulthood) and should not be influenced by ageing or atrophy. If the estimated TIV were significantly biased by atrophy, it would render it unsuitable for use as a constant variable to correct for interindividual premorbid brain size, especially in neurodegenerative conditions. However, cross-sectional studies are unable to gauge whether changing proportions of gray matter, white matter, and cerebrospinal fluid (CSF) could bias TIV estimations. To our knowledge, only 2 case reports of longitudinal stability of TIV estimation in dementia exist[12,18] and no study has directly compared TIV estimation methods longitudinally in a cohort of patients with progressive neurodegeneration.

The present study was designed to systematically contrast several methods in the same data set. In particular, the major aim was to evaluate the effect of atrophy on TIV by studying longitudinal data from patients with a progressive neurodegenerative disease. Semantic dementia patients were chosen because their imaging morphology is particularly challenging as it combines areas of preservation with areas of extreme focal (temporal lobe) atrophy. Although comparative reliability in TIV methods is hampered by the absence of a "ground truth" comparison, we hypothesized that an important criterion would be the absence of the influence of atrophy within subjects. Five manual methods, including a novel mid-cranial TIV estimation technique, and 2 automated methods were directly compared in order to establish the fastest and most reliable method that was least susceptible to the influence of progressive brain atrophy.

Methods

Subjects

Longitudinal MRI data from 11 patients (7 males, 4 females) who met consensus criteria for semantic dementia[21,22] were studied (see Table 1 for demographics). All patients included had longitudinal neuropsychological assessments and imaging. The mean interscan time interval was 19.4 months (range 12-36). Ethical approval was obtained by the local research ethics committee. All subjects gave informed consent.

Imaging

MR images were acquired with a 1.5-T GE Signa MRI scanner (GE Medical Systems, Milwaukee, WI). Volumetric T1-weighted images were coronally acquired using a spoiled gradient-echo (SPGR) technique, with an in-plane dimension of .86 mm^2 and a slice thickness of 1.8-2.2 mm. Although there was variation in slice thickness of volumetric T1-weighted scans among patients, each within-patient scan pair had identical voxel dimensions and was performed on identical machines. In addition, PD and T2-weighted axial dual-echo sequence images (matrix 256 × 256 × 40, slice thickness 7.0 mm) were acquired.

Table 1. Demographic Data

Male:female	7:4
Mean age at time of first MRI, yrs (range)	62.8 (54-79)
Mean duration of symptoms at time of first MRI, yrs (range)	3.3 (1-7)
Mean interval between the 2 MRI scans, months (range)	19.4 (12-36)
Mean MMSE at time point 1, /30 (range)	25 (21-30)
Mean ACE at time point 1, /100 (range)	68 (50-92)
Mean MMSE at time point 2, /30 (range)	23 (16-29)
Mean ACE at time point 2, /100 (range)	57 (32-84)

MMSE = mini-mental state examination; ACE = Addenbrooke's cognitive examination.

Image Analysis

All measurements/segmentations were performed blinded to subject details and the results of any other measurements.

Manual Intracranial Volume Estimation Methods

All manual measurements were performed with the ANALYZE 6.0 software package (Biomedical Imaging Resource, Mayo Clinic, Rochester, MN). No realignment of acquired images was performed. Volumetric analysis was performed by a single observer (GP), who received equivalent training in the manual estimation methods described below. The time taken to estimate the TIV per scan for each method was recorded (as a measure of labor intensity), but no time limit was imposed.

Full-Cranial T1 (FCT1)-Weighted Method. All T1-weighted images were preprocessed in ANALYZE 6.0 by isometrically resizing the voxels (ie, .86 × .86 × .86). Following this, the method as described by Whitwell et al.[12] was followed. Briefly, a standard gray-level threshold of 33% was applied to the images to help outline the outer border of the dura (Fig 1A). Autotracing was performed using a semi-automated "seed and expand" function, but extensive manual editing was necessary because the dural limit was often poorly defined. Every tenth axial slice was included,[15] starting from the most caudal slice containing cerebellum until the superior-most limit of the dura. The TIV was then estimated as the volume of the selected voxels multiplied by the interval between slices (ie, 10). Linear interpolation of the areas was not performed–there is, therefore, a systematic difference between the FCT1 method and the

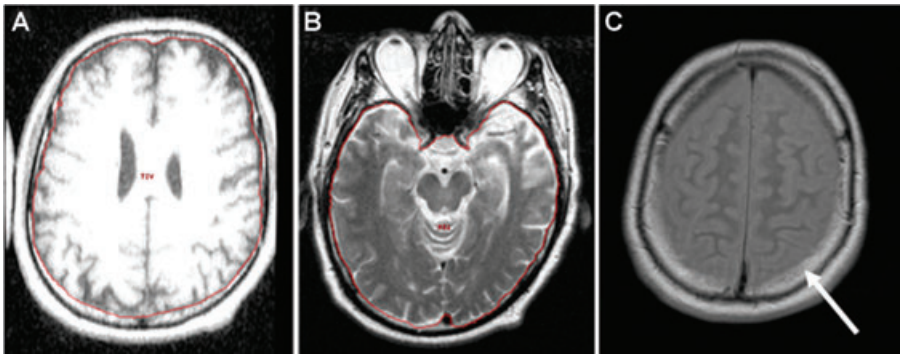

Fig 1. (A) ANALYZE image example of a T1-weighted sequence slice, at 33% threshold, illustrating the difficulty of establishing the dural margin. (B) Example of a T2 sequence slice at 60% threshold. Note the obvious temporal pole atrophy. (C) Example of a PD sequence slice close to the skull vertex, illustrating the partial volume effects: the distinction between brain tissue (GM, WM, and CSF) and skull has become indistinct (white arrow).

method described by Whitwell et al.[12] However, as all TIV estimates were paired (longitudinal measurements), this systematic error is cancelled out.

