Psychiatry Research: Neuroimaging 224 (2014) 14–21

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

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

The SIENA/FSL whole brain atrophy algorithm is no more reproducible at 3 T than 1.5 T for Alzheimer's disease

Keith S. Cover [a,*], Ronald A. van Schijndel [b], Veronica Popescu [b], Bob W. van Dijk [a], Alberto Redolfi [c], Dirk L. Knol [d], Giovanni B. Frisoni [c], Frederik Barkhof [b,e], Hugo Vrenken [a,b,e], neuGRID[1], the Alzheimer's Disease Neuroimaging Initiative[2]

[a] Department of Physics and Medical Technology, VU University medical center, Amsterdam, The Netherlands
[b] Department of Radiology, VU University medical center, Amsterdam, The Netherlands
[c] Laboratory of Epidemiology & Neuroimaging, IRCCS San Giovanni di Dio Fatebenefratelli, Via Pilastroni 4, 25125 Brescia, Italy
[d] Department of Epidemiology and Biostatistics, VU University medical center, Amsterdam, The Netherlands
[e] MS Center Amsterdam and Alzheimer Center, VU University medical center, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form
3 July 2014
Accepted 4 July 2014
Available online 14 July 2014

Keywords:
Brain atrophy
Alzheimer's
Back-to-back
Reproducibility'
FSL

ABSTRACT

The back-to-back (BTB) acquisition of MP-RAGE MRI scans of the Alzheimer's Disease Neuroimaging Initiative (ADNI1) provides an excellent data set with which to check the reproducibility of brain atrophy measures. As part of ADNI1, 131 subjects received BTB MP-RAGEs at multiple time points and two field strengths of 3 T and 1.5 T. As a result, high quality data from 200 subject-visit-pairs was available to compare the reproducibility of brain atrophies measured with FSL/SIENA over 12 to 18 month intervals at both 3 T and 1.5 T. Although several publications have reported on the differing performance of brain atrophy measures at 3 T and 1.5 T, no formal comparison of reproducibility has been published to date. Another goal was to check whether tuning SIENA options, including -B, -S, -R and the fractional intensity threshold (f) had a significant impact on the reproducibility. The BTB reproducibility for SIENA was quantified by the 50th percentile of the absolute value of the difference in the percentage brain volume change (PBVC) for the BTB MP-RAGES. At both 3 T and 1.5 T the SIENA option combination of "-B $f=0.2$", which is different from the default values of $f=0.5$, yielded the best reproducibility as measured by the 50th percentile yielding 0.28 (0.23–0.39)% and 0.26 (0.20–0.32)%. These results demonstrated that in general 3 T had no advantage over 1.5 T for the whole brain atrophy measure – at least for SIENA. While 3 T MRI is superior to 1.5 T for many types of measurements, and thus worth the additional cost, brain atrophy measurement does not seem to be one of them.

1. Introduction

Brain volume loss has become accepted as an objective marker in diseases such as Alzheimer's disease (AD) and multiple sclerosis (MS) (Barkhof et al., 2009; Filippi et al., 2004; De Stefano et al., 2003; Fox and Freeborough, 1997; Fox et al., 1996). SIENA (Smith et al., 2002, 2007) – which is part of the FSL software package (FSL, http://www.fmrib.ox.ac.uk/fsl) – is one of several techniques that is used to measure whole-brain atrophy from structural MRI sequences, such as MP-RAGE (Fox et al., 2000; Freeborough and Fox, 1997; Leung et al., 2010; Rudick et al., 1999). MP-RAGE is a heavily T1 weighted 3D sequence that was adopted for Alzheimer's Disease Neuroimaging Initiative (ADNI). SIENA is a registration based software algorithm that calculates the percentage brain volume change (PBVC) between two structural MRI sequences, such as MP-RAGEs, of the same subject acquired at two different time points. Major advantages of using SIENA for this study include it is widely used, it runs fully automatically and can usually calculate a PBVC for a pair of MP-RAGEs in under one hour of core time on a computer.

For subjects in the age range prone to Alzheimer's disease (AD), the PBVC per year is roughly −0.6%/y for healthy controls (HC) (Cover et al., 2011; Enzinger et al., 2005; Evans et al., 2010; Sluimer

* Correspondence to: VU University Medical Center, Department of Physics and Medical Technology, Postbus 7057, 1007MB Amsterdam, The Netherlands.
Tel.: +31 20 444 0677; fax: +31 20 444 4816.
E-mail address: Keith@kscover.ca (K.S. Cover).

[1] see http://neuGRID.eu

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf).

http://dx.doi.org/10.1016/j.pscychresns.2014.07.002
0925-4927/

et al., 2009; Smith et al., 2002), −1.1%/y for mildly cognitively impairment (MCI) (Cover et al., 2011; Evans et al., 2010) and −1.6%/y for AD (Cover et al., 2011; Evans et al., 2010). For a drug to be classified as effective, it should slow the additional progress of atrophy in MCI patients over and above that occurring in healthy aging by about half. Therefore any brain atrophy measure needs to be sensitive to changes of about 0.25% in the PBVC.

There have been several publications discussing the performance of SIENA, in particular its reproducibility (Smith et al., 2007; Cover et al. 2011). Smith et al. scanned 68 subjects (45 AD and 23 HC) at weeks 0, 6 and 38 to assess the reproducibility of SIENA at 1.5 T. With research studies in AD and MS shifting to 3 T (de Jong et al., 2008; Watson et al. 2010) an assessment of the performance of atrophy methods at 3 T, and a comparison between 3 T and 1.5 T are warranted. A direct comparison between 3 T and 1.5 T has been performed for a cross sectional method to extract the intracranial cavity (Keihaninejad et al. 2010), but not for a longitudinal measure like SIENA. While the ADNI1 BTB MP-RAGE data set can be used to compare the reproducibility of a wide range of brain atrophy measures at two differing field strengths, we have limited this first application to a single atrophy measure to ensure that we fully understand all the issues involved. SIENA is a good choice for a first application of a technique because it is widely used, fully automatic and conveniently yields a single number simplifying the reproducibility analysis.

