Published in final edited form as:

Neuroimage. 2009 May 15; 46(1): 177–192. doi:10.1016/j.neuroimage.2009.02.010.

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

Jorge Jovicich[1], **Silvester Czanner**[2], **Xiao Han**[3], **David Salat**[4,5], **Andre van der Kouwe**[4,5], **Brian Quinn**[4,5], **Jenni Pacheco**[4,5], **Marilyn Albert**[8], **Ronald Killiany**[9], **Deborah Blacker**[7], **Paul Maguire**[10], **Diana Rosas**[4,5,6], **Nikos Makris**[4,5,11], **Randy Gollub**[4,5], **Anders Dale**[12], **Bradford Dickerson**[4,6,7,13,*], and **Bruce Fischl**[4,5,14,*]

[1]Center for Mind-Brain Sciences, Department of Cognitive and Education Sciences, University of Trento, Italy

[2]Warwick Manufacturing Group, School of Engineering, University of Warwick, United Kingdom

[3]CMS, Inc., St. Louis, MO, USA

[4]Athinoula A. Martinos Center for Biomedical Imaging

[5]Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[6]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[7]Gerontology Research Unit/Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[8]Department of Neurology, Johns Hopkins University School of Medicine, USA

[9]Departments of Anatomy and Neurobiology, Boston University School of Medicine, USA

[10]Pfizer Global Research & Development, Groton, CT, USA

[11]Center for Morphometric Analysis, Massachusetts General Hospital, Boston, MA, USA

[12]University of California San Diego, CA, USA

[13]Division of Cognitive and Behavioral Neurology, Department of Neurology, Brigham & Women’s Hospital, Boston, MA, USA

[14]CSAIL/HST, MIT, Cambridge, MA, USA

Abstract

Corresponding author: Jorge Jovicich, Assistant Professor, Center for Mind/Brain Sciences, University of Trento, Italy, Phone: +39-0461-88 3064, Fax: +39-0461-88 3066, jorge.jovicich@unitn.it.

[*]Authors contributed equally to this work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Automated MRI-derived measurements of in-vivo human brain volumes provide novel insights into normal and abnormal neuroanatomy, but little is known about measurement reliability. Here we assess the impact of image acquisition variables (scan session, MRI sequence, scanner upgrade, vendor and field strengths), Freesurfer segmentation preprocessing variables (image averaging, B1 field inhomogeneity correction) and segmentation analysis variables (probabilistic atlas) on resultant image segmentation volumes from older (n=15, mean age 69.5) and younger (both n=5, mean ages 34 and 36.5) healthy subjects. The variability between hippocampal, thalamic, caudate, putamen, lateral ventricular and total intracranial volume measures across sessions on the same scanner on different days is less than 4.3% for the older group and less than 2.3% for the younger group. Within-scanner measurements are remarkably reliable across scan sessions, being minimally affected by averaging of multiple acquisitions, B1 correction, acquisition sequence (MPRAGE vs. multi-echo-FLASH), major scanner upgrades (Sonata-Avanto, Trio-TrioTIM), and segmentation atlas (MPRAGE or multi-echo-FLASH). Volume measurements across platforms (Siemens Sonata vs. GE Signa) and field strengths (1.5T vs. 3T) result in a volume difference bias but with a comparable variance as that measured within-scanner, implying that multi-site studies may not necessarily require a much larger sample to detect a specific effect. These results suggest that volumes derived from automated segmentation of T1-weighted structural images are reliable measures within the same scanner platform, even after upgrades; however, combining data across platform and across field-strength introduces a bias that should be considered in the design of multi-site studies, such as clinical drug trials. The results derived from the young groups (scanner upgrade effects and B1 inhomogeneity correction effects) should be considered as preliminary and in need for further validation with a larger dataset.

Keywords

Subcortical brain volumes; Reproducibility; Reliability; Structural MRI; Cerebral cortex; Morphology

Introduction

Techniques that enable the *in vivo* MRI-derived quantitative characterization of the human brain, such as subcortical brain volumes (including for this purpose the archicortical hippocampal formation and the ventricular system), are beginning to demonstrate important potential applications in basic and clinical neuroscience. Alterations in subcortical brain volumes are manifested in normal aging (Mueller et al., 2006; Jack, et al. 2005; Szentkuti et al., 2004), Alzheimer's disease (Kantarci and Jack, 2004; Anstey and Maller, 2003), Huntington's disease (Douaud et al., 2006; Peinemann et al., 2005, Kipps et al., 2005; Kassubek et al., 2005; Kassubek et al. 2004; Rosas et al., 2003; Thieben et al., 2002), and schizophrenia (Makris et al., 2006; Koo et al., 2006; Kuroki et al.; 2006, Shenton et al., 2001). Cross-sectional and longitudinal imaging-based biomarkers of disease will likely be of great utility in better understanding brain disorders and in evaluating therapeutic efficacy (Dickerson and Sperling, 2005; DeKosky and Marek, 2003).

The accurate and reliable measurement of subcortical brain volumes from MRI data is a non-trivial task. Manual measurements are difficult and time consuming. It can take a trained anatomist several days to manually label a single high-resolution set of structural MR brain images. In addition, manual measurements are susceptible to rater bias. To facilitate efficient, operator-independent subcortical region-of-interest (ROI) quantification, several automated and semi-automated algorithms have been proposed, including atlas-based methods (Haller et al., 1997; Collins et al. 1999, Fischl et al., 2002; Magnota et al., 2002; Fischl et al., 2004; Alemán-Gómez et al. 2006), voxel-based morphometry using Statistical Parametric Mapping (Ashburner and Friston, 2000), tensor-based morphometry (Studholme et al., 2001; Leow et al., 2005) and boundary shift integral methods (Smith et al., 2002; Camara et al., 2007; Barnes et al., 2007; Anderson et al., 2007).

Although the accuracy validation of automated segmentation methods has been performed against regional manual measurements derived from both *in vivo* and postmortem brain scans (Fischl and Dale, 2000, Fischl et al., 2002), the influence of image acquisition and data analyses parameters on the reliability of the derived measures has received relatively little systematic investigation (Leow et al., 2006; Ewers et al., 2006, Smith et al., 2002). Furthermore, the measurement of reliability also provides a means for assessing the impact of measurement error on sample size requirements. Defining the reliability of subcortical morphometric methods is therefore important.

Reliability in MRI-derived automated morphometric measures can be influenced by several sources of variance, including subject-related factors, such as hydration status (Walters et al., 2001), instrument-related factors, such as field strength, scanner manufacturer, imaging magnetic gradients (Jovicich et al., 2006), pulse sequence, and data processing-related factors, including not only software package and version but also the parameters chosen for analysis (Senjem et al, 2005; Han et al. 2006). All of these factors may affect the ability to detect morphometric differences between groups in typical cross-sectional studies (e.g., morphometric differences between two subject groups, where each subject is scanned once and all subjects are scanned on the same scanner). Longitudinal studies of normal development, aging, or disease progression face additional challenges associated with both subject-related factors as well as instrument-related factors (e.g., major scanner upgrades, across-session system instabilities). For studies that combine data acquired from multiple sites it is critical to understand and adjust for instrument-related differences between sites, such as scanner manufacturer, field strength, and other hardware components. Thus, detailed quantitative data regarding the degree to which each of the factors outlined above contributes to variability in morphometric measures would be helpful for both study design and interpretation. Recent publications discuss such studies with regard to the reproducibility of cortical thickness measures (Han et al., 2006) and tensor based-morphometry (Leow et al., 2006).

The goal of the present study is to extend the work from Han et al., 2006 by adding a new dataset and focusing the analysis on the evaluation of the test-retest reliability of subcortical volume measurements (as opposed to cortical thickness reproducibility) in the context of mapping brain morphometry changes using the FreeSurfer software package, which is an automated method for full brain segmentation (Fischl et al., 2002, Fischl et al., 2004). To keep a manageable number of variables this study uses a constant segmentation method (Freesurfer). Comparisons with other subcortical volume measurement methods have been recently reported (Pengas et al., 2008; Tae et al., 2008) and are beyond the scope of this work. We study a group of 15 older healthy subjects (older than 65, Han et al. 2006) to assess anatomic variability related to atrophy and age-related MRI signal changes. Each subject was scanned four times (twice in a Siemens Sonata 1.5T, once in a GE Signa 1.5T and once in a Siemens Trio 3T) using the same protocol in separate sessions within a time period of two weeks, to include both subject hydration and scanner variability effects. We also study two smaller groups of 5 young subjects (mean ages 34 and 36.5) to investigate reproducibility effects across major scanner upgrades (Siemens Sonata to Avanto, Han et al. 2006 and Siemens Trio to Tim Trio, new dataset), with two separate acquisitions before and after the upgrade. Each dataset is treated independently to derive the volume of various brain structures (hippocampus, thalamus, caudate, putamen, pallidum, amygdala, lateral ventricles, inferior lateral ventricles, intracranial), which for summary purposes will be referred to as ‘subcortical structures’ when not defined explicitly. We study how subcortical volume reproducibility depends on MRI acquisition sequences, scanner upgrade, data pre-processing, segmentation analyses methods, MRI system vendor and field strength.

Materials and methods

MRI data acquisition and subject groups

We acquired and analyzed three datasets to characterize how reliability of subcortical volume estimation is affected by various image acquisition parameters (see Table 1 for summary), by pre-processing choices (data averaging and B1 inhomogeneity correction) and by segmentation analysis methods (choice of probabilistic atlas). All participants were healthy volunteers with no history of major psychiatric, neurological or cognitive impairment, and provided written informed consent in accordance with the Human Research Committee of Massachusetts General Hospital. For all subject groups the scans were randomized over days over a period of approximately two months.

Dataset 1—This dataset, previously used in Han et al. 2006 to study cortical thickness reproducibility, was analyzed to evaluate how the reliability of subcortical volume estimates in older healthy subjects depends on T1-weighted MRI acquisition sequence (MPRAGE and multi-echo FLASH), scan session, data averaging, segmentation atlas, scanner platform and field strength.

Fifteen healthy older subjects participated in this dataset (age between 66 – 81 years; mean: 69.5 years; std: 4.8 years). Each subject underwent 4 scan sessions at approximately two-week intervals, including two sessions on a Siemens 1.5T Sonata scanner (Siemens Medical Solutions, Erlangen, Germany), one on a Siemens 3T Trio scanner, and one on a GE 1.5T Signa scanner (General Electric, Milwaukee, WI). The Siemens scanners are located at the Martinos Center for Biomedical Imaging at Massachusetts General Hospital, and the GE scanner is located at the Brigham and Women’s Hospital.

In each Siemens scan session, the acquisition included two MPRAGE volumes (bandwidth=190 Hz/pixel, flip angle= 7°, TR/TE/TI=2.73s/3.44ms/1s), and two multi-echo multi flip angle (30° and 5°) fast low-angle shot (FLASH) volumes (bandwidth=651 Hz/pixel, TR=20ms, TE=(1.8 + 1.82*n) ms, n=0, …,7). In each GE scan session, a custom MPRAGE sequence was programmed with parameters as similar as possible, and two volumes were acquired. The total scanning time for each MPRAGE and each multi-echo FLASH (MEF) volume was roughly the same and about 9 minutes. All structural scans were 3D sagittal acquisitions with 128 contiguous slices (imaging matrix = 256×192, in-plane resolution = $1\times1\text{mm}^2$, and slice thickness = 1.33mm). In each Siemens session, the acquisitions were automatically aligned to a standardized anatomical atlas to ensure consistent slice prescription across scans (van der Kouwe et al. 2005; Benner et al., 2006).

Dataset 2—This dataset, previously used in Han et al. 2006 to study cortical thickness reproducibility, was analyzed to evaluate how a 1.5T MRI system upgrade (Siemens Sonata to Avanto) affects the reproducibility of subcortical volume estimates. This dataset is also used to compare the reproducibility between the young and older groups.

Five healthy volunteers (age between 29 and 37 years; mean: 34 years; std.: 3 years) were each scanned in four sessions, two before and two after an MRI scanner upgrade (within one week for the repeated scans on the same scanner, and the total time span is about 6 weeks). The upgrade was from a Siemens Magnetom Sonata to a Magnetom Avanto, which included the following major changes: a) main magnet (both are 1.5T, Avanto’s length is 150cm, Sonata’s is 160 cm), (b) gradient system (Avanto coils are more linear, Sonata 40 mT/m @ 200 T/m/s, Avanto 45mT/m @ 200 T/m/s), c) head RF coil (circularly polarized on Sonata, 12 channels in Avanto), and d) software upgrade.

The acquisition protocol before and after upgrade was kept the same and consisted of two sets of 3D acquisitions: two MPRAGE and two multi-echo FLASH scans, with the same parameters as used in Group 1. The head RF coil set-ups are different: circularly polarized for Sonata, and 12-channel for Avanto (used in 4 channel mode). Also, for both platforms brains were automatically aligned to an atlas in each scanning session for setting the slice prescription in approximately AC-PC orientation (van der Kouwe et al. 2005; Benner et al., 2006).

Dataset 3—This dataset, an extension to Han et al. 2006, was acquired and analyzed to evaluate how a 3T MRI system upgrade (Siemens Trio to Tim Trio) affects the reproducibility of subcortical volume estimates. This is a major hardware and software upgrade, similar to the one described above, where essentially the only thing that does not change across the upgrade is the main static magnet.

Five healthy volunteers (age between 30 and 40 years; mean: 36.5 years; std.: 3 years) were each scanned in four sessions, two before and two after the upgrade. The MPRAGE data acquisition protocol was essentially the same as for Dataset 2; no MEF data was collected. The vendor’s birdcage head RF coil was used for both pre- and post-upgrade scans.

