Alzheimer's & Dementia 10 (2014) 430–438

Robustness of automated hippocampal volumetry across magnetic resonance field strengths and repeat images

Robin Wolz[a,b], Adam J. Schwarz[c], Peng Yu[c], Patricia E. Cole[c], Daniel Rueckert[b], Clifford R. Jack, Jr.,[d], David Raunig[e], Derek Hill[a,*], for The Alzheimer's Disease Neuroimaging Initiative

[a]*IXICO Plc, London, UK*
[b]*Department of Computing, Imperial College London, London, UK*
[c]*Eli Lilly and Company, Indianapolis, IN, USA*
[d]*Department of Radiology, Mayo Clinic and Foundation, Rochester, MN, USA*
[e]*Icon Medical Imaging, Warrington, PA, USA*

Abstract

Background: Low HCV has recently been qualified by the European Medicines Agency as a biomarker for enrichment of clinical trials in predementia stages of Alzheimer's disease. For automated methods to meet the necessary regulatory requirements, it is essential they be standardized and their performance be well characterized.

Methods: The within-image and between-field strength reproducibility of automated hippocampal volumetry using the Learning Embeddings for Atlas Propagation (or LEAP) algorithm was assessed on 153 Alzheimer's Disease Neuroimaging Initiative subjects.

Results: Tests/retests at 1.5 T and 3 T, and a comparison between 1.5 T and 3 T, yielded average unsigned variabilities in HCVs of 1.51%, 1.52%, and 2.68%. A small bias between field strengths (mean signed difference, 1.17%; standard deviation, 3.07%) was observed.

Conclusions: The measured reproducibility characteristics confirm the suitability of using automated magnetic resonance imaging analyses to assess HCVs quantitatively and to represent a fundamental characterization that is critical to meet the regulatory requirements for using hippocampal volumetry in clinical trials and health care.

Keywords: Alzheimer's disease; Hippocampus; Segmentation; Test/retest; Clinical trials; Reproducibility

1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia and one of the major health care issues of the future [1]. The most widely established biomarkers for AD are derived from cerebrospinal fluid, amyloid imaging, or structural magnetic resonance imaging (MRI) [2]. Hippocampal volume (HCV) is an extensively evaluated and established biomarker from structural MRI, correlating with disease stage [2,3], cognitive performance [4–6, 35], postmortem Braak stage [8], and local neuronal density [9]. In the AD pathological cascade [10], hippocampal atrophy accelerates before the transition to clinical dementia [11]. Indeed, low HCV has been shown to be a predictive marker of AD in subjects with mild cognitive impairment (MCI) [12–14]. Although quantitative MRI has been primarily a research technique applied to research studies and clinical trials [15], recent regulatory approval of algorithms for the automated analysis of structural MRI as medical devices [16,17] are enabling more widespread use in clinical practice.

After a submission by a public–private consortium led by the Critical Path Institute's Coalition against Major

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 the ADNI and/or provided data, but did not participate in analysis or writing of this article. A complete listing of ADNI investigators can be found at www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf.

*Corresponding author.
E-mail address: derek.hill@ixico.com

1552-5260/$ - see front matter © 2014 The Alzheimer's Association. All rights reserved.
http://dx.doi.org/10.1016/j.jalz.2013.09.014

Diseases, HCV was qualified by the European Medicines Agency in late 2011 as a biomarker to enrich clinical trials of AD during the predementia phase (European Medicines Agency/document ID EMACHMP/SAWP/809208/2011) [18]. The Critical Path Institute's Coalition against Major Diseases is currently preparing data to support a similar qualification with the U.S. Food and Drug Administration (FDA) in accordance with their guidance (FDA) 2010 [19,20]. The current article addresses one key question around test/retest variability posed by the FDA to support that qualification submission.

Manual hippocampal volumetry is a time-consuming task [21], requiring automated methods to enable routine use [7,22–26]. A fundamental performance characteristic is test/retest reproducibility (i.e., how the outcome measure varies when computed from two repeat images acquired in the absence of plausible biological variability or across different scanners from different vendors and/or at different field strengths). In a clinical trial context, although within-subject repeat images to measure atrophy are performed on the same imager, it remains an open issue whether (and how) single-point measurements of HCV might be combined across field strengths for use as an inclusion criterion.