Although not originally intended for reproducibility studies, the MP-RAGEs in the MRI component of the ADNI (Jack et al., 2008) provide a readily available benchmark for the reproducibility of brain atrophy measures (Cover et al. 2011). The ADNI1 data set consists of more than 800 subjects – 200 HC, 400 MCI and 200 AD – most of which have received MR scans at least 4 times over a two year period. In spite of the fact that many of the MP-RAGEs have received additional processing during acquisition, such as gradient distortion (GD) correction while others may not have, these MP-RAGEs can be used to study reproducibility – as was done by Cover et al. (2011) – because each patient had the identical protocol in the same MRI scanner for their 4MP-RAGEs used in this study. Also, the MP-RAGEs used in this study had no other processing than that included in the MRI acquisition sequence such as 2D or 3D GD correction.

Specifically, as part of the ADNI1 protocol, each subject on each visit had a pair of MP-RAGEs acquired with the second MP-RAGE starting within seconds of the completion of the first. These back-to-back (BTB) MP-RAGEs were originally included in the protocol to ensure that at least one MP-RAGE was of high quality (Jack et al., 2008) but in practice both MP-RAGEs were of high quality for most subject visits. Smith et al. (2007) had also used BTB acquisition to assess the reproducibility of brain atrophy measures including SIENA but using another data set. The key point that makes the differences of BTB PBVCs so useful is that processing steps, such as GD correction, will have little impact on the BTB differences even though they can substantially change the individual PBVCs. The BTB differences are relatively immune to processing steps such as GD corrections because nearly identical corrections are applied to each of the BTB PBVCs for a patient's MP-RAGEs and thus the BTB differences are usually relatively small.

In addition, Cover et al. (2011) has checked for correlation between BTB differences and the change of position of each ADNI1 subject between the two MP-RAGES in a BTB scan. While there was large change of position in a few cases, somewhat surprisingly, these cases did not lead to large absolute BTB differences in atrophy rates.

In addition to the MP-RAGEs acquired at 1.5 T for all subjects in the ADNI1 study, at least 118 subjects also received BTB MP-RAGEs at 3.0 T during their visits (Jack et al., 2008). This is believed to be the first large multi-center BTB MP-RAGE study acquired at both 1.5 T and 3 T for the same subjects. Thus it provides an invaluable data set for comparing the reproducilibity of MP-RAGEs acquired at 3 T and 1.5 T.

There has been discussion in the literature over which combination of SIENA parameters yields the best performance with the measure of performance depending on the study. A recent extensive accuracy study by Popescu et al. (2012) compared various options of BET, the brain extraction tool used by SIENA, against "gold standard" of manually outlined brains of subjects from the MAGNIMS data set. They found the option combination "-B $f=0.1$" yielded the best results. In the current study a range of SIENA parameter options were compared to see how they affected the BTB differences and how the results were compared to Popescu et al. (2012).

Thus, the goal of this study was to use the ADNI1 BTB MP-RAGEs to compare the reproducibility of the SIENA brain atrophy measurement at 3 T and 1.5 T. In addition to the standard default settings for SIENA, a variety of settings for SIENA's *–f* parameter, which controls the fractional intensity threshold in the brain extraction tool (BET), were also assessed for their affect on reproducibility. The fractional intensity threshold varies from 0 to1 and smaller values give larger brain outline estimates. The Gaussianity of the BTB difference distributions was checked if to see, as was found by Cover et al. (2011), that the shoulder of the BTB distributions violated the Gaussianity assumption.

2. Methods

A table was downloaded from the ADNI web site that contained all the MP-RAGEs that had been acquired as part of the ADNI1 study before February 2010. The table was used to determine the MP-RAGEs to be included in each "subject-visit-pair" for a subject. A complete set for a subject-visit-pair contained 8 MP-RAGEs–2 MP-RAGEs for each BTB pair, a BTB pair each for the first and second visits and 2 field strengths. To maximize the number of subject-visit-pairs and thereby the amount of information to be gained through this study, we used subject-visit-pairs of 0–12 months and 6–24 months. By selecting these specific time points, we ensured that no subject-visit-pairs shared any MP-RAGEs. Only subject-visit-pairs with all 8 MP-RAGEs were included in this study.

These inclusion criteria yielded a total of 118 subject-visit-pairs for 0–12 months, and 82 subject-visit-pairs for 6–24 months, giving a total of 200 subject-visit-pairs. As the BTB reproducibility has been shown to be independent of the subject (Cover et al. 2011) a total of 69 subjects that had subject-visit-pairs at both 0–12 months and 6–24 months included in this study. Thus, the data set had 131 unique subjects (41 HC, 64 MCI and 26 AD). The Spearman rank correlation coefficient (SRCC) of the BTB differences between 12 month and 18 month intervals for 69 subjects was calculated at each of the field strengths to check the independence of 12 month and 18 month subject-visit-pairs.

The BTB MP-RAGEs used in the current study had no post-acquisition processing by the ADNI consortium. Therefore, the pixel values used in the current study were exactly the same as those generated by the MR scanner at the ADNI1 acquisition sites. This is an important point as the ADNI consortium only selected one of each BTB pair of MP-RAGEs for ADNI processing. The particular steps employed in the ADNI processing varied with acquisition site and scanner but had no impact on the analysis performed in the current study because only MP-RAGEs prior to ADNI processing were used.

Another key point of the ADNI protocol in the design of this reproducibility study is the 4 MP-RAGEs acquired over 2 visits for each subject using the identical acquisition sequence. The use of the identical sequence also means any in-scanner processing, such as the type of gradient wrap correction applied, was the same for all 4 MP-RAGEs for each patient. Thus any changes introduced by any in-scanner processing would largely cancel out in the reproducibility calculation.

The PBVC measurements were performed using SIENA from version 4.1.4 of FSL. SIENA was run with a range of options on 200 subject-visit-pairs. These options included one of four flag options for the brain extraction tool (BET) – the default, -B, -S or -R. The -B option implements a correction for spatial signal inhomogeneity and robust brain center estimation for optimized brain extraction. The -S option cleans residual eye and optic nerve voxels which BET can sometimes leave behind. And the -R option attempts to reduce image bias, and residual neck voxels. For convenience, "–" is used in the current publication to represent the default when none of -B, -S or -R are set. In addition the fractional intensity threshold parameter (*f*) was set to one of 11 values ranging from 0.0 to 1.0 in steps of 0.1. SIENA's default *f* value is 0.5.

