NeuroImage 83 (2013) 472–484

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Brain morphometry reproducibility in multi-center 3 T MRI studies: A comparison of cross-sectional and longitudinal segmentations

Jorge Jovicich [a,*,1], Moira Marizzoni [b,1], Roser Sala-Llonch [c], Beatriz Bosch [q], David Bartrés-Faz [c], Jennifer Arnold [d], Jens Benninghoff [d], Jens Wiltfang [d], Luca Roccatagliata [e,f], Flavio Nobili [g], Tilman Hensch [h], Anja Tränkner [h], Peter Schönknecht [h], Melanie Leroy [i], Renaud Lopes [r], Régis Bordet [i], Valérie Chanoine [j], Jean-Philippe Ranjeva [j], Mira Didic [k,l], Hélène Gros-Dagnac [m,n], Pierre Payoux [m,n], Giada Zoccatelli [o], Franco Alessandrini [o], Alberto Beltramello [o], Núria Bargalló [p], Olivier Blin [s], Giovanni B. Frisoni [b], The PharmaCog Consortium

[a] Center for Mind Brain Sciences, University of Trento, Trento, Italy
[b] LENITEM Laboratory of Epidemiology, Neuroimaging, & Telemedicine — IRCCS San Giovanni di Dio-FBF, Brescia, Italy
[c] Department of Psychiatry and Clinical Psychobiology, Universitat de Barcelona and IDIBAPS, Barcelona, Spain
[d] LVR-Clinic for Psychiatry and Psychotherapy, Institutes and Clinics of the University Duisburg-Essen, Essen, Germany
[e] Department of Neuroradiology, IRCCS San Martino University Hospital and IST, Italy
[f] Department of Health Sciences, University of Genoa, Italy
[g] Department of Neuroscience, Ophthalmology and Genetics University of Genoa, Genoa, Italy
[h] Department of Mental Health, Clinic for Psychiatry and Psychotherapy, University Hospital Leipzig, Leipzig, Germany
[i] Department of Pharmacology, EA 1046, University of Lille Nord de France, 59045 Lille Cedex, France
[j] CRMBM-CEMEREM UMR 7339 Aix Marseille Université - CNRS, Marseille, France
[k] APHM, CHU Timone, Service de Neurologie et Neuropsychologie, 13005 Marseille, France
[l] Aix-Marseille Université, INSERM U 1106, 13005 Marseille, France
[m] INSERM; Imagerie cérébrale et handicaps neurologiques UMR 825; F-31059 Toulouse, France
[n] Université de Toulouse; UPS; Imagerie cérébrale et handicaps neurologiques UMR 825; CHU Purpan, Place du Dr Baylac, F-31059 Toulouse Cedex 9, France
[o] Department of Neuroradiology, General Hospital, Verona, Italy
[p] Department of Neuroradiology and Image Research Platform, Hospital Clínic de Barcelona, IDIBAPS, Barcelona, Spain
[q] Alzheimer's Disease and Other Cognitive Disorders Unit, Department of Neurology, Hospital Clínic, and IDIBAPS, Barcelona, Spain
[r] Department of Neuroradiology, CHRU Roger Salengro, Lille, France
[s] CIC-CPCET, Ap-Hm and UMR 7289 Aix-Marseille University-CNRS, France

ARTICLE INFO

Article history:
Accepted 1 May 2013
Available online 11 May 2013

Keywords:
Brain morphometry
Reproducibility
Reliability
Structural MRI
Multi-center
Multi-site MRI

ABSTRACT

Large-scale longitudinal multi-site MRI brain morphometry studies are becoming increasingly crucial to characterize both normal and clinical population groups using fully automated segmentation tools. The test–retest reproducibility of morphometry data acquired across multiple scanning sessions, and for different MR vendors, is an important reliability indicator since it defines the sensitivity of a protocol to detect longitudinal effects in a consortium. There is very limited knowledge about how across-session reliability of morphometry estimates might be affected by different 3 T MRI systems. Moreover, there is a need for optimal acquisition and analysis protocols in order to reduce sample sizes. A recent study has shown that the longitudinal FreeSurfer segmentation offers improved within session test–retest reproducibility relative to the cross-sectional segmentation at one 3 T site using a nonstandard multi-echo MPRAGE sequence. In this study we implement a multi-site 3 T MRI morphometry protocol based on vendor provided T1 structural sequences from different vendors (3D MPRAGE on Siemens and Philips, 3D IR-SPGR on GE) implemented in 8 sites located in 4 European countries. The protocols used mild acceleration factors (1.5–2) when possible. We acquired across-session test–retest structural data of a group of healthy elderly subjects (5 subjects per site) and compared the across-session reproducibility of two full-brain automated segmentation methods based on either longitudinal or cross-sectional FreeSurfer processing. The segmentations include cortical thickness, intracranial, ventricle and subcortical volumes. Reproducibility is evaluated as absolute changes relative to the mean (%), Dice coefficient for volume overlap and intraclass correlation coefficients across two sessions. We found that this acquisition and analysis protocol gives comparable reproducibility results to previous studies that used longer acquisitions without acceleration. We also show that the longitudinal processing

* Corresponding author. Fax: +39 0461 88 3066.
E-mail address: jorge.jovicich@unitn.it (J. Jovicich).
[1] Authors contributed equally to this work.

1053-8119/$ – see front matter
http://dx.doi.org/10.1016/j.neuroimage.2013.05.007

is systematically more reliable across sites regardless of MRI system differences. The reproducibility errors of the longitudinal segmentations are on average approximately half of those obtained with the cross sectional analysis for all volume segmentations and for entorhinal cortical thickness. No significant differences in reliability are found between the segmentation methods for the other cortical thickness estimates. The average of two MPRAGE volumes acquired within each test–retest session did not systematically improve the across-session reproducibility of morphometry estimates. Our results extend those from previous studies that showed improved reliability of the longitudinal analysis at single sites and/or with non-standard acquisition methods. The multi-site acquisition and analysis protocol presented here is promising for clinical applications since it allows for smaller sample sizes per MRI site or shorter trials in studies evaluating the role of potential biomarkers to predict disease progression or treatment effects.

Introduction

Methods that enable the characterization of human brain morphometry from MRI data are demonstrating important applications in neuroscience. Several reviews describe how morphometry tools have been applied to investigate a variety of populations, including, but not limited to, normal development (Silk and Wood, 2011), normal aging (Mueller et al., 2007), Alzheimer's disease (Drago et al., 2011; Fjell and Walhovd, 2012; Frisoni et al., 2010; Jack, 2011), Parkinson's disease (Kostić and Filippi, 2011), autism (Chen et al., 2011), bipolar disorders (Selvaraj et al., 2012), epilepsy (Bernasconi et al., 2011) and schizophrenia (Levitt et al., 2010). One particular example of a successful contribution of brain morphometry to the field of neurodegenerative diseases is the fact that hippocampal volume has been recently approved as biomarker to enrich the population selection in clinical trials that study early stages of Alzheimer's disease (EMA/CHMP/SAWP/809208/2011).

There are several methods to obtain brain morphometry estimates from MRI data. Manual segmentation of specific brain structures on MRI made by trained raters, with its high inter-rater reliability, is considered as the gold standard by many neuroimaging studies (Rojas et al., 2004; Whitwell et al., 2005). However, due to its time-costs, manual segmentations are not practically applicable for large studies involving many subjects and different brain structures. Various automated and semi-automated algorithms have been proposed, including atlas-based methods (Alemán-Gómez et al., 2007; Fischl et al., 2002; Lötjönen et al., 2010; Magnotta et al., 2002; Wolz et al., 2010), voxel-based morphometry with statistical parametric mapping (Ashburner and Friston, 2000), tensor-based morphometry (Leow et al., 2005; Studholme et al., 2001) and boundary shift integral methods (Camara et al., 2007; Smith et al., 2002). This list of brain morphometry analysis methods is by no means complete nor does this paper attempt to compare and contrast these methods.

Automated morphometric analysis is of particular interest in longitudinal studies aimed at characterizing disease progression or the effect of therapeutic treatments, both when using known and when searching for new useful biomarkers. In particular, longitudinal multi-center MRI studies are becoming an increasingly common strategy to collect large datasets while distributing the data acquisition load across multiple partners (Van Horn and Toga, 2009), and probably one of the largest examples is the Alzheimer's Neuroimage Initiative, or ADNI (Carrillo et al., 2012). One critical factor that limits the sensitivity to detect changes in any longitudinal study is the reproducibility of repeated measures. The test–retest reliability of MRI-derived morphometric estimates may be affected by a variety of factors (Jovicich et al., 2009), including hydration status of the subject (Walters et al., 2001), instrument related factors such as scanner manufacturer, field strength, head RF coil, magnetic gradients (Jovicich et al., 2006), pulse sequence and image analysis methods (Han et al., 2006). Repeated acquisitions within a single scan session without subject repositioning may be used to characterize the best attainable reproducibility conditions from an acquisition and analysis protocol. However, the reproducibility errors present in a longitudinal study are better described by repeated acquisitions obtained in different sessions several days apart. Such across-session differences will include additional sources of variance like MRI system instabilities, differences in head positioning within the RF coil, differences in automated acquisition procedures like auto shimming, as well as potential effects from how different operators follow instructions to execute the same acquisition protocol. Across-session reproducibility is even more challenging in multicenter neuroimaging clinical studies where comparable results are usually difficult to obtain due to the added variability from site differences in the MRI hardware, acquisition protocols and operators.

Despite the wide usage of automated morphometric techniques applied to 3 T MRI studies, across-site test–retest reliability of morphometry measures has not been thoroughly investigated and thus its impact on statistical analysis is not clearly defined. Table 1 outlines studies that, to the best of our knowledge, have reported across-session test–retest reproducibility measures of morphometric data derived from healthy volunteers using 3 T systems. Most studies were done on a single MRI system (Kruggel et al., 2010; Morey et al., 2010; Wonderlick et al., 2009), except for one study that evaluated major MRI system upgrade effects on reproducibility, therefore considering effectively two different systems (Jovicich et al., 2009). These studies have been performed on only two vendors (Siemens and GE), and three models (Trio, Trio TIM, GE Excite) that nowadays tend to be less common as the manufacturers develop newer versions. In addition, morphometry segmentation tools have also been evolving. Recently, a FreeSurfer longitudinal image processing framework has been developed (Reuter et al., 2012) showing a significant increase in precision and discrimination power when compared with tools originally designed for the FreeSurfer cross-sectional analysis. In that study the test–retest reliability of the longitudinal stream was evaluated at 3 T, but it was done for repeated acquisitions obtained during the same session and also when using a particular sequence, multi-echo 3D MPRAGE (van der Kouwe et al., 2008), that has interesting advantages relative to the standard 3D MPRAGE (Wonderlick et al., 2009) but that is not yet commonly available across all vendors. To date there are no studies evaluating the across-session test–retest reproducibility of this new longitudinal analysis at 3 T, for one or more MRI system vendors, while using an MRI acquisition that is standard across vendors.