We assess the repeatability and reproducibility of HCVs calculated using one of the four methods included in the qualification submission to the European Medicines Agency and the FDA—namely, Learning Embeddings for Atlas Propagation (LEAP) [7,27]. This algorithm is assessed in 153 subjects from the Alzheimer's Disease Neuroimaging Study (ADNI)-1 [28] who were scanned at both 1.5 T and 3 T.

We extend previous validation of LEAP against reference segmentations [7]. Specifically, we address the within-subject reproducibility of the extracted HCVs to support the requested analysis for the qualification submission[19,20], extending previous work substantially [29–31] by studying a much larger sample size across the full spectrum of AD (control subjects, and patients with MCI and AD), and assessing reliability both within-imager and across field strengths in a rigorous statistical framework. We characterize test/retest measures from three distinct within-subject data pairings—namely, (*i*) intraexamination (back-to-back) repeat images without subject repositioning at 1.5 T, (*ii*) intraexamination (back-to-back) repeat images without subject repositioning at 3 T, and (*iii*) repeat images acquired at different field strengths with minimal interimage temporal separation (1.5 T and 3 T). The independent evaluation of the different variabilities allows conclusions to be drawn on the interplay of instrument-induced noise, patient-induced noise, and the robustness of the used algorithm toward these different noise sources.

Although this article addresses test/retest performance for one particular HCV algorithm, our approach, based on the publically available ADNI data, can also serve as a template for similar analyses with other HCV algorithms or structural MRI biomarkers.

2. Materials and methods

2.1. Image data

In the ADNI study [33], brain magnetic resonance (MR) images were acquired from approximately 200 cognitively normal older subjects, 400 subjects with amnestic MCI (aMCI), and 200 subjects with early AD. For all subjects, T1-weighted Magnetization prepared rapid acquisition gradient echo (MP-RAGE volumetric MR images were acquired at 1.5 T at baseline and at regular follow-up intervals. For every subject and at every time point, two back-to-back (test/retest) repeat images were acquired during a single imaging session. For a subset of subjects, images were also acquired at 3 T [32]. Both unprocessed images are used in our study and preprocessed in house to allow a comparison of back-to-back images. A more detailed description of ADNI and its study design, as well as image acquisition parameters are included in Appendix A.

Our study aims to analyze all ADNI subjects for which a baseline and month 12 image is available at both field strengths. Inclusion/exclusion criteria as detailed in Appendix B led to 153 included ADNI subjects.

Although the study population allows measures of both baseline and longitudinal reproducibility, this work addresses baseline variability specifically. The 612 unprocessed baseline images for all 153 subjects were downloaded from the ADNI repository [33] and were used in this study. Table 1 summarizes the subject characteristics of the study population.

2.2. Preprocessing

Each unprocessed ADNI image was skull-stripped [46] individually and corrected for intensity inhomogeneity [34].

2.3. Hippocampal volumetry

After preprocessing, all used ADNI images were segmented automatically using LEAP [7]. Details about the hippocampus definition and resulting average volumes used are reported by Wolz and colleagues [7, 35].

We report results obtained using the hemispheric average HCV computed from each image. Findings with left and right HCVs independently were substantially similar as detailed in the Discussion.

2.4. Difference measures

Interimage differences in HCV were evaluated using both signed and unsigned (magnitude) difference measures. For each of these, both absolute (δ, measured in cubic millimeters) and relative (Δ, measured as a percent) differences were computed.

Table 1
Subject metadata

Clinical group	n, female (%)	Age, y; mean ± SD	MMSE score, pt; mean ± SD	*APOE* status (ε2-3/ε2-4/ε3-3/ε3-4/ε4-4), %
AD	28 (60.7)	74.6 ± 8.7	22.8 ± 2.1	3.6/0.0/28.6/35.7/32.1
aMCI	74 (35.1)	74.9 ± 7.8	26.4 ± 3.7	1.4/2.7/45.9/32.4/17.6
Normal	51 (64.7)	75.7 ± 4.9	29.3 ± 0.9	13.7/0.0/54.9/29.4/2.0

Abbreviations: MMSE, Mini-Mental State Examination; *APOE*, apolipoprotein E; SD, standard deviation; AD, Alzheimer's disease; aMCI, amnestic mild cognitive impairment.

NOTE. Number, female fraction, age in years, MMSE scores, and *APOE* status are shown for the AD, aMCI, and normal cohorts. For *APOE* status, the prevalence for appearing combinations of both *APOE* alleles is shown.