Measures of brain structure volumes

Segmentation of brain structures from T1-weighted 3D structural MRI data and estimation of structure volumes was performed using the FreeSurfer toolkit, which is freely available to the research community (http://surfer.nmr.mgh.harvard.edu/). This suite of methods uses a probabilistic brain atlas of choice, which was initially proposed in 2002 (Fischl et al., 2002), and has undergone several important improvements over the years (Fischl et al., 2004, Han et al. 2006). With these updates, the current subcortical segmentation method is fully automated and has been recently described for application with MPRAGE and MEF data, using specific atlases for each of these image acquisition protocols (Han et al. 2006). Briefly regarding MEF, each multi-echo FLASH acquisition gives 8 image brain volumes (one volume for each gradient echo). The 16 volumes that are available from the 30 degree and the 5 degree MEF acquisitions are combined using a weighted linear average approach to create a single T1-weighted volume (Han et al, 2006). The combination of the data is obtained by applying a linear discriminant analysis technique in order to find an optimal set of weights (a projection vector) such that the weighted average volume has the best contrast to noise ratio between white and gray matter.

The procedure automatically labels each voxel in the brain as one of 40 structures (Fischl et al., 2002). Here we focus on only a subset of them which are of interest in neurodegenerative diseases and are by far the largest in volume: hippocampal formation, amygdala, caudate nucleus (caudate), putamen, globus pallidus (pallidum), thalamus, lateral ventricles, inferior lateral ventricles and total intracranial volume. For each of these structures (except the intracranial volume) the right and left hemisphere volumes are estimated separately. Reproducibility errors in cortical surface structures have been previously discussed in a separate manuscript (Han et al, 2006).

The segmentation atlas used for the automated full brain Freesufer segmentation was generated from the following subjects: health young (10, 6 females, age 21±2 years), healthy middle age (10, 6 females, age 50±6 years), healthy old (8, 7f, age 74±7 years), demented old (11, 6 females, age 77±6 years). The atlas demographics’ covers well the demographics of the three datasets considered in this work (summarized in Table 1) so it is not expected to bias the segmentation results.

All data was visually inspected for quality assurance prior to analyses. All data was analyzed using the same and latest public Freesurfer software version (4.0). No manual edits were done, the segmentation results were inspected visually prior to the volume analysis.

Measures of volume reliability

We examined the test-retest reliability of volume estimates under two general conditions: when the volumes compared are derived from repeated identical acquisitions and analyses methods (i.e., volume repeatability, for example across scan sessions) and when the volumes compared are derived from different acquisition or analysis methods (i.e., volume agreement, for example across different scanners). For both assessments we used a Bland-Altman analysis (Bland and Altman, 1986). In short, for each pair-wise comparison of volumes (repeatability or agreement), the volume differences are plotted (y-axis) against the volume means (x-axis) for each subject. From this we obtain two metrics, each with its 95% confidence interval: the mean volume difference (± t_{n-1} SD root(1/n), where t_{n-1} is the t-statistics for a two-tailed test with 95% power and n-1 degrees freedom) and the limits of agreement (2 SD ± t_{n-1} SD root(3/n)) of the volume differences (SD: standard deviation of the differences, n: number of subjects, Bland and Altman, 1986). The plots show the spread of data, the mean difference and the limits of agreement. For excellent reproducibility the mean difference should be ideally zero with a narrow distribution of data around zero across the range of volume measurements. These two metrics (with their 95% confidence interval) summarize the reliability of a comparison for each structure and are then used to compare test-retest reliability across different conditions (next section). An improvement in test-retest reliability will be detected as a significant reduction of the mean volume difference and/or as a significant reduction of the standard deviation of the volume differences. The significance test across conditions is done using a t-test (two-tailed, p<0.05).

To investigate the sensitivity of our results to the number of subjects, a Jacknife method was used to estimate the bias in the confidence interval of the volume reproducibility due to the number of subjects used (Efron and Tibshirani, 1993). In other words, we estimate how sensitive the confidence interval of the mean reproducibility is to the number of subjects used by comparing the group confidence interval with the mean confidence interval when one subject is left out. The confidence interval for the group of n subjects is $CI_{group} = t_{n-1} * SD_{group} / (n-1)$, SD_{group} is the standard deviation of the reproducibility within the group. In the Jacknife analysis, one at a time each subject is left out and the confidence interval is calculated from the set of n-1 remaining subjects. For example, when subject j is left out, confidence interval is $CI_j = t_{n-2} * SD_j / (n-2)$, where SD_j is the standard deviation of the reproducibility without including subject j. The bias is then estimated as:

$$bias = \frac{1}{n-1}\sum_{j=1}^{n} CI_j - CI_{group}$$

This method gives a non parametrical estimate of bias (i.e. no assumptions on the probabilistic distribution of the volumes and hence reproducibility test statistics is used).

Investigation of variables that affect test-retest volume reliability

There are several factors that may affect the test-retest reliability of subcortical volumes segmented from structural MRI data acquired at a single site. These factors include the scan session specific variables such as subject positioning, shim settings, hydration status, etc…, the choice of image acquisition sequence, data analysis methods (which sometimes may be related to specific additional acquisitions for correction purposes) and unavoidable system hardware upgrades. First we considered the situation that is expected to give the best reproducibility (i.e., multiple scans acquired within the same scan session) with standard processing and no other corrections. This within-session reproducibility was used as best case scenario reference when exploring the effects of the other variability factors, namely, brain segmentation pre-processing (data averaging, correction of intensity inhomeogeneities), image acquisition sequence (MPRAGE, MEF), choice of brain atlas for the segmentation, and MRI system upgrades (Sonata-Avanto). Finally, we looked at volume reproducibility effects when using data derived from different MRI system vendors (Siemens vs. GE at 1.5T) and different field strengths (1.5T vs. 3T).

Effect of scan session on test-retest subcortical volume reproducibility—The best possible reproducibility is expected to be obtained from multiple scans acquired within the same session on young subjects (elderly subjects typically move more, and hence data is suboptimal). This situation minimizes variability in the segmentation results that could come from changes in the acquisition sequence, scanner hardware/instability and/or the subject (head position, physiology, etc.). Ideally, within and across session reproducibility should be comparable, and the interest is in investigating whether manipulations on the processing can further improve this reproducibility. We examined the test-retest repeatability of the two scans acquired within each of the two sessions (test_scan1 vs. test_scan2, retest_scan1 vs. retest_scan2), and also the other four combinations obtained across sessions (test_scan1 vs. retest_scan1, test_scan1 vs. retest_scan2, test_scan2 vs. retest_scan1 and test_scan2 vs. retest_scan2). Subcortical volumes were derived from single MPRAGE acquisitions segmented with the standard method (MPRAGE atlas).

Effect of number of acquisitions on subcortical volume reliability—Averaging of several image acquisitions has the effect of improving tissue signal-to-noise ratio by canceling out random signal fluctuations from the subject and the MR electronics (assuming that all scans averaged have equally good quality, e.g. negligible motion artifacts). By reducing the noise in the averaged image the tissue contrast-to-noise ratio may also increase, thus improving the accuracy of the tissue segmentation algorithm and potentially the reproducibility of subcortical volumes derived from the segmentation. The cost of averaging is increased image acquisition time. An additional potential cost is that averaging may reduced image contrast due to the resampling done when the images are co-registered prior to averaging (to minimize motion effects across scans) or also if there is a differential degree of head motion during one of the two scans averaged (blurring is introduced). Here we tested whether there was a significant improvement in test-retest volume reliability for the average of two within-session acquisitions as compared to a single acquisition from the same session. The MPRAGE scans from dataset 1 (Sonata, older group, n=15) and dataset 2 (Sonata, young group, n=5) were used to study the effect of different number of image acquisitions on the subcortical volume reliability. For this purpose, subcortical segmentation and volume estimations were computed starting from each one of the two MPRAGE acquisitions and also from the average of the two within-session acquisitions, for each subject, for each test-retest session. The subcortical volume reliabilities, for each structure, were then compared between the case of single and averaged (which is the recommended default approach for use with FreeSurfer) acquisitions.

Effects of B1 RF inhomogeneity correction on subcortical volume reliability—Image intensity distortions from radiofrequency (RF) B1 field inhomogeneity may affect the reliability of MRI tissue segmentation (Leow et al., 2006, Alecci, et al. 2000). To study the effect of B1 inhomogeneity we used the MPRAGE data from dataset 2 (Sonata, young group, n=5) in which a B1 correction profile was obtained as described in Leow et al. (2006). Briefly, the sensitivity profile of the receive RF coil was estimated by dividing an image volume obtained with the head RF coil by a corresponding image volume obtained with the body coil on a voxel-by-voxel basis. With this sensitivity profile all subsequent volumes can be corrected by dividing each voxel's intensity by the estimated sensitivity value at that location[1]. To evaluate the effects of B1 inhomogeneity correction on test-retest subcortical volume reproducibility we compared the reproducibility derived from two single MPRAGE scans on separate sessions, both scans either with or without the B1 correction. We also investigated the effects of the B1 correction when the MPRAGE scan of each session was the average of two within-session scans. In this latter case, the B1 correction was applied to each single MPRAGE scan prior to averaging. Note that this method for correcting intensity inhomogeneity from B1 imperfections only works for MRI systems in which the body RF coil is used as transmit and the head RF coil as receive. The method will not work in systems that use multichannel transmit and receive head RF coils.

Effects of MRI acquisition sequence on subcortical volume reliability—Structural 3D T1-weighted imaging protocols (typically MPRAGE) are commonly used for segmentation of cortical gray, white and subcortical gray matter structures. MEF is an attractive alternative T1-weighted method with the following advantages: A) In the multi-echo approach, the alternating gradient-echoes are acquired with opposite readout directions, resulting in less distortions related to the case in which echoes are always collected in the same direction. Standard high bandwidth MPRAGE or single echo FLASH scans could also be acquired with alternate readout directions, but this would have to be setup by hand in the protocol and is prone for operator's mistakes, while it happens automatically in MEF. Most importantly, multiple single echo acquisitions would take N times (N = number of echoes) as much acquisition time as multiecho acquisitions to achieve the same SNR). B) Each MEF acquisition gives extra information that can be used: a T2* map, and with sufficient closely- or unevenly-spaced echoes the B0 field offset can be calculated when the phase information is available. C) In the MEF approach time is used to acquire two different flip angles from which additional information can be obtained: T1 and PD tissue maps. D) The combination of the previous properties has been shown to allow for sequence-independent segmentation (Fischl et al. 2004). Recently a new multi-echo MPRAGE sequence has been demonstrated to give improved results than the standard MPRAGE (van der Kouwe et al 2008).

The acquisition time of any one scan of these protocols (MPRAGE or MEF) is essentially the same (approximately 9 minutes) at the same resolution and with no acceleration from parallel imaging.

The effects of acquisition sequence on subcortical volumes were studied in two steps. First we used a Bland-Altman analysis to assess the within session agreement of the volumes derived from both sequences. Second we compared the across-session test-retest reproducibility of each sequence. For the comparison we kept the total acquisition time approximately constant across image sequences: subcortical volumes were derived from two averaged MPRAGE volumes and from two MEF volumes with different flip angles (Fischl et al., 2004).

Effects of segmentation atlas choice on subcortical volume reliability—The brain probabilistic atlas used for the subcortical segmentation and volume estimation is typically defined by the same imaging acquisition method (i.e., for segmenting MPRAGE data an MPRAGE atlas is created from a different manually segmented MPRAGE dataset). Since creating the probabilistic atlas is time consuming, it would be convenient if the segmentation results did not depend strongly on image acquisition differences between the data to be

[1] It is important to note that this procedure only corrects for inhomogeneities in the receive RF field, not the transmit field. This latter inhomogeneity results in change in the effective flip angle, and thus of image contrast and is much more difficult to account for. Fortunately using body transmit coils at 1.5T this effect is negligible, although at 3T it becomes visible, and at higher field strengths such as 7T it is a dramatic effect. The correction of transmit inhomogenieties is beyond the scope of this paper, but will become increasingly important as ultra high field scanners become more routinely available.

segmented and the data used to build the atlas. To evaluate this, we used MPRAGE and MEF atlases derived from different manually labeled datasets (Fischl et al., 2004) to segment MPRAGE and MEF acquisitions from datasets 1 and 2. We compared the test-retest subcortical volume reliability derived from MPRAGE data segmented with MPRAGE atlas, versus the one derived from MEF data segmented with the MEF atlas and then separately with the MPRAGE atlas.

Effects of MRI system upgrades on subcortical volume reliability—MRI system upgrades that involve major hardware and software changes may introduce reliability changes (Han et al., 2006; Jovicich et al., 2005; Czanner et al., 2006). Measurement of these reliability effects is especially important for longitudinal studies. We used two system upgrades (dataset 2: Sonata to Avanto upgrade; dataset 3: Trio to Tim Trio upgrade) as opportunities to measure and compare the test-retest reliability of subcortical volumes before, after and across the upgrade (i.e., using as test one scan acquired before the upgrade and as retest a scan acquired after the upgrade). We also use dataset 2 to investigate how subcortical volume reliability depends on MRI sequence choice (3D T1-weighted acquisitions were MPRAGE and MEF) before and after the upgrade.

Effects of MRI vendor platform and field strength on the reproducibility of subcortical volumes—Multi-center neuroimaging studies usually have to consider combining data acquired from different MRI vendors and field strengths, which may add some variance to the data if the measures across systems are not reproducible. To study this, we used Dataset 1 to compare the subcortical volume test-retest reproducibility from two averaged MPRAGE scans derived from a single system (Sonata-Sonata) with those from mixed systems (Siemens Sonata- GE Signa, both 1.5T) and mixed vendors and field strengths (Siemens 1.5T Sonata- Siemens 3.0T Trio, GE 1.5T Signa – Siemens 3.0T Trio). We also used a linear regression analysis to investigate whether there are biases in the volumes derived from different systems and field strengths. The MPRAGE atlas used for all segmentations was constructed from Siemens Sonata data.