All of these issues are relevant to the PharmaCog project, a new industry-academic European project aimed at identifying biomarkers sensitive to symptomatic and disease modifying effects of drugs for Alzheimer's disease (http://www.alzheimer-europe.org/FR/Research/PharmaCog). One of the objectives of the PharmaCog project is to investigate potential biomarkers derived from human brain structural and functional MRI, in particular brain morphometry. Within this context, the goals of the present PharmaCog study were the following: i) implement a multi-site 3 T MRI data acquisition protocol for morphometry analysis, ii) acquire across-session test–retest data from a population of healthy elderly subjects, and iii) evaluate and compare the across-session reproducibility of the cross-sectional and longitudinal FreeSurfer segmentation analyses within and across MRI sites. This work is therefore an extension of previous work (Reuter et al., 2012), evaluating the across-session reproducibility of the segmentation results (cortical thickness, intracranial, ventricular and subcortical volumes) on a variety of 3 T MRI scanning platforms (Table 1). To keep a manageable number of

Table 1
Summary of studies that evaluated within-scanner across session test–retest reproducibility of 3 T MRI brain morphometry results on healthy subjects. Abbreviations: FreeSurfer cross-sectional (CS) or longitudinal (LG) segmentations, intra-class correlation coefficient (ICC).

Study	3 T MRI scanners for test–retest within scanner (number)	Subjects (number), age (mean ± SD)	Analysis tool	Reproducibility metrics (days between test–retest)
This study	Siemens Allegra (1), TIM Trio (2), Verio (1), Skyra (1); GE HDxt (1); Philips Achieva (2)	Healthy N = 40 (5 per/scanner), (63.2 ± 8.1) years	FreeSurfer v5.1.0 (CS/LG)	Test–retest absolute % differences and ICC of volume and thickness structures. Across-session tests (14–31 days)
Morey et al. (2010)	GE Excite (1)	Healthy N = 23, (23.4 ± 3.3) years	FreeSurfer v4.5 and FIRST v1.2 (CS/LG)	Test–retest ICC and absolute % difference of volume structures. Across-session both within-day (1 h apart) and a week apart (7–9 days)
Kruggel et al. (2010)	Siemens Trio (1)	ADNI (normal 3, MCI 9, D 3) (74.6 ± 7.0)	FANTASM	Global volumes[a]. Across-session tests (30 days)
Wonderlick et al. (2009)	Siemens Trio TIM (1)	Healthy N = 5, (21.4 ± 3.8) years N = 6, (64.3 ± 12.2) years	FreeSurfer v4.0.1 (CS)	Test–retest ICC[a] of volume and thickness structures. Across-session tests (14 days)
Jovicich et al. (2009)	Siemens Trio TIM (1) Siemens Trio (1)	N = 5, (36.5 ± 3) years	FreeSurfer (CS)	Test–retest absolute % and signed differences of volume structures. Across-session (7–42 days)

[a] Total brain volume of white matter, gray matter, cerebral spinal fluid.

variables in this study we do not manipulate the acquisition sequence other than trying to implement a target common protocol across all sites following in great part ADNI recommendations. The study is focused on the comparison of the test–retest reproducibility of morphometric results derived from two variants of the FreeSurfer segmentation, comparisons with other segmentation methods are beyond the scope of this work.

Methods

Subjects

Nine clinical sites participated in this study across Italy (Brescia, Verona, and Genoa), Spain (Barcelona), France (Marseille, Lille, and Toulouse) and Germany (Leipzig and Essen). The Brescia site was responsible for the coordination and analysis of the whole study and did not acquire MRI data. Each MRI site recruited 5 local volunteers within an age range of 50–80 years. The subject's age range corresponds to the same one of the clinical population that will be studied with the protocols tested in this reproducibility study. Each subject underwent two MRI sessions completed at least 7 days (but no more than 60 days) apart at the site, to minimize biological changes that could affect the reliability of the measures. Table 2 summarizes information about age, gender and test–retest interval times of the subjects recruited at each site. All participants were volunteers with no history of major psychiatric, neurological or cognitive impairment (referred to as healthy in this study), and provided written informed consent in accordance with the "classification" of the study as regards to the national regulations and laws in the different participating countries. In France, the study received an authorization from the national drug regulatory agency (Agence Nationale de Sécurité du Médicament et des produits de santé) and an approval from the Comité de Protection des Personnes Sud-Méditerranée 1 (Marseille), for the three French sites (Marseille, Lille, and Toulouse). In Germany, Spain and Italy the study obtained authorization from one Ethics Committee relevant to each institution: Essen (Ethik-Kommission des Universitätsklinikums Essen), Leipzig (Ethik-Kommission der Universität Leipzig), Barcelona (Comité de Etica e Investigación Clínica Hospital Clínic de Barcelona), Verona (Comitato Etico Istituzioni Ospedaliere Cattoliche, CEIOC) and Genoa (Comitato Etico IRCCS-Azienda Ospedaliera Universitaria San Martino-IST). All subjects signed informed consent.

MRI acquisitions

The eight 3 T MRI sites that participated in this study used different MRI system vendors and models (Siemens, GE, and Philips). Table 2 summarizes the main MRI system differences across sites. Each MRI scanning session consisted of several acquisitions using only vendor-provided sequences, including: anatomical T2*, anatomical FLAIR, resting state fMRI, B0 map, DTI and two anatomical T1 scans (without repositioning the subject), with a total acquisition time of approximately 35 min. For this work, we utilized only the two anatomical T1 scans (MPRAGE on Siemens and Philips, IR-SPGR on GE), which were used for brain morphometry analysis (3D sagittal acquisition, square FOV = 256 mm, $1 \times 1 \times 1$ mm^3, TR/TI = 2300/900 ms, flip angle = 9°, no fat suppression, full k-space, no averages). These parameters were largely based on the MPRAGE recommendations from ADNI 2 (http://adni.loni.ucla.edu/research/protocols/mri-protocols/) except for two factors: nominal spatial resolution (we used isotropic 1 mm^3 instead of $1 \times 1 \times 1.2$ mm^3) and image acceleration (when allowed by the RF coil we used an acceleration factor in the range of 1.5–2, instead of no acceleration). The choice for using accelerated MPRAGE acquisitions was motivated by several factors: most modern 3 T scanners allow for it, the reduction of scanning time

Table 2
Summary of demographic, MRI system and acquisition differences across MRI sites.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
MRI site location	Verona	Barcelona	Marseille	Lille	Toulouse	Genoa	Leipzig	Essen
Subjects' age: mean ± SD, (range) years	67.8 ± 9.9 (26)	74.6 ± 2.7 (6)	66.0 ± 8.3 (20)	64.2 ± 5.3 (13)	59.2 ± 4.5 (12)	58.2 ± 2.2 (5)	62.8 ± 2.6 (6)	52.4 ± 1.5 (3)
Test–retest time interval (days)	28 ± 23	10 ± 3	23 ± 22	15 ± 11	14 ± 10	24 ± 17	13 ± 3	11 ± 5
Gender, (females/N)	2/5	5/5	4/5	3/5	3/5	2/5	3/5	2/5
3 T MRI scanner	Siemens Allegra	Siemens TrioTim	Siemens Verio	Philips Achieva	Philips Achieva	GE HDxt	Siemens TrioTim	Siemens Skyra
MR system software version	VA25A	B17	B17	3.2.2	3.2.2	15 M4A	B17	D11
TX/RX coil	Birdcage	Body/8-chan.	Body/12-chan.	Body/8-chan.	Body/8-chan.	Body/8-chan.	Body/8-chan.	Body/20-chan.
Parallel imaging: method, acceleration	None	GRAPPA 2	GRAPPA 2	SENSE 1.5	SENSE 1.5	ASSET 2	GRAPPA 2	GRAPPA 2
TE (ms, shortest)	2.83	2.98	2.98	3.16	3.16	2.86	2.98	2.03
MPRAGE volume acquisition time (min:sec)	9:50	5:12	5:12	6:50	6:50	4:43	5:12	5:12

is expected to reduce the sensitivity to head motion artifacts even at an expense of some loss in signal, and previous studies have reported no test–retest reproducibility costs when accelerating relative to non-accelerated acquisitions, both when using 3 T (Wonderlick et al., 2009) and 1.5 T (Jovicich et al., 2009) MRI systems. The parallel acquisition methods were different across sites, the choices were made based on the optimal or possible options available at the different platforms (see Table 2). Default options for geometric distortion corrections were kept at each scanner. All images from multi-channel coils were reconstructed by the scanner as the sum of the squares across channels. When allowed by the MRI system, images were reconstructed and saved without additional filtering options that could differ across scanners introducing different degrees of smoothing.

Data preparation

Imaging data were initially anonymized at each site by replacing the subject name with a unique identifier using the free DicomBrowser tool (http://hg.xnat.org/dicombrowser). Anonymized dicom data were then compressed and uploaded on to a data sharing system accessible to all member sites, from where they were subsequently downloaded for analysis at the central site (Brescia).

Downloaded anonymized dicom data were converted to nifti format using the free dcm2nii software (http://www.mccauslandcenter.sc.edu/mricro/mricron/dcm2nii.html, output format FSL — 4D NIFTI nii) from which the original dicom converted to nifti files were used. All data were visually inspected for quality assurance prior to analyses to check that there were no major visible artifacts, including motion, wrap around, RF interference and signal intensity or contrast inhomogeneities. Each subject had a total of four anatomical scans, two from the test session and two from the retest session. No within-session averaging was done.

Brain segmentations

Each MPRAGE anatomical volume was analyzed in FreeSurfer (Dale et al., 1999, Fischl et al., 1999) to automatically generate subject-specific cortical thickness (Fischl et al., 2004, Desikan et al., 2006) and subcortical volume (Fischl et al., 2002) estimates in regions-of-interest (ROIs). For each subject we used two FreeSurfer analyses: the cross-sectional (CS) and the longitudinal (LG) streams. Detailed explanations of the differences between these two FreeSurfer segmentations can be found both in a recent study (Reuter et al., 2012) as well in the distribution site (http://freesurfer.net/fswiki/LongitudinalProcessing). Briefly, in the FreeSurfer cross-sectional analysis each time point is processed independently for each subject. These cortical and subcortical segmentation and parcellation procedures involve solving many complex nonlinear optimization problems that are typically calculated using iterative methods. Such methods need starting conditions that may introduce biases in the final results. The FreeSurfer longitudinal analysis is designed to minimize such biases with respect to any time point in a subject. The longitudinal analysis uses results from the cross-sectional analysis and consists of two main steps: i) creation of a template for each subject using all time points to build an average subject anatomy and ii) analysis of each time point using information from the template and the individual cross-sectional runs to initialize several of the segmentation algorithms. This procedure of using the repeated measures as common information from the subject to initialize the processing in each time point can reduce variability compared to independent processing, as has been shown recently (Reuter et al., 2012).

Our study is focused on a subset of the automatically segmented regions which are of interest in neurodegenerative diseases. The volumetric ROIs included the hippocampal formation, amygdala, caudate nucleus (caudate), putamen, globus pallidus (pallidum), thalamus, lateral ventricles and total intracranial volume. The cortical thickness ROIs included the parahippocampus gyrus, fusiform gyrus, superior temporal gyrus, precuneus, superior parietal gyrus, supramarginal gyrus, lateral occipital gyrus, lingual gyrus, superior frontal gyrus and entorhinal cortex (Han et al., 2006). For each of these structures (except the intracranial volume) the right and left hemisphere volumes are estimated separately on each anatomical scan. The segmentation results were visually inspected prior to the volume and thickness analysis to confirm that no major errors were present. No manual edits were done. All analyses were done using FreeSurfer version 5.1, running on a Linux workstation (Ubuntu 10.04) equipped with Intel CPU 8 × 3.07 GHz processors and 7.9 GB of RAM.

