SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Neurodegenerative diseases

» Brain imaging

» Neuroscience

» Magnetic resonance imaging

Reliability of brain volume measurements: A test-retest dataset

Julian Maclaren[1], Zhaoying Han[1], Sjoerd B. Vos[1,2], Nancy Fischbein[1] & Roland Bammer[1]

Received: 21 March 2014

Accepted: 02 September 2014

Published: 14 October 2014

Evaluation of neurodegenerative disease progression may be assisted by quantification of the volume of structures in the human brain using magnetic resonance imaging (MRI). Automated segmentation software has improved the feasibility of this approach, but often the reliability of measurements is uncertain. We have established a unique dataset to assess the repeatability of brain segmentation and analysis methods. We acquired 120 T1-weighted volumes from 3 subjects (40 volumes/subject) in 20 sessions spanning 31 days, using the protocol recommended by the Alzheimer's Disease Neuroimaging Initiative (ADNI). Each subject was scanned twice within each session, with repositioning between the two scans, allowing determination of test-retest reliability both within a single session (intra-session) and from day to day (inter-session). To demonstrate the application of the dataset, all 3D volumes were processed using FreeSurfer v5.1. The coefficient of variation of volumetric measurements was between 1.6% (caudate) and 6.1% (thalamus). Inter-session variability exceeded intra-session variability for lateral ventricle volume ($P < 0.0001$), indicating that ventricle volume in the subjects varied between days.

Design Type(s)	observation design ● reference design ● replicate design ● time series design
Measurement Type(s)	structural magnetic resonance imaging
Technology Type(s)	Magnetic Resonance Imaging
Factor Type(s)	session
Sample Characteristic(s)	*Homo sapiens* ● brain

[1]Center for Quantitative Neuroimaging, Department of Radiology, Stanford University, Stanford, California 94305, USA. [2]Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands. Correspondence and requests for materials should be addressed to J.M. (email: julian.maclaren@stanford.edu).

Background & Summary

Quantitative measurement of the volume of human brain structures using magnetic resonance imaging (MRI) may have application in the monitoring of neurodegenerative disease progression or the assessment of general structural abnormalities[1–3]. Traditionally, volumetric data for this purpose have been obtained using time-consuming manual segmentation methods. Developments in freely available automated segmentation software, however, have greatly improved the practicality of the process[4–8]. A further recent development has been the introduction of standardized imaging protocols and online databases containing structural MRI brain data from both patients and normal controls. One notable example is the database and set of protocols made available by the MRI core of the Alzheimer's Disease Neuroimaging Initiative (ADNI)[9].

The combination of freely available data and automatic segmentation software has the potential to make volumetric brain MRI an indispensible tool for both research and clinical medicine. It is now possible to compare quantitative brain volume measurements from a patient with control data from databases of normal subjects within only a few minutes[10]. Since brain volume measurements are a known biomarker for neurodegenerative diseases, such as Alzheimer's disease[2,11], this general approach may have great clinical value as a screening tool. Performing volumetric measurements at multiple time points in the same patient may also be a powerful means of measuring changes in the brain. Recent longitudinal results have shown that atrophy in medial temporal lobe (MTL) structures is predictive of subsequent memory decline in elderly subjects[12]. Additionally, performing longitudinal MRI studies improves predictive power in assessing the prognosis of individuals with mild cognitive impairment (MCI)[13], and predicting the development of Alzheimer's disease in non-demented subjects[14] as well as MCI patients[15].

We have begun to investigate the feasibility of employing the above approach at our institution. However, clinical implementation requires thorough validation, and one existing cause for concern regarding this methodology is that the repeatability (i.e., test-retest reliability) of individual measurements is still largely unknown. This prevents robust comparison of a single patient measurement against those in a database of normal subjects. It also confounds the analysis of longitudinal measurements from an individual, since it is crucial to know whether any apparent trend is simply due to measurement variation. That is, without knowing the fluctuations of our measurements in a single subject, we cannot be certain what disease effects we can detect. A number of confounding factors can be controlled, such as the scanner, imaging protocol, and software version used for data processing. MRI data are inherently noisy, however, and effects such as patient positioning and field of view selection are not constant between scans. There is also evidence of possible day-to-day variations in the brain structures of interest, for example due to hydration levels[16,17] or the menstrual cycle[18].