Power analysis

Statistical power calculations can help approximate the number of subjects needed to detect a percent volumetric change with a given estimation of the measurement error. In the results section, we focus on sample-size estimates for detecting hippocampal volume effects of a hypothetical treatment that successfully slows atrophy in Alzheimer patients. Specifically, assuming that in an untreated Alzheimer's group hippocampal volume atrophy rate is approximately 4.9%/year, and with the liberal assumption that a disease modifying therapy may be able to slow it by 50% (Jack et al. 2003), we are interested in detecting a difference between 4.9%/year (in an untreated group) vs. 2.45%/year (in a treated group) to characterize potentially clinically meaningful treatment effects. Our goal was, from the estimation of variance in hippocampus volume measures, to determine how large a sample would be needed to detect this net 2.45% effect between the two groups 1) within the same scanner with no upgrade, 2) within the same scanner after an upgrade, 3) between scanners of different manufacturers, and 4) between scanners of different field strengths. To keep the focus on the reproducibility variables of this study, the standard deviation of the measurement error was estimated from the raw hemispheric hippocampal volumes as derived from the automatic segmentation procedure, without adding any other normalizations or adjustments like for total intracranial volume, age or gender (Jack et al., 2003). This type of sample-size determination is an important step in planning a multi-center clinical trial.

The sample-size is generally estimated by setting the chosen significance level (the probability of type I error), desired statistical power (one minus the probability of type II error), the effect size, and the standard deviation of measurement error (Cohen, 1988). For all sample-size estimates, a significance level of 0.05 (one-sided) and a statistical power of 0.9 were assumed, as has been done previously (Jack et al., 2003). The computation was performed using the standard formula as implemented in an online Java software algorithm developed by David A. Schoenfeld at Massachusetts General Hospital (http://hedwig.mgh.harvard.edu/sample_size/size.html). The sample size calculated in this way assumes that there are no losses in the patients followed up and that for all subjects all image pairs can be used. More conservative sample size estimations can be done as needed from the basic estimates, for example increasing the sample size by 10% for subject dropout and 10% increase for inadequate scans (Fox et al, 2000).

Spatial reproducibility

Spatial reproducibility was examined using the co-registered test-retest volumes segmented in Dataset 1 to study volume reproducibility effects with no system or sequence changes (Sonata MPRAGE: test and retest), with sequence changes (Sonata: MPRAGE and MEF), with vendor changes (1.5T MPRAGE: Siemens Sonata and GE Signa) and with field strength changes (Siemens MPRAGE: 1.5T and 3T). Dice coefficients where computed for the volume overlap (van Rijsbergen, 1979). In particular, given two different labels (test and retest sessions) of a structure from the same subject, denoted by W1 and W_2, and a function V(W), which takes a label and returns its volume, the Dice coefficient is given by (van Rijsbergen, 1979):

$$D(W_1, W_2)=\frac{V(W_1 \cap W_2)}{(V(W_1)+V(W_2))\ /2}$$

For identical spatial labels, $D(W_1, W_2)$ achieves its maximum value of one, with decreasing values indicating less perfect overlap. For each subject the Dice coefficients were calculated for hippocampus, thalamus, caudate, putamen, pallidum, amygdale and lateral ventricle taking an average across the right and left hemispheres. The group results where generated by averaging the Dice coefficients across subjects for each structure.

RESULTS

We analyzed how the test-retest reliability of subcortical volumes derived from structural T1-weighted MRI data is affected by the following factors: scan session (within and across session), brain segmentation pre-processing (data averaging, correction of intensity inhomogeneities), image acquisition sequence (MPRAGE, MEF), brain segmentation analyses (brain atlas), MRI system upgrade, MRI platform (vendor and field strength). Fig. 1 shows sample subcortical segmentation results (right hemisphere only) color coded according to structure: hippocampus (yellow), thalamus (green), caudate (light blue), putamen (pink), pallidum (dark blue) and amygdala (turquoise). The subcortical labels are generated by FreeSurfer and shown in three orthogonal views (Fig. 1a). The corresponding surface models for each structure (Fig. 1b) are derived using the freely available 3D Slicer package (http://www.slicer.org/).

To investigate how image quality features varied as function of the main parameters manipulated the mean intensity in each ROI with the standard deviation of the voxels within the ROI were calculated (Supplementary Figure 1). These intensity means and standard deviations are the signal-to-noise ratio (SNR) from a segmentation standpoint. The measures where done using the normalized images used for the final automated segmentation. Supplementary Figure 1 has two parts: i) a sample dataset from the old group (Dataset 1) is used to show signal changes as function of vendor, field strength and pulse sequence and ii) a sample dataset from the young group (Dataset 2) is used to show signal changes as function of system upgrade and pulse sequence. For each structure right and left hemispheres intensities and standard deviations were averaged because they did not show mayor differences between. The main observation from these results is that for each structure there are no mayor differences in intensity across conditions. The results also illustrate the challenges in distinguishing structures based on signal intensity only. In particular, we found no substantial differences between 1.5T and 3T data. This may be due to the fact that in the 3T data an 8-channel surface coil was used, which at the brain center gives comparable SNR to the 1.5T birdcage coil.

Table 2 lists the mean volumes of the brain structures studied for Datasets 1 and 2 with their test-retest reproducibility errors and confidence interval bias estimates. Here the reproducibility error considered is within scanner (Siemens Sonata system for Dataset 1 and Siemens Avanto for Dataset 2), across two separate scan sessions and derived from two-averaged MPRAGE scans with no B1 inhomogeneity correction. We found that the volumes of the structures in the old and young groups investigated were not significantly different except only for the lateral ventricles (two-sample t-test, $p<0.01$, smaller lateral ventricle volumes for the young group).

Table 2 shows the group reproducibility errors for each structure derived by averaging the reproducibility errors across subjects, where for each subject the error is estimated as the absolute test-retest volume percent change relative to the mean test-retest volume. Supplementary Figure 2 shows the percent volume reproducibility for each structure listed in Table 2, for both groups. Reproducibility of older subjects is denoted by star symbols, and that of younger subjects with full circle symbols. Results from each young subject are connected with reference lines to facilitate distinction with the reproducibility from the older subjects. We found that the FreeSurfer software generated consistent outcome measures for the hippocampus, thalamus, caudate, putamen, lateral ventricles and total intracranial volume with reproducibility errors 4.3% or less for the older subjects and 2.3% or less for the younger subjects. For the smaller structures (pallidum, amygdala and inferior lateral ventricles) the reproducibility errors were larger, 10.2% or less for older subjects and 10.4% or less for younger subjects.

Table 2 shows also the Jacknife bias estimates of the volume reproducibility confidence interval for each structure. We found that across all structures the mean proportion of the bias to the volume reproducibility was 1.8% for the old group (n=15) and 46% for the young group (n=5). These results suggest that the volume reproducibility estimates from the old group have a relatively low bias and that the number of subjects used was adequate. The fewer number of subjects in the young group gave a higher bias, thus suggesting that the results derived from this group should be considered as preliminary.

Effects of scan session scan averaging and B1 correction

The goals in this study were to compare: a) within-session vs. across-session repeatability of single scans, b) single scan vs. two-averaged repeatability across sessions, and c) the across-session repeatability with or without image intensity correction from B1 inhomogeneities.

Figure 2 shows an example of within-session volume repeatability analysis: Bland-Altman plots of mean volume difference versus mean volume for the two MPRAGE scans acquired in the test session on the group of older subjects (Siemens Sonata, n=15). To help visualization brain hemispheres are differentiated by colors and symbols (left hemisphere: red crosses, right hemisphere: blue circles). For each hemisphere (and corresponding color) the plots show the mean difference volume (solid horizontal line) and the lower and upper boundaries of 95% confidence interval using ±2 SDs (interrupted lines) from the mean difference volume. For all structures the distribution of difference volumes is approximately symmetric around zero (black dotted reference line), with a small constant bias and with both hemispheres behaving similarly.

From Figure 2 we can extract, for each structure, the mean difference volume and the limits of the agreement, each with its 95% confidence interval. This analysis was then extended to the other conditions: within and across session combinations as well as for the case when the two scans of each test and retest sessions are averaged and then compared across sessions. The results are summarized in Figure 3 for all the conditions, with the mean difference volumes in blue and the limits of agreement in red (left hemisphere: crosses, right hemisphere: circles, error bars denote the 95% confidence intervals). A black dotted line connects each of the mean difference volume and limits of agreement across conditions to facilitate the visualization of changes of the two metrics across conditions. As can be seen, neither the volume differences nor the limits of agreement change significantly across conditions, indicating that within-session reproducibility is maintained across sessions and not significantly improved when two scans are averaged in the older group. Similar results were obtained for the younger group (Figure 4, dataset 2, Sonata, n=5). For the young group we also assessed the effects of correction of B1 image intensity inhomogeneities (B1 correction applied to each single MPRAGE prior to averaging within each session), which gave no reproducibility changes.

Effects of image acquisition sequence and probabilistic segmentation atlas

The goals in this study were twofold: a) assess within-session agreement of volumes derived from MPRAGE and MEF acquisitions, and b) compare the across-session test-retest repeatability of the two sequences. MEF segmentations were computed using both the MEF atlas and the MPRAGE atlas.

Figure 5 shows the Bland-Altman agreement results between the subcortical volumes derived from two averaged MPRAGE scans and those derived from the two-flip angle MEF scans (both segmentations using the MPRAGE atlas, older group, n=15, Sonata), both sequences acquired within the same scan session. Similar results where obtained if the MEF atlas was used to segment the MEF acquisitions. For some structures there is a clear offset from zero in the mean volume differences (putamen, lateral ventricles, inferior lateral ventricles and intracranial). For some of these structures (e.g. putamen and inferior lateral ventricles) there seemed to be a systematic linear relationship between the difference volumes and the mean volumes across the measurement range. To investigate this we made linear regression fits for each brain structure grouping the data from both hemispheres and calculated the slopes, which are unitless because both axes have mm^3 units (green lines in Figure 5): hippocampus (0.084±0.001), amygdala (0.015±0.002), caudate (0.133±0.002), putamen (0.216±0.001), pallidum (−0.155 ±0.002), thalamus (−0.176±0.001), lateral ventricles (0.009±0.0004), inferior lateral ventricles (0.200±0.001) and intracranial (0.022±0.0001). The effects of the estimated slopes are relatively small when considering the measurement range and the overall variability. The mean difference volume was taken as the mean across the range of measurements ignoring the slope, which will give an overestimation of the limits of agreement for the structures with the biggest slopes (putamen and inferior lateral ventricles, slope approximately 0.2).

Figure 6 shows the within-session agreement results between MPRAGE and MEF as function of the segmentation atlas used for the MEF data (MEF: MEF atlas, MEF_MP: MPRAGE atlas). Volume differences (blue) and limits of agreement (red), are shown for each structure (left hemisphere: cross symbol, right hemisphere: circle symbol) with their respective 95% confidence interval. For all structures the atlas used to segment the MEF data (MPRAGE or MEF) did not significantly affect the mean volume difference (bias) between the two sequences. For hippocampus, amygdala, caudate, pallidum and thalamus the bias between MPRAGE and MEF derived-volumes was not significantly different than zero. For putamen, lateral ventricles, inferior lateral ventricles and intracranial volumes the mean volume difference was significantly different than zero ($p<0.05$), indicating that a bias correction would be necessary if volumes from these two sequences are to be used in a single analyses. The across-session repeatability results show that the test-retest reproducibility of MPRAGE and MEF (regardless of segmentation atlas) are comparable.

MRI system upgrade effects on subcortical volume reproducibility

One goal of this study was to evaluate whether the reproducibility of volumes can change significantly as consequence of a major MRI system upgrade. Figure 7 shows MRI system upgrade effects (Siemens Sonata-Avanto, 1.5T, dataset 2, n=5 subjects) on volume reproducibility when the volumes are derived from within-session averaged MPRAGE scans and reproducibility is evaluated: before upgrade across sessions (Sonata test- retest: StSrt), across upgrade sessions (Sonata test vs. Avanto test, StAt, Sonata retest vs. Avanto retest, SrtArt) and after upgrade across sessions (Avanto test-retest, ArArt). Other combinations across upgrade were evaluated giving similar results. In Figure 7 we plot the Bland-Altman results of mean volume difference (blue) and limits of agreement (2SD, red) for both brain hemispheres (left: crosses, right: circles) are shown with their respective 95% confidence intervals (error bars) for the various repeatability conditions. We found that in terms of mean volume differences (blue), brain hemispheres behave similarly and with no significant reproducibility changes across conditions ($p<0.05$). We find that for most structures neither the mean volume difference nor its variance changes significantly between the pre-upgrade and the across-upgrade conditions. For some structures there is a reduction in the variance of the volume differences post-upgrade (caudate and thalamus).

A similar analysis was done for the Trio-TIM Trio upgrade (dataset 3, young group, n=5, Supplementary Fig. 3). We found that for most structures the variance of the mean volume differences did not change across conditions. For some structures (lateral ventricles, intracranial) there was a significant reduction of the variance of the reproducibility in the post-upgrade condition. For several structures there was a small bias in the mean volume difference in the across-upgrade conditions (caudate, putamen, pallidum).

Overall, within the limitations of these small datasets, the results suggest that when mixing subcortical, ventricular and intracranial volumes derived from data acquired across a major scanner upgrade, the variance does not significantly change but some structures may show a slight bias.

MRI system vendor and field strength effects on subcortical volume reproducibility

One goal of this study was to compare within-scanner across-session reproducibility (1.5T Siemens Sonata), with the reproducibility obtained in the following conditions of mixed MRI systems (always two-averaged MPRAGE acquisitions from the same session): different scanner vendors at the same field strength (1.5T, Siemens Sonata – GE Signa), different field strengths same vendor (Siemens, 1.5T Sonata – 3T Trio), different field strength and vendor (1.5T GE Signa – 3T Siemens Trio).

Figure 8 shows the Bland-Altman agreement results of the Sonata-Trio condition, showing volume difference vs. volume mean (group of older subjects, n=15). For each brain hemisphere (left: red crosses, right: blue circles) the mean volume difference (solid horizontal line) and the limits of agreement (±2 standard deviations, interrupted horizontal lines) are shown. For reference, zero volume difference is shown as a black dotted line. Green lines show the linear regression fits for each brain structure (grouping data from both hemispheres). Some structures have a non zero bias in the mean volume difference (hippocampus, amygdala, pallidum, thalamus, lateral ventricles, where the absolute value of the mean difference volume is larger than one standard deviation of the volume differences). In the other structures the bias is within one standard deviation from zero (caudate, putamen, inferior lateral ventricles, intracranial). Most structures show a constant bias across the range of volume measurements (regression slope less than 0.1), except for the amygdala and pallidum (regression slopes −0.129 ± 0.003 and −.0526 ± 0.002, respectively). The results for the comparison Sonata-Signa and Signa-Trio give similar results (Supplementary Figure 4 and Supplementary Figure 5, respectively).