For two HCVs, v_1 and v_2, the absolute unsigned and signed differences, respectively, are defined as

$$\delta_{\text{unsigned}} = |v_1 - v_2| \tag{1}$$

$$\delta_{\text{signed}} = v_1 - v_2 \tag{2}$$

Relative differences Δ_{unsigned} and Δ_{signed} were defined as the volumetric difference normalized by the average of both volumes:

$$\Delta_{\text{unsigned}} = 2\left|\frac{v_1 - v_2}{v_1 + v_2}\right| \tag{3}$$

$$\Delta_{\text{signed}} = 2\frac{v_1 - v_2}{v_1 + v_2} \tag{4}$$

The relative unsigned difference is closely related to the coefficient of variance (CV):

$$\text{CV} = \frac{\Delta_{\text{unsigned}}}{\sqrt{2}} = \sqrt{2}\left|\frac{v_1 - v_2}{v_1 + v_2}\right| \tag{5}$$

The distributions of the signed differences provide information on both the variance (spread) and potential bias in the measurements, and were summarized using parametric (Gaussian) statistics. The unsigned differences present a skewed, positive-definite distribution and are commonly used to assess the accuracy of a measurement against a known ground truth value. These distributions were summarized using nonparametric statistics. We also calculated the root mean square of the CV as a robust measure of central tendency in the unsigned measure. Of the two back-to-back images at 1.5 T and 3 T, the one acquired first was used for interfield strength comparison, and the 1.5 T volume was subtracted from the 3 T volume (i.e., v_1 corresponds to 3 T and v_2 corresponds to 1.5 T).

2.5. Intraclass correlation coefficient

The intraclass correlation coefficient (ICC) was evaluated to test the agreement of test/retest volumes and between volumes acquired at different field strengths. A measure of absolute agreement (criterion-references reliability, ICC [2,1] [36]), was used to establish correlations between a set of two measurements. To model a possible field strength difference in the intrafield strength comparison, ICC(3,1) was also computed to measure the degree of consistency between the two sets of volumes [37].

2.6. Establishing significance among different measures

Our analysis was performed independently for different subgroups of the analyzed cohort. In particular, the diagnosis, the scanner model, and the time interval between acquisition of 1.5-T and 3-T images were used to define subgroups of interest. To establish significance, a one-way analysis of variance (ANOVA) [38], with the subgroup variable as a factor, was carried out for the signed measurements. For the unsigned measurements, a nonparametric Kruskal-Wallis test [39] was conducted. In addition, to compare two groups (e.g., healthy subjects and patients with AD), unpaired, two-tailed, unequal-variance t tests and Wilcoxon's rank tests were carried out between the signed and unsigned differences of the subgroups of interest respectively.

3. Results

3.1. Test/retest reliability at 1.5 T and 3 T

Fig. 1 presents Bland-Altman plots (difference vs. average) and scatterplots (image 2 vs. image 1) of the paired HCVs measured from the within-field strength repeat images at 1.5 T and 3 T, respectively. Mean HCVs and differences for the three diagnostic groups (normal, aMCI, and AD) are presented in Table 2 for images acquired at 1.5 T and 3 T. Both Bland-Altman plots show a mean difference close to zero between HCVs obtained from back-to-back images acquired during the same imaging session. In addition, Table 2 shows ICCs for the comparison of test/retest images (ICC[2,1]) at 1.5 T and 3 T, and the coefficient of determination (R^2).

A slightly greater variation can be observed for the 3-T images. The scatterplots show a strong correlation between measurements at both field strengths. Absolute HCV does not influence measured variability substantially.

3.2. Interfield strength variability

In Fig. 2, a Bland-Altman plot together with a scatterplot is presented for the comparison of 1.5-T and 3-T HCV measurements. Although within 2 standard deviations, both plots indicate a slightly greater volume, on average, at 3 T compared with 1.5 T (significant on a two-tailed t test with