Mid-Cranial Proton Density (MCPD) Method. This method was developed in response to our perceived concerns about whole-brain measurements using PD and T2-weighted acquisitions. These were: variability in the level at which the lowest slice includes cerebellum between acquisitions and partial volume effects at the vertex that make it difficult to delineate the intracranial boundary (Fig 1C). Axial slices were used to outline the intracranial area. This was defined as the outer border of the brain and CSF; the semi-automated "seed and expand" method was adopted with the autotrace function in ANALYZE. This produced the best results if the seed consisted of a bright, CSF-intensity voxel. Minor manual editing was needed in some scans to exclude the superior sagittal venous sinus, blood vessels (eg, internal carotid arteries), nerves (eg, optic), and pituitary fossa. The most inferior supratentorial slice (defined as the slice that contained more cerebrum than cerebellum) was taken as the inferior border and the superior border was defined as 10 slices (inclusive) superior to the inferior border. This yields an intracranial volume estimation based on 10 supratentorial slices (Fig 2); the average volume estimated by this method was approximately 70% of the full-cranial estimation methods.

Full-Cranial T2 (FCT2)-Weighted Method. The method as described by Jenkins et al.[13] was followed. The gray-level threshold was set to 60%. The semi-automated autotracing technique was used to delineate the outer CSF boundary. The inferior plane of the TIV was defined as the most caudal slice containing cerebellum in the axial plane. Every slice superior to this, up to and including the most superior slice containing brain or CSF (at the vertex), was incorporated in the TIV estimation (Fig 1B).

Mid-Cranial T2 (MCT2)-Weighted Method. To investigate the interaction of the image acquisition method (ie, PD vs. T2) against the defined limits of TIV estimation (ie, full-cranial vs. mid-cranial), the craniocaudal limits described above (MCPD method) were applied to T2-weighted images.

Full-Cranial PD (FCPD) Method. For the same reason as described above (MCT2-weighted method), a FCPD method was also included using the method described for MCPD with the limits being those described for FCT2.

Automated TIV Estimation Methods.

All automated methods made use of the T1-weighted volumetric sequences.

SPM5 (http://www.fil.ion.ucl.ac.uk/spm/)[23] The SPM method was based on the subjects' segmented images in native space: gray matter, white matter, and CSF. In order to gain more sensitivity, the a priori SPM template maps (standard ICBM templates, default in SPM5) used in the segmentation step were undersampled so as to have 1 mm^3 isotropic voxels. The SPM procedure employed the "Segment" button in SPM5 to extract the tissue maps in *native space* (no normalization). Knowing that the segmentation in SPM yields a map, which is an estimate of the belonging probability distribution for each tissue class, and that it makes use of 4 clusters or tissue classes (grey matter, white matter, CSF, and other),[23] an estimation of the TIV can be obtained by summing the first three. A hard threshold was then used in order to exclude from the volume any voxel whose probability of belonging to any of the first three classes was less than .5. Finally, in order to calculate the TIV, the number of surviving voxels was obtained and multiplied by the volume of a single voxel. All these calculations were performed using MATLAB (Mathworks, Inc., Natick, MA).

FreeSurfer (FS) (Version FreeSurfer-Linux-centos4_× 86_64-stable-pub-v3.0.2, http://surfer.nmr.mgh.harvard.edu/)[24,25]. FS TIV estimates are derived from the notion that when a brain is normalized to a standard space atlas, the determinants of the affine transformations that connect each subject to the atlas space contain information about the contraction/expansion required to perform the registration.[18] Therefore, knowing (1) the TIV of the atlas and (2) the determinant of the affine transformation, the TIV of the subject can be calculated by dividing the former by the latter, thus yielding what is referred to as estimated TIV (eTIV). This method[18] has subsequently been altered in FS by iteratively adjusting the scale factor in reference to manually derived TIVs. The reference atlas used for this was FS's default, composed of 40 subjects spanning from young to old and including 10 Alzheimer's disease (AD) cases. TIV estimates were