Evaluation of reliability

To evaluate the reliability of the brain segmentation results we analyzed their variability, or reproducibility error, across the test–retest sessions for each site. There are several sources of variability for a fixed scanner, which include variability from hydration status (expected to be small if scans are repeated within a short time interval), variability due to slightly different acquisitions in the two sessions (head position change in the scanner, motion artifacts, scanner instability, etc.), and finally variability due to the imaging processing methods themselves. In addition, in a multi-center study there is also the added variability from the different MRI systems (vendor, model, acquisition parameters). In this study the goal was to evaluate the across session reliability of FreeSurfer brain segmentations, within each site and across sites, both for the CS and LG processing streams. The main hypothesis we wanted to test here is whether the LG processing stream can reduce across session variability, both within and across sites, relative to the CS segmentation stream.

Since every subject had segmentation results derived separately from each of the two test and the two retest MPRAGE volumes, we used these four possible test–retest comparisons across sessions to estimate a mean across-session variability error per subject. As variability error we used the dimensionless measure of absolute percent change of volume (or thickness) of a structure with respect to its average. In other words, for each subject, for each volumetric or thickness structure, and for each analysis stream (LG or CS), the across-session variability error was estimated as follows:

$$\varepsilon_{ij} = 100 \times \frac{\left|\mathrm{Vretest}_i - \mathrm{Vtest}_j\right|}{\left(\mathrm{Vretest}_i + \mathrm{Vtest}_j\right)/2}$$

$$\varepsilon = (\varepsilon_{11} + \varepsilon_{12} + \varepsilon_{21} + \varepsilon_{22})/4$$

where ε is the mean across-session variability error and the indices *i* and *j* can take values 1 or 2 to refer to the first or second MPRAGE volume in each of the test (Vtest) and retest (Vretest) sessions. The group variability error for every MRI site and brain structure was then averaged across subjects, within each analysis stream separately. Such estimation of variability can be interpreted as the mean measurement error. The measure was chosen because it is intuitive and because the estimation of the means is more robust than the estimation of the variance from the signed differences, in particular for low number of subjects.

The distributions of volume (or thickness) differences plotted against volume (or thickness) means across sessions were examined with a Bland–Altman analysis (Bland and Altman, 1986). These plots show the spread of data, the mean difference and the limits of agreement, and were used to confirm that the distributions were approximately symmetric around zero and to check for possible outliers.

An additional evaluation of variability was done by computing the spatial reproducibility of the segmented subcortical and ventricular volumes. Spatial reproducibility was examined by computing the Dice coefficients for the volume overlap (van Rijsbergen, 1979)

on the co-registered test–retest volumes segmented with both FreeSurfer streams. In particular, given two different labels (test and retest sessions) of a structure from the same subject, denoted by Vtest and Vretest, and a function Vol (V), which takes a label and returns its volume or of the intersection of two volumes, the Dice coefficient is given by van Rijsbergen (1979):

$$D_{ij} = \frac{\mathrm{Vol}\left(\mathrm{Vretest}_j \cap \mathrm{Vtest}_i\right)}{\left(\mathrm{Vol}\left(\mathrm{Vretest}_j\right) + \mathrm{Vol}(\mathrm{Vtest}_i)\right)/2}$$

$$D = (D_{11} + D_{12} + D_{21} + D_{22})/4.$$

For identical spatial labels $\mathrm{Vretest}_i$ and Vtest_j, D_{ij} achieves its maximum value of one, with decreasing values indicating less perfect spatial overlap. For each subject the Dice coefficients were calculated as an average across the right and left hemispheres. The group results for each site were generated by averaging the Dice coefficients across subjects for each structure.

The intraclass correlation coefficient (ICC) was used as an additional measure of test–retest absolute agreement across sessions, ICC (2,1) (Rajaratnam, 1960). The ICC analysis (SPSS, version 13.0) was computed separately for both the volumetric and thickness estimates, for each MRI site and each analysis stream. The mean ICC value for each site was the mean across subjects, and the ICC of each subject was the mean of the four possible across-session test–retest combinations, as described for the other reliability measures in this study.

Statistical analysis

The following statistical analyses were done, using MATLAB and SPSS (v.13.0):

- To test for MRI site effects of the subject's distributions of age, segmentation volume, cortical thickness, across-session reproducibility error (of volumes and thickness) and across-session spatial overlap, one-way Kruskall–Wallis tests (non-parametric version of ANOVA) were used with MRI site as factor, with a significance threshold of $p < 0.05$.
- To test for differences between the mean reproducibility errors of the two FreeSurfer streams (LG vs. CS), for each cortical or volumetric brain structure and site, the two-tailed Wilcoxon rank sum test was used (non-parametric version of the paired Student's t-test), with a significance threshold of $p < 0.05$.

Sample size comparisons

It is of interest to estimate the degree to which a potential improvement in test–retest variability can affect the design in a multi-site longitudinal study, for example in terms of reducing the number of subjects that need to be recruited or reducing the length of a trial aimed at detecting longitudinal changes. The formulation that describes longitudinal sample size calculations (Diggle et al., 2002) can be used to compare the longitudinal and cross-sectional segmentation methods in terms of the percent of subjects (SS_{frac}) needed when processing the data with the LG as opposed to the CS segmentation method (Reuter et al., 2012):

$$SS_{frac} = 100 \times \frac{\sigma^2_{LG}(1-\rho_{LG})}{\sigma^2_{CS}(1-\rho_{CS})}$$

where σ^2 and ρ are the variance and correlation, respectively, of the across-session test–retest estimates of a structure (thickness or volume) for the LG and CS segmentation methods. The stability of these results can be estimated via bootstrapping (1000 resamples).

Results

In this study, we estimate the test–retest reliability of morphometry measures derived from structural T1-weighted 3 T MRI data and evaluate how their reproducibility errors are affected by FreeSurfer processing stream (CS, LG) and MRI site (eight 3 T MRI scanners from different vendors: GE, Siemens, Philips) on healthy elderly volunteers scanned in two separate sessions at least one week apart. This short period between the test and retest sessions was chosen to minimize biological changes that could affect the reliability of the measures and to mimic the variability expected from separate sessions, as measured in longitudinal studies. The 40 subjects enrolled (5 for each center, see Table 2 for summary of demographic information) had similar age distribution except for site 2 (older group, mean age 74.6 ± 2.7 years, significantly different from sites 5–8, Kruskall–Wallis, $p < 0.05$) and site 8 (younger group, mean age 52.4 ± 1.5 years, significantly different from sites 1–4, Kruskall–Wallis, $p < 0.05$). There were no age distribution differences between the other MRI sites. The time interval between test and retest scans ranged from 7 to a maximum of 55 days, with a mean and standard deviation of 17 ± 14 days.

Our initial goal was to compute and evaluate the segmentations of a total of 320 brain volumes: 8 MRI sites, 5 subjects per site, 4 acquisitions per subject (two tests, two retests), and 2 FreeSurfer segmentation analysis protocols. In practice we had 3 missing volumes: two subjects of site 5 had missing MPRAGE volume repetitions during the test session, and one MPRAGE from site 1 was discarded because it required manual edits to complete the segmentation. Visual inspection of FreeSurfer segmented images showed a high similarity of result quality across sites (Fig. 1).

Estimation of brain morphometric volumes across MRI sites

Table 3 summarizes the group mean volumetric results (subcortical, ventricle and intracranial), averaged across hemispheres and across the test–retest sessions, for each MRI site as derived from the FreeSurfer LG segmentation stream. A Kruskall–Wallis test for MRI site effect on the hemispheric volumes showed that there were significant site-effects ($p < 0.05$) for only 2 of the 15 structures evaluated: the left putamen and right pallidum. This variability of morphometric results across sites is consistent with the fact that the groups of subjects were different at the various sites, and might simply reflect anatomical variability.

Estimation of volume reproducibility: effects of MRI sites and segmentation analyses

Fig. 2 shows an example of a Bland–Altman plot for a single site on two sample structures: the hippocampus (left) and the amygdala (right). The plot shows, for site 2, the distribution of across-session volume differences relative to the volume means for the two analysis streams, CS (top) and LG (bottom). 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. The 20 data points in each plot correspond to the 5 subjects and their respective 4 test–retest possible comparisons. As it can be seen the volume differences are symmetrically distributed around zero. The signed difference means were not significantly different from zero, indicating no biases between the across-session measures. Similar results were found for all other sites and structures. In this example it is also possible to see how the spread of the data appears reduced in the LG relative to the CS analysis.

Table 4 summarizes the across-session test–retest reproducibility errors of the various segmented volumes for each site, for both analysis streams (CS and LG). In each site the mean reproducibility error is computed as a mean across subjects, across the four test–retest segmentations and across the two brain hemispheres where relevant (intracranial volume is the only exception). No significant MRI site effects

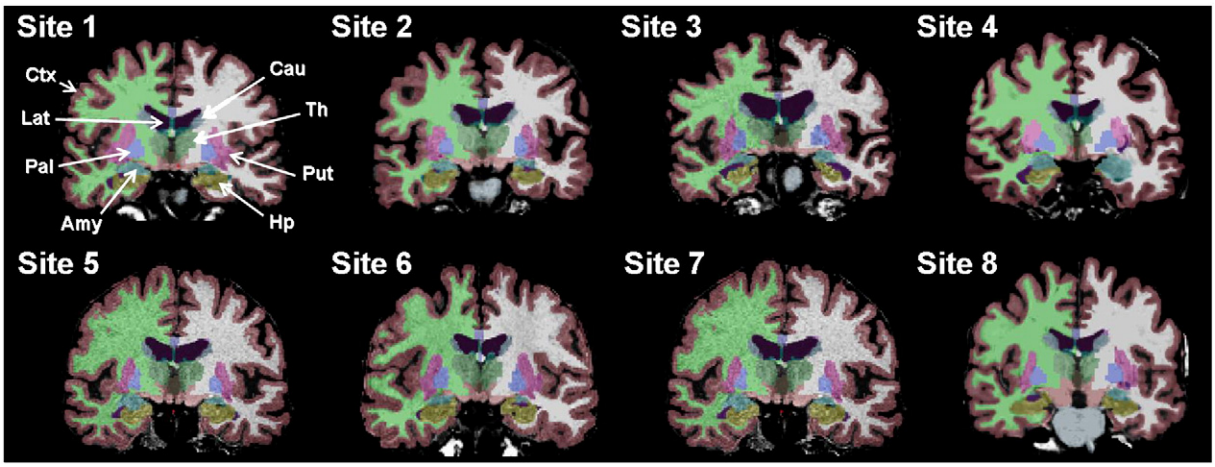

Fig. 1. Sample MPRAGE images and FreeSurfer segmentation results across different 3 T MRI sites for qualitative comparison. Abbreviations: Hp = hippocampus, Amy = amygdala, Cau = caudate, Put = putamen, Pal = pallidum, Thal = thalamus, Lat = lateral ventricle, Ctx = cerebral cortex. See Table 2 for MRI site characteristics.

were found on the reproducibility error, regardless of structure and analysis stream used for the brain segmentations. Averaging the reproducibility errors across sites allows summarizing the effects of analysis on the various structures (Table 4, last column). For all structures the LG stream showed a significantly lower reproducibility error relative to the CS stream (Wilcoxon test, $p < 0.01$), except for the lateral ventricle volumes, which gave no significant differences between analysis streams. When considering the separate hemispheric volumes within each site and test across all structures, we also found that in all sites the LG analysis gave significantly lower reproducibility errors relative to the CS analysis (Wilcoxon test, $p < 0.05$).