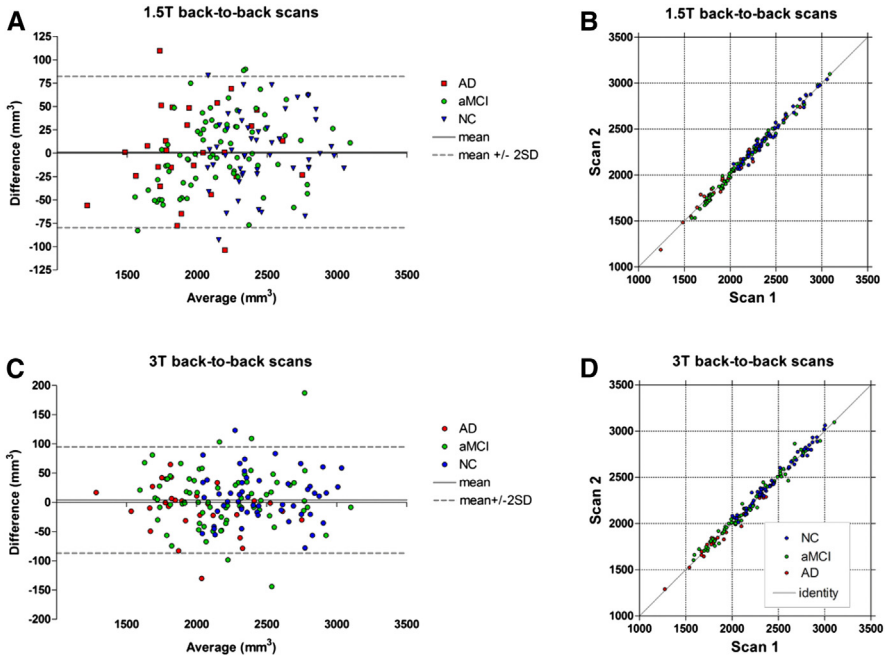

Fig. 1. (A–D) Test/retest variability. The top row shows variability at 1.5 T, the bottom row at 3 T. Bland-Altman plots (A, C) and scatterplots (B, D) of the hippocampal volumes measured in the 1.5-T and 3-T back-to-back images, respectively, are shown. (A, C) The differences are calculated as image 2 less image 1. AD, Alzheimer's disease; aMCI, amnestic mild cognitive impairment; NC, normal control subject; SD, standard deviation.

$P < .001$). Because the 3-T images were acquired systematically after their 1.5-T equivalent, this trend is *against* potential volumetric changes resulting from atrophy. This systematic bias was approximately one-third the standard deviation of the differences. Quantitative results for this comparison are presented in Table 3 [40]. Both signed and unsigned differences are presented for volumetric and percentage differences. ICCs (ICC[2,1] and ICC[3,1]) and the coefficient of determination (R^2) are presented for the interfield strength comparison. No significant difference was observed between ICC values measured on 1.5 T and 3 T when applying Fisher's Z test [41].

3.3. Influence of diagnostic group

The difference in absolute HCV between diagnostic groups was significant on both field strengths (ANOVA, F = 23.3 for 1.5 T, F = 20.5 for 3 T; $P < .001$).

For all presented comparisons, both the unsigned and signed variability measured according to Eqs. 1 to 4 are not significantly different between diagnostic groups on a Kruskal-Wallis test or ANOVA respectively.

3.4. Influence of imager vendor and model

To detect a potential bias of measurement variability to a particular MR imager type, the test/retest analysis for 1.5 T and 3 T was repeated independently for different imager vendors or models. No significant difference in signed or unsigned variability could be observed between images acquired on different imager models. Seventy-six subjects were imaged on imagers from different manufacturers for their 1.5-T and 3-T image acquisition. The measured interfield strength variabilities in this group and in the group of the remaining 77 subjects was not statistically significant ($P = .05$).

3.5. Influence of time interval between the 1.5-T image and the 3-T image

To establish the influence that the time interval between the acquisition of two images can play, the interfield strength variation was measured independently for thresholds of 30 days and 45 days. Results for time intervals of 3-30 days were compared to 31-103 days and results for 3-45 days were compared to 46-103 days. The only group that shows significantly different variation in the signed difference measure is the one with a time interval of 46-103 days. No significance was observed for the difference in unsigned variabilities.