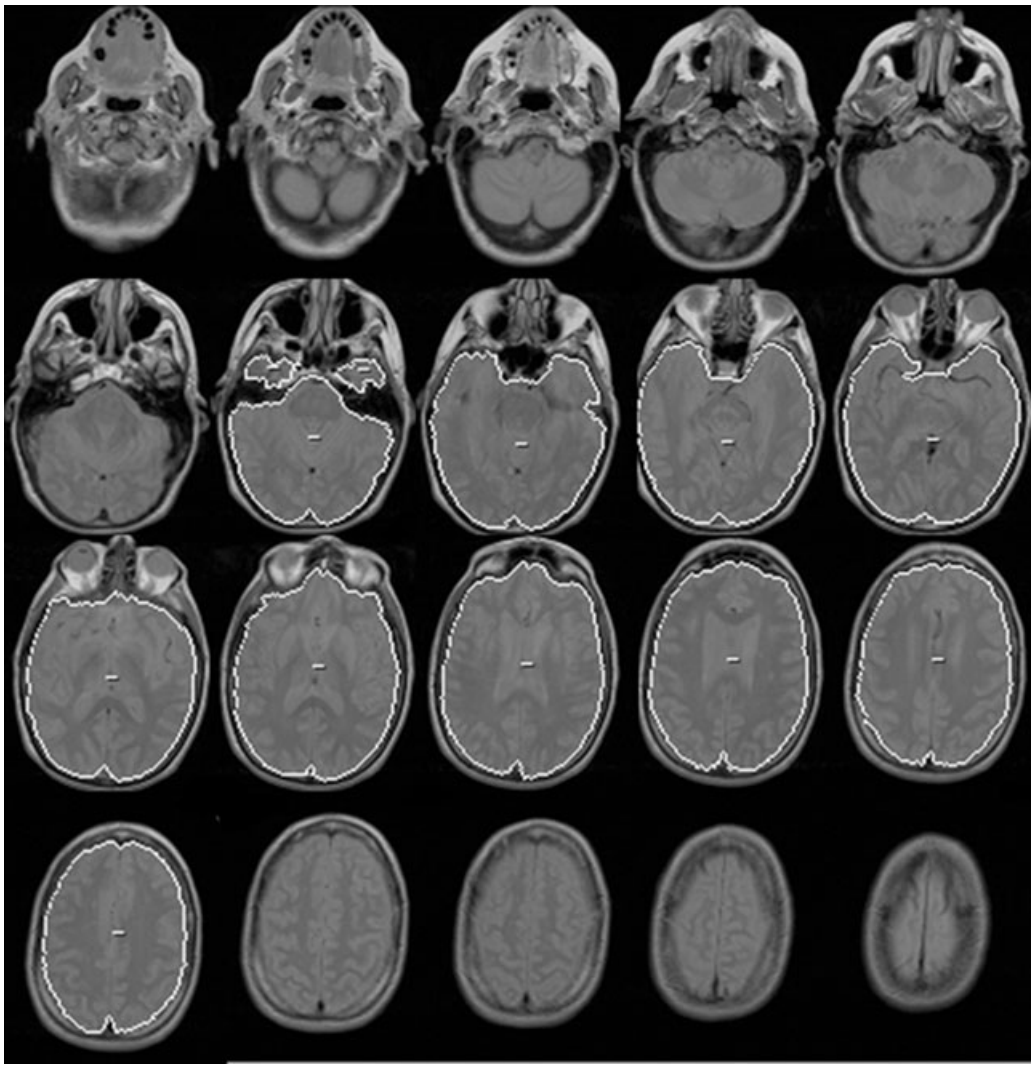

Fig 2. Panel of all 20 axial slices from 1 subject used to define MCPD in ANALYZE. The most inferior supratentorial slice was defined as the first slice in a caudorostral direction that contained more cerebrum than cerebellum. The volume of MCPD is delineated by a white line. Note the exclusion of the sagittal sinus and pituitary fossa.

obtained using the FS fully automated (FS_{AUTO}) default and FS after correction of transform (FS_{ACT}).

FS_{AUTO}. FS was allowed to run the full processing cycle (autorecon1 and autorecon2), at the end of which the value of the TIV was extracted from one of the output files (aseg.stats). The TIV thus obtained was referred to as FS_{AUTO} TIV because it requires no human intervention.

FS_{ACT}. Being fully automated, it is possible that FS_{AUTO} results may be adversely affected by suboptimal performance of the affine transformation. To address this issue, we reran FS after correcting the transform (FS_{ACT}). This method implies the evaluation and eventual alteration in the affine transform output at the end of autorecon1. The final cost function value indicates how optimally the transformation was performed; adjustments can be made in order to attain a heuristic threshold of under .1. The adjustments to achieve this value include, in order of execution, (1) rerunning the affine transform algorithm using a white matter-normalized intensity image; if this fails, (2) rerunning the affine transform algorithm using the skull-stripped scan; and, if both previous adjustments fail, (3) manually adjusting the affine parameters with a user interface to inspect and alter the affine transform output parameters (tkregister2).

Total Brain Volume (TBV) Estimation

The brain segmentation derived from SPM5 was also used to calculate TBV by summation of the gray matter and white matter voxels, with the threshold set at .5 as above (section Automated TIV estimation method: SPM5). The amount and rate of atrophy were then calculated. The rate of atrophy was assumed to be constant (within each individual) and was computed as the gradient of the line between TBV measurements at 2 time points, per individual, expressed as percentage change per year.

Reproducibility

A repeat measurement of TIV by each of the 5 manual methods was performed for all 11 patients on their initial (time 1) images,

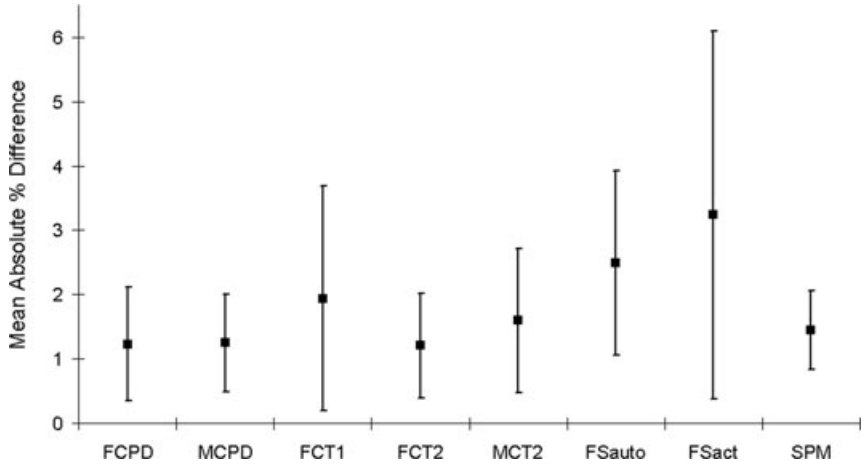

Fig 3. MAPD of TIV estimation by each method. This shows most methods are equally precise (MAPD 1.2–1.9%) except the FreeSurfer methods, which have a MAPD $\geq$ 2.5%. Error bars represent the 95% confidence intervals (CI).

at least 1 week after the original measurements. Three methods of assessing intrarater reliability were used for each method.

The Intraclass Correlation Coefficient[26,27]

This was calculated from a two-way mixed model analysis of variance (ANOVA) (carried out using SPSS for Windows, version 13.0; SPSS Inc., Chicago, IL).

The Bland and Altman Method[20]

This method can illustrate graphically the difference between the 2 measurements against the average of the 2 measurements. It also derives a coefficient of repeatability, which is the 95% confidence interval (CI) of the differences (calculated as twice the standard deviation of the differences) and is expressed in meaningful units (mm^3 in this case).