Combining the four possible flag options with 11 different values of the f parameter yielded a total of 44 different option combinations. These 44 combinations were run at both filed strengths over 200 subject-visit-pairs and each of the two BTB pairs yielding 35,200 SIENA jobs. All the calculations for this study took about 12 core-years of computation time. In other words, it would have taken 12 year of computation time if all the calculations had been run on a single desktop computer with a single core.

2.1. BTB differences

For each field strength the BTB difference was defined for a subject-visit-pair to be the PBVC of the second MP-RAGEs of the BTB pairs ($PBVC_2$) from the subject visits minus the PBVC of the first MP-RAGEs of the BTB pairs ($PBVC_1$). As an equation the definition of a BTB difference is

$$\text{BTB difference} = PBVC_2 - PBVC_1 \quad (1)$$

In rare instances where SIENA failed to yield a PBVC a statistically conservative approach was used to handle the failure. Instead of excluding the value from the statistics, extreme values of PBVC were inserted in place of the failed PBVC. Thus the statistics would better reflect the failure. The actual number of failures was also listed. For the first of the BTB pair the PBVC was set to −100%. If the second of the BTB pair failed to yield a PBVC the PBVC was set to +100%. Thus if SIENA did not yield a PBVC for either of the BTB pairs the BTB was 200%, several times larger than any BTB difference calculated in this study. As non-parametric statistics were used throughout this manuscript, as long as the failures were assigned PBVC were larger than any of the non-failed PBVC values, the actual values did not matter.

Several basic statistics of BTB differences at both 3 T and 1.5 T were calculated. These included the minimum, maximum, median, mean and standard deviation. In addition, the SRCC between 3 T and 1.5 T BTB differences was also calculated (Altman 1991). This non-parameteric measure of correlation was used because it is valid for both Gaussian and non-Gaussian distributions. Cover et al. (2011) demonstrated the non-Gaussian nature of the BTB differences and thus supports the cautious choice to use non-parametric statistics. To estimate the statistical significance of the SRCC, bootstrapping (Mooney & Duval, 1985) was used to simulate a distribution for the SRCC as if no correlation was present. The distribution of BTB differences with no correlation was generated using bootstrapping by randomly ordering the subjects at 1.5 T while keeping the order of the subjects at 3 T constant.

Several measures of the reproducibility of the SIENA brain atrophy measure are included as part of this study. The most commonly used measure of the reproducibility of the SIENA brain atrophy measure is the 50th percentile (also known as the median) of the absolute value of the BTB difference (Smith et al., 2002; 2007). In addition to the 50th percentile, the 90th percentile of the absolute value of the BTB difference distributions was also included as a reproducibility measure to check how close the BTB difference distributions were to Gaussian (Cover et al. 2011). A third reproducibility measure, the standard deviation of the BTB difference (without taking the absolute value), was included for completeness.

As part of this study, 95% confidence intervals for 3 reproducibility measures were also calculated. To avoid any complications that might be introduced by BTB differences having non-Gaussian distributions, the confidence intervals were calculated using bootstrapping (Mooney and Duval, 1985). The bootstrapping for this study was implemented in line with standard practices. For each reproducibility measure and subject-visit-pair the set of the BTB difference used to calculate the reproducibility measure subjects was collected. Then for each primary set 10,000 secondary sets, with the same number of elements as the primary set, were created by selecting with replacement the required number of elements from the primary set. The reproducibility measure was then calculated for each of the 10,000 secondary sets. Finally, 2.5 and 97.5 percentile of distribution of 10,000 reproducibility measures were extracted yielding a 95% confidence interval for each reproducibility measure for each interval grouping.

2.2. ADNI data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as $60 million, five-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research – approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

3. Results

The output of SIENA for a single slice of an MCI patient for the BTB MPRAGEs acquired at 3 T and 1.5 T is shown in Fig. 1. The areas of growth (red) and atrophy (blue) of the brain in four images roughly match. For example, the posterior regions of the ventricles are blue in all the four images indicating that the brain has atrophied – and thus the ventricles have enlarged – over the year. For 3 T MPRAGEs, SIENA yielded PBVCs of −1.82 and −0.50. At 1.5 T it yielded −2.22 and −1.46. The slices within the field strengths matched well, because a patient remained in the scanner for acquisitions at the same field strength. However, the slice alignment between the field strengths is not as good as patients had to change scanners to change field strength.

Table 1 shows basics statistics for the two BTB difference distributions calculated before taking their absolute values. Both had the option combination "-B f=0.5" – the one most commonly used in the literature. Both the median and mean are near zero, as expected.

For 69 subjects that had visit pairs in both 12 and 18 month intervals the SRCC for the BTB differences at "-B f=0.5" between the two time intervals at 3 T was 0.018 and at 1.5 T was 0.048, respectively. The respective p values were 0.86 and 0.67. Thus, as expected, there was no correlation between the BTB differences over the two time intervals. This result justifies the choice of including the two different time intervals in the combined data set and the time intervals are statistically independent.

The SRCCs between 3 T and 1.5 T BTB differences for "-B f=0.5" were 0.094 (0–12 M) and 0.092 (6–24 M). From bootstrapping at 3 T and 1.5 T the p values for the SRCCs were 0.12 and 0.28 indicating no statistically significant correlation of the BTB differences between the two field strengths.

Table 2 shows the statistics for three reproducibility measures of the BTB difference distributions for "-B f=0.5" at both 3 T and 1.5 T for 0–12 M, 6–24 M and combined group. A key assumption for the combined group, confirmed above, is that the BTB differences of 12 and 18 month time intervals are not statistically correlated. An example from Table 2 is the 50th percentile of the BTB difference. The combined subjects at 3 T had 0.40% with a 95% confidence interval of 0.36–0.44% and 1.5 T has 0.31% with a 95% confidence interval of 0.25–0.36%. Since both 95% confidence intervals exclude zero both 3 T and 1.5 T BTB differences are significantly different from zero. Also, since 95% confidence intervals do not overlap, the difference between the BTB differences for the two field strengths is statistically significant. But it should be kept in mind that this poorer performance of 3 T as compared to 1.5 T has only been demonstrated for the default SENIA setting of f=0.5.