The goal of this study was to generate a publicly available test dataset to assist in the validation of the repeatability of current and future segmentation methods. We believe that this will be of great utility to other researchers who may not have the imaging facilities or resources to collect such a dataset. To demonstrate the utility of this dataset, we use it to obtain basic test-retest reliability information about the volume data corresponding to brain regions of clinical interest. This includes an analysis to test the hypothesis that test-retest reliability (repeatability) of measurements can be measured using multiple scans in a single session (intra-session) and that changes from day to day (inter-session) do not affect repeatability.

Methods

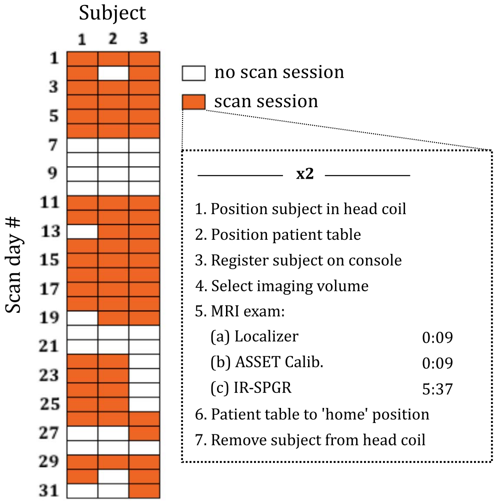
An experiment was designed to allow separate calculation of intra- and inter-session test-retest reliability (Figure 1). Human data collection was performed with the approval of the Stanford University Institutional Review Board (IRB) and all subjects gave their written informed consent. A total of 120 T1-weighted volumes were acquired from 3 subjects (40 scans/subject). Each subject was scanned twice on twenty different days within a 31-day period. Subjects were repositioned between the two scans in each session, and reregistered on the scanner console, so that all scans were treated as separate measurements (with a resulting break of ~5 min between scans). This ensured that confounding effects such as subject positioning were consistent between and within sessions.

We used a GE MR750 3T scanner (software version DV22.0_V02_1122.a, XRMB gradient set) and the ADNI-recommended T1-weighted imaging protocol for this system (accelerated sagittal 3D IR-SPGR, 27 cm FOV, 256 x 256 matrix slice, 1.2 mm slice thickness, standard 8-channel phased array head coil, TR: 7.3 ms, TE: 3 ms, TI: 400 ms, flip angle: 11 deg., acquisition time: 5 min 37 s). We chose to use the ADNI protocol[9] because it exhibits excellent contrast between gray and white matter and has become standardized, so it is relevant to the research community.

Other data recorded throughout the study included subject weight, time of day, and phantom data for quality assurance and scanner stability. Both the ADNI phantom[19] and the ACR (American College of Radiology) phantom were used for this purpose.

Data Records

Following the guidelines of our IRB, written informed consent was obtained from each subject allowing their data to be shared online. The test-retest dataset is available through FigShare (Data Citation 1). It consists of a single zip file containing the following:



x2

1. Position subject in head coil
2. Position patient table
3. Register subject on console
4. Select imaging volume
5. MRI exam:
   (a) Localizer 0:09
   (b) ASSET Calib. 0:09
   (c) IR-SPGR 5:37
6. Patient table to 'home' position
7. Remove subject from head coil

Figure 1. Experimental protocol. Data were acquired in 60 scan sessions (20 per subject); each session comprised two back-to-back scans ('intra-session') where the subject was removed from the scanner and re-registered on the system between scans, in order to replicate effects between days ('inter-session').

- A folder containing the brain imaging data, comprising 120 data files, separated by subject number into three folders. Each file contains an anonymized MRI brain volume in NifTI-1 format. Note that these files have not been processed by FreeSurfer in any way and are not corrected for B1 inhomogeneity or otherwise modified. Conversion from DICOM format was performed using a publicly available and often-used converter, known as 'dcm2nii', which is available online at http://www.mccauslandcenter.sc.edu/mricro/mricron/dcm2nii.html, together with extensive documentation. More information about the NIfTI file format is available from the Neuroimaging Informatics Technology Initiative webpage at http://nifti.nimh.nih.gov/.
- A text file ('anonymized_dicom_header.txt') containing the anonymized DICOM header from Subject 2. Researchers can use this file to obtain any extra information about the scan protocol that is not adequately described in this manuscript.
- Three files in Microsoft Excel format, indicating the day and time each image volume was acquired from each subject, the weight of each subject at the time of each scan, and the age and gender of each subject.
- DICOM and NifTI files corresponding to 3D scans of the ADNI phantom and ACR phantom.