Figure 9 summarizes the Bland-Altman results of mean volume difference (blue) and limits of agreement (two standard deviations, red) for the conditions of interest: Sonata-Sonata (Son_Son), Sonata-Signa (Son_Sig), Sonata-Trio (Son-Tri) and Signa-Trio (Sig_Tri). Data is shown for both brain hemispheres (left: crosses, right: circles). Error bars represent the 95% confidence interval. Overall the results show that when combining data of different vendors and/or field strength the standard deviation of the volume differences does not significantly change across conditions relative to the test-retest reproducibility within a fixed MRI system, but the mean volume difference bias does change. These sign and magnitude change of the bias is not systematic; it depends on the specific brain structure and on the MRI vendor-field strength condition. For example, some structures (amygdala and thalamus) show negligible bias in comparisons within the same field (Sonata-Sonata and Sonata-Signa), but significant bias when combining 1.5T with 3T data (Sonata-Trio and Signa-Trio). For other structures (hippocampus and lateral ventricles) the Sonata-Sonata and Signa-Trio conditions give comparable biases, with larger bias effects in the other two conditions (Sonata-Signa and Sonata-Trio).

Power analyses

One goal of this study was to use the test-retest variance of hippocampal volume measures to estimate the sample size that would allow detection of a difference between an untreated group of Alzheimer’s patients (4.90 %/year reduction) vs. an hypothetically treated group with half that atrophy rate (2.45 %/year) to characterize treatment effects (Jack et al., 2003). In the older group of subjects (dataset 1, n=15) we found that the standard deviation of the hemispheric hippocampal volume measurement error (test-retest within the same scanner across sessions) was approximately 3.5%, with mean hemispheric volume of 3400 mm^3 (Table 2). The variability error of this structure was not different in the comparisons across field or platforms. This gives a sample size of 36 subjects in each treatment arm to detect a difference in hippocampal volume decrease rate of 2.45 %/year (90% power, 5% significance).

Spatial reproducibility

One goal of this study was to evaluate the spatial reproducibility of segmentation labels derived from data acquired in separate test-retest sessions, either with simple repetitions of exactly the same protocol (Sonata MPRAGE), or changing sequence (Sonata system, MPRAGE and MEF), or changing MR vendor (1.5T MPRAGE, Siemens Sonata and GE Signa) or changing field strength (Siemens MPRAGE: 1.5T Sonata and 3T Trio). The group-wise mean and standard deviation of the Dice coefficients are shown in Table 3 for seven individual structures and the average of all structures. There were no significant differences between the Dice coefficients computed in the different conditions, indicating that the spatial reproducibility was as good (0.88 in average for all structures) as for the case in which data is acquired from the same MR system and sequence.

DISCUSSION

In this paper, we show that human subcortical volume estimates derived from brain structural MRI data are remarkably reproducible for a variety of data acquisition and analysis factors when using the publicly available FreeSurfer automated segmentation tool. Specifically, using a group of healthy older (mean age 69.5 years, n=15) and two different groups of young subjects (n=5 for both, mean ages 34 and 36.5 years) we examined how the volume test-retest reproducibility of hippocampus, thalamus, caudate, putamen, pallidum, amygdala, ventricular and intracranial structures are affected by scan session, structural MRI acquisition sequence, data preprocessing, subcortical segmentation analyses, major MRI system upgrades, vendor and field effects. We identified a number of factors that contribute little to within- or across-session variability, and other factors that contribute potentially important variability to within- and across-session variability.

The segmentation errors reported in this work represent the best estimate we can give for the error of the method under the reported measurement conditions. The main factors that introduce errors in the final segmentation results are image quality factors (signal-to-noise ratio and contrast-to-noise ratio) and brain anatomical variability relative to the probabilistic atlas. These factors are intermingled. Realistic brain anatomical simulations with pre-defined characteristics for subcortical structures and their spatial arrangements could be attempted to separate the contribution of segmentation errors from image quality and segmentation atlas factors. These issues are important but are beyond the scope of this manuscript. The closest to a ground-truth that can be currently used to assess the accuracy of the FreeSurfer segmentation method is the comparison with manual segmentations by a neuroanatomist, as validated in Fischl et al. 2002.

The segmentation results (Table 2) are comparable with previously reported results (The Internet Brain Volume Database, http://www.cma.mgh.harvard.edu/ibvd/). For most structures, there's a fairly wide range of estimates of normal volume, and ours are within the typical range.

We expect the best possible volume reproducibility from data acquired within the same scanning session using identical acquisition sequences. For both the old and the young groups we find that within-session reproducibility was comparable to across session reproducibility when data was acquired with the same MRI system. For the hippocampus, thalamus, caudate, putamen, lateral ventricles and intracranial volumes, reproducibility error across sessions in the same scanner were less than 4.3% in the older group and less than 2.3% in the young group. This difference is most likely due to the fact that older subjects tend to move more during scans, hence giving suboptimal image quality (gray-white matter contrast to noise ratio) relative to the younger subjects. Smaller structures (pallidum, amygdala and inferior lateral ventricles) gave higher reproducibility errors (under 10.2% for the old group and under 10.4% for the young group). The reproducibility error is derived as (100*SD/MEAN) where SD is the standard deviation of the test-retest volume differences and MEAN is the mean volume within the group. For small structures MEAN decreases and therefore for a similar or worse SD the reproducibility error increases. The result that the reproducibility of the young and older group becomes more similar for smaller volumes indicates that the size effect (MEAN volume) in the reproducibility dominates the SD differences between groups.

Having an adequate number of subjects is very important to minimize biases in the results, yet it can be challenging for reproducibility studies like the one described here because of the significant cost in scanner time usage. Each subject in each of the three datasets (dataset 1: 15 subjects, dataset 2: 5 subjects, dataset 3: 5 subjects) was scanned in four different 1-hour sessions, so the effective ‘hourly costs’ were 60, 20 and 20 for datasets 1, 2 and 3, respectively. The acquisition of the scanner upgrade data (datasets 2 and 3) has additional practical challenges: scans have to be acquired within a short time before/after the upgrade. In particular right before an upgrade scanner availability tends to be lower than normal because of the need that many projects have for completing acquisitions before the upgrade. For this reason datasets 2 and 3 resulted with fewer subjects, with gender biases that were hard to avoid. The Jacknife bias analysis indicated that the number of subjects used in the older dataset gave a relatively low mean proportional bias across the structures investigated (1.8%), whereas the same value was substantially higher for the young group (46%). This indicates that the results derived from the young groups (scanner upgrade effects and B1 inhomogeneity correction effects) should be considered as preliminary and in need for further validation with a larger dataset.

In agreement with results obtained in a cortical thickness reproducibility study (Han et al., 2006), we found that averaging two acquisitions made relatively minor contributions to improvement in the reproducibility of subcortical volumes. The acquisition of two MPRAGE volumes is still recommended mainly for practical reasons. If both scans are good they can either be averaged or the best quality scan selected for the segmentation. If one volume is bad (e.g. due to motion artifacts) then the other can still be used for segmentation without averaging. Furthermore, the data acquired and analyzed in this study were collected under ideal circumstances, with cooperative volunteer participants and highly skilled scanner operators, and both of these factors may reduce the apparent added value of averaging multiple acquisitions. In addition, as the signal-to-noise ratio of a single acquisition diminishes (e.g., with parallel acquisition acceleration protocols), the added value of volumes averaged from multiple acquisitions may increase.

In the small sample of young subjects we found that the B1 inhomogeneity correction method tested did not significantly improve volume reproducibility, suggesting that the extra calibration-related scans and inhomogeneity correction pre-processing step can be avoided when only data acquired with the same MRI system will be considered. Further, the standard automated Freesurfer segmentation includes an intensity normalization step (Non-parametric Non-uniform intensity Normalization, N3), so our results suggest that the effects of the N3 correction are stronger than the corrections introduced by our B1 corrections. We did not have data to evaluate whether B1 correction improves reproducibility across MRI system vendors or field strength, but will be critical for large N phased arrays or small coils in general.

The choice of imaging sequence (MPRAGE or multi-echo FLASH) with the corresponding brain atlas used for the automated segmentation analyses did not show significant differences in volume repeatability. This suggests that the segmentation algorithm is robust across a variety of similar image contrast properties, thus alleviating the need to create manually labeled probabilistic atlases for different acquisition methods, consistent with recent work (Han & Fischl, 2007). The comparison between subcortical volumes derived from MPRAGE and MEF sequences showed that for some structures (putamen, lateral ventricles, inferior lateral ventricles, and intracranial) there were significant biases in the mean volume difference given by the two methods. These differences may be due to the differential sensitivity (acquisition bandwidth) that the sequences have to signal T2* (signal loss, geometric distortions). The MPRAGE sequence has the advantage of being currently more standard than multi-echo FLASH, thus it is easier to implement it consistently in multi-center studies. It is also important to recognize that the MPRAGE and multi-echo FLASH sequences have very similar contrast properties, which may not apply to T1 sequences with different contrast properties (e.g., SPGR) or non T1 sequences.

Within the limits of our small sample size we find that in major MRI system upgrades (Sonata-Avanto and Trio-TrioTIM) combining pre- and post-upgrade data does not significantly worsen the variance but may introduce a bias in the mean volume differences. Combining this with our segmentation atlas results suggests that it is safe to use the same brain atlas after a system upgrade, which is very convenient. For longitudinal studies we believe that it is appropriate to plan a system upgrade calibration study as part of the design, particularly with samples from the population under study scanned shortly prior to and immediately after the upgrade for a correct estimation of potential biases. An important practical issue is to know about the upgrade sufficiently far in advance to plan for the calibration study, optimally complete the study prior to the upgrade. If the longitudinal study will continue after the upgrade it should ideally be balanced across relevant study groups with respect to number of acquisitions before and after the upgrade, since subtle effects of interest in longitudinal studies may in fact be within the small range of variance identified in this study (e.g., hippocampal volume differences of 2–5%).

We found that when data from different MRI systems are combined (same field different vendors, same vendor different fields, or different vendor and fields) then the variance of the volume differences doesn't significantly change relative to the test-retest reproducibility from data acquired in a fixed MRI system, but biases of the mean volume differences may become significant. All data were segmented using an atlas from a single MRI system suggesting that image contrast differences arising from differences in hardware and field strength were strong enough to be detected by the segmentation algorithm. The spatial reproducibility results showed constant and high spatial consistency of the segmentation volumes (average Dice coefficient of 0.88 ± 0.04) for a variety of test-retest conditions, ranging from no MR system and sequence changes to changes of system, sequence or field strength. The spatial overlap results are also in good agreement with a previous study (Han et al., 2007) that compared Siemens Sonata segmentations of the same structures with manual segmentations, suggesting that both spatial accuracy and reproducibility and accuracy are high.

One extension of this work would be to test if reproducibility differences across MRI platforms can be reduced by doing the subcortical segmentations with a probabilistic atlas that is constructed from manual segmentations of data acquired with the various MRI systems under consideration. Alternatively, statistical models could be used given that they have been proved successful in combining data that are sufficiently different in acquisition sequence to fail pooling straight out (Fennema-Notestine et al, 2007). Further, our cross-vendor comparison (Siemens Sonata MPRAGE – GE Signa MPRAGE) did not include potentially significant sources of variation found when each vendor uses its own product sequence, which can lead to image contrast differences. Therefore our results might underestimate the variance seen with cross vendor switches that introduce strong sequence changes, as may occur in practice.

The fact that the rest-retest reproducibility variance of the segmented volumes does not significantly change across platforms and field strengths (particularly in the hippocampus) implies that a multicenter study with these MRI systems does not necessarily require a much larger sample dataset to detect a specific effect. Of course, this is under ideal circumstances with highly motivated cognitively intact older adults. These conclusions may not generalize to other brain structures or to patient populations with cognitive impairment if there are any reductions in raw data quality related to movement or other issues. Our power calculations for detecting a net 2.45% hippocampal volume reduction rate difference between hypothetical non-treated and treated AD groups resulted in an estimate of 49 subjects per group, which but was not appreciably worsened by scanner upgrades or differences in scanner platform or field strength. These results differ from previous calculations (Jack et al 2003) which estimate, for the same treatment effect, 21 subjects per treatment arm with a 2.1% standard deviation. The differences may be due to the fact that Jack et al used various data adjustments that were not applied in our analysis, including normalization for total intracranial volume, adjustments for age and gender and corrections for skew data distributions. These adjustments might help reducing the reproducibility error thereby reducing the sample size.

In addition to their volume, subcortical structures have started to be characterized also by their 3D shapes (Munn, 2007; Wang, 2007; Patenaude, 2006; Miller, 2004). Combining both volume and shape metrics might improve the power of detecting cross-sectional differences across populations or longitudinal changes. An important extension of the reproducibility study here presented would be to examine the reproducibility of shape metrics. An important limitation of this study is the lack of quantification of spatial differences in voxel labeling; that is, different voxels may be labeled as the same structure in two different sessions, but if the volumes do not differ, our analyses would not demonstrate this potentially important variance effect.

Knowledge of the degree to which different MRI instrument-related factors affect the reliability of metrics that characterize subcortical structures is essential for the interpretation of these measures in basic and clinical neuroscience studies. Furthermore, the knowledge of reproducibility is critical if these metrics are to find applications as biomarkers in clinical trials of putative treatments for neurodegenerative or other neuropsychiatric diseases, particularly with the growth of large sample multi-center studies (Jack et al., 2003; Mueller et al., 2005; Murphy et al., 2006; Belmonte et al., 2007).