Fig. 3 is a graphical example of some of the findings reported in Table 4, showing the distribution of volumetric reproducibility errors (%) across the eight MRI sites for just two structures, the hippocampus (left) and the amygdala (right). Each point represents an MRI site, with the longitudinal error on the vertical axis and the cross sectional error along the horizontal axis, with corresponding within-site standard deviations. The diagonal unity line corresponding to perfect agreement between the two measures is plotted as a thin reference line. The fact that the overall reproducibility error is smaller with the longitudinal line can be easily seen by having all MRI site points under the unity line. The vertical and horizontal dotted lines mark the maximum range of the spread in absolute errors. It can be seen that the spread of errors for the longitudinal stream (range along vertical axis) is smaller than the spread of errors given by the cross-sectional stream (range along horizontal axis). Overall this means that the across-session test–retest errors in volumetric estimates of the longitudinal stream give lower reproducibility errors and also lower variability across MRI sites.

The ICC results for absolute volumetric agreement across sessions are summarized in Supplementary Table 1. Overall the results are consistent with the previous volumetric reliability analysis, showing that the test–retest reliability is consistently higher for the longitudinal stream than for the cross-sectional segmentation (Wilcoxon test, $p < 0.01$) with the only exception of the lateral ventricles, which showed no significant volume reliability differences.

Estimation of spatial reproducibility of volumetric segmentations: effects of MRI sites and segmentation analyses

The across-session test–retest spatial overlaps for both analysis streams are reported in Table 5, which shows that for each site, structure and analysis the mean Dice coefficient of spatial overlap averaged across subjects, across the 4 test–retest scans and across hemispheres. There were no significant MRI site effects of the Dice coefficients, regardless of analysis stream and structure. When averaged across MRI sites, the LG analysis showed significantly higher spatial reproducibility relative to the CS analysis, for all brain structures evaluated (Wilcoxon test, $p < 0.01$). When grouping hemispheric structures within each site separately we also found that the spatial reproducibility of the LG analysis was significantly higher than that obtained with the CS analysis (Wilcoxon test, $p < 0.02$). Overall this means that the LG analysis stream not only gives higher test–retest volume reproducibility than the CS analysis, but also higher spatial consistency, both within each independent MRI site and across sites when these are grouped.

Estimation of cortical thickness across sites

Table 6 summarizes the group mean cortical thickness results, averaged across hemispheres and across the test–retest sessions, for each MRI site as derived from the FreeSurfer LG segmentation stream. The Kruskall–Wallis test for MRI site effect on the hemispheric volumes showed that there were significant site-effects ($p < 0.01$) for only 3 of the 18 cortical structures evaluated: the right/left fusiform and the right superior frontal gyrus. This variability of morphometric results across sites is consistent with different degrees of anatomical variability from the different groups scanned at the different sites.

Effects of site and analysis on thickness reproducibility

Fig. 4 shows, similar to Fig. 2, an example of a Bland–Altman plot for a single site on the across-session thickness reproducibility of two sample cortical structures: the supramarginal gyrus (left) and the entorhinal cortex (right). In this example it is possible to see how the spread of thickness variability data is very similar in the LG and CS analyses for the supramarginal gyrus, but visibly reduced with the LG for the entorhinal cortex.

Table 7 summarizes the mean across-session test–retest reproducibility errors in the cortical thickness estimates. For each site the mean error is averaged across subjects, across the four test–retest scans and across brain hemispheres. No significant MRI site effects were found on the reproducibility error, regardless of structure and analysis stream used for the brain segmentations. The LG stream gave a significant reduction of the reproducibility error in the entorhinal cortex relative to the CS analysis (Wilcoxon test, $p < 0.01$). For all other evaluated cortical structures there were no significant reproducibility differences between the LG and CS analyses. This null effect on reproducibility differences was confirmed for the thickness of additional areas not reported in Table 6: cuneus, pre-central, inferior parietal and caudal middle frontal.

Fig. 5 is similar to Fig. 3, and it used to illustrate in a plot an example of the cortical thickness findings reported in Table 7. The figure shows the distribution of cortical thickness reproducibility errors (%) across

Table 3
Volume estimates across sites. Within-site group means and standard deviation (across subjects, scanner sessions and hemispheres) of subcortical, ventricle and intracranial volumes derived from the FreeSurfer longitudinal segmentation stream. Abbreviations for the segmented volumes: Hp = hippocampus, Amy = amygdala, Cau = caudate, Put = putamen, Pal = pallidum, Thal = thalamus, Lat = lateral ventricle volume, ICV = intracranial volume. See Table 2 for MRI site characterization.

Structure volume	MRI sites: volumetric estimates (mm^3)							
	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
Hp	3652 ± 394	3643 ± 399	3831 ± 382	4170 ± 251	4018 ± 241	4182 ± 209	3716 ± 396	3859 ± 589
Amy	1573 ± 239	1330 ± 229	1580 ± 160	1804 ± 320	1735 ± 263	1576 ± 179	1485 ± 211	1567 ± 248
Cau	3577 ± 293	2707 ± 305	3428 ± 265	3235 ± 295	3172 ± 207	4099 ± 434	3453 ± 295	3736 ± 992
Put	5152 ± 525	4458 ± 410	5051 ± 364	5441 ± 897	4950 ± 241	6713 ± 1008	4681 ± 260	5321 ± 622
Pal	1579 ± 167	1292 ± 118	1570 ± 248	1578 ± 248	1342 ± 342	2209 ± 409	1476 ± 169	1551 ± 324
Thal	6157 ± 525	5302 ± 458	6279 ± 745	6290 ± 813	6305 ± 646	7735 ± 731	5827 ± 424	6178 ± 827
Lat	14055 ± 7296	7579 ± 2465	11478 ± 5577	9150 ± 3252	6990 ± 2427	11509 ± 3965	13081 ± 6244	8018 ± 4397
ICV	1332302 ± 19476	1187475 ± 47990	1426996 ± 112925	1233636 ± 213199	1135640 ± 94246	1544876 ± 227217	1387241 ± 107959	1367857 ± 209535

the eight MRI sites for just two structures, the supramarginal gyrus (left) and the entorhinal cortex (right). As can be seen, the distribution of errors falls above and below the unity line, and the spread of errors of both analysis streams is comparable for the supramarginal gyrus, yet they appear greatly reduced for the entorhinal cortex. In other words, relative to the cross-sectional stream the longitudinal analysis shows significant improved reliability in the cortical thickness estimates of entorhinal cortex while offering comparable reliability for all other cortical areas investigated.

The ICC results for absolute thickness (not shown) were consistent with the absolute error analysis (difference relative to the mean), giving no significant differences between the thickness reproducibility errors from LG and CS analyses.

Effects of segmentation method on sample size

Fig. 6 shows the percent of subjects needed when using the longitudinal segmentation with respect to those needed by the cross-sectional segmentation to obtain the same power at same p-value to detect the same effect size. The longitudinal analysis offers a clear reduction in sample size, less than 40% as many subjects are required for most structures. A few of the structures showed smaller effects in sample size reductions (caudate volume, left entorhinal thickness) because the correlation of the estimates across sessions was high and similar for the two segmentation methods.

Effects of within session MPRAGE averaging

The two within session MPRAGE volumes acquired during the test and retest sessions were co-registered, averaged and segmented with the longitudinal segmentation analysis to test if the across-session reproducibility errors of volume and cortical thickness estimates would be reduced relative to those obtained with single MPRAGE acquisitions. We found no systematic and clear advantages when using two averaged MPRAGE volumes. The absolute reproducibility errors did not significantly differ in most structures between the two cases. Supplementary Fig. 1 shows summary results that compare the power analysis advantages (similar to Fig. 6) of the longitudinal analysis relative to the cross-sectional analysis for both the averaged and non-averaged MPRAGE volumes. It can be seen how for several structures averaging does not change the relative power to the cross-sectional analysis (hippocampus, putamen, thalamus), for a few structures averaging increases errors (amygdala, right hemisphere entorhinal and pallidum) and for a few other structures averaging reduces errors (right hemisphere caudate, left hemisphere entorhinal).

The global cortical gray matter signal intensity was also evaluated to investigate how image quality features varied across MRI sites for the averaged and non-averaged MPRAGE scans. For each subject the cortex intensity mean divided its standard deviations across the brain represents the signal-to-noise ratio (SNR) from a segmentation standpoint. The measures were done from the normalized images used for the final automated segmentation. Only the test-session was considered, the results from the retest session were similar. Supplementary Fig. 2 shows, for each MRI site, the cortical gray matter SNR (mean and standard deviation across subjects) for the first MPRAGE volume and the two averaged MPRAGE volumes. The Kruskall–Wallis test on global gray matter SNR gave significant MRI site effects ($p = 0.004$) on the averaged MPRAGE but no site effects on this single MPRAGE ($p > 0.05$). The effect was driven by lower signal from Site 1 (Siemens Allegra) and Site 6 (GE HDxt). Paired t-tests showed no significant group differences between the cortical gray matter SNR of the averaged and single acquisitions, at none of the sites ($p > 0.05$). There are two main observations from these results. One is that there were slight SNR differences across sites, most likely due to a combination of several reasons including differences in subject groups, differences in MRI hardware (Site 1 is the only one using

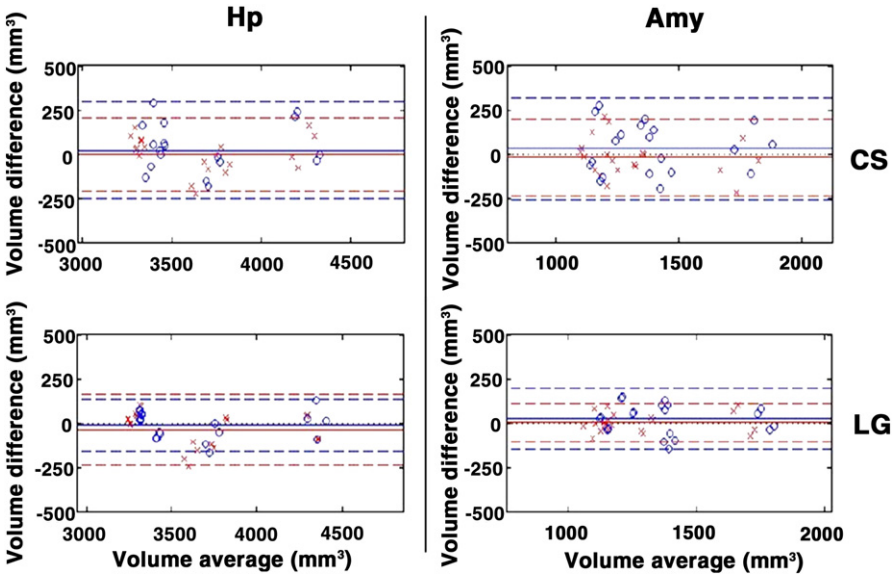

Fig. 2. Sample distribution of cross-sectional (CS) and longitudinal (LS) volume reproducibility results (Site 2) in hippocampus (Hp) and amygdala (Amy). Bland–Altman plots showing volume difference versus volume mean (two single MPRAGE acquisitions per session, subjects, n = 5). For each brain hemisphere (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.

a birdcage RF coil) and differences in MRI acquisition sequences (Site 6 is the only site using an IR-SPGR sequence). These SNR differences could be potentially reduced with further adjustments in the acquisition protocol. The second observation is that the SNR differences did not affect the across-session reproducibility of the morphometry measures studied, which gave no significant MRI site effects.