Coefficient of Variation (CV)

This was calculated as CV = (standard deviation of the TIV differences/mean of the TIV differences) $\times$ 100%.[28]

The Effect of Atrophy

The acceptance of the null hypothesis (that the TIV does not significantly change with time) was tested using 2 methods proposed by Lew[29]–the confidence interval method and the trade-off method. The *confidence interval* method is a representation of the mean difference between the 2 time intervals along with their 90% CI, ie, TIV at time 1 minus TIV at time 2. This implies that a positive result (TIV at time 2 < TIV time 1) indicates that the method underestimates TIV as atrophy progresses with time, while a negative result (TIV at time 2 > TIV at time 1) indicates overestimation of TIV with time. A zone of indifference is superimposed on these, ie, an effect size that is considered too small to be of biological interest, and if the method's mean difference and its 90% CI lie wholly within this zone, the method is deemed to accept the null hypothesis. Because the mid-cranial methods yield smaller volumes and have, therefore, smaller absolute errors than the full-cranial methods, we converted our results into relative percentage differences between the 2 time points as measured by each method. The zone of indifference was arbitrarily set at $\pm$ 2%. This means that we are willing to accept any method whose mean percentage difference and its 90% CI lie wholly within the zone of –2 and +2%. It is important to state that this value is arbitrary and is a trade-off between precision and feasibility, always guided by the size of effect that is deemed to be biologically significant.

The *trade-off* method is derived from the notion that the implication of type 1 and type 2 errors is reversed where one is trying to uphold the null hypothesis. Thus, setting a low β (to reduce the chance of incorrectly accepting a null hypothesis) can be achieved in a trade-off where the level of α is relaxed. The limit of indifference was again set at 2% and the false success rate, β, was set at .05. The method determines a P value (P_{critical}) based on the limit of indifference, the standard deviation of the differences (expressed as relative percentages), the number of subjects, and the false success rate chosen, which needs to be *exceeded* by a paired Student's t-test (P_{observed}) in order to fail to reject (ie, uphold) the null hypothesis.

The relationship between the change in TIV (TIV at time 1 – TIV at time 2) and the change in TBV, measured using SPM5 (TBV at time 1 – TBV at time 2), was assessed using Pearson's product moment correlation coefficient in SPSS 13.0 for each of the methods described. This was done to investigate the presence of a systematic error (degree of atrophy) influencing the method outcome (TIV).

In order to assess the validity of using a mid-cranial sampling method as a measure of (whole-brain) TIV, Pearson's correlation coefficients were derived for each set of paired data, eg, FCPD TIV compared to MCPD TIV.

Results

Mean Absolute Percentage Difference (MAPD)

The average of the differences in TIV between the 2 time points, expressed as absolute values, divided by the TIV at time point 1 and multiplied by 100%, is shown in Figure 3 and Table 2.

The Effect of Atrophy

The TBV was estimated by SPM5 in order to calculate the rate of atrophy in each individual over the 2 time points. The

Table 2. Results

	FCPD	MCPD	FCT1	FCT2	MCT2	FS_{AUTO}	FS_{ACT}	SPM
MAPD, %	1.23	1.25	1.94	1.21	1.6	2.49	3.24	1.45
Relative % difference	.85	−.54	1.48	.9	−.07	−2.08	−2.22	.4
Trade-off ($\beta = .05$)								
$P_{critical}$	.265	.265	.812	.265	.728	.812	.901	.265
$P_{observed}$	.126	**.328**	.119	.080	**.925**	.034	.206	**.464**
Power	.735	.735	.188	.735	.272	.188	.099	.735
Correlations with TBV								
Pearson's R	−.078	−.197	.287	−.011	−.19	.515	−.282	.496
1-tailed P	.41	.28	.2	.49	.48	.05	.2	.06
Reproducibility								
ICC	.998	.997	.965	.994	.999	N/A	N/A	N/A
RC, mm^3	19,900	15,000	91,200	35,300	11,300	N/A	N/A	N/A
CV, %	.42	.38	2.22	.65	.31	N/A	N/A	N/A
Labor, min	9.9	5.6	20.7	12.2	6.4	>38 hrs	20 hrs	15
Slices included								
Median slice number	19	10	17	19	10	N/A	N/A	N/A
Slice range	17-20	N/A	16-19	16-20	N/A	N/A	N/A	N/A
Estimated TIV								
Mean TIV, dm^3	1.5	1.09	1.57	1.55	1.11	1.63	1.64	1.81
TIV 95% CI, dm^3	1.43-1.57	1.05-1.12	1.50-1.65	1.48-1.62	1.07-1.15	1.54-1.71	1.54-1.74	1.73-1.89

Bold denotes significant result (ie, the method fails to reject the null hypothesis).
Note that the relative percentage difference corresponds to the 90% CI method.
MAPD = mean absolute percentage difference; ICC = intraclass correlation coefficient; RC = Bland & Altman repeatability coefficient; CV = coefficient of variation; CI = confidence interval; N/A = not applicable.

average percentage rate of atrophy was 1.99% per year, as illustrated in Figure 4B.

The change in TIV estimation from time 1 to time 2, depending on method, is illustrated in Figure 4A using the *confidence interval* method (see also Table 2 for values). This showed that the MCPD, MCT2, and SPM methods were the most stable in representing TIV independent of atrophy. Applying a zone of indifference of ±2% (Fig 4A, bold horizontal lines) shows that the FCPD, MCPD, FCT2, MCT2, and SPM methods fail to reject the null hypothesis (ie, there is no significant difference in TIV with time).