To conclude, our results suggest that, for the purpose of designing morphometric longitudinal studies at a single site, one structural MPRAGE acquisition segmented with the corresponding MPRAGE atlas can be optimal. Subcortical volumes derived from T1-weighted structural imaging data acquired at a single 1.5T site are reliable measures that can be pooled even if there are differences in image acquisition sequence and major system upgrades. However, MRI-instrument specific factors should be considered when combining data from different MRI systems (vendors and/or fields). It should be noted that we do not report a random effects study, therefore the results should not be extrapolated to pulse sequences or scanners not included in this study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the following grants: NCRR Morphometry Biomedical Informatics Research Network (U24 RR021382), NCRR P41-RR14075 and RO1-RR16594-01A1, Pfizer Inc., the NIA (K23-AG22509 & P01-AG04953) and the MIND Institute. We thank Gabriela Czanner, Mark Vangel and David Raunig for useful discussions on the statistical analyses of the data.

References

Alecci M, Zhang Y, Brady JM, Jezzard P, Smith S. Image-based evaluation of a-priori B1 field correction and its effect on MRI tissue segmentation. Proc. Int. Soc. of Magnetic Resonance in Medicine 2000:109.

Alemán-Gómez, Y.; Melie-García, L.; Valdés-Hernandez, P. IBASPM: Toolbox for automatic parcellation of brain structures; Human Brain Mapping, 12th Annual Meeting; Florence, Italy. 2007.

Anstey KJ, Maller JJ. The role of volumetric MRI in understanding mild cognitive impairment and similar classifications. Aging Ment Health 2003;7:238–250. [PubMed: 12888435]

Ashburner J, Friston KJ. Voxel-based morphometry - The methods. NeuroImage 2000;11:805–821. [PubMed: 10860804]

Barnes J, Lewis EB, Scahill RI, Bartlett JW, Frost C, Schott JM, Rossor MN, Fox NC. Automated measurement of hippocampal atrophy using fluidregistered serial MRI in AD and controls. J. Comput. Assist. Tomogr 2007;31:581–587. [PubMed: 17882036]

Belmonte MK, Mazziotta JC, Minshew NJ, Evans AC, Courchesne E, Dager SR, Bookheimer SY, Aylward EH, Amaral DG, Cantor RM, Chugani DC, Dale AM, Davatzikos C, Gerig G, Herbert MR, Lainhart JE, Murphy DG, Piven J, Reiss AL, Schultz RT, Zeffiro TA, Levi-Pearl S, Lajonchere C, Colamarino SA. Offering to share: how to put heads together in autism neuroimaging. J Autism Dev. Disord 2008;38(1):2–13. [PubMed: 17347882]

Benner T, Wisco JJ, van der Kouwe A, Fischl B, Vangel MG, Hochberg FH, Sorensen AG. Comparison of manual and automatic slice positioning of brain MR images. Radiology 2006;239:246–254. [PubMed: 16507753]

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–310. [PubMed: 2868172]

Camara O, Scahill RI, Schnabel JA, Crum WR, Ridgway GR, Hill DL, Fox NC. Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal data. Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv 2007;10:785–792.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2nd edn.. New York: Academic Press; 1988.

Collins, DL.; Zijdenbos, AP.; Barré, WFC.; Evans, AC. ANIMAL+INSECT: Inproved cortical structure segmentation. In: Kuba, A.; Samal, M.; Todd-Pokropek, A., editors. Proc. of the Annual Symposium on Information Processing in Medical Imaging. Berlin: Springer; 1999. p. 210-223.1613 of LNCS

Czanner, S.; Han, X.; Pacheco, J.; Wallace, S.; Busa, E.; van der Kouwe, A.; Fischl, B.; Jovicich, J. Test-retest reliability assessment for longitudinal MRI studies: effects of MRI system upgrade on morphometric analysis of structural MRI data. Human Brain Mapping; 12th Annual Meeting; Florence, Italy. 2006.

DeKosky ST, Marek K. Looking Backward to Move Forward: Early Detection of Neurodegenerative Disorders. Science 2003;302:830–834. [PubMed: 14593169]

Dickerson BC, Sperling RA. Neuroimaging biomarkers for clinical trials of disease-modifying therapies in Alzheimer's disease. NeuroRx 2005;2:348–360. [PubMed: 15897955]

Douaud G, Gaura V, Ribeiro MJ, Lethimonnier F, Maroy R, Verny C, Krystkowiak P, Damier P, Bachoud-Levi AC, Hantraye P, Remy P. Distribution of grey matter atrophy in Huntington's disease patients: A combined ROI-based and voxel-based morphometric study. Neuroimage 2006;32:1562–1575. [PubMed: 16875847]

Efron, B.; Tibshirani, RJ. An Introduction to the Boostrap. New York: Chapman & Hall/CRC; 1993.

Ewers M, Teipel SJ, Dietrich O, Schonberg SO, Jessen F, Heun R, Scheltens P, Pol L, Freymann NR, Moeller HJ, Hampel H. Multicenter assessment of reliability of cranial MRI. Neurobiol Aging 2006;27:1051–1059. [PubMed: 16169126]

Fennema-Notestine C, Gamst AC, Quinn BT, Pacheco J, Jernigan TL, Thal L, Buckner R, Killiany R, Blacker D, Dale AM, Fischl B, Dickerson B, Gollub RL. Feasibility of Multi-site Clinical Structural Neuroimaging Studies of Aging Using Legacy Data. Neuroinformatics 2007;5:235–245. Nov 13 [Epub ahead of print]. [PubMed: 17999200]

Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from Magnetic Resonance Images. Proceed. Nat. Acad. Sciences 2000;97:11044–11049.

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–355. [PubMed: 11832223]

Fischl B, Salat DH, van der Kouwe AJW, Makris N, Ségonne F, Dale AM. Sequence-Independent Segmentation of Magnetic Resonance Images. NeuroImage 2004;22:1060–1075. [PubMed: 15219578]

Fox NC, Freeborough PA. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer’s disease. J. Magn. Reson. Imaging 1997;7:1069–1075. [PubMed: 9400851]

Fox NC, Cousens S, Scahill R, Harvey RJ, Rossor MN. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects. Arch. Neurol 2000;57:339–344. [PubMed: 10714659]

Haller JW, Banerjee A, Christensen GE, et al. Threedimensional hippocampal MR morphometry with high-dimensional transformation of a neuroanatomic atlas. Radiology 1997;202:504–510. [PubMed: 9015081]

Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R, Maguire P, Rosas D, Makris N, Dale AM, Dickerson B, Fischl B. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. Neuroimage 2006;32:180–194. [PubMed: 16651008]

Han X, Fischl B. Atlas Renormalization for Improved Brain MR Image Segmentation Across Scanner Platforms. Medical Imaging, IEEE Transactions 2007;26:479–486.

Jack CR Jr, Slomkowski M, Gracon S, Hoover TM, Felmlee JP, Stewart K, Xu Y, Shiung M, O'Brien PC, Cha R, Knopman D, Petersen RC. MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. Neurology 2003;60:253–260. [PubMed: 12552040]

Jack CR Jr, Shiung MM, Weigand SD, O'Brien PC, Gunter JL, Boeve BF, Knopman DS, Smith GE, Ivnik RJ, Tangalos EG, Petersen RC. Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. Neurology 2005;65:1227–1231. [PubMed: 16247049]

Jack CR Jr, Slomkowski M, Gracon S, Hoover TM, Felmlee JP, Stewart K, Xu Y, Shiung M, O'Brien PC PC, Cha R, Knopman D, Petersen RC. MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. Neurology 2003;60:253–260. [PubMed: 12552040]

Jovicich, J.; Czanner, S.; Greve, D.; Pacheco, J.; Busa, E.; van der Kouwe, A.; Morphometry, BIRN.; Fischl, B. International Society of Magnetic Resonance in Medicine. Miami, USA: 2005. Test-retest reproducibility assessments for longitudinal studies: quantifying MRI system upgrade effects.

Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, Macfall J, Fischl B, Dale AM. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 2006;30:436–443. [PubMed: 16300968]

Kassube J, Juengling FD, Kioschies T, Henkel K, Karitzky J, Kramer B, Ecker D, Andrich J, Saft C, Kraus P, Aschoff AJ, Ludolph AC, Landwehrmeyer GB. Topography of cerebral atrophy in early Huntington's disease: a voxel based morphometric MRI study. J Neurol Neurosurg Psychiatry 2004;75:213–220. [PubMed: 14742591]

Kantarci K, Jack CR Jr. Quantitative magnetic resonance techniques as surrogate markers of Alzheimer's disease. NeuroRx 2004;1:196–205. [PubMed: 15717020]

Kassubek J, Juengling FD, Ecker D, Landwehrmeyer GB. Thalamic atrophy in Huntington's disease co-varies with cognitive performance: a morphometric MRI analysis. Cereb Cortex 2005;15:846–853. [PubMed: 15459079]

Kipps CM, Duggins AJ, Mahant N, Gomes L, Ashburner J, McCusker EA. Progression of structural neuropathology in preclinical Huntington's disease: a tensor based morphometry study. J Neurol Neurosurg Psychiatry 2005;76:650–655. [PubMed: 15834021]

Koo MS, Levitt JJ, McCarley RW, Seidman LJ, Dickey CC, Niznikiewicz MA, Voglmaier MM, Zamani P, Long KR, Kim SS, Shenton ME. Reduction of caudate nucleus volumes in neuroleptic-naive female subjects with schizotypal personality disorder. Biol Psychiatry 2006;60:40–48. [PubMed: 16460694]

Kouwe van der A, Benner T, Fischl B, Schmitt F, Salat D, Harder M, Sorensen AG, Dale AM. On-line automatic slice positioning for brain MR imaging. NeuroImage 2005;27:222–230. [PubMed: 15886023]

Kuroki N, Kubicki M, Nestor PG, Salisbury DF, Park HJ, Levitt JJ, Woolston S, Frumin M, Niznikiewicz M, Westin CF, Maier SE, McCarley RW, Shenton ME. Fornix integrity and hippocampal volume in male schizophrenic patients. Biol Psychiatry 2006;60:22–31. [PubMed: 16406249]

Leow AD, Yu CL, Lee SJ, Huang SC, Nicolson R, Hayashi KM, Protas H, Toga AW, Thompson PM. Brain structural mapping using a novel hybrid implicit/explicit framework based on the level-set method. NeuroImage 2005b;24:910–927. [PubMed: 15652325]

Leow AD, Klunder AD, Jack CR Jr, Toga AW, Dale AM, Bernstein MA, Britson PJ, Gunter JL, Ward CP, Whitwell JL, Borowski BJ, Fleisher AS, Fox NC, Harvey D, Kornak J, Schuff N, Studholme C, Alexander GE, Weiner MW, Thompson PM. ADNI Preparatory Phase Study. Longitudinal stability of MRI for mapping brain change using tensor-based morphometry. Neuroimage 2006;31:627–640. [PubMed: 16480900]

Magnotta VA, Harris G, Andreasen NC, O'Leary DS, Yuh WT, Heckel D. Structural MR image processing using the BRAINS2 toolbox. Comput Med Imaging Graph 2002;26:251–264. [PubMed: 12074920]

Makris N, Goldstein JM, Kennedy D, Hodge SM, Caviness VS, Faraone SV, Tsuang MT, Seidman LJ. Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr Res 2006;83:155–171. [PubMed: 16448806]

Miller MI. Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms. Neuroimage 2004;23:S19–S33. [PubMed: 15501089]

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L. The Alzheimer’s disease neuroimaging initiative. Neuroimaging Clin. N. Am 2005;15:869–877. xi-xii. [PubMed: 16443497]

Mueller SG, Stables L, Du AT, Schuff N, Truran D, Cashdollar N, Weiner MW. Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. Neurobiol Aging 2007;28:719–726. [PubMed: 16713659]

Munn MA, Alexopoulos J, Nishino T, Babb CM, Flake LA, Singer T, Ratnanather JT, Huang H, Todd RD, Miller MI, Botteron KN. Amygdala volume analysis in female twins with major depression. Biol Psychiatry 2007;62:415–422. [PubMed: 17511971]

Murphy SN, Mendis ME, Grethe J, Gollub R, Kennedy D, Rosen BR. A Web portal that enables collaborative use of advanced medical image processing and informatics tools through the Biomedical Informatics Research Network (BIRN). AMIA Annu. Symp. Proc 2006:579–583. [PubMed: 17238407]

Patenaude, BM.; Smith, S.; Kennedy, D.; Jenkinson, M. Automated subcortical segmentation using statistical shape models; Twelfth Annual Meeting of the Organization for Human Brain Mapping; 2006.

Peinemann A, Schuller S, Pohl C, Jahn T, Weindl A, Kassubek J. Executive dysfunction in early stages of Huntington's disease is associated with striatal and insular atrophy: a neuropsychological and voxel-based morphometric study. J Neurol Sci 2005;239:11–19. [PubMed: 16185716]

Pengas G, Pereira JMS, Williams GB, Nesto PJ. Comparative Reliability of Total Intracranial Volume Estimation Methods and the Influence of Atrophy in a Longitudinal Semantic Dementia Cohort. J Neuroimaging. 2008 (in press).

Rosas HD, Koroshetz WJ, Chen YI, Skeuse C, Vangel M, Cudkowicz ME, Caplan K, Marek K, Seidman LJ, Makris N, Jenkins BG, Goldstein JM. Evidence for more widespread cerebral pathology in early HD: an MRI-based morphometric analysis. Neurology 2003;60:1615–1620. [PubMed: 12771251]

Senjem ML, Gunter JL, Shiung MM, Petersen RC, Jack CT Jr. Comparison of different methodological implementations of voxel-based morphometry in neurodegenerative disease. Neuroimage 2005;26:600–608. [PubMed: 15907317]

Shenton ME, Dickey CC, Frumin M, McCarley RW. A review of MRI findings in schizophrenia. Schizophr Res 2001;49:1–52. [PubMed: 11343862]

Smith, SM.; De Stefano, N.; Jenkinson, M.; Matthews, PM. Measurement of brain change over time, FMRIB Technical Report TR00SMS1. 2002. http://www.fmrib.ox.ac.uk/analysis/research/siena/siena/siena.html

Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, De Stefano N. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 2002;17:479–489. [PubMed: 12482100]

Studholme C, Cardenas V, Schuff N, Rosen H, Miller B, Weiner MW. Detecting spatially consistent structural differences in Alzheimer’s and frontotemporal dementia using deformation morphometry. MICCAI 2001:41–48.