Also, from Table 2, the 90th percentile for both groups was about 4.2 times larger at 3 T and 5.7 times larger at 1.5 T than the 50th percentile. Thus the 90th percentile for the BTB difference distributions was about twice as large as would be expected for a Gaussian distribution with the same 50th percentile. This large shoulders of the BTB difference distributions are inline with those found by Cover et al. (2011) at 1.5 T.

For comparison, a simulation was performed for 200 subjects assuming a BTB differences with a Gaussian distribution. After 1,000,000 random simulations, the maximum 90th percentile from the simulation was 3.20 times the 50th percentile. This is smaller than the factors found in the current study. Thus the departure of the distributions of the BTB differences from Gaussian

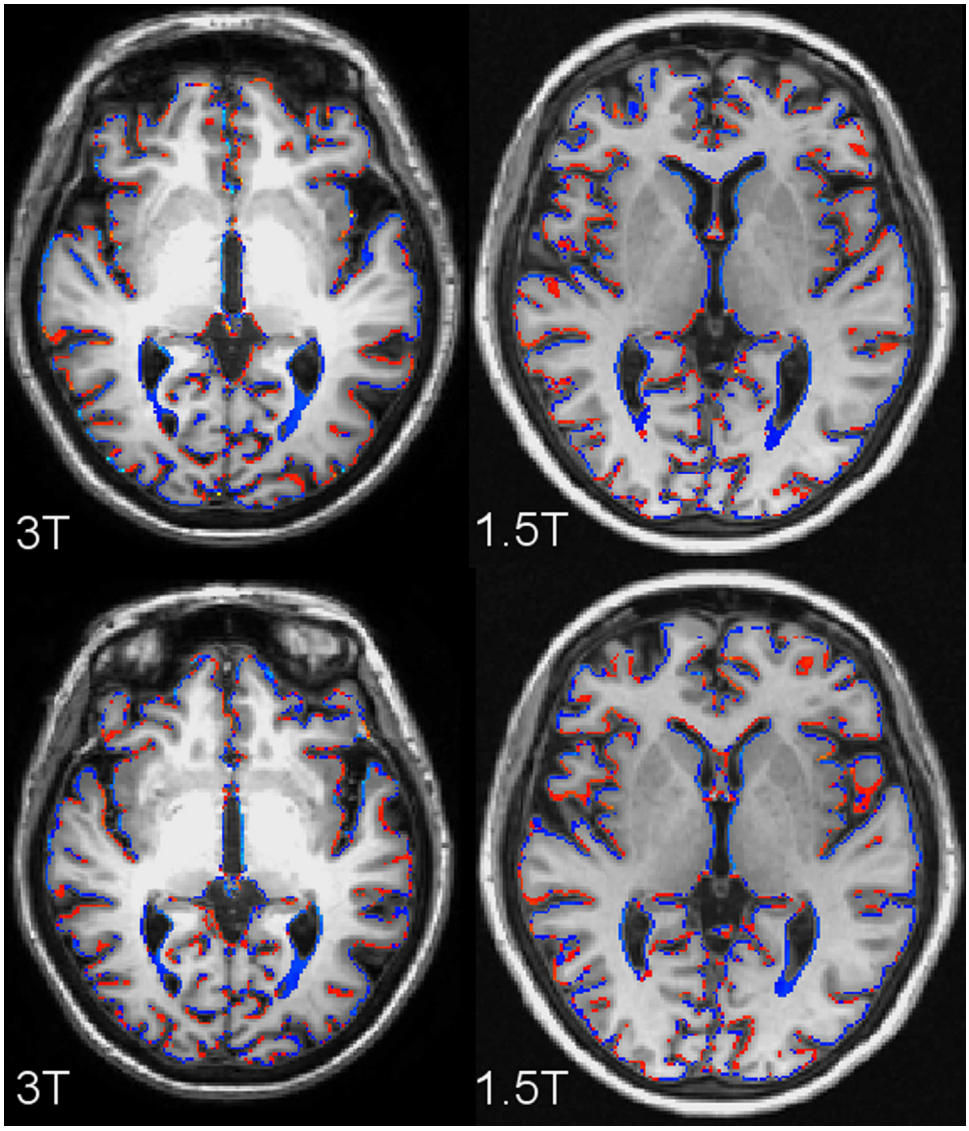

Fig. 1. Examples of changes detected by SIENA for a single subject. Each image shows the changes in the slice between baseline and one year. Red indicates brain volume increase while blue means brain volume decrease ("atrophy"). Each of the four images is a SIENA "halfway" image – an image calculated to be midway between the baseline and one year image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Basic statistics of the BTB difference distributions (before taking the absolute values) for two of the interval groups ("-B f=0.5"). Units are percentage points.

Interval (M)	Strength (T)	Maximum	Minimum	Median	Mean
0–12	3.0	9.59	−3.88	0.04	−0.04
	1.5	3.52	−4.56	0.16	−0.10
6–24	3.0	3.43	−4.81	0.10	0.00
	1.5	1.89	−3.89	0.05	−0.03

is highly statistically significant. Therefore, to ensure reliable results from our analysis in the current paper, only statistical techniques that were valid for both Gaussian and non-Gaussian distributions have been used. The scatter plots shown in Fig. 2 that are discussed below illustrate this point.

Table 3 shows 20 most reproducible – as measured by the 50th percentile – of the SIENA option combinations and field strengths of the 88 evaluated. The 95% confidence interval on the 50th percentile was calculated using bootstrapping. The 90th percentile measure of reproducibility was also included but no confidence interval was included as only 20 subject-visit-pairs were above the 90th percentile.

In Table 3, 1.5 T option combinations are the fourth best with a 3 T combination as the fifth best. The 95% confidence intervals shows overlap between 1.5 T and 3.0 T reproducibilities except when $-f$=0.5. Thus for the cases when $-f$ is not equal to 0.5, there is no significant different between the 3 T and 1.5 T reproducibilities. In addition 11 of the top 20 are for 1.5 T. Thus 3 T is no better than 1.5 T at whole brain atrophy measures - at least for SIENA - and perhaps slightly worse.