However, the FCT1 and both FS methods showed significant perturbation of TIV with time (and therefore atrophy), as shown by their 90% CI crossing out of the zone

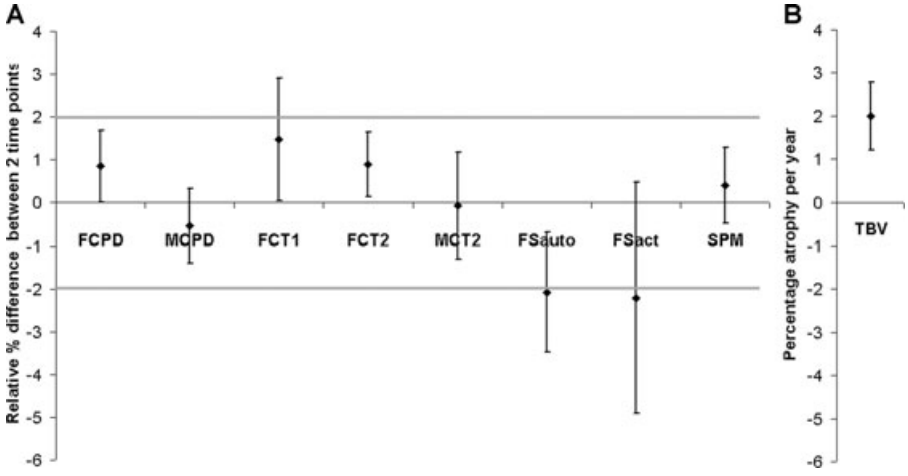

Fig 4. (A) The average difference between time point 1 and time point 2, in chronological order, according to method used, is represented. Zero difference represents absolute agreement, a positive value indicates that TIV at time 2 < TIV at time 1 (underestimation at time 2), while a negative value indicates that TIV at time 2 > TIV at time 1 (overestimation at time 2). Within the horizontal lines lies the zone of indifference (+2 to –2%). Error bars represent the 90% CI. (B) The percentage rate of atrophy (loss in TBV) per year as estimated by SPM. Error bars represent the 95% CI.

of indifference. The FCT1 method underestimated the TIV and FS overestimated the TIV as atrophy progressed. Interestingly, when the FS transform was adjusted to correct the final cost function value to <.1 (FS_{ACT}), the mean error increased.

When the *trade-off* method was applied,[29] only the MCPD, MCT2, and SPM methods succeeded in upholding the null hypothesis (see Table 2). The results remain unchanged if β is relaxed to .1.

Pearson's product moment correlation coefficients of TIV change (time 1 – time 2) against TBV change were derived for each method (see Table 2): for the mid-cranial and full-cranial PD and T2 methods, R^2 values were small (all <.05) and 1-tailed P values were high (all >.25), indicating that increasing atrophy did not influence TIV estimation. The FCT1 method was slightly worse, but probably acceptable. However, results of both FS_{AUTO} and SPM indicated a trend toward correlation of atrophy with TIV. In spite of the greater mean error with FS_{ACT}, the correlation was diminished compared to FS_{AUTO}.

There was high correlation between the full-cranial and mid-cranial measurements within each imaging modality (Pearson correlations: FCPD and MCPD $R^2 = .732$, $P < .001$; FCT2 and MCT2 $R^2 = .715$, $P < .001$) (see Fig 5).

Reproducibility of Manual Methods

The Intraclass Correlation Coefficient

The intraclass correlation coefficients are shown in Table 2.

The Bland and Altman Repeatability Coefficient

The repeatability coefficient values derived by this method are shown in Table 2. The interpretation of these results is that, for example, by the MCPD method, the next time the TIV is estimated, it has a 95% chance of being $\pm 15{,}000$ mm^3 different from the "true" value. This clearly shows that the most reproducible methods are in order: MCT2 > MCPD > FCPD > FCT2 > FCT1. Note the large difference between FCT1 and the other methods. When graphically represented, this analysis shows that the FCT1 method includes an outlier, yet even without including that in the analysis, the spread of the remaining values for FCT1 is still greater than for any of the other methods (results not shown).

Coefficient of Variation

The CV values are shown in Table 2. The same order of reproducibility as derived by the Bland and Altman method was also reflected by the CV.

Labor Intensity

The average time taken per scan for TIV estimation, for each manual method, is shown in Table 2. The automated methods only depend on the processing time. The SPM method requires the segmentation of the scans, which can take up to 15 minutes per scan on a DELL PowerEdge 1850 workstation, equipped with 2 Jewell 2.8 GHz Xeon processors and 2 GB RAM, running a 64-bit Linux operating system (Debian 3.1; SPI Inc., Indianapolis, IN). FS requires approximately 20 hours (FS_{AUTO}, autorecon1 and 2) or more to process a subject on the same workstation. FS_{ACT} requires around 15 minutes for autorecon1 and an additional time of between 1 and 5 minutes of intervention depending on the type of adjustment required; nonetheless, after such adjustments, another 18-hour run is required (autorecon1 and 2).

Discussion

By definition, TIV estimation should be independent of atrophy; yet, to our knowledge, there have been only 2 case reports in the literature where a patient was longitudinally followed with serial TIV and TBV estimations[12,18] and no study has compared the different TIV methods longitudinally. In this study, we directly contrasted several methods for estimating TIV. In addition to within-scan reliability measures, the main novelty of the work was that we used longitudinal scan data to specifically address the important issue of whether TIV methods are influenced by brain atrophy.

From the methods evaluated, the MCPD method performed very well. It showed high precision across time that was not affected by progressive atrophy. Its reproducibility was very high and it was both fast and user-friendly. The FCT2 method was also within our predefined zone of indifference. Its reliability was adequate, but, notably, the Bland and Altman repeatability

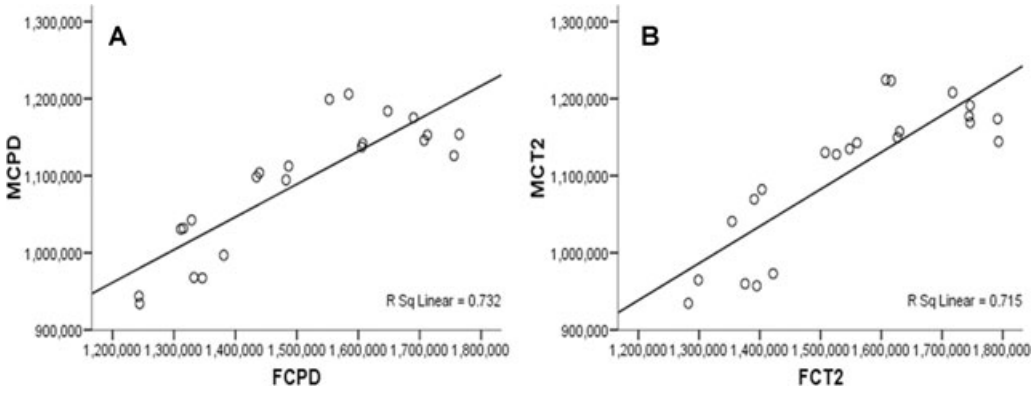

Fig 5. Correlation of full-cranial with corresponding mid-cranial manual methods using proton-density (A) and T2-weighted (B) sequences. R^2 linear = linear interpolation, Pearson's correlation coefficient squared (the coefficient of determination).