Szentkuti A, Guderian S, Schiltz K, Kaufmann J, Munte TF, Heinze HJ, Duzel E. Quantitative MR analyses of the hippocampus: unspecific metabolic changes in aging. J Neurol 2004;251:1345–1353. [PubMed: 15592730]

Tae WS, Kim SS, Lee KU, Nam EC, kim KW. Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. Diagnostic Neuroradiology. 2008 DOI 10.1007/s00234-008-0383-9.

Thieben MJ, Duggins AJ, Good CD, Gomes L, Mahant N, Richards F, McCusker E, Frackowiak RS. The distribution of structural neuropathology in pre-clinical Huntington’s disease. Brain 2002;125:1815–1828. [PubMed: 12135972]

van der Kouwe AJW, Benner T, Salat DH, Fischl B. Brain morphometry with multiecho MPRAGE. Neuroimage 2008;40:559–569. [PubMed: 18242102]

van Rijsbergen, CJ. Information Retrieval. 2nd ed. London, U.K: Butterworths; 1979.

Walters RJC, Fox NC, Crum WR, Taube D, Thomas DJ. Hemodialysis and cerebral edema. Nephron 2001;87:143–147. [PubMed: 11244309]

Wang L, Beg F, Ratnanather T, Ceritoglu C, Younes L, Morris JC, Csernansky JG, Miller ML. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE Trans Med Imaging 2007;26:462–470. [PubMed: 17427733]

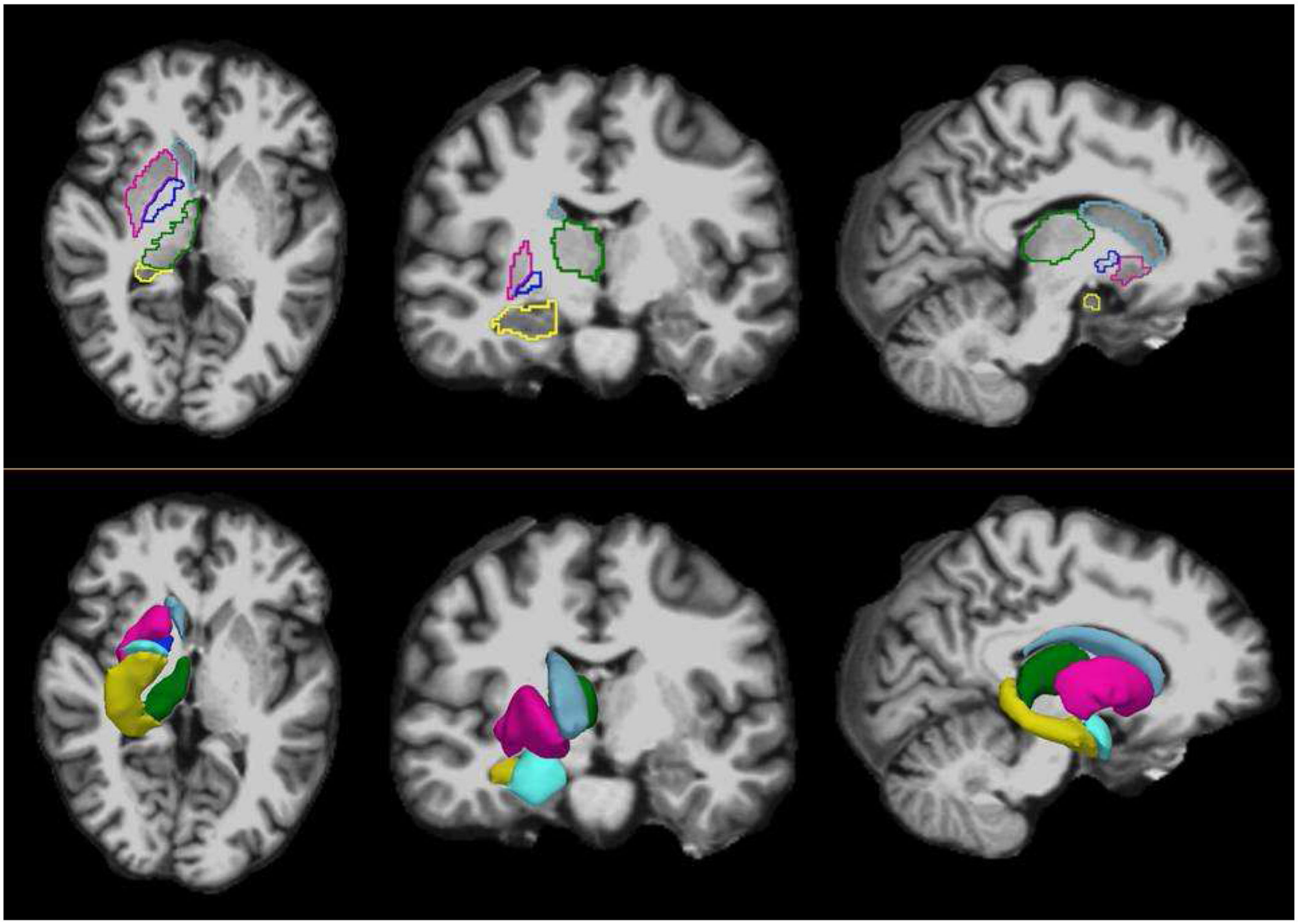

Figure 1.

Sample color-coded subcortical segmentation results (right hemisphere only): hippocampus (yellow), thalamus (green), caudate (light blue), putamen (pink), pallidum (dark blue) and amygdala (turquoise). Top: Freesurfer derived subcortical labels, from a two-averaged MPRAGE in axial (left), coronal (center) and sagittal (right) views. Bottom: 3D surface models created with 3D Slicer derived from the Freesurfer subcortical segmentations.

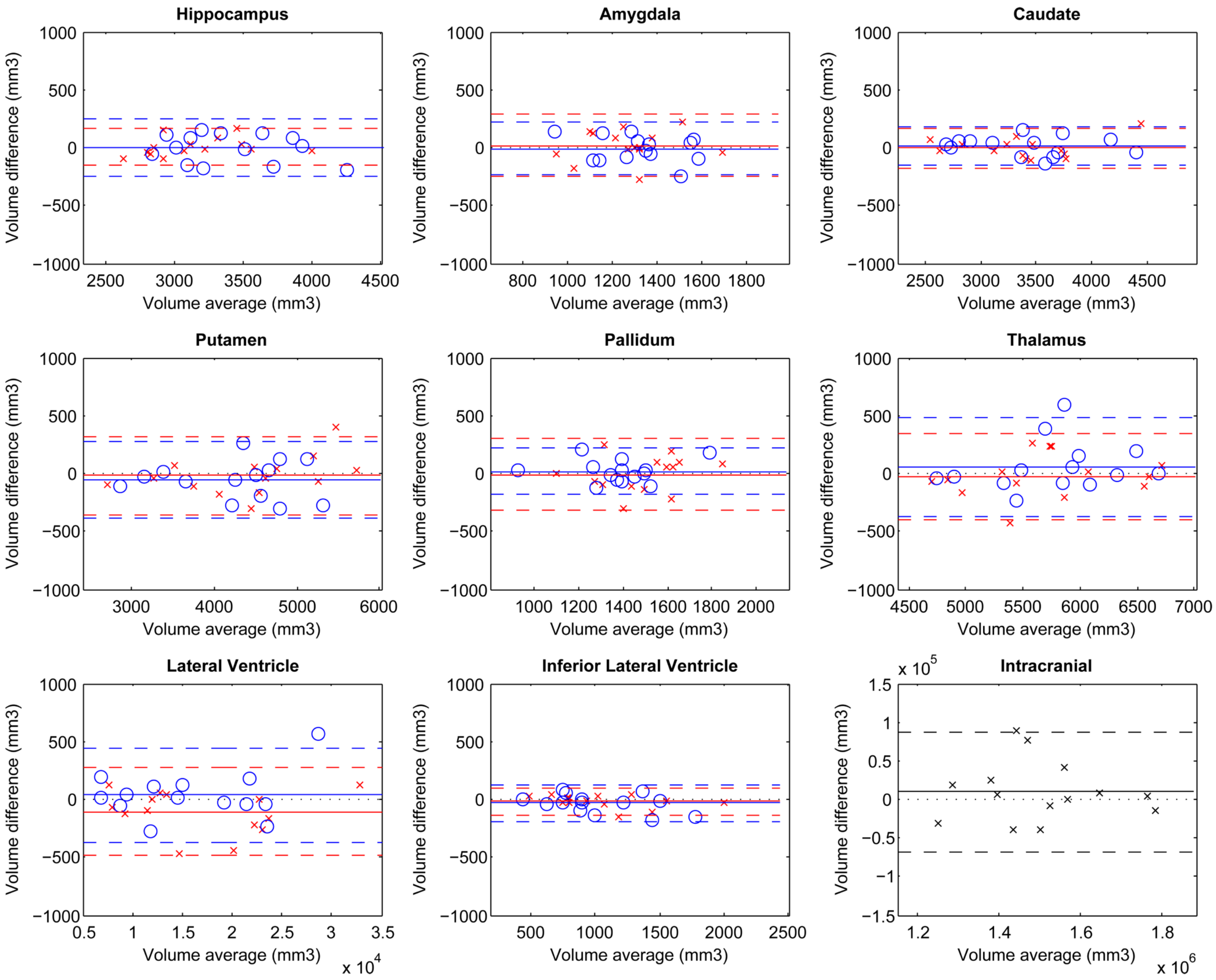

Figure 2.

Within-session repeatability of volumes: Bland-Altman plots showing volume difference vs. volume mean (single MPRAGE acquisitions, Siemens Sonata, group of older subjects, n=15). For each brain hemisphere (left: red crosses, right: blue circles) the mean volume difference (solid horizontal line) and the limits of agreement (±2 standard deviations, interrupted horizontal lines) are shown. For reference, zero volume difference is shown as a black dotted line.

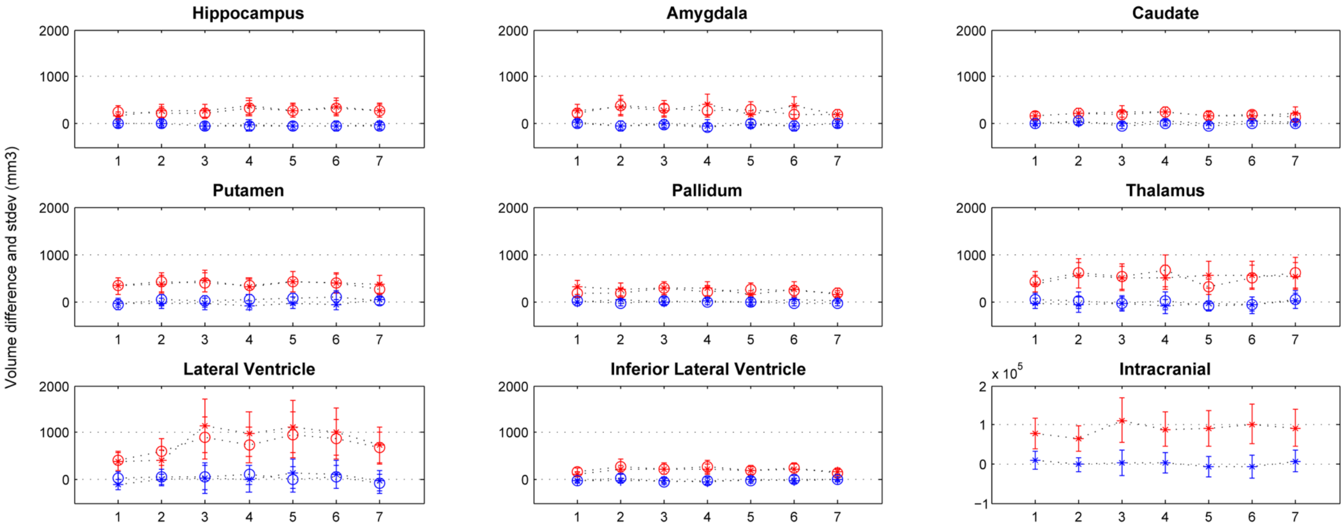

Figure 3.

Effects of scan session (within session and across sessions) and data averaging on volume repeatability in the older group (MPRAGE acquisitions, Siemens Sonata, n=15). Bland-Altman results of mean volume difference (blue) and limits of agreement (red) for both brain hemispheres (left: crosses, right: circles) are shown with their respective 95% confidence intervals for the various repeatability conditions: test session, (1: variability across the two acquisitions); retest session (2: variability across the two acquisitions); single acquisition data mixed from the test-retest sessions (3: first test with first retest acquisitions; 4: first test with second retest acquisitions; 5: second test with first retest acquisition and 6: second test with second retest acquisitions); and average scans from each session (7: two test scans averaged with two retest scans averaged). Black dotted lines connect the volume differences and the limits of agreement across conditions.