4. Discussion

We presented a detailed analysis to characterize the performance of automated volumetry of brain structures across (*i*) intraexamination (back-to-back) images without subject repositioning and (*ii*) images acquired at different field strengths within a few days or weeks. This dual approach measures the robustness of the extracted volumes with respect to both instrumental and patient noise within a single 30-minute examination (see [*i*]), and to significant changes to instrument (including vendor, field strength, and coil type), patient positioning, and patient changes over days or weeks

Table 2
Test/retest variability

	Combined	AD	aMCI	Normal
1.5 T				
Volume, mm^3	2221.7 ± 363.7	1964.3 ± 340.7	2162.6 ± 340.9	2447.8 ± 270.0
$\delta_{unsigned}$, mm^3	32.2 ± 24.4	36.6 ± 29.6	32.4 ± 22.3	29.6 ± 24.3
(25%/50%/75%)	(13.19/26.25/48.52)	(13.19/29.51/51.82)	(13.71/28.30/48.38)	(12.66/22.50/45.62)
δ_{signed}, mm^3	−2.56 ± 40.4	−3.92 ± 47.4	−4.94 ± 39.1	1.63 ± 38.5
$\Delta_{unsigned}$, %	1.51 ± 1.22	1.92 ± 1.62	1.54 ± 1.12	1.23 ± 1.04
(25%/50%/75%)	(0.57/1.21/2.25)	(0.72/1.56/2.75)	(0.67/1.24/2.30)	(0.54/0.93/1/91)
Δ_{signed}, %	−0.19 ± 1.93	−0.24 ± 2.53	−0.34 ± 1.93	0.06 ± 1.62
RMS, CV	1.70 (0.40/0.86/1.59)	1.97 (0.51/1.10/1.95)	1.79 (0.47/0.88/1.63)	1.40 (0.38/0.66/1.35)
(25%/50%/75%)				
R^2	0.988	0.982	0.987	0.980
ICC(2,1) [CI]	0.994 [0.992–0.996]	0.991 [0.980–0.996]	0.994 [0.990–9.996]	0.990 [0.983–0.994]
3 T				
Volume, mm^3	2245.1 ± 366.5	1992.2 ± 334.9	2192.2 ± 334.9	2461.6 ± 292.6
$\delta_{unsigned}$, mm^3	33.6 ± 30.7	29.7 ± 30.7	35.4 ± 33.6	33.1 ± 26.3
(25%/50%/75%)	(10.52/27.05/48.31)	(6.75/21.3/42.4)	(9.12/29.70/47.84)	(13.80/26.42/49.11)
δ_{signed}, mm^3	−4.55 ± 45.3	−11.08 ± 41.6	−4.59 ± 48.8	−0.90 ± 42.5
$\Delta_{unsigned}$, %	1.52 ± 1.37	1.51 ± 1.53	1.63 ± 1.47	1.36 ± 1.12
(25%/50%/75%)	(0.49/1.14/2.20)	(0.37/1.06/2.38)	(0.45/1.32/2.24)	(0.55/0.89/1.97)
Δ_{signed}, %	−0.21 ± 2.50	−0.53 ± 2.10	−0.19 ± 2.20	−0.05 ± 1.78
RMS, CV	1.45 (0.35/0.81/1.56)	1.51 (0.27/0.75/1.68)	1.55 (0.32/0.94/1.59)	1.24 (0.39/0.63/1.40)
(25%/50%/75%)				
R^2	0.985	0.984	0.980	0.980
ICC (2,1) [CI]	0.992 [0.990–0.994]	0.993 [0.984–0.997]	0.991 [0.985–0.994]	0.989 [0.980–0.993]

Abbreviations: AD, Alzheimer's disease; aMCI, amnestic mild cognitive impairment; RMS, root mean square; CV, coefficient of variance; ICC, intraclass correlation coefficient; CI, confidence interval.

NOTE. The following measures are shown for 1.5 T (top part) and 3 T (bottom part) volumes: volumetric difference ($\delta_{unsigned}$, δ_{signed}) and percentage difference ($\Delta_{unsigned}$, Δ_{signed}) for the combined cohort and the different clinical groups. The 25%, 50%, 75% percentiles are presented in brackets for the unsigned and CV measurements. Furthermore, the RMS of the CV and the coefficient of determination (R^2) are presented. Also, ICC(2,1) is presented together with its 95% CI. The observed volumetric differences and variabilities are not significantly different statistically.

(see [*ii*]). Analyzing HCVs measured from the ADNI data set allowed the reproducibility characteristics and the sensitivity to different variance components to be assessed in a relatively large cohort spanning normal elderly control, aMCI, and AD diagnostic categories. This approach complements current research aimed at standardizing the automated extraction of HCV and validating it as a biomarker [45,21]. In this context, a quantitative measurement of the variability of automated volumetry is essential.