Discussion

The main goal of this study was to investigate the effects on reliability of two variants of the automated FreeSurfer brain segmentation analysis when used in a 3 T MRI consortium. The choices of MRI data acquisition and data analysis protocols can affect reproducibility errors and are therefore crucial in longitudinal studies aimed at evaluating MRI-derived biomarkers for disease progression and/or treatment efficacy. In this brain morphometry study we show for the first time the across-session test–retest reproducibility advantages of the fully automated longitudinal FreeSurfer segmentation analysis relative to the cross-sectional analysis, when tested in a consortium of different 3 T MRI scanners using different vendors (Siemens, Philips, GE). Specifically, cortical, subcortical and ventricular segmentations were obtained from a group of 40 healthy elderly subjects (mean age 63.2 ± 8.1 years, 5 different subjects per MRI site) who were scanned in two separate sessions (mean time interval of 17 days), using two standard 3D MPRAGE acquisitions per session (with parallel imaging when possible, no averaging) on eight different 3 T MRI scanners (Table 2). Our study confirms the hypothesis that the longitudinal FreeSurfer segmentation offers an overall improvement of morphometry reproducibility relative to the cross-sectional segmentation, both at the single site level and also in the overall consortium when the data from all sites are pooled. These results were consistently derived from three different across-session reliability evaluations: absolute percent change relative to the mean, Dice coefficient for spatial overlap and intraclass correlation coefficients.

Table 4
Brain volumetric reproducibility errors for the various 3 T MRI sites derived from the cross-sectional (CS) and longitudinal (LG) FreeSurfer segmentations. Within each site the mean reproducibility errors (percent absolute difference relative to the mean) are computed across subjects, across the four test–retest acquisitions and across brain hemispheres. There are no significant MRI site effects, regardless of analysis (Kruskall–Wallis test, $p < 0.01$). The last column shows the reproducibility errors for each site and analysis when averaged across sites. Except for the lateral ventricles, for all other structures the reproducibility errors of LG are significantly lower than those from CS analysis (Wilcoxon test, $p < 0.01$). Abbreviations for the segmented volumes: Hp = hippocampus, Amy = amygdala, Cau = caudate, Put = putamen, Pal = pallidum, Thal = thalamus, Lat = lateral ventricle volume. See Table 2 for MRI site characterization.

Structure and analyses		MRI sites: volumetric reproducibility errors (%)								Mean error across MRI sites (%)
		Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	
Hp	CS	3.50 ± 2.84	2.58 ± 2.02	3.56 ± 3.52	1.99 ± 1.59	2.40 ± 1.71	4.93 ± 4.53	3.34 ± 2.36	3.79 ± 2.58	3.26 ± 0.93
	LG	1.95 ± 1.77	1.92 ± 1.57	1.96 ± 1.44	0.91 ± 0.71	1.80 ± 1.31	2.07 ± 1.99	1.94 ± 1.40	1.76 ± 1.27	1.79 ± 0.37
Amy	CS	7.38 ± 7.04	8.02 ± 5.80	4.84 ± 3.73	4.26 ± 4,54	6.76 ± 6,80	8.40 ± 9.10	7.13 ± 5.37	9.46 ± 8.95	7.03 ± 1.75
	LG	4.59 ± 3.64	4.57 ± 3.15	3.56 ± 2.29	2.49 ± 1,96	3.48 ± 3,27	3.68 ± 2,63	2.91 ± 3.05	5.17 ± 5.64	3.81 ± 0.91
Cau	CS	2.76 ± 1.65	2.78 ± 2.26	3.19 ± 4.07	2.27 ± 1,49	2.37 ± 1,73	2.76 ± 2,07	2.16 ± 1.88	2.28 ± 2.01	2.57 ± 0.36
	LG	1.35 ± 1.07	1.69 ± 1.27	2.45 ± 3.47	1.64 ± 1,38	2.03 ± 1,35	2.46 ± 1,91	1.56 ± 1.26	1.51 ± 0.88	1.84 ± 0.43
Put	CS	5.38 ± 3.91	5.47 ± 4.94	3.14 ± 3.00	3.70 ± 3,52	4.32 ± 4,54	5.51 ± 3,86	4.98 ± 7.45	4.34 ± 5.31	4.61 ± 0.88
	LG	3.24 ± 2.96	2.09 ± 1.63	1.88 ± 1.26	2.07 ± 1,75	1.70 ± 1,44	2.82 ± 2,21	1.66 ± 1.41	1.52 ± 0.99	2.12 ± 0.60
Pal	CS	6.28 ± 5.23	5.54 ± 5.17	5.71 ± 4.70	6.11 ± 7,40	8.82 ± 10,70	11.21 ± 7,82	6.28 ± 7.13	8.34 ± 8.14	7.44 ± 1.95
	LG	4.93 ± 5.41	3.15 ± 3.16	3.63 ± 2.37	2.23 ± 1,30	4.46 ± 4,29	4.99 ± 4,16	2.67 ± 1.91	2.99 ± 2.44	3.76 ± 1.27
Thal	CS	4.15 ± 3.37	3.65 ± 3.18	4.09 ± 3.07	3.69 ± 3,21	5.52 ± 7,30	7.29 ± 5,19	5.40 ± 7.02	5.94 ± 7.01	4.97 ± 1.29
	LG	2.27 ± 1.71	1.78 ± 1.60	1.51 ± 1.17	1.79 ± 1,21	1.88 ± 1,38	2.11 ± 1,87	1.52 ± 1.44	1.42 ± 1.45	1.78 ± 0.30
Lat	CS	3.43 ± 2.64	1.88 ± 1.43	2.50 ± 1.69	2.35 ± 1,66	2.73 ± 2,68	2.36 ± 2,62	1.67 ± 1.28	1.90 ± 2.07	2.35 ± 0.56
	LG	2.37 ± 2.30	2.49 ± 1.49	2.70 ± 1.39	2.00 ± 0,98	2.47 ± 1,27	1.54 ± 1,24	2.17 ± 1.56	2.73 ± 2.73	2.31 ± 0.40

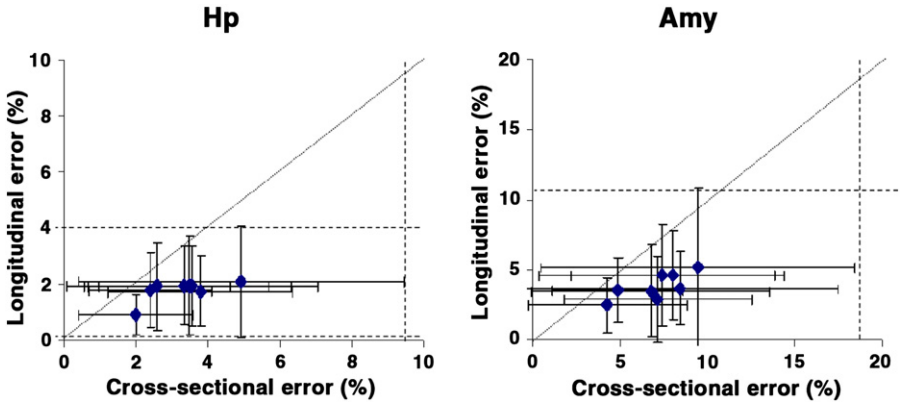

Fig. 3. Across-session test–retest reproducibility errors of hippocampus (Hp) and amygdala (Amy) volume estimates, effects of MRI site and processing stream. The plots show the reproducibility errors from the longitudinal and cross-sectional segmentations for each one of the eight 3 T MRI sites, with their respective within-site standard deviations. Data derived from Table 4.

Given the high resemblance of our MPRAGE protocol with the one proposed by ADNI for 3 T MRI systems, the multi-site implementation of this study was greatly facilitated by the detailed acquisition information that ADNI has made publicly available (http://adni.loni.ucla.edu/research/protocols/mri-protocols/). Using ADNI's sample protocols relevant to our MR systems and adding the few variations adopted in our study (isotropic voxels $1 \times 1 \times 1$ mm^3 and accelerated acquisitions when possible), it was possible to implement the target protocol. Our goal was to use a protocol that was as uniform as possible across sites while using the standard sequences made available by the various MRI system vendors. In particular, our target was to use an acceleration factor of 2 for all vendors with parallel imaging possibilities. The fact that in two sites this was instead set to 1.5 was an oversight, and introduced a slightly longer acquisition at those sites yet with no detected effects in reproducibility. The international nature of the study meant that the coordination and follow up of processes related to Ethical Committee approvals took considerable effort and time. In our experience it is highly advised to start with such procedures as soon as possible.

The brain segmentation results of volume (Table 3) and thickness (Table 6) are comparable to previous studies reporting similar metrics measured on elderly subjects (Fennema-Notestine et al., 2009; Han et al., 2006; Jovicich et al., 2009; Reuter et al., 2012; Velayudhan et al., 2013; Wonderlick et al., 2009). For most structures, there's a fairly wide range of estimates reported in the literature and the values found in this study are within the reported ranges.

The across-session reliability of the volumetric estimates was found to be overall consistent across the eight 3 T MRI sites for each structure and segmentation analysis tool (Table 4). In most structures, with only one exception, we found that for all sites the longitudinal analysis resulted in significantly improved volumetric reliability relative to the cross-sectional analysis, in average reducing the reproducibility error by half. Only in the lateral ventricle volume we found that there were no reliability differences between the two segmentation methods. We found that the smaller structures (pallidum and amygdala) yielded the highest absolute volume reproducibility errors, approximately 3.8% (average across sites), whereas all other structures had errors in the range 1.8–2.2% (average across sites), with the longitudinal segmentation analysis. Our absolute % errors in test–retest volumetric estimates are comparable to those reported by previous studies (Kruggel et al., 2010; Morey et al., 2010; Reuter et al., 2012). The spatial reproducibility of the segmented volumes was fairly constant and already good using the cross-sectional stream, with a mean Dice coefficient range across sites from 0.84 to 0.88 (Table 5). The spatial reproducibility was significantly improved with the longitudinal pipeline (mean Dice coefficient range across sites from 0.90 to 0.95). Spatial overlap results are also in good agreement with a previous within-session test–retest study (Reuter et al., 2012).

The thickness reproducibility results of the various structures were largely consistent across sites and vendors, with errors in the range 0.8–5.0% for the longitudinal analysis (Table 7). There was a trend

Table 5
Spatial reproducibility of volume segmentations. Within-site group mean volume overlap (Dice coefficient) and standard deviation (across subjects, scanner sessions and hemispheres) derived from the FreeSurfer cross-sectional (CS) and longitudinal (LG) segmentation streams. There are no significant MRI site effects, regardless of structure and analysis (Kruskall–Wallis test, $p < 0.01$). The last column shows the spatial reproducibility for each site and analysis when averaged across sites. For all structures the spatial reproducibility was significantly higher with the LG analysis relative to the CS analysis (Wilcoxon test, $p < 0.01$). Abbreviations for the segmented volumes: Hp = hippocampus, Amy = amygdala, Cau = caudate, Put = putamen, Pal = pallidum, Thal = thalamus, Lat = lateral ventricle volume. See Table 2 for MRI site characterization.