coefficient was twice as high as that for the MCPD method. Also, it was moderately more labor-intensive. The discrepancy between MCPD and FCT2 methods was explained by the difference in slice samplings (mid- vs. full-cranial) rather than by the different acquisition methods (PD vs. T2). The results of the FCPD and MCT2 methods indicated that the reproducibility improves and the effect of atrophy diminishes when 10 supratentorial slices are used to estimate TIV, and that the MCT2 method is very similar to the MCPD method. In other words, the difference between the FCT2 and MCT2 (or the FCPD and MCPD) methods lie in increased error in defining the intracranial border at the base and vertex. These results confirm that our prior concerns regarding full-cranial T2 and PD methods were justified. There were significant variations in cross-sectional area of the inferior-most cerebellar slice and excluding the lower cerebellum removed this random error. The superior-most slices at the vertex of the brain displayed partial volume effects due to the high slice thickness (7 mm), leading to inconsistencies in defining a clear intracranial boundary. It was for these reasons that the mid-cranial sampling approach was developed.

It is worth noting that neither of these issues was present in the T1 volumetric scans. The fine slice thickness meant that the caudal limit of the cerebellum was consistently defined and that there were minimal partial volume effects at the vertex. Yet despite these advantages, FCT1 was outperformed. Even though the MAPD for the FCT1 method was adequate (1.94%), it consistently underestimated TIV in progressive atrophy. This has clear implications in dementia research and suggests that TIV estimation by this method is less desirable compared to MCPD and MCT2 methods for brain volume correction. Also, it was the most labor-intensive method, on average 20 minutes per scan, often requiring a considerable degree of manual editing. We propose that the inferior results for the FCT1 method were due to problems with this sequence in consistently delineating an intracranial dural boundary. The dura in T1-weighted images is often indistinct and this leads to extensive manual editing, which is inherently less accurate and less reproducible. In contrast to the current results, Whitwell et al.[12] reported considerably better reliability using T1-weighted data with the same sampling protocol and intensity threshold (CV .16% vs. 2.22% in the current study). However, it must be acknowledged that different image-processing software were used in the two studies: ANALYZE in this study and MIDAS in Whitwell et al.'s study. It is conceivable that this might have affected the results obtained by the FCT1 method.

Turning to the correlation between change in TIV estimation and change in TBV, the MCPD, MCT2, FCPD, FCT2, and FCT1 methods, all yielded results suggesting that change in atrophy did not influence change in TIV measurement. Our interpretation is that atrophy did not introduce a systematic error in TIV estimation by these methods–the influence of atrophy, as, for example, seen in FCT1, must therefore reflect an increase in random error. This accords with our experience in using FCT1 that increased manual editing would be expected to increase random error.

The automated methods produced mixed results. In terms of MAPD, the SPM5 analysis was more precise than the FCT1, MCT2, and FS methods and very close to the most accurate manual methods. As SPM5 is fully automated, it does not suffer with human rater reliability or labor intensity issues. In fact, if data are being prepared for an SPM analysis, the information required to calculate TIV is virtually derived en passant during the segmentation step, making it a particularly attractive method as it requires only some very basic scripting. In terms of the influence of atrophy in TIV estimation, the CI and the trade-off methods[29] (see Fig 4 and Table 2) showed that SPM5, with a .5 threshold, consistently failed to reject the null hypothesis (ie, not influenced by atrophy), but the correlation analysis of change in TBV with change in TIV suggested a trend toward influence of atrophy. This interpretation should be considered with caution, however, since the TBV used in this analysis was obtained from the same SPM-derived measurements as those used to derive the TIV; hence, at least part of the increased correlation with TBV by the SPM-derived TIV can be explained by this fact. Overall, SPM5 with a .5 threshold can provide an automated and reliable method of deriving TIV, which may be acceptable in situations where the highest level of precision is deemed unnecessary. It should be noted that the threshold of .5 was chosen arbitrarily. As the threshold is relaxed, the agreement between 2 TIV estimations improves, but the TIV derived is not anatomically accurate as more soft tissue (tissue class: other) voxels are included. The converse is true if the threshold is too strict (more anatomical accuracy at the cost of increased variability between measurements at 2 time points). Related to this is the observation that the mean estimated TIV was significantly greater for SPM than for other methods. This is partly due to the fact that the manual methods model the TIV as consecutive thick slices (without interpolation) and, therefore, systematically underestimate TIV compared to SPM (see also Methods: Manual intracranial volume estimation methods: Full-cranial T1 (FCT1)-weighted method). However, visual inspection of the segmented volumes showed that a small amount of "other" tissue class voxels was included, eg, sagittal venous sinus, optic nerves, and dura. These voxels and the resultant TIV were reduced as the threshold was increased, but at the cost of increased variability in TIV estimation between time intervals (data not shown). We propose that a threshold of .5 is an acceptable compromise between anatomical accuracy and precision of TIV estimation in time.

As an aside, an interesting point for consideration in voxel-based morphometric studies is that a voxel-by-voxel measure, and not TIV, could be a more desirable means of removing interindividual variability of noninterest from the signal of interest. However, voxel-based morphometry is a mass univariate approach, and hence a voxel-by-voxel measure would require nontrivial changes to the statistical model used in the current software packages (eg, SPM). TIV is a gross (global) measure and the authors acknowledge that it fails to provide a detailed account of intersubject nuisance variability, but its ease of calculation and insertion into the statistical model make it a desirable covariate.