Previous work has been published that measures the variability of automated and manual brain volumetry [29], which measured the variability in HCV when traced manually in repeat images acquired within 2 to 4 weeks. A median CV of 0.28% was observed in 10 subjects. Jovicich and colleagues [30] evaluated, on cohorts of 15 elderly and 5 young subjects, how different effects influence the measurement of brain volumes with FreeSurfer [22]. The reported mean volume difference [42] for HCVs on two test/retest images (same imager, different imaging session) is 3.1%. ICCs of 0.793 and 0.844 were reported for HCVs when using the SRI24 atlas [31] and automated extraction [22], respectively. Volumes extracted from 1.5 T were, on average, 5.5% larger than on 3 T with the SRI atlas and 4% smaller on 1.5 T than 3 T when using FreeSurfer.

With 153 subjects imaged at 33 different ADNI sites, our analysis was applied to a much larger, much more diverse

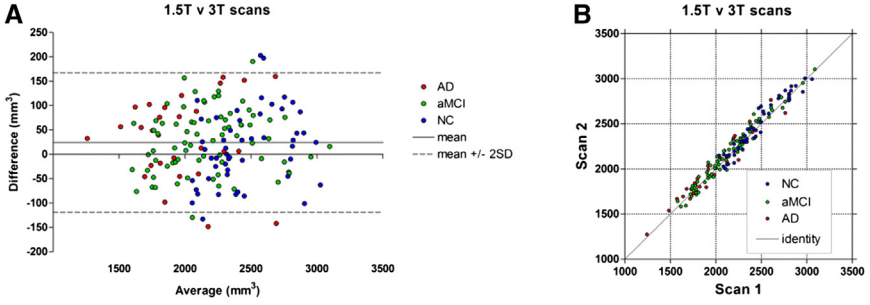

Fig. 2. (A, B) Interfield strength variability. Bland-Altman plot (A) and scatterplot (B) of the hippocampal volumes measured in the 3-T images vs. 1.5-T images. (A) The differences are calculated as 3 T less 1.5 T, for the first image in each imaging session. AD, Alzheimer's disease; aMCI, amnestic mild cognitive impairment; NC, normal control subject; SD, standard deviation.

Table 3
A 1.5-T vs. a 3-T comparison

	Combined	AD	aMCI	Normal
$\delta_{unsigned}$, mm^3 (25%/50%/75%)	59.5 ± 42.9 (22.44/53.78/88.13)	70.7 ± 49.3 (36.59/55.61/99.32)	56.5 ± 42.0 (16.36/55.46/78.85)	57.7 ± 40.2 (23.68/50.63/84.31)
δ_{signed}, mm^3	26.4 ± 68.5	31.5 ± 81.1	29.9 ± 64.0	18.6 ± 68.2
$\Delta_{unsigned}$, % (25%/50%/75%)	2.68 ± 1.89 (1.00/2.46/3.87)	3.51 ± 2.16 (1.88/3.70/5.35)	2.59 ± 1.87 (0.85/2.42/3.87)	2.35 ± 1.65 (0.99/2.11/3.46)
Δ_{signed}, %	1.17 ± 3.07	1.67 ± 3.82	1.32 ± 2.93	0.67 ± 2.81
RMS, CV (25%/50%/75%)	2.32 (0.71/1.74/2.74)	2.90 (1.33/2.61/3.79)	2.26 (0.60/1.72/2.74)	2.02 (0.70/1.50/2.45)
R^2	0.960	0.938	0.958	0.939
ICC(2,1) [CI]	0.979 [0.971–0.986]	0.960 [0.915–0.982]	0.984 [0.974–0.990]	0.961 [0.934–0.978]
ICC(3,1) [CI]	0.978 [0.967–0.984]	0.960 [0.915–0.982]	0.981 [0.966–0.989]	0.960 [0.930–0.977]

Abbreviations: AD, Alzheimer's disease; aMCI, amnestic mild cognitive impairment; RMS, root mean square; CV, coefficient of variance; ICC, intraclass correlation coefficient; CI, confidence interval.