In Table 3 for both 3T and 1.5T the most reproducible SIENA option combination was "-B f=0.2". However, the upper bound of the confidence intervals of the 50th percentiles of both field strengths included all the top twenty combinations. The second best option combination was 1.5T with "-B f=0.1" which is the one found to be the best by Popescu et al. 2012 using a different methodology and data set.

In Table 3, the 90th percentile is inline with Cover et al. (2011) in that they indicate distributions with large shoulders as

Table 2
Statistics of the two BTB distributions for the three interval groups at 3 T and 1.5 T ("-B f=0.5"). Below each statistic are 2.5 and 97.5 percentiles for the 95% confidence interval calculated by bootstrapping. The 50th and 90th percentiles are based on the absolute values of the BTB differences while the standard deviation is calculated before taking the absolute value. Units are percentage points.

	50th percentile			90th percentile			Standard deviation		
	Combined	0–12 (M)	6–24 (M)	Combined	0–12 (M)	6–24 (M)	Combined	0–12 (M)	6–24 (M)
N	200	118	82	200	118	82	200	118	82
3.0T	0.40	0.41	0.40	1.69	1.71	1.69	1.23	1.34	1.06
	0. 36–0.44	0.38–0.47	0.33–0.48	1.52–2.01	1.37–2.08	0.95–2.01	0.85–1.55	0.82–1.81	0.80–1.32
1.5T	0.31	0.33	0.29	1.77	1.80	1.77	1.04	1.10	0.97
	0.25–0.36	0.23–0.40	0.21–0.35	1.50–2.05	1.03–2.11	1.22–2.35	0.89–1.19	0.95–1.25	0.78–1.17

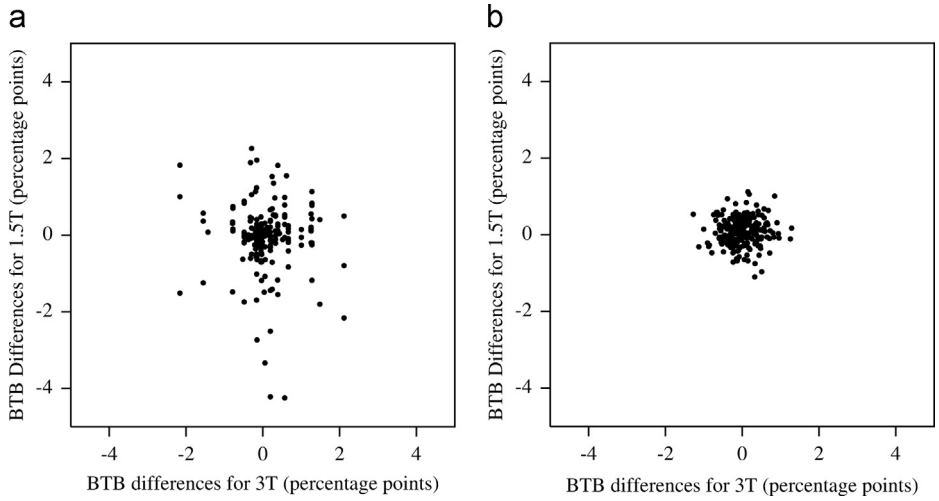

Fig. 2. Scatter plots of 3 T and 1.5 T BTB differences for 200 subject-visit-pairs for the (a) measured and (b) simulated BTB distributions. The simulated results have the same 50th percentiles as that of the measured distributions but assume that the BTB distributions are Gaussian. Comparison of (a) and (b) show how large shoulders of the BTB distributions are as compared to a Gaussian distribution.

Table 3
The best 20 of the most reproducible combination of SIENA options are measured by the 50th percentile. The 95% confidence interval for the 50th percentile is also provided.

Options	Reproducibility 50th percentile	Reproducibility 90th percentile	Number of failures
-B 0.2 1.5T	0.20–0.26–0.32	1.49	2
-B 0.1 1.5T	0.21–0.27–0.31	1.50	2
-R 0.6 1.5T	0.24–0.27–0.37	1.61	2
-B 0.3 1.5T	0.22–0.27–0.37	1.41	2
-B 0.2 3.0T	0.23–0.28–0.39	1.44	2
-R 0.4 1.5T	0.23–0.28–0.35	1.34	0
-R 0.5 1.5T	0.24–0.28–0.33	1.29	0
-R 0.4 3.0T	0.24–0.28–0.36	1.74	2
– 0.4 3.0T	0.23–0.28–0.38	1.45	2
– 0.6 1.5T	0.24–0.28–0.35	1.43	0
-B 0.1 3.0T	0.24–0.29–0.34	1.22	0
-R 0.6 3.0T	0.24–0.30–0.41	3.18	7
-S 0.4 3.0T	0.25–0.30–0.36	1.48	0
– 0.0 1.5T	0.16–0.30–0.68	1.06	0
-R 0.5 3.0T	0.24–0.30–0.37	1.40	0
-S 0.5 3.0T	0.24–0.31–0.43	1.80	2
-B 0.5 1.5T	0.25–0.31–0.38	1.58	0
-S 0.4 1.5T	0.28–0.31–0.38	1.56	0
-R 0.7 1.5T	0.25–0.32–0.39	1.76	2
-R 0.7 3.0T	0.26–0.32–0.39	1.52	2

Table 4
The worst 20 of the most reproducible combination of SIENA options are measured by the 50th percentile.