Technical Validation

All 3D volumes were visually inspected at the time of acquisition to check for severe motion corruption or any other potentially confounding artifacts. No significant or unexpected artifacts were observed, and so no scans were repeated. Regular QA tests on the scanner did not indicate any performance issues during the 31-day period of the study. To further validate the technical quality of the dataset, and to demonstrate its utility, we processed and analyzed the data using the methods described in this section.

Data processing

All 3D volumes were processed using FreeSurfer[6–8] (freely available online at http://surfer.nmr.mgh.harvard.edu/), which provides quantitative volume data for a range of brain structures. We ran FreeSurfer without user intervention ('*recon-all –i data1.dcm –all*'), since this is the mode of operation that would be used in an automated pipeline processing patient data. The computing hardware, operating system (Ubuntu 12.04.1) and FreeSurfer version (v5.1) were kept constant for the duration of this study, in line with recommendations made by Gronenschild *et al.*[20], who showed that these factors can affect results. We chose to restrict our analysis to the lateral ventricle volumes as well as a select group of subcortical structures of particular interest to us, including the hippocampus, amygdala, putamen, pallidum, caudate and thalamus. In all cases, volumes from the left and right hemispheres were summed together to generate a single value for each structure.

Data analysis

We applied the following statistical analysis to the FreeSurfer output to assess variability for measurements obtained together on the same day (intra-session) and measurements from day to day (inter-session). Paired acquisitions allowed the intra-session variability to be computed using the expression for standard deviation from paired data[21]. The following analysis was applied to each brain structure of interest and for each subject. Denoting the standard deviation of intra-session measurements by σ_s, then

$$\sigma_s = \sqrt{\sum \left(x_i{'} - x_i^{''}\right)^2 / 2m},$$

where $x_i{'}$ and $x_i{''}$ are the ith paired measurements over the m pairs (or sessions), evaluated separately for each subject and assuming that the variance is the same for all samples. Intra-session variability reflects manual repositioning differences, noise and segmentation errors, but not biological variations occurring from day to day. The total standard deviation, σ_t, is computed across all 40 measurements from each subject.

In each case, the standard deviation is then converted to coefficient of variation (CV) expressed as a percentage, namely

$$CV_s = \frac{\sigma_s}{\bar{x}} \times 100$$

$$CV_t = \frac{\sigma_t}{\bar{x}} \times 100$$

where CV_s is the intra-session coefficient of variation, CV_t is the total coefficient of variation, and $\bar{x}$ is the arithmetic mean of all 40 volume values for the given structure and subject. We use the coefficient of variation since the normalization by the mean volume is somewhat more intuitive than absolute standard deviations. This is particularly the case when comparing between the segmentation repeatability of different structures, which may have very different volumes. Finally, CV_s and CV_t data from each subject were pooled by taking the root-mean-square average, as proposed by Gluer *et al.*[22], since taking the arithmetic mean directly is not valid.

A Monte Carlo permutation test[23] was used to evaluate observed differences between CV_t and CV_s for each structure under the null hypothesis of 'no difference', i.e., that $CV_s = CV_t$. In brief, the scan order of all 40 volumes for each subject is randomly permuted N times. The above analysis is repeated for each permutation and the value obtained for $|CV_s - CV_t|$ is recorded. The two-sided P-value is then calculated as the proportion of the N permutations where the absolute difference $|CV_s - CV_t|$ was equal to or greater than the value obtained with the measured data. In this work, we used $N = 100{,}000$.

Data analysis results

Figure 2 shows typical segmentation results obtained using FreeSurfer. The volumetric data segmented in this example were acquired from Subject 2 on Day 1 of the study. Despite generally good segmentation accuracy, minor differences are visible between the two scans from one session. This is to be expected, since the input data differ slightly due to noise, subject positioning, and other confounding factors.

