Test-Retest Reliability of Brain Volume Measurements

Julian Maclaren[1], Zhaoying Han[1], Sjoerd B Vos[1,2], Christoph Seeger[1,3], Alexander Brost[1], Nancy Fischbein[1], and Roland Bammer[1]

[1]*Center for Quantitative Neuroimaging, Dept. of Radiology, Stanford University, Stanford, CA, United States,* [2]*Image Sciences Institute, University Medical Center Utrecht, Utrecht, Netherlands,* [3]*Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany*

Target audience: Researchers using brain volumetry and those involved in the Alzheimer's Disease Neuroimaging Initiative (ADNI) project.

Introduction and purpose: The monitoring of neurodegenerative disease progression may be assisted by quantification of the volume of structures in the human brain, particularly medial temporal lobe structures, such as the hippocampus [1]. Recent developments in automated segmentation software have improved this process, but it is still unclear how applicable these tools are for clinical routine. Often, the reliability of measurements is uncertain. The aims of this work were twofold:

(a) To determine whether test-retest reliability (repeatability) of measurements can be measured in a single session (intra-session), or whether changes from day to day (inter-session) also affect repeatability.

(b) To generate a publically available test dataset to assist in the validation of the repeatability of current and future segmentation methods.

Methods: An experiment was designed to allow separate calculation of intra- and inter-session test-retest reliability (Fig.1). A total of 120 3D T1-weighted volumes were acquired from 3 subjects (40 scans/subject) over one month. Importantly, each subject was scanned twice within each session, and repositioned between the two scans, so that all scans were treated as separate measurements (with a break of ~5min). This ensured that confounding effects such as subject positioning were consistent between and within sessions. Other data recorded throughout the study included subject weight, exercise, caffeine and alcohol intake, time of day, and phantom data for quality assurance and scanner stability.

We used the ADNI-recommended T1w imaging protocol for our GE MR750 3 T scanner (accel. sagittal IR-SPGR, 1x1x1.2 mm res., 8-channel head coil, 5min 37s acq. time). All 3D volumes were processed using FreeSurfer [2], which provides quantitative volume data for a range of brain structures. Statistical analysis of the FreeSurfer output was used to assess variability for measurements obtained 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, i.e.,

$\sigma(\text{intra-session}) = \sqrt{\sum\left(x_i' - x_i''\right)^2 / 2m}$, where x_i' and x_i'' are the ith paired measurements over the m pairs, evaluated separately for each subject. Intra-session variability reflects manual repositioning differences, noise and segmentation errors, but not biological variations occurring from day to day.

Results: Fig. 2 shows results for one of the seven brain regions analyzed. It is visually apparent that inter-session variance exceeds intra-session variance; however, the ventricles are the only structures where this increase was statistically significant ($p < 0.005$) for all three volunteers. Quantitative results are shown in Table 1, reported as standard deviations expressed as a percentage of the mean (i.e., coefficient of variation). The confidence intervals shown are computed directly from the total variance.

Discussion: In all brain regions analyzed, the total variance is greater (i.e., the repeatability is worse) than for the intra-session variance alone. In the case of the ventricular volume, inter-session variance exceeds intra-session. We hypothesize that this is due to subject hydration effects, since dehydration can affect brain volume (3); however, we did not observe a significant correlation between subject weight and ventricular volume. This is perhaps due to other confounding factors affecting water balance in the brain.

The 95% confidence intervals shown in Table 1 indicate the repeatability of our volumetric measurements. Some values exceed 10%, which indicates the need for caution when interpreting brain volume data, as this could potentially mask the effects of disease. By scanning each subject 40 times in only 31 days, we have created a unique data set, which will assist other researchers (available online at http://neuroquant.stanford.edu/data/test-retest).

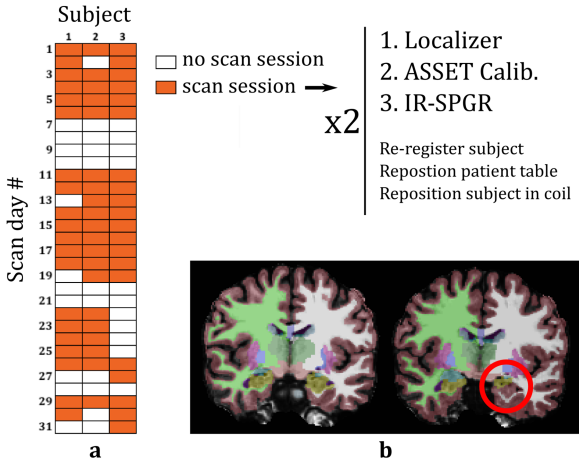

Fig. 1: (a) Data were acquired in 60 scan sessions (20 per subject); each session comprised two back-to-back scans ('intra-session') where the subject was removed from the scanner between scans to replicate effects between days ('inter-session'). (b) 3D volumes were segmented using FreeSurfer. One factor affecting repeatability is segmentation error: here the left hippocampus is incorrectly segmented (red circle) in one of the two volumes acquired in the same session.

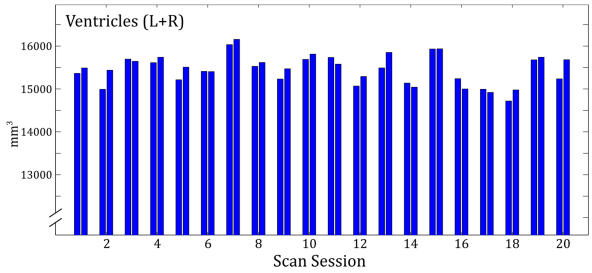

Fig. 2: Example segmentation results for a single subject and brain region. Each of the 40 bars for each region represents a single scan. The bars are paired to indicate which measurements were obtained within one session.

Table 1: Intra- and inter-session variation were decoupled from each other and are shown as standard deviations, expressed as percentages of the estimated 'true' value. Values are computed as the mean over all subjects.

Structure	σ, intra-session	σ, total	95% confidence interval of volume
Hippocampus	2.6%	2.9%	+/- 5.7%
Ventricles (L+R)	1.5%	3.4%	+/- 6.7%
Amygdala	4.7%	5.2%	+/- 10.2%
Putamen	3.8%	3.9%	+/- 7.7%
Pallidum	5.0%	5.4%	+/- 10.6%
Caudate	1.5%	1.6%	+/- 3.1%
Thalamus	5.5%	6.1%	+/- 11.9%

Conclusion: Fluctuations in brain volume measurements between days mean that the conventional means of assessing repeatability by using multiple scans within a single session underestimates the true variance.

Acknowledgements: NIH (2R01 EB00271108-A1 , 5RO1 EB008706, 5R01 EB01165402-02), the Center of Advanced MR Technology at Stanford (P41 EB015891), Lucas Foundation, Oak Foundation. **References:** [1] Jack, et al. Neurology 49:786-794. [2] Dale, et al. Neuroimage 9:179-194. [3] Duning, et al. Neurology 2005;64:548–550