Options	Reproducibility 50th percentile	Reproducibility 90th percentile	Number of failures
-S 0.8 3.0T	0.66–0.78–0.89	2.61	0
—0.9 3.0T	0.62–0.78–0.95	3.41	0
—0.8 3.0T	0.71–0.80–0.94	3.85	0
-S 0.9 3.0T	0.65–0.83–0.97	2.72	0
—0.8 1.5T	0.70–0.85–1.11	105.58	51
-R 0.9 1.5T	0.72–0.86–1.11	3.90	2
-B 0.7 1.5T	0.73–0.91–1.11	4.41	2
-S 1.0 3.0T	0.86–0.99–1.24	3.76	0
—1.0 3.0T	0.86–1.03–1.24	3.92	0
-B 0.7 3.0T	0.76–1.03–1.26	4.55	0
-B 0.0 3.0T	0.88–1.13–1.37	3.65	0
-B 0.8 1.5T	0.97–1.18–1.48	5.79	3
-R 1.0 3.0T	0.71–1.19–1.91	4.97	0
-B 0.0 1.5T	0.71–1.24–1.91	3.42	2
-B 0.8 3.0T	1.19–1.32–1.58	5.59	0
-B 0.9 1.5T	1.66–1.92–2.22	8.76	2
-R 1.0 1.5T	1.64–1.95–2.22	6.10	2
-B 0.9 3.0T	1.62–1.96–2.32	7.13	0
-B 1.0 3.0T	1.96–2.47–3.34	12.24	2
-B 1.0 1.5T	3.81–4.56–6.33	40.50	23

compared to the Gaussian distribution. The occasional failure of SIENA to yield a PBVC is not unexpected as it is inline with previous experience.

Table 4 shows 20 least reproducible option combinations of SIENA as measured by the 50th percentile. The message from this table is to avoid f values greater than 0.7 or equal to 0.0. This is inline with the best option combination found above "-B f=0.2". Two option combinations gave very high failure rates but these

are of little concern since the option combinations are far from recommends values.

Fig. 2(a) shows the scatter plots for the measured BTB difference distributions at 3 T and 1.5 T for the two grouping intervals for "-B f=0.2". Visual inspection shows little if any correlation between 3 T and 1.5 T BTB distributions consistent with the SRCC results discussed above.

Fig. 2(b) shows the same scatter plot as Fig. 2(a) but using simulated Gaussian distributions with the same 50th percentiles as 1(a) (0.28 for 3 T and 0.26 for 1.5 T). The much more compacted nature of scatter plot 1(b) indicates the significance of the large shoulders of the measured BTB distributions as compared to Gaussian distributions.

4. Discussion

The purpose of the current study was to assess the reproducibility of the SIENA whole brain atrophy measure between 1.5 T and 3 T using the ADNI1 BTB MP-RAGEs. In particular, if one field strength is clearly more reproducible than the other.

4.1. Correlation

Cover et al. (2011) found no correlation between BTB differences and subjects for two different but related brain atrophy algorithms. This lack of correlation motivated the decision to increase the number of data points for analysis in the current study by including subject-visit-pairs for 69 subjects for both 0–12 months and 6–24 months. The results of the current study confirmed a lack of correlation for the BTB differences for 69 subjects at two different time intervals. The results also found no correlation between all the subject-visit-pairs at two field strengths. Visual inspection of Fig. 2(a) illustrates the latter's lack of correlation. This lack of correlation supports the hypothesis that BTB differences are providing us with important information about the brain atrophy algorithms and not the subjects – a valuable property for studying the reproducibility of brain atrophy algorithms.

4.2. Gaussianity

Another property of the BTB differences for SIENA that was first reported by Cover et al. (2011) and confirmed by the current study is that the large shoulders of the BTB distribution deviate far from the Gaussian assumption. The current study provides more evidence that distributions of the BTB differences have large shoulders. For a Gaussian distribution the 90th percentile is 2.44 times larger than the 50th percentile. Tables 2 and 3 have 90th percentiles 4 to 6.5 times larger than their 50th percentile for the SIENA for a range of parameter settings. As demonstrated above these factors demonstrate the BTB differences departure from a Gaussian distribution is highly statistically significant. Examination of the simulated scatter plot in Fig. 2(b) using a Gaussian distribution as compared to the measured BTB differences shown in Fig. 2(a) visually illustrates the size of the shoulders.

The large shoulders of the distributions of BTB differences are important for a couple of reasons. As can be seen from Fig. 2, if two BTB distributions have the same 50th percentile but substantially different 90th percentile the distribution with the substantially smaller 90th percentile will be more sensitive to brain atrophy changes. While the current practice in the literature is only to compare the 50th percentile, the non-Gaussian nature means the 90th percentile should be considered as well.

A second problem is the use of statistical tests that assume a Gaussian distribution such as Student's t-test or a parametric correlation coefficient. If such tests are applied to non-Gaussian distributions the results can be misleading because of their sensitivity to large shoulders in the distributions. Fortunately, this matter is easily dealt with by using non-parametric statistical tests since they do not assume a Gaussian distribution.

4.3. Best parameters

The results shown in Table 3 are inline with the recently published results of Popescu et al. (2012). While the best in Table 3 is "-B f=0.2 for 1.5 T", the second best is "-B f=0.1 for 1.5 T" which is recommended by Popescu et al. (2012). For 1.5 T difference methodology and data set were used. The top 20 also includes "-B f=0.5" for 1.5 T which is commonly used in the literature but is probably not as good as "-B f=0.2" as the 50th percentile is 0.05 percentage points higher and the 90th percentile is slightly higher.

4.4. 3 T or 1.5 T

A key question for the current study is whether MP-RAGEs acquired at 3 T perform better than those acquired at 1.5 T for brain atrophy measures. MRI scanners at 3 T, although more expensive than 1.5 T, are routinely used in research because their signal to noise ratio is often better. Hence many research projects are now conducted only at 3 T.

A quick glance at Table 2 indicates that 3 T is actually significantly worse than 1.5 T judging by the 50th percentile and their non-overlapping 95% confidence intervals for the default value of "-f". However, Table 3, where the f value is varied, does not support this conclusion. While the first four best combinations in Table 3 are 1.5 T, of the top 20 only 11 are 1.5 T. Also the confidence intervals of the 50th percentile, even for the best combination, include 20 most reproducible combinations. Therefore, the statistically significant result in Table 2 is likely a statistical fluke.

A key question is what 3 T acquisition of the MP-RAGEs could have caused similar reproducibility to that of 1.5 T when many would expect 3 T to be superior. There are a few publications that discuss the performance of SIENA or other brain atrophy measures as a function of MRI sequence acquisition parameters. Smith et al. (2002) briefly mentioned a PBVC reproducibility for the 50th percentile of 1.2% between several MRI scanners, including a 3 T, for a single subject. Clark et al. (2006) has some limited discussions on the effect of acquisition protocols on tissue segmentation and noted that the optimal acquisition parameters may vary with tissue type. However there are still many questions left open.