On the other hand, both FS methods, FS_{AUTO} and FS_{ACT}, were very imprecise. This can be seen from the MAPD values in Figure 3 (average of 2.5% for FS_{AUTO} and 3.2% for FS_{ACT}) and from Figure 4, in which it is clear that neither method fell within

the zone of indifference of the CI method. It was also observed that both methods consistently overestimated the TIV for the second scan (Fig 4), indicating that the method is susceptible to atrophy in that precision declines with increasing brain volume loss. FS_{AUTO} also showed a trend toward a significant relationship between TIV and TBV. Although the relationship was weak (1-tailed, $P = .05$), given that TIV measurement should be independent of atrophic change, we feel that this trend raises grave concerns about the use of this method in degenerative diseases.

Another interesting observation was that optimization through correction of the affine transform used to obtain the FS_{ACT} TIV was not helpful; in fact, when compared to FS_{AUTO}, the output was less precise in terms of MAPD (Fig 3) and the overestimation phenomenon worsened (Fig 4). This may be related to how the correction was performed: by attempting to reduce the final cost function value, which can be regarded as a similarity coefficient between the atlas and the image being fitted, it is possible that the algorithm fails by overfitting the atrophic brain to the atlas, hence achieving a good numerical result, but a poor actual fit. Interestingly, although the mean error was worse with FS_{ACT}, there was no correlation trend of TIV and TBV changes. These findings suggest that the inferior results with FS_{ACT}, compared to FS_{AUTO}, were a consequence of greater random error rather than a systematic worsening secondary to the influence of atrophy.

In considering the poor performance of FS eTIV, certain caveats need to be discussed. Firstly, FS is optimized for Magnetisation Prepared Rapid Gradient Echo (MPRAGE)-acquired images and, although this is of greater relevance for segmentation, we cannot exclude that our use of SPGR-acquired scans may have adversely affected the results. Another factor is atlas dependency: given that FS TIV estimation hinges on the affine transformation, the target space atlas is a critical determinant, ie, it should be a good representation of the subjects being studied. As previously stated, the FS atlas uses 40 subjects spanning a wide age range, including 10 AD cases, whereas our cohort was limited to a narrow age range (54-79 years), each with marked, but nonuniform, atrophy. This may help explain the difference between the current results and those of Buckner et al.,[18] who reported a maximum absolute percentage difference of 1.11% between the 2 most discrepant points in an AD patient who was scanned repeatedly over 4 years. The current results obtained with FS_{AUTO} show an absolute percentage difference range from .3% to 6.6%, thus the reported 1.11% value falls within the present range. This suggests that FS eTIV may not be as invulnerable to atrophy as initially described. One final qualifier is that the FS algorithm has been altered since Buckner et al.'s report, although this change, ironically, was intended as an optimization. In summary, although there are numerous caveats to the current, disappointing FS results, they indicate that FS eTIV estimation may not be accurate in all circumstances and that validation in any given cohort of interest is necessary. This leads to perhaps the most critical issue with respect to the FS methods: there is a floor effect in the error incurred that is derived from the atlas measurement of TIV. This means that both FS_{AUTO} and FS_{ACT} have their performances limited by the best possible (currently manual) TIV measurement methods and it is assumed that extra variance (error) will be added by their implementation. In other words, FS eTIV can approach the accuracy of the measured TIV for a given atlas, but cannot exceed it. In addition, introduction of manual TIV atlas measurements in order for FS to produce adequate results increases the labor intensity and, given that its performance is limited by the manual method chosen for estimating the atlas TIV, there are no clear benefits to using FS for TIV estimation.

There are several comments on this study that warrant mention. Although ours is the first longitudinal study in a group (ie, beyond single-case studies) of patients with semantic dementia comparing the precision and reliability of different TIV methods, we acknowledge that our small group size ($n = 11$) is a limitation. Similar studies with larger group sizes and in other neurodegenerative conditions are required to validate our findings. Another limitation is that all measurements were performed by a single rater and, therefore, only intrarater reliability could be assessed. Assessing interrater reliability would be an important aspect of future work. Furthermore, our study was focused on methods of deriving a measure of TIV, but another important topic for future work is the mathematical use of TIV to correct volumes of interest as different methods have been used in the literature. Another issue is that because PD and T2 sequences were not volumetric, no correction of head position was made to any scans acquired. Therefore, there is a theoretical effect of head position that may impact more on mid-cranial than full-cranial methods. Despite this concern, the mid-cranial methods have performed best in terms of precision across time, suggesting that this potential added variation is not having a significantly detrimental effect. Another potential limitation of mid-cranial methods is that subjects with larger heads would have less TIV sampled if only 10 slices are included than subjects with smaller heads. As already discussed, the mid-cranial methods sample approximately 70% of full-cranial TIV. When MCPD was compared to FCPD and MCT2 to FCT2 by gender, we found that MCPD was 71% of FCPD and MCT2 was 71% of FCT2 in males, whereas in females this was 75% and 73%, respectively (ie, a difference of 2-4%), even though the female TIVs were 13-18% (depending on the method) smaller than male TIVs. This suggests that there is no gross oversampling of TIV by mid-cranial methods.

In conclusion, this study found differences among TIV measurement techniques in terms of reproducibility and, more importantly, in terms of the impact of atrophy. Taking all evaluation procedures into account, the best methods, overall, were the MCPD and MCT2 methods. Their levels of precision were high and not influenced by atrophy, they showed excellent reproducibility, and they can be performed within 5-6 minutes. The results suggest that the novel mid-cranial limits described are the most important determinants of precision. The PD sequences are particularly easy and fast to trace as they require only one "seed and expand" action per slice and very little manual editing. Although the precision results were similar with mid-cranial T2 sequences, this method is more labor-intensive as the subarachnoid CSF is often not continuous on axial images and, therefore, may require multiple "seed and expand" actions. Of the 2 fully automated methods tested, SPM5 performed well and, in certain circumstances where the highest

level of precision is not deemed necessary, it is an attractive option as it has perfect reproducibility and does not suffer from the labor-intensity issues attendant on manual methods.