Quantitative differences between segmentation volumes are shown in Figure 3 for several brain structures from Subject 2. For the thalamus, caudate, pallidum and hippocampus, the differences between each scan visually appear to be random. However, in the case of the lateral ventricles it is evident that inter-session variance exceeds intra-session variance, since values obtained on the same day (neighboring bars in the plot) are typically closer in magnitude to each other than values obtained on different days.

Results from the numerical analysis show total coefficient of variation values between 1.6% (caudate) and 6.1% (thalamus) for the structures analyzed (Table 1). The mean structure volume appears to have little influence on the coefficient of variation value. The lateral ventricles are the only structures showing a

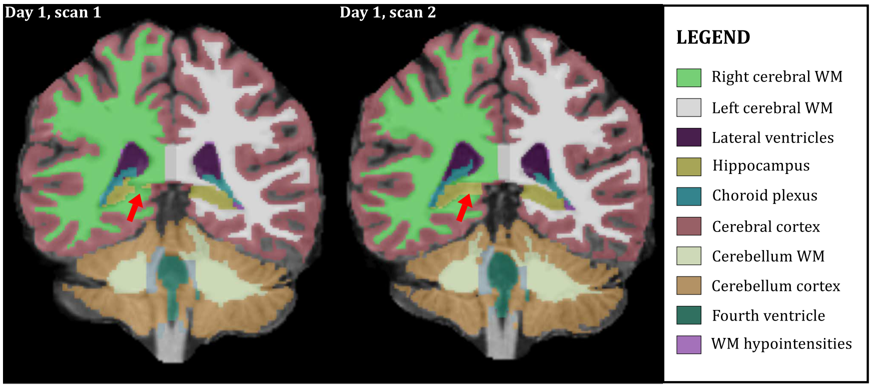


Figure 2. Example FreeSurfer segmentation. 3D volumes were segmented using FreeSurfer v5.1. Segmentation differences, due to effects such as image noise, affect repeatability. Here, voxels at the anterior end of the right hippocampus (red arrows) are classified differently in two volumes from Subject 2 acquired in the same session (Day 1).

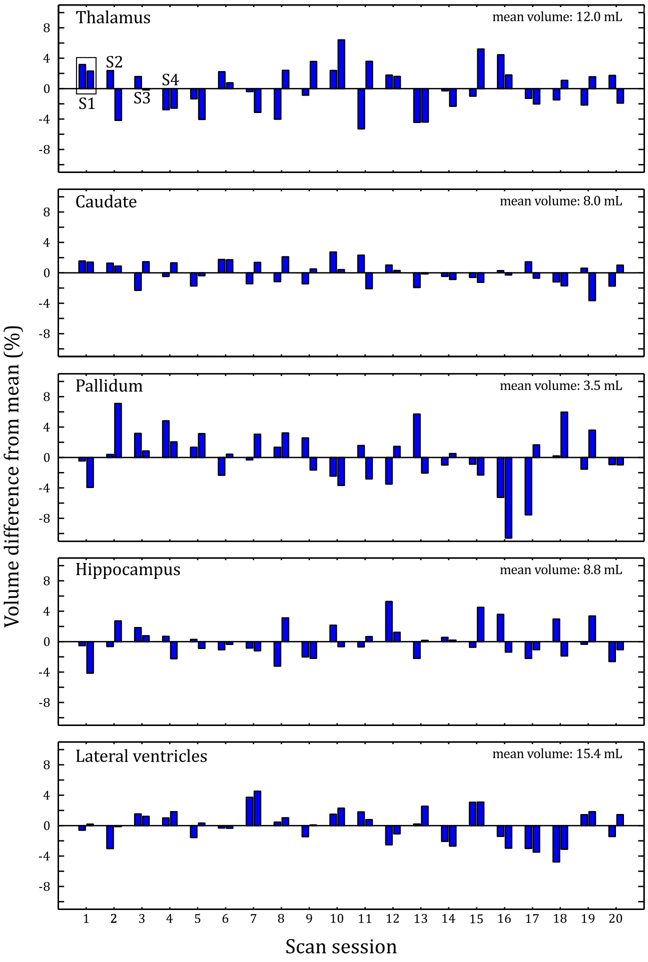


Figure 3. Scan-to-scan volume differences. Segmentation results for five brain regions in a single subject (Subject 2). Each of the 40 bars for each region represents a single scan. The bars are paired to indicate which measurements were obtained within the same session, as shown for the thalamus for the first four sessions (top left corner of plot).