There are several possibilities of why 3 T does not perform substantially better than 1.5 T for brain atrophy measures using SIENA. One possible source of problems with the reproducibility at 3 T is the higher field inhomogeneities that may lead to distortions in the image and contrast. While a prime suspect, there is insufficient evidence to make any definitive statements.

Another possibility is suggested by the examination of the ADNI MP-RAGEs used in the current study. It was found that most, if not all, of the MP-RAGEs for Philips or Siemens MRI scanners did not receive GD correction while most, if not all, of the MP-RAGEs from General Electric scanners did receive 2-dimensional GD correction. It is impossible to make a more definitive statement than "most, if not all" since information on GD correction in the DICOM headers varied with the manufacture and version. Thus it was impossible to separate out the effects of the GD correction from the manufacture of the MRI scanner used. Therefore a role for the GD corrections in the reproducibility of SIENA at 3 T cannot be completely ruled out from the MP-RAGEs used in this study but judged unlikely based on the design of the ADNI study.

Another possible reason why 3 T and 1.5 T BTB reproducibility was similar, in spite of 3 T superior signal to noise ratio, is what the ADNI study was designed to detect – changes related to the differences between 3 T and 1.5 T field strengths. A much less likely but possible compared to 1.5 T are the MRI parameters that were not well controlled in the ADNI study such as the difference in voxel volumes at 3 T versus 1.5 T, manufacture or scanners GD correction algorithm. However, there are no detailed studies in the literature presenting the SIENA reproducibility versus any of these parameters. The closest research is that by Caramanos et al. (2010) that discusses the effects of gradient distortion on SIENA. In addition, for example, the voxel volumes in the ADNI study, 1.5 T ranges from 17% larger to 36% smaller than 3 T volumes for the same subjects. Thus there is no clear relationship between voxel volume and field strength. Therefore no definitive statements can be made regarding the relationship between SIENA reproducibility and other MRI parameters.

The ADNI1 data set was only designed to test differences due to scanner field strength. However, it is possible to design a study that at least conceptually would test many more of them. An ideal study would have 3 T and 1.5 T scanner for each manufacturer side by side. Each subject would then be scanned in each scanner, in random order, at a variety of pixel sizes, again in random order. All MP-RAGEs would be acquired twice for each pixel size, with the second MP-RAGE starting within seconds of the completion of the first to provide BTB images. All the MP-RAGEs would be acquired without a GD correction. This study design would allow the pixel size, field strength and manufacture variables to be tested independently while the others are controlled for. The GD correction could then be applied post-acquisition, to find its effects on reproducibility. However, such a well controlled study would likely be too expensive and time consuming to be practical.

4.5. Comparison to other literature values

As shown in Table 3, the current study yielded the best reproducibility as measured by the 50th percentile of 0.26 (0.32–0.20)% for "-B f=0.2 1.5 T". For "-B f=0.5 1.5 T" the current study also yielded 0.31(0.38–0.26)%. Other BTB values in the literature for SIENA include 0.35% as reported by Cover et al. (2011) over 385 subjects from ADNI1 BTB MP-RAGESs but for "-B f=0.5 1.5 T". This is well within the confidence interval for the same parameter options of the current study.

Smith et al. (2007) provided a 50th percentile of 0.16% for SIENA in a same day scan-rescan study on a 1.5 T scanner while we observed 0.28%. It is difficult to rectify the factor of nearly two discrepancy in these numbers. One difference between Smith et al. and the current study may be different MRI scanners. While further study is required to resolve this issue perhaps a BTB study using only phantoms over a range of MRI scanners would be sufficient.

4.6. ADNI post-processing

As it was not considered in the current study, it would also be interesting to study the effects of ADNI's post-acquisition processing on the reproducibility of SIENA and other brain atrophy measures. Currently the ADNI post-processing is only applied to one of each pair of BTB MP-RAGEs. With some extra effort, it should be possible for ADNI to process both of the BTB MP-RAGEs. Then a reproducibility study similar to the one presented in the current paper could be performed on the MP-RAGEs, but with full post-processing including the correction for GD. Then it would be possible to determine whether the ADNI post-processing improves the reproducibility of SIENA applied to MP-RAGEs.

While the current study only compares 3 T and 1.5 T for one particular version of FSL/SIENA, the similar reproducibilities of the two field strengths suggests from an atrophy point of view that there is little to differentiate the two field strengths in spite of 3 T superior signal to noise ratio. While definitive statements can only be made about superior reproducibility based on the results of reproducibility studies comparing different algorithms and versions of algorithms, whatever factor is causing the similar results between 3 T and 1.5 T may not change with algorithm or version. Thus other algorithms may find no difference between the reproducibilities at 3 T and 1.5 T.

5. Conclusions

The current study has again shown the usefulness of the ADNI1 BTB MP-RAGEs in assessing the reproducilibity of algorithms for measuring brain atrophy. For the SIENA default parameter setting of "f=0.5" the reproducibility at 1.5 T was statistically significantly better than 3 T but only slightly. However, when the "-f" parameter was varied the best reproducibility at both 3 T and 1.5 T was around "f=0.2" but there was no longer any statistically significant difference between 3 T and 1.5 T. The demonstration of the most reproducible choice of option combination for SIENA at both 3 T and 1.5 T is that "-B f=0.2" is inline with what has been previously reported by Popescu et al. (2012) for an accuracy study.

The results of the current study also re-enforce the importance of using statistical methods that do not assume Gaussian distributions for distribution of BTB differences since the large shoulders of the BTB distributions for SIENA depart quite far from the Gaussian assumption as originally demonstrated in Cover et al. (2011).

The current study also extents previous results (Cover et al. 2011) by demonstrating a surprising lack of correlation between the BTB differences of subjects to differing field strengths. This is an important point as it supports the hypothesis (Cover et al. 2011) that the BTB differences are primarily determined by the algorithm performing the whole brain atrophy measure and not the subject.