Structure and analyses		MRI sites: Dice coefficients for spatial overlap								Mean Dice across MRI sites
		Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	
Hp	CS	0.88 ± 0.02	0.89 ± 0.02	0.87 ± 0.03	0.89 ± 0.02	0.88 ± 0.02	0.86 ± 0.04	0.88 ± 0.06	0.84 ± 0.04	0.87 ± 0.02
	LG	0.92 ± 0.02	0.94 ± 0.02	0.91 ± 0.06	0.95 ± 0.03	0.93 ± 0.03	0.91 ± 0.07	0.95 ± 0.01	0.88 ± 0.06	0.92 ± 0.02
Amy	CS	0.83 ± 0.04	0.84 ± 0.03	0.85 ± 0.03	0.87 ± 0.03	0.85 ± 0.03	0.81 ± 0.02	0.91 ± 0.05	0.90 ± 0.03	0.86 ± 0.03
	LG	0.89 ± 0.03	0.91 ± 0.02	0.90 ± 0.05	0.94 ± 0.02	0.91 ± 0.04	0.89 ± 0.04	0.92 ± 0.01	0.87 ± 0.06	0.92 ± 0.03
Cau	CS	0.88 ± 0.02	0.87 ± 0.01	0.86 ± 0.03	0.87 ± 0.02	0.87 ± 0.02	0.84 ± 0.03	0.86 ± 0.12	0.87 ± 0.02	0.86 ± 0.01
	LG	0.93 ± 0.02	0.93 ± 0.02	0.91 ± 0.04	0.94 ± 0.01	0.93 ± 0.03	0.89 ± 0.04	0.94 ± 0.01	0.92 ± 0.04	0.92 ± 0.02
Put	CS	0.86 ± 0.03	0.88 ± 0.03	0.88 ± 0.02	0.89 ± 0.03	0.88 ± 0.03	0.86 ± 0.02	0.86 ± 0.11	0.87 ± 0.02	0.87 ± 0.01
	LG	0.91 ± 0.02	0.94 ± 0.01	0.92 ± 0.02	0.95 ± 0.01	0.94 ± 0.02	0.92 ± 0.03	0.94 ± 0.01	0.92 ± 0.03	0.93 ± 0.01
Pal	CS	0.80 ± 0.14	0.78 ± 0.15	0.81 ± 0.09	0.81 ± 0.16	0.75 ± 0.21	0.81 ± 0.04	0.77 ± 0.21	0.74 ± 0.19	0.78 ± 0.03
	LG	0.90 ± 0.05	0.90 ± 0.05	0.90 ± 0.05	0.94 ± 0.03	0.89 ± 0.07	0.89 ± 0.04	0.92 ± 0.02	0.89 ± 0.05	0.90 ± 0.02
Thal	CS	0.91 ± 0.01	0.92 ± 0.01	0.91 ± 0.02	0.92 ± 0.01	0.91 ± 0.03	0.89 ± 0.02	0.83 ± 0.07	0.82 ± 0.04	0.89 ± 0.04
	LG	0.95 ± 0.01	0.96 ± 0.01	0.95 ± 0.02	0.97 ± 0.01	0.96 ± 0.02	0.94 ± 0.02	0.96 ± 0.01	0.95 ± 0.02	0.94 ± 0.03
Lat	CS	0.92 ± 0.02	0.90 ± 0.03	0.91 ± 0.03	0.90 ± 0.03	0.88 ± 0.04	0.89 ± 0.02	0.92 ± 0.05	0.85 ± 0.06	0.90 ± 0.02
	LG	0.95 ± 0.01	0.95 ± 0.03	0.94 ± 0.03	0.95 ± 0.02	0.93 ± 0.05	0.92 ± 0.03	0.96 ± 0.01	0.89 ± 0.07	0.94 ± 0.02

Table 6
Cortical thickness estimates across sites. Within-site group means and standard deviation (across subjects, scanner sessions and hemispheres) of cortical thickness derived from the FreeSurfer longitudinal segmentation stream. Abbreviations: Fus = fusiform gyrus, LatOc = lateraloccipital gyrus, Ling = lingual gyrus, Parahp = parahippocampal gyrus, Prec = precuneus, SupFr = superiorfrontal gyrus, SupPar = superiorparietal gyrus, SupTem = superiortemporal gyrus, Supra = supramarginal gyrus, Ent = entorhinal cortex. See Table 2 for MRI site characterization.

Structure thickness	MRI sites: cortical thickness estimates (mm)							
	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
Fus	2.55 ± 0.13	2.84 ± 0.15	2.74 ± 0.11	2.75 ± 0.08	2.82 ± 0.12	3.08 ± 0.11	2.75 ± 0.09	2.77 ± 0.08
LatOc	2.22 ± 0.15	2.31 ± 0.10	2.33 ± 0.12	2.14 ± 0.09	2.26 ± 0.18	2.44 ± 0.10	2.30 ± 0.09	2.30 ± 0.09
Ling	2.01 ± 0.07	1.97 ± 0.07	2.11 ± 0.11	1.96 ± 0.07	2.09 ± 0.15	2.15 ± 0.10	2.11 ± 0.08	2.01 ± 0.09
Parahp	2.79 ± 0.24	2.95 ± 0.25	2.88 ± 0.27	2.80 ± 0.26	3.04 ± 0.26	3.06 ± 0.24	2.89 ± 0.23	2.76 ± 0.32
Prec	2.26 ± 0.09	2.38 ± 0.05	2.38 ± 0.11	2.33 ± 0.09	2.40 ± 0.17	2.41 ± 0.19	2.34 ± 0.05	2.30 ± 0.07
SupFr	2.67 ± 0.11	2.71 ± 0.05	2.61 ± 0.13	2.63 ± 0.09	2.79 ± 0.10	2.64 ± 0.12	2.59 ± 0.08	2.59 ± 0.11
SupPar	2.14 ± 0.10	2.27 ± 0.05	2.27 ± 0.15	2.18 ± 0.07	2.22 ± 0.11	2.27 ± 0.09	2.26 ± 0.05	2.20 ± 0.07
SupTem	2.63 ± 0.10	2.70 ± 0.13	2.73 ± 0.12	2.80 ± 0.09	2.81 ± 0.10	2.86 ± 0.15	2.70 ± 0.16	2.77 ± 0.10
Supra	2.41 ± 0.09	2.51 ± 0.09	2.50 ± 0.13	2.46 ± 0.13	2.59 ± 0.13	2.61 ± 0.10	2.53 ± 0.12	2.53 ± 0.09
Ent	3.46 ± 0.28	3.62 ± 0.33	3.72 ± 0.33	3.62 ± 0.19	3.78 ± 0.32	3.89 ± 0.36	3.75 ± 0.39	3.54 ± 0.30

for Site 1 to show higher reproducibility errors on thickness estimates relative to all other sites, especially for the cross-sectional analysis. This may be due to the fact that Site 1 was the only one not using a multi-channel RF coil, which potentially leads to lower image quality on the cortex and in addition a longer image acquisition without parallel imaging that is more susceptible to signal degradation from head motion during the acquisition. Only on one structure, the entorhinal cortex, we found that the longitudinal segmentation gave consistently improved thickness reliability across sites relative to the cross-sectional segmentation. For the other cortical thickness structures investigated we found no significant differences in the across-session test retest reliability of the two segmentation streams. This is in contrast to previous studies that have shown in elderly subjects that the longitudinal analysis can improve test–retest thickness reproducibility (Han et al., 2006; Reuter et al., 2012). Several study differences may explain this discrepancy. The study of Han et al. used a 1.5 T system for the across-session test–retest, they used a larger number of subjects (N = 15), and the value reported is global mean thickness across the whole brain cortex while we use the standard FreeSurfer outputs of mean thickness for several gyri. In the case of the study of Reuter et al., although this is a 3 T study, several other factors may account for differences relative to our study, including increased sensitivity from their population size (N = 115), improved across-session co-registrations by using multi-echo MPRAGE sequence (van der Kouwe et al., 2008; Wonderlick et al., 2009) and reduced variability given that only within-session acquisitions were acquired and analyzed. One disadvantage of the multi-echo MPRAGE sequence is that it is not yet available on all MRI vendor platforms.

In agreement with two multi-site 1.5 T reproducibility studies, one focused on cortical thickness reproducibility (Han et al., 2006) and one focused on subcortical, ventricular and intracranial volume reproducibility (Jovicich et al., 2009), we found that averaging two MPRAGE acquisitions acquired within a session made relatively minor contributions to improvement in the across-session reproducibility. The acquisition of two MPRAGE volumes is still recommended mainly for practical reasons: if one volume is bad (e.g. due to motion artifacts) then the other can still be used for segmentation without averaging.

To minimize biases a multi-site reproducibility study should ideally use a large sample of volunteers who are all scanned repeated times at all sites within a short time period. Such a study is extremely challenging for multiple reasons, including costs and coordination, particularly in the case of a consortium distributed internationally. Our study has several limitations relative to this ideal scenario: i) each MRI site scanned a different set of subjects but with consistent recruitment

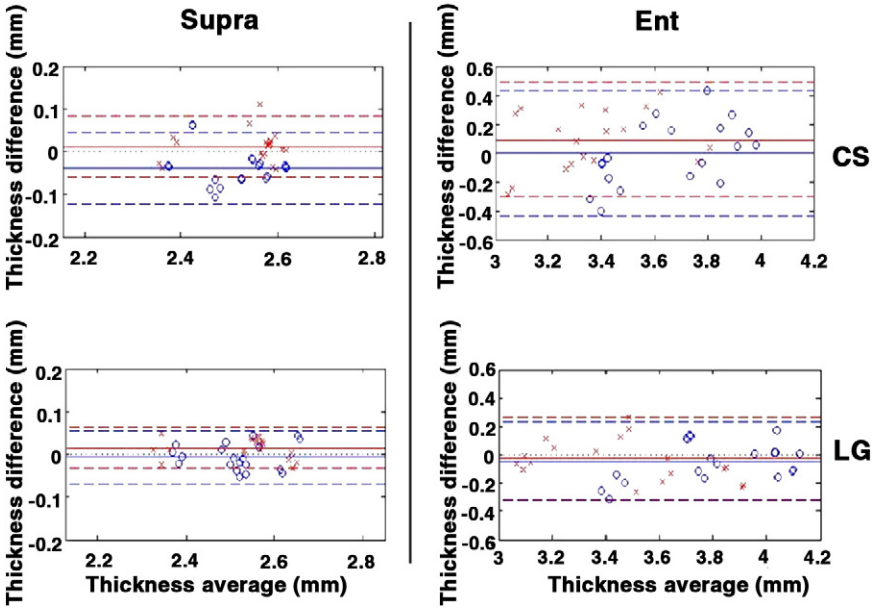

Fig. 4. Sample distribution of cross-sectional (CS) and longitudinal (LG) thickness reproducibility results (Site 2) in supramarginal gyrus (Supra) and entorhinal cortex (Ent). Bland–Altman plots showing thickness difference versus thickness mean (two single MPRAGE acquisitions per session, subjects, n = 5). 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.

Table 7
Effects of MRI site and processing stream on thickness reproducibility. Within-site group mean reproducibility error (percent absolute difference relative to the mean) and standard deviation (across subjects, scanner sessions and hemispheres) derived from the FreeSurfer cross-sectional (CS) and longitudinal (LG) segmentation streams. There are no significant MRI site effects, regardless of structure and analysis (Kruskall–Wallis test, $p < 0.01$). The last column shows the spatial reproducibility for each site and analysis when averaged across sites. No significant differences were found between the thickness reproducibility errors from LG and CS analyses when grouped across sites (Wilcoxon test, $p < 0.01$). Abbreviations: Fus = fusiform gyrus, LatOc = lateraloccipital gyrus, Ling = lingual gyrus, Parahp = parahippocampal gyrus, Prec = precuneus, SupFr = superiorfrontal gyrus, SupPar = superiorparietal gyrus, SupTem = superiortemporal gyrus, Supra = supramarginal gyrus, Ent = entorhinal cortex. See Table 2 for MRI site characterization.