Structure	Mean volume (ml)	CV_s (intra-session)	CV_t (total)	$\vert CV_t – CV_s \vert$	Two-sided *P*-value
Hippocampus	8.9	2.77%	2.92%	0.16%	0.41
Lateral ventricles	10.1	1.58%	3.40%	1.82%	< 0.0001
Amygdala	3.8	4.69%	5.21%	0.53%	0.12
Putamen	11.6	4.04%	3.92%	0.13%	0.68
Pallidum	3.2	5.25%	5.42%	0.17%	0.67
Caudate	7.4	1.54%	1.58%	0.04%	0.67
Thalamus	12.9	5.98%	6.06%	0.08%	0.88
Cerebral white matter	496.6	0.88%	0.87%	0.00%	0.96

Table 1. Measurement variation by structure. Intra-session (CV_s) and total variation (CV_t) shown as standard deviations, expressed as percentages of the mean value for each subject (i.e., the coefficient of variation). Values are averaged over all three subjects. The *P*-values shown indicate the likelihood of obtaining the observed difference, or greater, between CV_s and CV_t, when in fact there is no underlying difference between them.

statistically significant difference between intra- and inter-session variability. This difference is also substantial: accounting for variance between sessions more than doubles the total coefficient of variation for lateral ventricle volume from 1.6% to 3.4%.

Based on the results shown in Table 1, one might expect that true ventricle size changes are occurring between days. To investigate this further, registered volumes were compared directly. Figure 4 shows three orthogonal slices from Subject 2, acquired at the time of the maximum (Day 7, Measurement 2, shown in top row) and minimum (Day 18, Measurement 1, shown in bottom row) lateral ventricle volume. The reduced ventricle size is visible, particularly when the images are viewed dynamically (see video available through FigShare (Data Citation 1)).

In order to explore the apparent change in ventricle size, we investigated the relationship between the lateral ventricle volumes and (a) subject weight at the time of each scan, (b) image signal-to-noise ratio (SNR), and (c) time of day of scan (Figure 5). No correlation with weight or image SNR (Supplementary Figure 1) was observed, but in the case of time of day, a statistically significant positive correlation ($r = 0.66$, $P = 0.0016$) was observed for Subject 3, but not for Subjects 1 or 2.

Usage Notes

The analysis and results presented in this work demonstrate how the test-retest dataset can be used to determine the reliability of brain volume measurements. The coefficient of variation (CV) values shown in Table 1 indicate the repeatability of the volumetric measurements obtained using our FreeSurfer-based