It should be kept in mind that all these results have only been determined for the FSL/SIENA brain algorithm. However, it would be interesting to assess the reproducibility of a range of brain atrophy algorithms and see if they yield similar results.

Acknowledgments

Thanks to Gabriela Spulber for her input. We would also like to thank the reviewers for their invaluable questions and comments. The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007–2013 under Grant Agreement no. 211714 (neuGRID: A GRID-based e-infrastructure for data archiving/communication and computationally intensive applications in the medical sciences) and No. 283562 (N4U: NeuGRID for you: expansion of neuGRID services and outreach to new user Communities). The DAS3 cluster at the Computer Science Department of the VU University was used for the extensive computation required by this reanalysis. MirageGRID was developed with funding from Netherlands' Virtual Laboratory for e-Science VL-e Project. Additional funding for this project was received from the Image Analysis Center at the VU University Medical Center, Amsterdam. BWvD also received funding from NeuroScience Campus Amsterdam.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the

National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514.

References

Altman, D.G., 1991. Practical Statistics for Medical Research. Chapman and Hall, London.

Barkhof, F., Calabresi, P.A., Miller, D.H., Reingold, S.C., 2009. Imaging outcomes for neuroprotection and repair in multiple sclerosis trials. Nature Reviews Neurology 5, 256–266.

Caramanos, Z., Fonov, V.S., Francis, S.J., Narayanan, S., Pike, G.B., Collins, D.L., Arnold, D.L., 2010. Gradient distortions in MRI: Characterizing and correcting for their effects on SIENA-generated measures of brain volume change. Neuroimage 49, 3498 (-3498).

Clark, K.A., Woods, R.P., Rottenberg, D.A., Toga, A.W., Mazziotta, J.C., 2006. Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. Neuroimage 29, 185–202.

Cover, K.S., van Schijndel, R.A., van Dijk, B.W., Redolfi, A., Knold, D.L., Frisoni, G.B., Barkhof, F., Vrenken, H., 2011. Assessing the reproducibility of the SIENAX and SIENAbrain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. Psychiatry Research: Neuroimaging 193 (3), 182–190.

de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J.J., Westendorp, R.G.J., Bollen, E.L.E.M., de Bruin, P.W., Middelkoop, H.A.M., van Buchem, M.A., van der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimers disease: an MRI study. Brain 131, 3277–3285.

De Stefano, N., Matthews, P.M., Filippi, M., Agosta, F., De Luca, M., Bartolozzi, M.L., Guidi, L., Ghezzi, A., Montanari, E., Cifelli, A., Federico, A., Smith, S.M., 2003. Evidence of early cortical atrophy in MS. Neurology 60, 1157–1162.

Enzinger, C., Fazekas, F., Matthews, P.M., Ropele, S., Schmidt, H., Smith, S., Schmidt, R., 2005. Risk factors for progression of brain atrophy in aging: six-year follow-up of normal subjects. Neurology 64, 1704–1711.

Evans, M.C., Barnes, J., Nielsen, C., Kim, L.G., Clegg, S.L., Blair, M., Leung, K.K., Douiri, A., Boyes, R.G., Ourselin, S., Fox, N.C., 2010. Volume changes in Alzheimer's disease and mild cognitive impairment: cognitive associations. European Radiology 20, 674–682.

Filippi, M., Rovaris, M., Inglese, M., Barkhof, F., De Stefano, N., Smith, S., Comi, G., 2004. Interferon beta-1a for brain tissue loss in patients at presentation with syndromes suggestive of multiple sclerosis: a randomised, double-blind, placebo-controlled trial. Lancet 364, 1489–1496.

Fox, N.C., Freeborough, P.A., 1997. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. Journal of Magnetic Resonance Imaging 7, 1069–1075.

Fox, N.C., Freeborough, P.A., Rossor, M.N., 1996. Visualisation and quantification of rates of atrophy in Alzheimer's disease. Lancet 348, 94–97 (July).

Fox, N.C., Jenkins, R., Leary, S.M., Stevenson, V.L., Losseff, N.A., Crum, W.R., Harvey, R. J., Rossor, M.N., Miller, D.H., Thompson, A.J., 2000. Progressive cerebral atrophy in MS: a serial study using registered, volumetric MRI. Neurology 54, 807–812.

Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Transactions on Medical Imaging 16, 623–629.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27 (4), 685–691.

Keihaninejad, S., Heckemann, R.A., Fagiolo, G., Symms, M.R., Hajnal, J.V., Hammers, A., 2010. A robust method to estimate the intracranial volume across MRI field strengths (1.5 T and 3 T). Neuroimage 50, 1427–1437.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clegg, S., Jack, C.R., Weiner, M.W., Fox, N.C., Ourselin, S., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51, 1345–1359.

Mooney, C.Z., Duval, R.D., 1985. Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage Publications, Newbury Park, CA.

Popescu, V., Battaglini, M., Hoogstrate, W.S., Verfaillie, S.C.J., Sluimer, I.C., van Schijndel, R.A., van Dijk, B.W., Cover, K.S., Knol, D.L., Jenkinson, M., Barkhof, F., de Stefano, N., Vrenken, H., MAGNIMS Study Group, 2012. Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. NeuroImage 61, 1484–1494.

Rudick, R.A., Fisher, E., Lee, J.C., Simon, J., Jacobs, L., 1999. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Neurology 53, 1698–1704.

Sluimer, J.D., van der Flier, W.M., Karas, G.B., van Schijndel, R., Barnes, J., Boyes, R.G., Cover, K.S., Olabarriaga, S.D., Fox, N.C., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Accelerating regional atrophy rates in the progression from normal aging to Alzheimer's disease. European Radiology 19, 2826–2833.

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.

Smith, S.M., Rao, A., De Stefano, N., Jenkinson, M., Schott, J.M., Matthews, P.M., Fox, N.C., 2007. Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: cross-validation of BSI, SIENA and SIENAX. NeuroImage 36, 1200–1206.

Watson, P., Head, K., Pitiot, A., Morris, P., Maughan, R.J., 2010. Effect of exercise and heat-induced hypohydration on brain volume. Medicine and Science in Sports and Exercise 42, 2197–2204.