Cortical structures		MRI sites: cortical thickness reproducibility error (%)								Average error across sites (%)
		Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	
Fus	CS	4.53 ± 3.24	2.03 ± 1.57	1.76 ± 1.35	1.73 ± 1.25	2.73 ± 2.07	1.92 ± 1.46	3.18 ± 2.47	2.62 ± 1.59	2.56 ± 0.95
	LG	4.31 ± 3.42	1.74 ± 1.49	1.68 ± 1.50	2.36 ± 1.56	3.07 ± 2.17	1.41 ± 1.12	1.55 ± 1.05	1.95 ± 1.21	2.26 ± 0.99
LatOc	CS	4.12 ± 3.93	2.01 ± 1.36	1.61 ± 1.27	1.25 ± 0.77	2.51 ± 2.10	2.47 ± 2.65	2.65 ± 2.08	2.00 ± 1.81	2.33 ± 0.87
	LG	2.18 ± 1.52	1.96 ± 0.99	1.69 ± 1.33	1.83 ± 1.56	2.34 ± 2.35	2.05 ± 1.59	2.21 ± 1.30	2.27 ± 1.41	2.07 ± 0.23
Ling	CS	5.14 ± 4.43	2.51 ± 1.27	2.35 ± 2.39	1.45 ± 1.19	2.25 ± 1.37	3.04 ± 2.19	2.55 ± 1.81	2.11 ± 1.92	2.67 ± 1.09
	LG	2.07 ± 1.45	1.88 ± 1.27	1.91 ± 1.62	1.92 ± 1.67	2.62 ± 2.28	2.05 ± 1.59	1.80 ± 1.50	2.37 ± 1.68	2.08 ± 0.28
Parahp	CS	5.51 ± 5.13	2.86 ± 1.97	2.60 ± 1.58	3.53 ± 2.19	3.15 ± 2.52	2.22 ± 1.97	2.47 ± 2.26	2.90 ± 2.23	3.15 ± 1.03
	LG	4.54 ± 3.59	1.76 ± 1.07	2.14 ± 2.02	2.25 ± 1.94	2.93 ± 2.48	1.52 ± 1.29	1.40 ± 1.24	2.49 ± 1.90	2.38 ± 1.01
Prec	CS	3.47 ± 3.45	1.57 ± 1.05	1.67 ± 1.08	1.55 ± 1.35	2.43 ± 1.86	2.36 ± 1.55	1.78 ± 1.60	2.53 ± 1.61	2.17 ± 0.66
	LG	3.18 ± 3.35	1.36 ± 0.98	1.02 ± 0.86	1.44 ± 0.99	2.23 ± 1.73	2.78 ± 2.03	1.72 ± 1.36	2.47 ± 1.56	2.02 ± 0.76
SupFr	CS	1.58 ± 0.93	1.93 ± 1.56	1.58 ± 1.48	4.48 ± 2.88	2.21 ± 1.96	2.83 ± 2.96	1.99 ± 1.87	3.37 ± 1.96	2.50 ± 1.01
	LG	1.57 ± 1.05	1.16 ± 0.81	1.57 ± 1.07	4.29 ± 2.98	1.78 ± 1.59	1.45 ± 1.02	1.53 ± 1.05	3.21 ± 3.21	2.07 ± 1.09
SupPar	CS	3.66 ± 4.07	1.68 ± 1.16	1.69 ± 1.37	1.50 ± 1.02	3.16 ± 2.28	3.19 ± 2.54	2.66 ± 2.59	3.01 ± 2.16	2.57 ± 0.83
	LG	2.27 ± 2.28	0.85 ± 0.59	1.34 ± 1.09	1.57 ± 1.65	2.20 ± 1.68	1.38 ± 0.93	1.55 ± 1.11	2.33 ± 1.54	1.69 ± 0.53
SupTem	CS	2.41 ± 2.10	1.38 ± 1.11	1.11 ± 0.73	1.55 ± 1.14	3.16 ± 2.28	2.76 ± 2.16	1.78 ± 1.57	1.53 ± 1.36	1.96 ± 0.73
	LG	2.58 ± 2.28	1.33 ± 0.78	1.03 ± 0.84	1.31 ± 0.89	1.35 ± 0.93	1.96 ± 1.79	1.17 ± 0.86	1.13 ± 1.22	1.48 ± 0.53
Supra	CS	2.63 ± 2.59	1.58 ± 1.09	1.64 ± 1.07	1.81 ± 1.10	1.88 ± 1.83	2.45 ± 2.14	1.91 ± 2.01	2.99 ± 2.83	2.11 ± 0.51
	LG	2.91 ± 2.47	1.00 ± 0.59	1.37 ± 1.24	1.75 ± 1.41	1.86 ± 1.35	1.92 ± 1.30	1.30 ± 0.86	2.23 ± 1.40	1.79 ± 0.60
Ent	CS	9.63 ± 7.77	5.12 ± 3.47	4.35 ± 3.83	5.30 ± 3.42	3.82 ± 2.46	6.66 ± 6.35	4.06 ± 3.45	4.60 ± 3.68	5.53 ± 3.07
	LG	5.01 ± 3.50	3.34 ± 2.25	2.35 ± 2.29	3.41 ± 2.69	2.80 ± 2.63	2.31 ± 1.65	2.07 ± 1.50	2.67 ± 1.99	3.01 ± 1.42

criteria, ii) the number of subjects studied per site was low, five, and iii) the number of test–retest across-session repetitions acquired was the absolute minimum, two. The rather large range of recruitment ages (50–80), which was chosen to be consistent with the follow up clinical study, combined with the first two limitations explains the MRI site effects found for mean age across sites. This also led to some anatomical differences across the sites, with MRI site effects in a few of the mean volume (left putamen and right pallidum) and thickness (right/left fusiform and right superior frontal gyrus) estimates. Altogether the anatomical differences across sites are in the order of 15%, considering that the MRI site effects were significant in 5 of 33 evaluated structures, volumetric and thickness measures combined. Since this study was focused on evaluating test–retest reproducibility we expect that these few across-site anatomical differences will not affect the main findings. The use of only two across-session repetitions will probably lead to a lower-limit estimate of the test–retest variance, assuming that higher number of repetitions may introduce higher variance from a variety of sources (including MRI scanner instabilities, subject positioning, subject hydration). An additional limitation of our reproducibility study is that we do not have a balanced distribution of 3 T MRI vendor platforms (Siemens: 5, Philips: 2, GE: 1), yet this limitation might be reduced as new clinical centers join the consortium. With these limitations it is hard to establish whether the lack of significant site-dependent reproducibility findings will remain had we studied more subjects. Pooling the data across vendors to test for MRI system effects (e.g. Siemens vs. Philips) would allow evaluating a larger population. This, however, has two main problems related to the limitations previously mentioned: unbalanced number of sites for each vendor and unbalanced heterogeneity of scanner models across vendors (the two Philips sites used identical models, Achieva, whereas the Siemens sites used four different models, TrioTIM, Skyra, Verio and Allegra). Lastly, we do not report a random effects study, therefore the results should not be extrapolated to acquisition protocols (pulse sequence, scanner) or subject populations not included in this study.

In addition to volumetric and cortical thickness estimates other morphometric measures can be used to study brain anatomy. As recently shown, the characterization of 3D shape of brain structures (Miller, 2004; Miller et al., 2009; Wang et al., 2007) may be used to investigate differences between subject populations (Frisoni et al., 2008, Cavedo et al., 2011). Therefore, the combination of both volume and shape metrics might improve the power of detecting cross-sectional differences across populations or longitudinal changes. An important

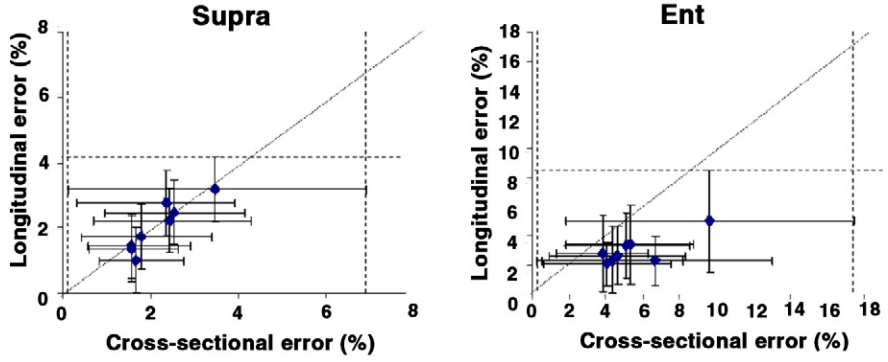

Fig. 5. Across-session test–retest reproducibility errors of supramarginal gyrus (Supra) and entorhinal cortex (Ent) thickness estimates, effects of MRI site and processing stream. The plots show the reproducibility errors from the longitudinal and cross-sectional segmentations for each one of the eight 3 T MRI sites, with their respective within-site standard deviations. Data derived from Table 7.

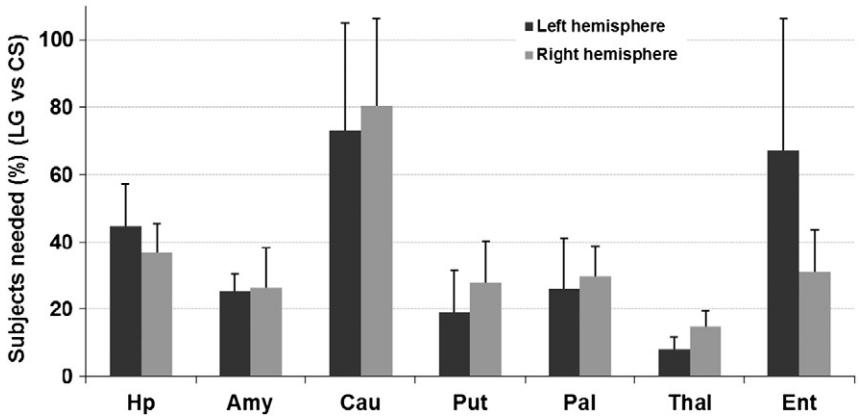

Fig. 6. Sample size ratio needed to have the same power analysis using LG instead of CS stream. The percentage of subjects needed is less than 50% for most of the structures considered. Left and right hemispheric structure labels are: Hp = hippocampus, Amy = amygdala, Cau = caudate, Put = putamen, Pal = pallidum, Thal = thalamus and Ent = entorhinal cortex.

extension of the reproducibility study here presented could be to examine the reproducibility of shape metrics.