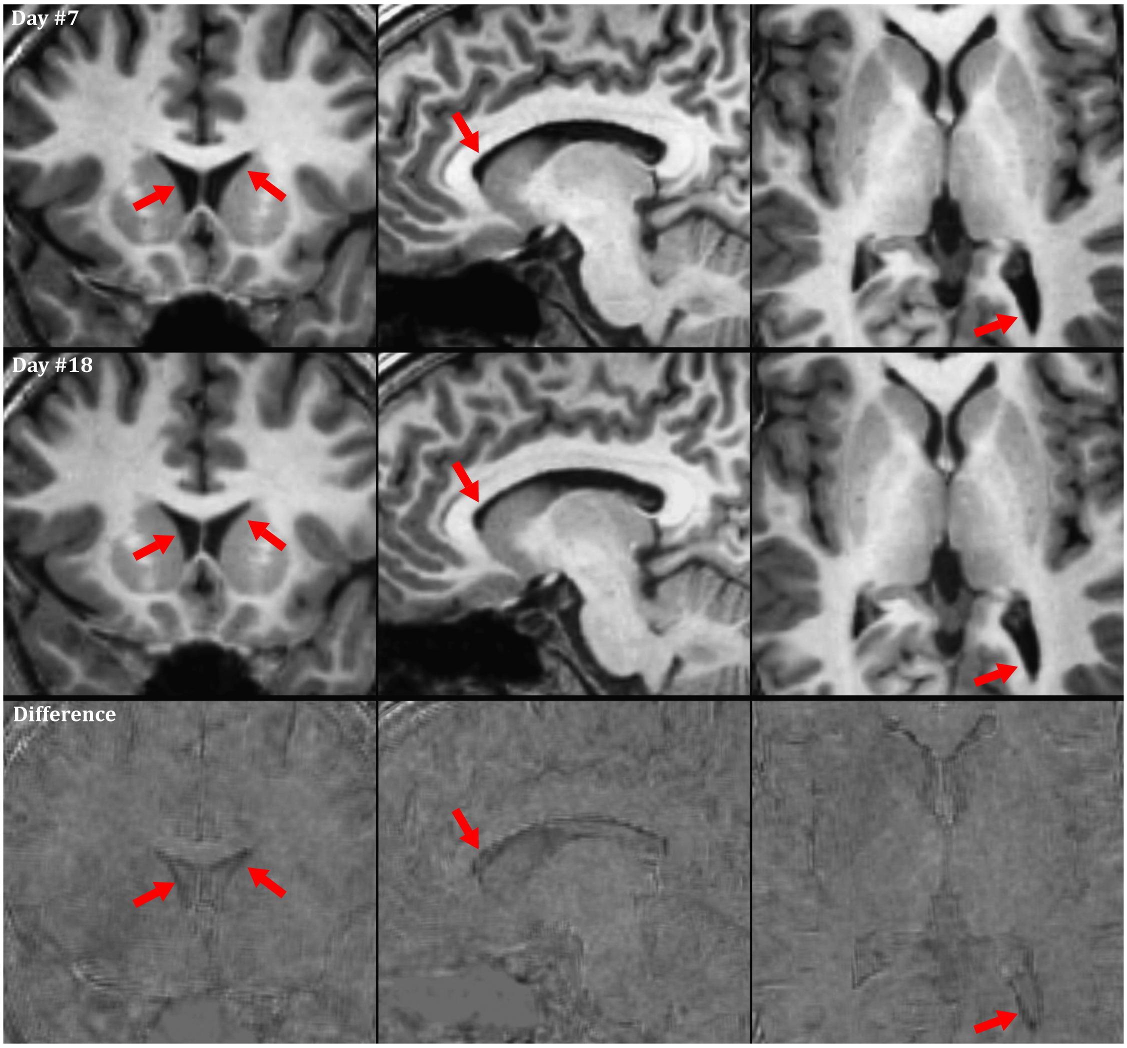


Figure 4. Apparent ventricle volume changes. The small, but statistically significant, inter-session ventricle volume changes observed in the FreeSurfer output are also visible in the structural MRI images (see red arrows, top row: larger ventricle volume, bottom row: lower ventricle volume). These volume changes are barely discernable to the naked eye in static images; therefore, this figure is available in an animated format in the supporting information.

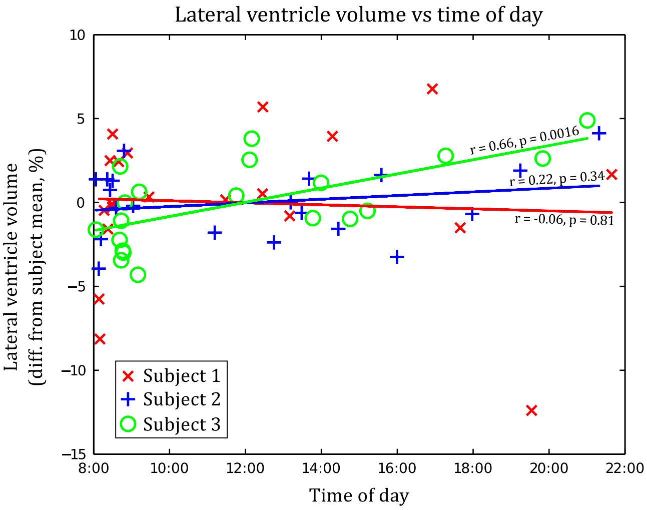


Figure 5. Lateral ventricle volume vs time of day. Lateral ventricle volume plotted against time of day, with a linear fit shown for each subject individually. A statistically significant correlation is present for Subject 3, but not for Subjects 1 or 2.

pipeline. The CV values shown (from 0.9% to 6.1%, depending on the structure) could be used for effect size calculations for volumes of individual structures in a study using the same processing pipeline. However, we did not attempt to perform a comprehensive valuation of FreeSurfer itself, or to compare its output to that from other software packages or manual segmentation, as the reliability of FreeSurfer has already been evaluated on numerous occasions[24–26]. We also limited our analysis to volumes of the lateral ventricles, cortical white matter, and the sub-cortical structures of particular interest to us. We hope that researchers developing alternative software packages will use the test-retest dataset to evaluate and compare their processing methods in brain regions of interest to them.