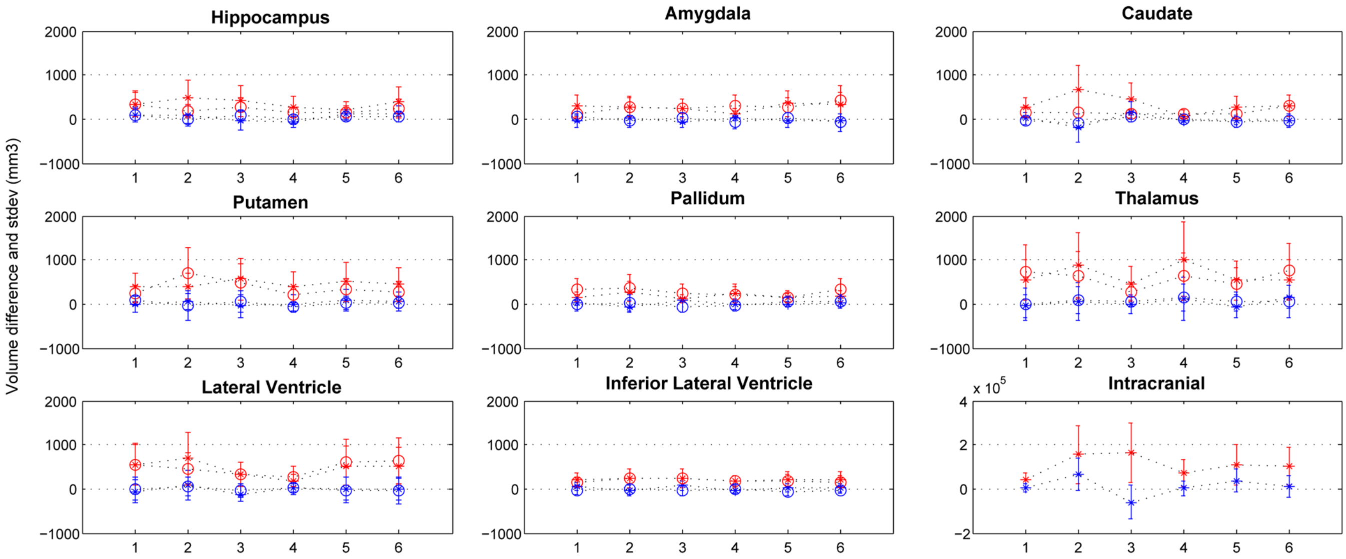

Figure 4.

Effects of scan session (within, across), data averaging and B1 inhomogeneity corrections in volume reproducibility in the younger group (MPRAGE acquisitions, Siemens Sonata, n=5). Bland-Altman results of mean volume difference (blue) and limits of agreement (red) for both brain hemispheres (left: crosses, right: circles) are shown with their respective 95% confidence intervals for the various repeatability conditions: test session, (1: variability across the two acquisitions); retest session (2: variability across the two acquisitions); single acquisition data mixed from the test-retest sessions (3: first test with first retest acquisitions; 4: second test with second retest acquisitions); average scans from each session (5: two test scans averaged with two retest scans averaged) and average B1 corrected scans (6: two B1 corrected test scans averaged with two B1 corrected retest scans averaged).

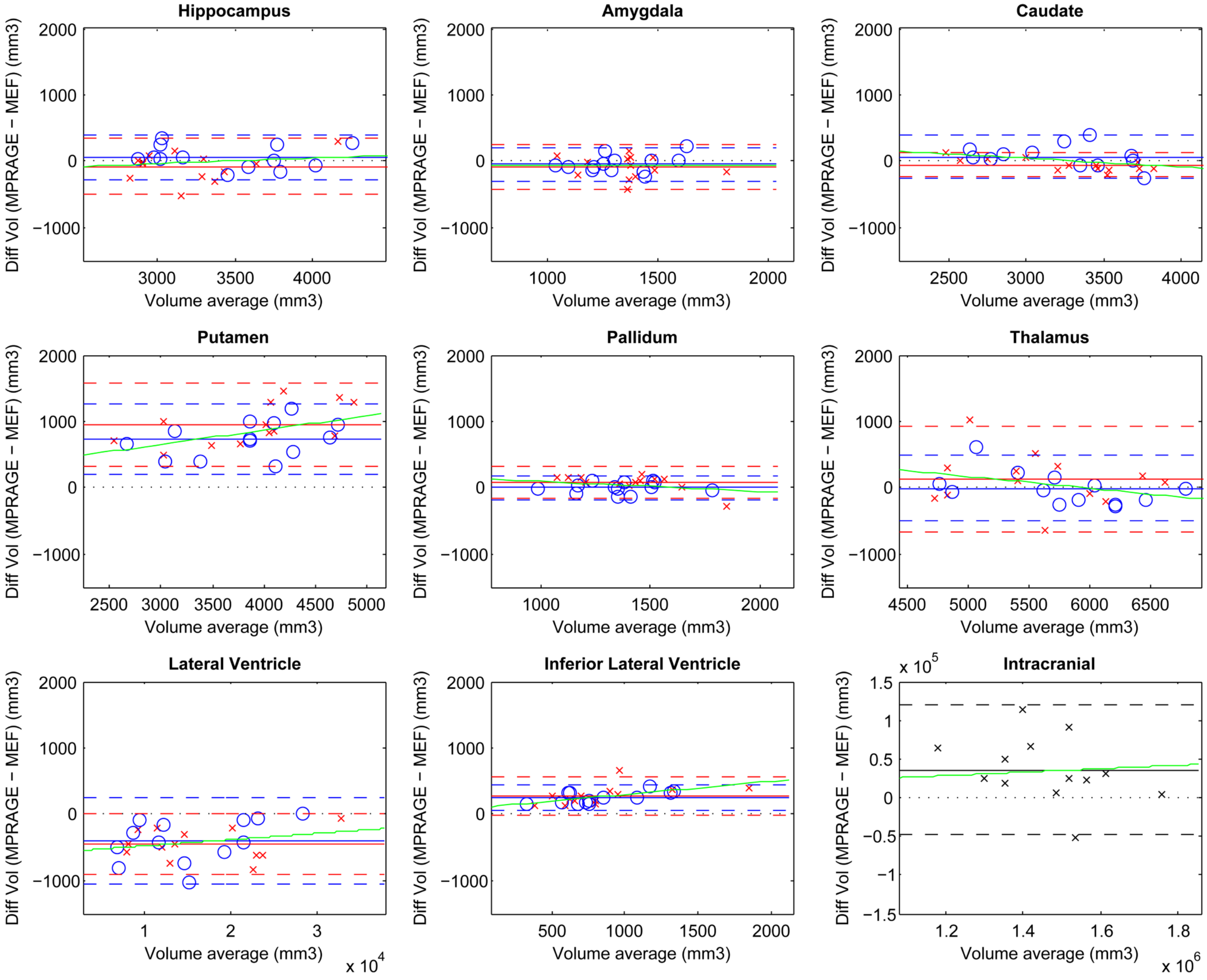

Figure 5.

Within-session agreement between volume estimates derived from MPRAGE and MEF acquisitions. Bland-Altman plots showing volume difference vs. volume mean (Siemens Sonata, group of older subjects, n=15). For each brain hemisphere (left: red crosses, right: blue circles) the mean volume difference (solid horizontal line) and the limits of agreement (±2 standard deviations, interrupted horizontal lines) are shown. For reference, zero volume difference is shown as a black dotted line. Green lines show the linear regression fits for each brain structure (grouping data from both hemispheres). See text for the regression slopes.

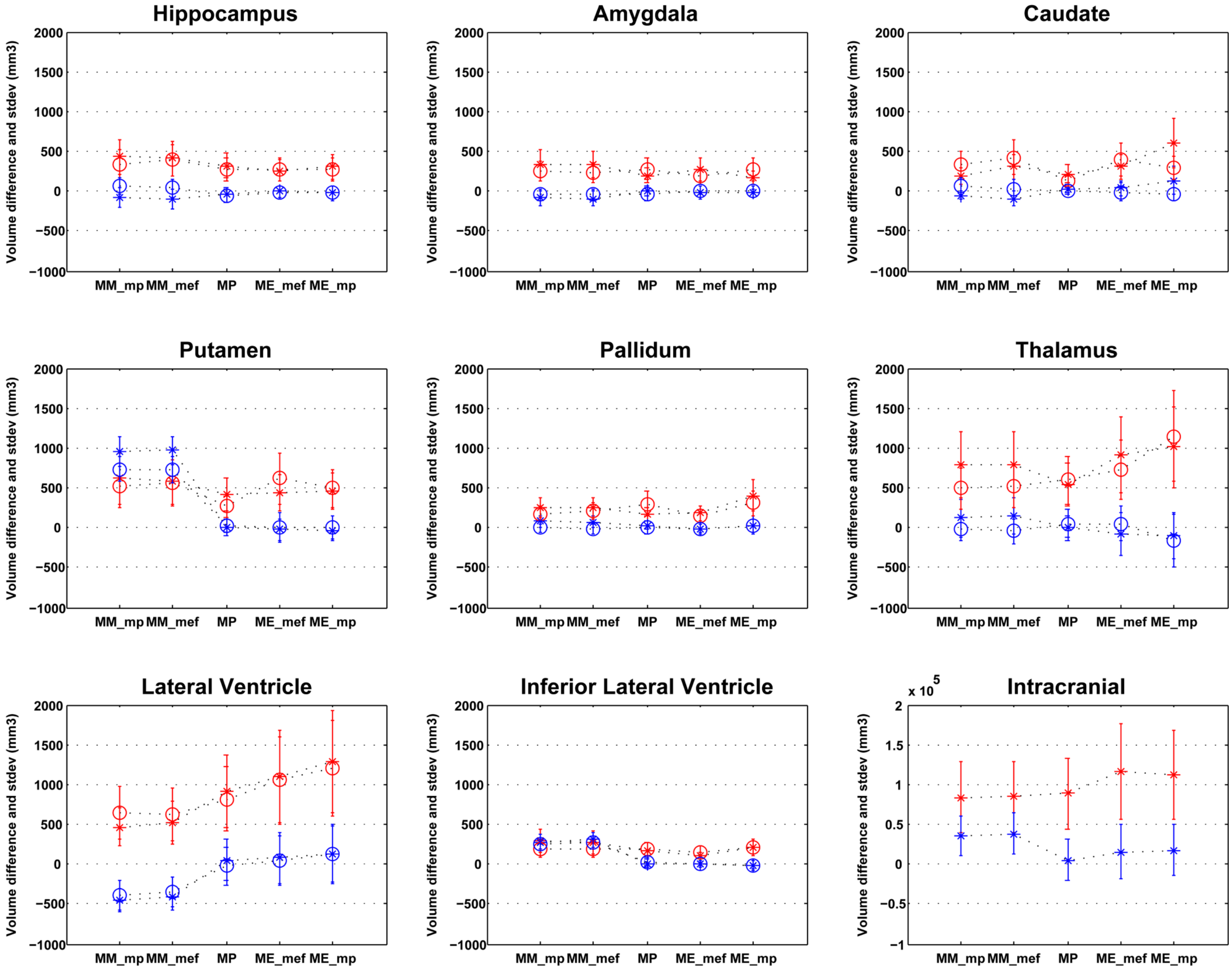

Figure 6.

Effects of acquisition sequence and segmentation atlas on volume reproducibility. Within-session agreement and across session test-retest repeatability of volumes derived from MPRAGE (MP) and multi-echo FLASH (MEF_mef: MEF atlas, MEF_mp: MPRAGE atlas). Bland-Altman results of mean volume difference (blue) and limits of agreement (red) for both brain hemispheres (left: crosses, right: circles) are shown with their respective 95% confidence intervals for the various repeatability conditions: within session agreement (MM_mp: MP vs. MEF_mp, MM_mef: MP vs. MEF_mef); across session test-retest reproducibility (MP, MEF_mef and MEF_mp). Black dotted lines connect the volume differences and the limits of agreement across conditions.

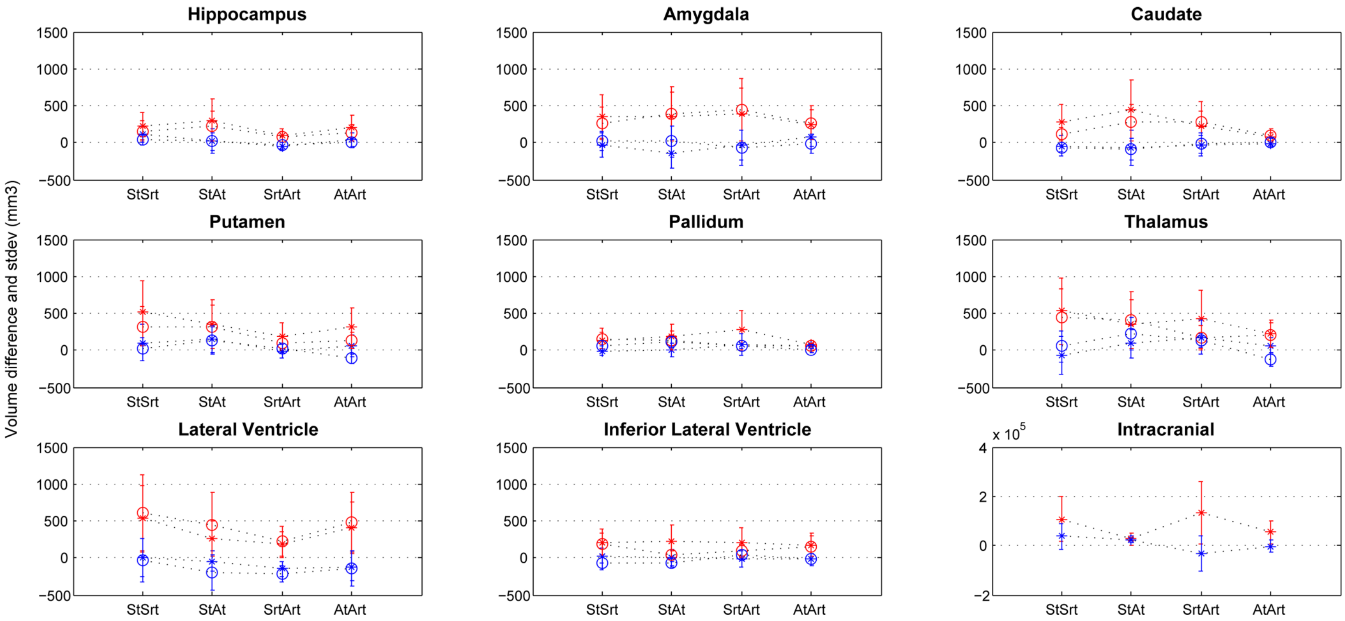

Figure 7.

Effect of MRI system upgrade on volume reproducibility (within-session averaged MPRAGE, Siemens Sonata-Avanto, 1.5T, young group, n=5). Bland-Altman results of mean volume difference (blue) and limits of agreement (red) for both brain hemispheres (left: crosses, right: circles) are shown with their respective 95% confidence intervals for the various repeatability conditions: before upgrade (Sonata test-retest: StSrt), across upgrade (Sonata test vs. Avanto test: StAt, Sonata retest vs. Avanto retest: SrtArt) and post upgrade (Avanto test-retest: AtArt). Black dotted lines connect the volume differences and the limits of agreement across conditions.