The optimization of protocols towards improving the reliability of metrics characterizing brain morphometry is crucial for longitudinal studies, both for the search of potentially new useful biomarkers as well as for the monitoring of a disease with known markers. In addition, biomarkers that are considered for validation by governmental organizations must be robust, indicating that the characterization of their reproducibility by means of multi-site MRI studies is important; not only for their general reliability but also for understanding how effectively they may be used to monitor disease progression. Our results show that, to detect the same effect size with same statistical power, the longitudinal segmentation analysis needs less than 40% of the subjects that would be needed with the cross-sectional segmentation. Such reduction in the number of subjects needed or the number of longitudinal acquisitions is the result of the higher across-session reliability and can translate into significant cost reductions in longitudinal studies such as for example drug trials. These results are very similar to those of a recent study that evaluated across-session test–retest data (two averaged non-accelerated MPRAGE acquisitions) obtained at a single 1.5 T scanner (Reuter et al., 2012). The study of Reuter et al. (2012) also showed how the refinement of the longitudinal stream is sufficient to improve the discrimination between patients in two longitudinal studies, one with dementia and one with Huntington's disease subjects. Based on these findings we believe that our confirmation of the improved reliability of the longitudinal stream in a multi-site 3 T MRI setting is not associated to a cost of sensitivity to detect changes related to neurodegeneration.

The multi-site anonymous 3D MPRAGE imaging data acquired in this study (158 brain volumes) will be made publicly available to promote the development and evaluation of brain segmentation tools (https://neugrid4you.eu/).

Conclusions

This study achieved the following three main goals: i) a structural MRI acquisition protocol for morphometry analysis was implemented across eight 3 T MRI sites (3D MPRAGE, most sites using mildly accelerated acquisitions) covering various vendors (Siemens, Philips, GE) and countries (Italy, Spain, Germany and France); ii) within- and across-session test–retest data were acquired from a group of 40 healthy elderly volunteers (5 different volunteers per MRI site), generating a dataset with a total of 158 brain MRI volumes (8 sites, 5 subjects per site, 2 within-session acquisitions and 2 across-session acquisitions at least a week apart, 2 missing volumes) and iii) two fully automated brain segmentation protocols were evaluated and compared in terms of the across-session reproducibility of their results: the cross-sectional and longitudinal FreeSurfer segmentation streams. The main result is that the longitudinal analysis yields a consistent improved reproducibility across the various sites relative to the cross-sectional segmentation, reducing the variability by about half in most volumetric estimates and in the entorhinal cortical thickness, while not significantly changing the variability in the rest of cortical structures studied. The average of two MPRAGE volumes acquired within each test–retest sessions did not result in a systematic reduction of the across-session reproducibility errors. To the best of our knowledge this is the first study that confirms the improved performance of the longitudinal analysis in a 3 T consortium with various MRI vendors using a population of healthy elderly subjects and a fairly standard acquisition protocol. In addition, within the limitations of the sample size and MRI sites tested, our study provides preliminary reference values for absolute percent test–retest variability errors for a variety of volumetric (subcortical, ventricle, intracranial) and cortical thickness structures. These errors may be used as instrumental error estimates for power analysis in the structures and measures of interest. Lastly, we make the raw anonymous MRI data of this study publicly available so that it can be used for studies evaluating other morphometric segmentation tools as well as for future developments of the analysis methods here tested.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.05.007.

Acknowledgments

PharmaCog is funded by the EU-FP7 for the Innovative Medicine Initiative (grant no. 115009). All members of the PharmaCog project deserve sincere acknowledgment for their significant efforts, but unfortunately, they are too numerous to mention. The authors would like to thank especially to people who contributed to the early phases of this study, including Luca Venturi, Genoveffa Borsci and Thomas Günther.

Conflict of interest

The authors have no conflict of interests to declare.

References

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

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Bernasconi, A., Bernasconi, N., Bernhardt, B.C., Schrader, D., 2011. Advances in MRI for 'cryptogenic' epilepsies. Nat. Rev. Neurol. 7, 99–108.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.

Camara, O., Scahill, R.I., Schnabel, J.A., Crum, W.R., Ridgway, G.R., Hill, D.L., Fox, N.C., 2007. Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal data. Med. Image Comput. Comput. Assist. Interv. 10, 785–792.

Carrillo, M.C., Bain, L.J., Frisoni, G.B., Weiner, M.W., 2012. Worldwide Alzheimer's disease neuroimaging initiative. Alzheimers Dement. 8, 337–342.
Cavedo, E., Boccardi, M., Ganzola, R., Canu, E., Beltramello, A., Caltagirone, C., Thompson, P.M., Frisoni, G.B., 2011. Local amygdala structural differences with 3T MRI in patients with Alzheimer disease. Neurology 76, 727–733.
Chen, R., Jiao, Y., Herskovits, E.H., 2011. Structural MRI in autism spectrum disorder. Pediatr. Res. 69, 63R–68R.
Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9, 179–194.
Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980.
Diggle, P.J., Heagerty, P.J., Liang, K.-Y., Zeger, S.L., 2002. Analysis of Longitudinal Data, 2nd edition.
Drago, V., Babiloni, C., Bartrés-Faz, D., Caroli, A., Bosch, B., Hensch, T., Didic, M., Klafki, H.W., Pievani, M., Jovicich, J., Venturi, L., Spitzer, P., Vecchio, F., Schoenknecht, P., Wiltfang, J., Redolfi, A., Forloni, G., Blin, O., Irving, E., Davis, C., Hårdemark, H.G., Frisoni, G.B., 2011. Disease tracking markers for Alzheimer's disease at the prodromal (MCI) stage. J. Alzheimers Dis. 26 (Suppl. 3), 159–199.
Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Jacobson, M.W., Dale, A.M., The Alzheimer's Disease Neuroimaging Initiative, 2009. Structural neuroimaging in the detection and prognosis of pre-clinical and early AD. Behav. Neurol. 21, 3–12.
Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195–207.
Fischl, B., Salat, D.H., 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, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.
Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14, 11–22.
Fjell, A.M., Walhovd, K.B., 2012. Neuroimaging results impose new views on Alzheimer's disease—the role of amyloid revised. Mol. Neurobiol. 45, 153–172.
Frisoni, G.B., Ganzola, R., Canu, E., Rüb, U., Pizzini, F.B., Alessandrini, F., Zoccatelli, G., Beltramello, A., Caltagirone, C., Thompson, P.M., 2008. Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. Brain 131, 3266–3276.
Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6, 67–77.
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, A., Dickerson, B., Fischl, B., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. NeuroImage 32, 180–194.
Jack, C.R., 2011. Alliance for aging research AD biomarkers work group: structural MRI. Neurobiol. Aging 32 (Suppl. 1), S48–S57.
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, A., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage 30, 436–443.
Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D., Maguire, P., Rosas, D., Makris, N., Gollub, R., Dale, A., Dickerson, B.C., Fischl, B., 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. NeuroImage 46, 177–192.
Kostić, V.S., Filippi, M., 2011. Neuroanatomical correlates of depression and apathy in Parkinson's disease: magnetic resonance imaging studies. J. Neurol. Sci. 310, 61–63.
Kruggel, F., Turner, J., Muftuler, L.T., Initiative, A.s.D.N., 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. NeuroImage 49, 2123–2133.
Leow, A., Yu, C.L., Lee, S.J., Huang, S.C., Protas, H., Nicolson, R., Hayashi, K.M., Toga, A.W., Thompson, P.M., 2005. Brain structural mapping using a novel hybrid implicit/explicit framework based on the level-set method. NeuroImage 24, 910–927.
Levitt, J.J., Bobrow, L., Lucia, D., Srinivasan, P., 2010. A selective review of volumetric and morphometric imaging in schizophrenia. Curr. Top. Behav. Neurosci. 4, 243–281.
Lötjönen, J.M., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., Initiative, A.s.D.N., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49, 2352–2365.
Magnotta, V.A., Harris, G., Andreasen, N.C., O'Leary, D.S., Yuh, W.T., Heckel, D., 2002. Structural MR image processing using the BRAINS2 toolbox. Comput. Med. Imaging Graph. 26, 251–264.
Miller, M.I., 2004. Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms. NeuroImage 23 (Suppl. 1), S19–S33.
Miller, M.I., Priebe, C.E., Qiu, A., Fischl, B., Kolasny, A., Brown, T., Park, Y., Ratnanather, J.T., Busa, E., Jovicich, J., Yu, P., Dickerson, B.C., Buckner, R.L., Birn, M., 2009. Collaborative computational anatomy: an MRI morphometry study of the human brain via diffeomorphic metric mapping. Hum. Brain Mapp. 30, 2132–2141.
Morey, R.A., Selgrade, E.S., Wagner, H.R., Huettel, S.A., Wang, L., McCarthy, G., 2010. Scan–rescan reliability of subcortical brain volumes derived from automated segmentation. Hum. Brain Mapp. 31, 1751–1762.
Mueller, S.G., Stables, L., Du, A.T., Schuff, N., Truran, D., Cashdollar, N., Weiner, M.W., 2007. Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4 T. Neurobiol. Aging 28, 719–726.
Rajaratnam, N., 1960. Reliability formulas for independent decision data when reliability data are matched. Psychometrika 25, 11.
Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61, 1402–1418.
Rojas, D.C., Smith, J.A., Benkers, T.L., Camou, S.L., Reite, M.L., Rogers, S.J., 2004. Hippocampus and amygdala volumes in parents of children with autistic disorder. Am. J. Psychiatry 161, 2038–2044.
Selvaraj, S., Arnone, D., Job, D., Stanfield, A., Farrow, T.F., Nugent, A.C., Scherk, H., Gruber, O., Chen, X., Sachdev, P.S., Dickstein, D.P., Malhi, G.S., Ha, T.H., Ha, K., Phillips, M.L., McIntosh, A.M., 2012. Grey matter differences in bipolar disorder: a meta-analysis of voxel-based morphometry studies. Bipolar Disord. 14, 135–145.
Silk, T.J., Wood, A.G., 2011. Lessons about neurodevelopment from anatomical magnetic resonance imaging. J. Dev. Behav. Pediatr. 32, 158–168.
Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17, 479–489.
Studholme, C., Cardenas, V., Schuff, N., Rosen, H., Miller, B., Weiner, M., 2001. Detecting Spatially Consistent Structural Differences in Alzheimer's and Frontotemporal Dementia Using Deformation Morphometry. MICCAI 41–48.
van der Kouwe, A.J., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. NeuroImage 40, 559–569.
Van Horn, J.D., Toga, A.W., 2009. Multisite neuroimaging trials. Curr. Opin. Neurol. 22, 370–378.
van Rijsbergen, C., 1979. Information Retrieval, 2nd ed. Butterworths, London, U.K.
Velayudhan, L., Proitsi, P., Westman, E., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Spenger, C., Hodges, A., Powell, J., Lovestone, S., Simmons, A., dNeuroMed Consortium, 2013. Entorhinal cortex thickness predicts cognitive decline in Alzheimer's disease. J. Alzheimers Dis. 33, 755–766.
Walters, R.J., Fox, N.C., Crum, W.R., Taube, D., Thomas, D.J., 2001. Haemodialysis and cerebral oedema. Nephron 87, 143–147.
Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J.C., Csernansky, J.G., Miller, M.I., 2007. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE Trans. Med. Imaging 26, 462–470.
Whitwell, J.L., Sampson, E.L., Watt, H.C., Harvey, R.J., Rossor, M.N., Fox, N.C., 2005. A volumetric magnetic resonance imaging study of the amygdala in frontotemporal lobar degeneration and Alzheimer's disease. Dement. Geriatr. Cogn. Disord. 20, 238–244.
Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., Initiative, A.s.D.N., 2010. LEAP: learning embeddings for atlas propagation. NeuroImage 49, 1316–1325.
Wonderlick, J.S., Ziegler, D.A., Hosseini-Varnamkhasti, P., Locascio, J.J., Bakkour, A., van der Kouwe, A., Triantafyllou, C., Corkin, S., Dickerson, B.C., 2009. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. NeuroImage 44, 1324–1333.