The main unique aspect of our test-retest dataset is the large number of scans performed on each subject over a short period (one month) and that two back-to-back scans were performed within each session, with the patient repositioned between each scan. This can be used to separate intra- from inter-session variation, using the methods demonstrated in this work. In particular, the data indicate that variance between days exceeds variance within a session for lateral ventricle volume and that this is due to actual volume changes of the ventricles between days, rather than measurement errors.

The observed ventricle volume changes may be due to subject hydration effects, since dehydration can cause shrinkage of brain tissue and enlargement of ventricle volume[16,17]. But unlike these previous studies, our subjects were not required to deliberately become dehydrated or to consume excess fluid, and we did not observe a reduction in body mass correlating with ventricular volume. We refrained from using a hydration protocol, as it is not standard practice to subject patients to such a measure for routine imaging. Despite this, the maximum observed ventricle volume changes in our study were of similar magnitude to that found in other studies where subjects underwent a ‘thermal-exercise dehydration protocol’[17,27]. It is therefore interesting to note that these effects also occur in normal subjects, and it is possible that other confounding factors are affecting water balance in the brain. The time of day did not appear to be a strong predictor of this effect (Figure 5). Regardless of the cause, this effect may need to be considered when looking for signs of ventricular enlargement using MR-based methods such as VIENA[28] . It is unclear whether the effect would be more pronounced in older subjects where ventricles are larger. While larger ventricles are potentially easier to segment, leading to less error in the volume estimates, absolute volume changes from day to day would lead to a smaller percentage change.

Although the lateral ventricles were the only structures where we were able to reject the null hypothesis of identical intra- and inter-session variance, a ‘conservation of volume’ assumption within the skull would suggest that volume changes should occur in other brain structures to compensate for the ventricle volume gain or loss. Perhaps future advances in segmentation methods will allow these changes to be detected in this dataset. Nonetheless, the data indicate that changes from day to day are likely to be very small for structures other than the ventricles.

Regular measurements of the QA phantoms and other standard scanner QA measurements gave no evidence of scanner-related variation over the 31-day period. We have also not observed any changes in the brain volume data that would indicate this was a limiting factor in the study. A more important limitation is that the subjects (healthy, age 26–31) are not representative of the general population or any specific patient group. With only three subjects, it is also possible that the inter-scan variability does not hold for a larger population. Nonetheless, the data are sufficient to perform basic test-retest reliability analysis, which was the goal of the work.

Our experimental protocol differed significantly from that used in previous studies examining test-retest reliability[29,30], since our subjects were imaged many times, rather than on only two occasions. While we specifically aimed at comparing intra- and inter-session variation, previous work has largely focused on the influence of confounding factors such as pulse sequence, imaging geometry, parallel imaging, scanner drifts and choice of segmentation software. We have collected a unique dataset, complimentary to existing studies, which we hope will become a valuable tool to assess repeatability of segmentation methods.