Peter J. Nestor is funded by the Medical Research Council, United Kingdom. We gratefully acknowledge Professor John R. Hodges for identifying patients as well as the patients themselves and their relatives for participating in the research studies.

References

1. Pfefferbaum A, Mathalon DH, Sullivan EV, Rawles JM, Zipursky RB, Lim KO. A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Arch Neurol* 1994;51(9):874-887.
2. Blatter DD, Bigler ED, Gale SD, Johnson SC, Anderson CV, Burnett BM, Parker N, Kurth S, Horn SD. Quantitative volumetric analysis of brain MR: normative database spanning 5 decades of life. *AJNR Am J Neuroradiol* 1995;16(2):241-251.
3. Watson C, Jack CR Jr, Cendes F. Volumetric magnetic resonance imaging. Clinical applications and contributions to the understanding of temporal lobe epilepsy. *Arch Neurol* 1997;54(12):1521-1531.
4. Dalton CM, Chard DT, Davies GR, Miszkiel KA, Altmann DR, Fernando K, Plant GT, Thompson AJ, Miller DH. Early development of multiple sclerosis is associated with progressive grey matter atrophy in patients presenting with clinically isolated syndromes. *Brain* 2004;127(Pt 5):1101-1107.
5. Galton CJ, Patterson K, Graham K, Lambon-Ralph MA, Williams G, Antoun N, Sahakian BJ, Hodges JR. Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology* 2001;57(2):216-225.
6. Fotenos AF, Snyder AZ, Girton LE, Morris JC, Buckner RL. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology* 2005;64(6):1032-1039.
7. Laakso MP, Partanen K, Riekkinen P, Lehtovirta M, Helkala EL, Hallikainen M, Hanninen T, Vainio P, Soininen H. Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia: an MRI study. *Neurology* 1996;46(3):678-681.
8. Lye TC, Grayson DA, Creasey H, Piguet O, Bennett HP, Ridley LJ, Kril JJ, Broe GA. Predicting memory performance in normal ageing using different measures of hippocampal size. *Neuroradiology* 2006;48(2):90-99.
9. Free SL, Bergin PS, Fish DR, Cook MJ, Shorvon SD, Stevens JM. Methods for normalization of hippocampal volumes measured with MR. *AJNR Am J Neuroradiol* 1995;16(4):637-643.
10. Hartley SW, Scher AI, Korf ES, White LR, Launer LJ. Analysis and validation of automated skull stripping tools: a validation study based on 296 MR images from the Honolulu Asia aging study. *Neuroimage* 2006;30(4):1179-1186.
11. Palm WM, Walchenbach R, Bruinsma B, Admiraal-Behloul F, Middelkoop HA, Launer LJ, van der Grond J, van Buchem MA. Intracranial compartment volumes in normal pressure hydrocephalus: volumetric assessment versus outcome. *AJNR Am J Neuroradiol* 2006;27(1):76-79.
12. Whitwell JL, Crum WR, Watt HC, Fox NC. Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging. *AJNR Am J Neuroradiol* 2001;22(8):1483-1489.
13. Jenkins R, Fox NC, Rossor AM, Harvey RJ, Rossor MN. Intracranial volume and Alzheimer disease: evidence against the cerebral reserve hypothesis. *Arch Neurol* 2000;57(2):220-224.
14. Jack CR Jr, Twomey CK, Zinsmeister AR, Sharbrough FW, Petersen RC, Cascino GD. Anterior temporal lobes and hippocampal formations: normative volumetric measurements from MR images in young adults. *Radiology* 1989;172(2):549-554.
15. Eritaia J, Wood SJ, Stuart GW, Bridle N, Dudgeon P, Maruff P, Velakoulis D, Pantelis C. An optimized method for estimating intracranial volume from magnetic resonance images. *Magn Reson Med* 2000;44(6):973-977.
16. Perry RJ, Graham A, Williams G, Rosen H, Erzinclioglu S, Weiner M, Miller B, Hodges J. Patterns of frontal lobe atrophy in frontotemporal dementia: a volumetric MRI study. *Dement Geriatr Cogn Disord* 2006;22(4):278-287.
17. Maller JJ, Reglade-Meslin C, Anstey KJ, Sachdev P. Sex and symmetry differences in hippocampal volumetrics: before and beyond the opening of the crus of the fornix. *Hippocampus* 2006;16(1):80-90.
18. Buckner RL, Head D, Parker J, Fotenos AF, Marcus D, Morris JC, Snyder AZ. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *Neuroimage* 2004;23(2):724-738.
19. Callen DJ, Black SE, Gao F, Caldwell CB, Szalai JP. Beyond the hippocampus: MRI volumetry confirms widespread limbic atrophy in AD. *Neurology* 2001;57(9):1669-1674.
20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307-310.
21. Hodges JR, Patterson K, Oxbury S, Funnell E. Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain* 1992;115(Pt 6):1783-1806.
22. Neary D, Snowden JS, Gustafson L, Passant U, Stuss D, Black S, Freedman M, Kertesz A, Robert PH, Albert M, Boone K, Miller BL, Cummings J, Benson DF. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. *Neurology* 1998;51(6):1546-1554.
23. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage* 2005;26(3):839-851.
24. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 1999;9(2):179-194.
25. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 1999;9(2):195-207.
26. Fleiss JL. Reliability of Measurement. In the Design and Analysis of Clinical Experiments. New York, NY: John Wiley & Sons, 1986;1-32.
27. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil* 1998;12(3):187-199.
28. Rudick RA, Fisher E, Lee JC, Simon J, Jacobs L. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Multiple Sclerosis Collaborative Research Group. *Neurology* 1999;53(8):1698-1704.
29. Lew MJ. Principles: when there should be no difference–how to fail to reject the null hypothesis. *Trends Pharmacol Sci* 2006;27(5):274-278.