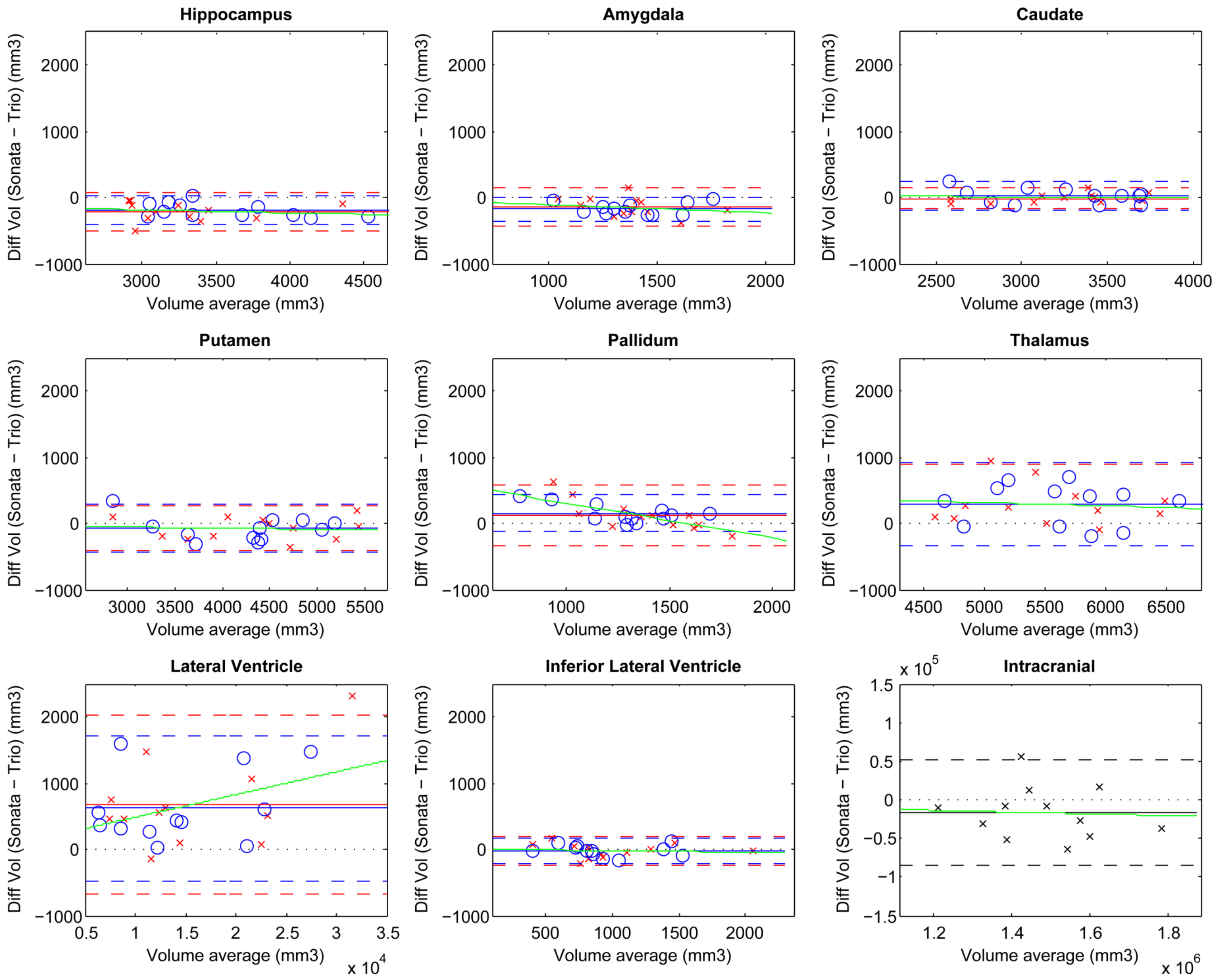

Figure 8.

Agreement of subcortical volume estimates from two scanners from the same vendor with different field strengths: Siemens Sonata and Siemens Trio. The Bland-Altman plots show volume difference vs. volume mean (Sonata - Trio, from each scanner volumes are segmented from a two-averaged MPRAGE acquisition, group of older subjects, n=15). For each brain hemisphere (left: red crosses, right: blue circles) the mean volume difference (solid horizontal line) and the limits of agreement (±2 standard deviations, interrupted horizontal lines) are shown. For reference, zero volume difference is shown as a black dotted line. Green lines show the linear regression fits for each brain structure (grouping data from both hemispheres). See text for the regression slopes.

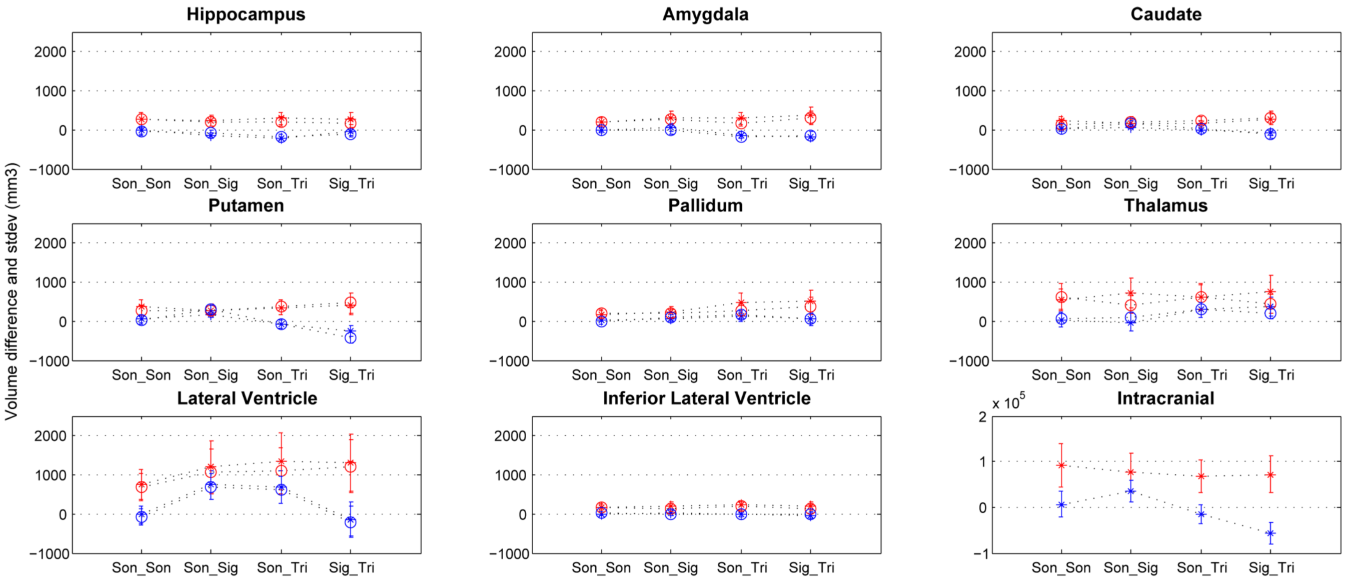

Figure 9.

Effects of scanner vendor and field strength on volume reproducibility (across-session, each session with a two-averaged MPRAGE, older group, n=15). Bland-Altman results of mean volume difference (blue) and limits of agreement (red) for both brain hemispheres (left: crosses, right: circles) are shown with their respective 95% confidence intervals for the following conditions: same MRI system (Sonata-Sonata, Son_Son), same field different vendor (Sonata-Signa, Son_Sig), same vendor different field (Sonata-Trio, Son_Tri) and different vendor and field (Signa-Trio, Sig_Trio). Black dotted lines connect the volume differences and the limits of agreement across conditions.

Table 1

Summary of datasets and acquisition variables used in this study. See text for more details.

		Vendor's RF Coil		3D T1-weighted MRI acquisition[4]								Healthy Volunteers	
Dataset	MRI Platform[1]	Birdcage	Phase Array[3]	MPRAGE				Multi-Echo FLASH				Mean Age (years)	N
				Session 1		Session 2		Session 1		Session 2			
				S1	S2	S1	S2	S1	S2	S1	S2		
1	Siemens Sonata[1]	Yes	-	√	√	√	√	√	√	√	√	69.5	15
	GE Signa[2]	Yes	-	√	√	-	-	-	-	-	-		
	Siemens Trio[1]	-	8-ch	√	√	-	-	√	√	-	-		
2	Siemens Sonata[1]	Yes	-	√	√	√	√	√	√	√	√	34.0	5
	Siemens Avanto[1]	-	12-ch	√	√	√	√	√	√	√	√		
3	Siemens Trio	Yes	-	√	√	√	√	-	-	-	-	36.5	5
	Siemens TrioTIM	Yes	-	√	√	√	√	-	-	-	-		

[1] Data acquired at the Massachusetts General Hospital, Boston, MA, USA.

[2] Data acquired at Brigham & Women's Hospital, Boston, MA, USA.

[3] Number of RF coil channels used.

[4] In each session two scans (S1 and S2) are acquired (√). In the case of MPRAGE the two scans are identical. In the case of MEF, each session has two scans (with flip angles of 30° and 5°). A dash (−) means that no data was acquired.

Table 2

Mean subcortical volumes (with standard deviation across subjects), test-retest reproducibility errors and confidence interval bias for the reproducibility errors. The mean subcortical values are derived from two averaged MPRAGE scans acquired in separate sessions on the same scanner with no B1 intensity inhomogeneity corrections (Siemens Sonata system for Dataset 1, Siemens Avanto for Dataset 2). The group reproducibility error for each structure is derived averaging the reproducibility errors across subjects, where for each subject the error is estimated as the absolute test-retest volume percent change relative to the mean test-retest volume.

Subcortical structures	**Dataset 1: Older Group (n=15, age=69.5 ± 4.8 years)**			**Dataset 2: Younger Group (n=5, age=34 ± 3 years)**		
	Volume (mm^3)	**Reprod. error (%)**	**Bias in confidence interval (%)**	**Volume (mm^3)**	**Reprod. error (%)**	**Bias in confidence interval (%)**
Hippocampus Left	3269 ± 419	3.61	0.11	3877 ± 395	2.34	0.87
Hippocampus Right	3508 ± 436	3.44	0.08	4060 ± 330	1.27	0.55
Thalamus Left	5498 ± 637	3.85	0.08	6147 ± 628	1.69	0.45
Thalamus Right	5587 ± 550	4.26	0.08	6233 ± 639	2.04	0.77
Caudate Left	3315 ± 479	2.43	0.06	3371 ± 648	0.55	0.40
Caudate Right	3407 ± 480	1.48	0.04	3455 ± 528	1.22	0.40
Putamen Left	4654 ± 848	3.56	0.09	5880 ± 833	1.93	1.01
Putamen Right	4357 ± 697	2.59	0.04	5641 ± 887	1.97	0.55
Pallidum Left	1585 ± 218	5.30	0.12	1703 ± 201	3.31	1.28
Pallidum Right	1470 ± 251	7.68	0.25	1593 ± 232	1.72	0.45
Amygdala Left	1324 ± 207	5.68	0.16	1447 ± 349	6.17	2.67
Amygdala Right	1311 ± 267	7.39	0.19	1413 ± 307	8.08	3.14
Lateral Ventricle Left**	16696 ± 7440**	2.37	0.07	7780 ± 4010**	1.16	0.63
Lateral Ventricle Right**	15942 ± 7045**	2.38	0.06	7514 ± 4788**	1.80	0.58
Inferior Lateral Ventricle Left	1054 ± 409	7.94	0.21	622 ± 354	10.45	3.58
Inferior Lateral Ventricle Right	1021 ± 382	10.23	0.42	702 ± 316	7.66	2.16

Subcortical structures	**Dataset 1: Older Group (n=15, age=69.5 ± 4.8 years)**			**Dataset 2: Younger Group (n=5, age=34 ± 3 years)**		
	Volume (mm^3)	**Reprod. error (%)**	**Bias in confidence interval (%)**	**Volume (mm^3)**	**Reprod. error (%)**	**Bias in confidence interval (%)**
Intracranial	1501100 ± 158900	2.56	0.05	1484200 ± 133700	1.34	0.62

[**] Structures with a significant mean volume difference (p<0.01) between the old and young groups are indicated with. The bias due to the limited sample sizes is estimated with a jackknife analysis on the reproducibility confidence interval (see text for more details).

Table 3

Comparison of average test-retest volume overlap (Dice coefficients) of segmented structures in a group of subjects (n=15, mean age 69.5) scanned on separate sessions under various conditions: same scanner and sequence (Sonata MPRAGE), same scanner but different sequences (MPRAGE and MEF), same field/sequence (1.5T/MPRAGE) but different scanner vendors (Siemens Sonata and GE Signa), and same vendor/sequence (Siemens/MPRAGE) different field strengths (1.5T and 3T)

	Sonata-Sonata (MPRAGE)	**MPRAGE-MEF (Sonata)**	**Sonata-GE (MPRAGE)**	**Sonata-Trio (MPRAGE)**
Hippocampus	0.87 ± 0.02	0.86 ± 0.04	0.81 ± 0.05	0.83 ± 0.04
Thallamus	0.92 ± 0.01	0.90 ± 0.01	0.91 ± 0.01	0.90 ± 0.02
Caudate	0.87 ± 0.03	0.83 ± 0.04	0.83 ± 0.05	0.84 ± 0.03
Putamen	0.89 ± 0.01	0.85 ± 0.02	0.86 ± 0.02	0.87 ± 0.01
Pallidum	0.85 ± 0.04	0.81 ± 0.04	0.82 ± 0.04	0.79 ± 0.11
Amygdala	0.84 ± 0.03	0.83 ± 0.05	0.78 ± 0.08	0.80 ± 0.04
Lateral Ventricle	0.94 ± 0.02	0.93 ± 0.03	0.93 ± 0.03	0.93 ± 0.02
All	0.88 ± 0.04	0.86 ± 0.05	0.83 ± 0.12	0.85 ± 0.07