References

1. Jack, C. R. Jr *et al.* Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* **49,** 786–794 (1997).
2. Dubois, B. *et al.* Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurology* **6,** 734–746 (2007).
3. Jack, C. R. Jr *et al.* Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology* **55,** 484–489 (2000).
4. Kovacevic, S., Rafii, M. S. & Brewer, J. B. High-throughput, fully automated volumetry for prediction of MMSE and CDR decline in mild cognitive impairment. *Alzheimer Dis. Assoc. Disord.* **23,** 139–145 (2009).
5. Brewer, J. B., Magda, S., Airriess, C. & Smith, M. E. Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer disease. *AJNR Am. J. Neuroradiol.* **30,** 578–580 (2009).
6. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9,** 179–194 (1999).
7. Fischl, B. FreeSurfer. *NeuroImage* **62,** 774–781 (2012).
8. Fischl, B. *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33,** 341–355 (2002).
9. Jack, C. R. Jr *et al.* The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* **27,** 685–691 (2008).
10. Brewer, J. B. Fully-automated volumetric MRI with normative ranges: translation to clinical practice. *Behavioural Neurology* **21,** 21–28 (2009).
11. Schuff, N. *et al.* MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain* **132,** 1067–1077 (2009).
12. Murphy, E. A. *et al.* Six-month atrophy in MTL structures is associated with subsequent memory decline in elderly controls. *Neuroimage* **53,** 1310–1317 (2010).
13. McEvoy, L. K. *et al.* Mild cognitive impairment: baseline and longitudinal structural MR imaging measures improve predictive prognosis. *Radiology* **259,** 834–843 (2011).
14. Sluimer, J. D. *et al.* Accelerating regional atrophy rates in the progression from normal aging to Alzheimer's disease. *European Radiology* **19,** 2826–2833 (2009).
15. Karas, G. *et al.* Amnestic mild cognitive impairment: structural MR imaging findings predictive of conversion to Alzheimer disease. *AJNR Am. J. Neuroradiol.* **29,** 944–949 (2008).
16. Duning, T. *et al.* Dehydration confounds the assessment of brain atrophy. *Neurology* **64,** 548–550 (2005).
17. Kempton, M. J. *et al.* Dehydration affects brain structure and function in healthy adolescents. *Hum. Brain Mapp.* **32,** 71–79 (2011).
18. Hagemann, G. *et al.* Changes in brain size during the menstrual cycle. *PLoS ONE* **6,** e14655 (2011).
19. Gunter, J. L. *et al.* Measurement of MRI scanner performance with the ADNI phantom. *Medical Physics* **36,** 2193–2205 (2009).
20. Gronenschild, E. H. *et al.* The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS ONE* **7,** e38234 (2012).
21. Massart, D. L. *et al. Handbook of Chemometrics and Qualimetrics: Part A* (Elsevier Science Pub Co., 1997).
22. Gluer, C. C. *et al.* Accurate assessment of precision errors: how to measure the reproducibility of bone densitometry techniques. *Osteoporosis Int.* **5,** 262–270 (1995).
23. Good, P. *Permutation, Parametric and Bootstrap Tests of Hypotheses* (Springer, 2005).
24. Tae, W. S., Kim, S. S., Lee, K. U., Nam, E. C. & Kim, K.W. Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. *Neuroradiology* **50,** 569–581 (2008).
25. Jovicich, J. *et al.* MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage* **46,** 177–192 (2009).
26. Lehmann, M. *et al.* Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. *NeuroImage* **49,** 2264–2274 (2010).
27. Kempton, M. J. *et al.* Effects of acute dehydration on brain morphology in healthy humans. *Hum. Brain Mapp.* **30,** 291–298 (2009).
28. Vrenken, H. *et al.* Validation of the automated method VIENA: An accurate, precise, and robust measure of ventricular enlargement. *Hum. Brain Mapp.* **35,** 1101–1110 (2013).
29. Wonderlick, J. S. *et al.* Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. *Neuroimage* **44,** 1324–1333 (2009).
30. Morey, R. A. *et al.* Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Hum. Brain Mapp.* **31,** 1751–1762 (2010).

Data Citation

1. Maclaren, J., Han, Z., Vos, S. B., Fischbein, N. & Bammer, R. *FigShare* http://dx.doi.org/10.6084/m9.figshare.929651 (2014).

Acknowledgements

We thank Christoph Seeger, Alexander Brost and Norbert Schuff for valuable discussions and Andrew Gentles for his advice concerning the statistical methods used in this work. We would also like to acknowledge the generosity of the authors of FreeSurfer (Anders Dale, Bruce Fischl and the entire development team) for making their software freely available to the public.

Author Contributions

J.M. co-designed the study, performed the imaging experiments, analyzed the data, and wrote the manuscript with contributions from all authors. Z.H. provided advice on the study design, performed the imaging experiments, processed the data in FreeSurfer, and contributed to the manuscript. S.B.V. co-designed the study, performed the imaging experiments, formatted the submitted data, and contributed to the manuscript. N.F. provided advice on the study design and contributed to the manuscript. R.B. provided advice on the study design, managed the project, and contributed to the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/sdata.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Maclaren, J. *et al.* Reliability of brain volume measurements: A test-retest dataset. *Sci. Data* 1:140037 doi: 10.1038/sdata.2014.37 (2014).