NOTE. Presented are signed and unsigned volumetric differences (δ_{signed}, $\delta_{unsigned}$), and signed and unsigned percentage differences (Δ_{signed}, $\Delta_{unsigned}$). The 25%, 50%, and 75% percentiles are presented in brackets for the unsigned and CV measurements. Furthermore, the RMS of the CV and the coefficient of determination (R^2) are presented. The unsigned difference ($\Delta_{unsigned}$) is significantly different among the three diagnostic groups ($P < .05$).

cohort than previous single-site studies. The performed analysis addressed directly the questions raised by the FDA in the ongoing qualification effort of low HCV as an enrichment biomarker in clinical AD trials [19]. Mean unsigned differences in HCV for 1.5-T and 3-T back-to-back test/retests were 1.51% and 1.52%, respectively. The mean absolute difference across field strengths was 2.68%. No significant difference was observed between variability in the left and right hippocampus. A strong ICC was measured for all three test/retest comparisons, with ICC(2,1) ranging from 0.979 to 0.994. The results presented in our study are substantially less variable than most of the previously published findings described earlier. The one study with smaller reported variability is that of Jack and colleagues [29] for manual volumetry, which reported a median CV of 0.28%, in comparison with median CVs in the current study of 0.86% and 0.81% for the 1.5-T and 3-T back-to-back test/retest, respectively. However, all the subjects investigated in the study by Jack and colleagues [29] were young and healthy, and all imaging was done on the same vendor platform at a single site. A slightly lower median CV (0.66% for 1.5 T, 0.63% for 3 T) was observed in the healthy group in our study. Also, because of their less diverse appearance, the segmentation of younger subjects is likely to be less challenging than that of elderly healthy subjects.

Our results show good agreement in HCVs across back-to-back repeat images but also across MR field strengths and imager manufacturers. The variabilities measured in the 1.5-T vs. the 3-T comparison contain noise resulting from change in instrument, change in field strength, subject positioning, and patient physiological variations. The results suggest that for the LEAP algorithm, these errors are on the order of the errors from instrumental and positioning noise within a single session as measured in both back-to-back experiments.

It is useful to place the variability measured in the current study in the context of differences and changes in HCVs relevant to the study of AD. First, the mean absolute signed measurement variabilities of 32.2 mm^3, 33.6 mm^3, and 59.5 mm^3, in the 1.5-T test/retest, 3-T test/retest, and interfield strength comparisons, respectively, are small compared with the mean difference between HCVs in the AD and normal cohorts—equating to 6.7%, 7.2%, and 12.5%, respectively, of the mean volumetric differences between hippocampi from AD and normal subjects (calculated from the 28 AD and 51 normal subjects in these data). (Variability for the 1.5-T test/retest (6.7%) is measured from volumes on 1.5-T data in our study. Variability for the 3-T test/retest (7.2%) is measured from volumes on 3-T data, and interfield strength variability (12.5%) is measured from averaged volumes between 1.5T and 3T). Second, the observed test/retest variabilities within diagnostic categories (1.92% and 1.54%) are less than the mean annualized change (∼3.85% per year and ∼2.34% per year) in the ADNI AD and aMCI populations, respectively, although that in the normal cohort (1.23%) is greater than the corresponding annualized change in this group (∼0.85% per year) [35]. Third, the mean variability of 32.2 mm^3 at 1.5 T is 9.4% of the overall standard deviation in the distribution of 1.5-T HCVs in the MCI cohort. This is relevant for the use of HCV measures as an inclusion criterion for clinical trials; although the precise implications of test/retest variability for enrichment are beyond the scope of this article, these results suggest that the test/retest variability is likely to be sufficiently tight to allow HCV measures to be used reliably for enrichment purposes.

HCVs measured on 3 T are on average 26.4 mm^3, or 1.17% larger than on 1.5 T on the defined signed difference measures. However, this is substantially less than the standard deviation (3.07%) in the 1.5-T vs. the 3-T comparison and also less than the standard deviations in the measured variability between back-to-back images at each field strength individually (1.22% and 1.37%, respectively). This suggests that for cross-sectional applications in which

only a single image is required (e.g., for enrichment based on a HCV inclusion criterion), combining measures derived from both 1.5-T and 3-T images will result in only a small power loss when LEAP is used. Additional work is also needed to identify whether the field strength bias in volume is related to the difference in spatial resolution in the 1.5-T and 3-T images (see Appendix B [32], for details).

Knowing the variability of an automated algorithm for hippocampal volumetry allows modeling the performance of a biomarker and its power to predict a certain outcome. More detailed analyses of the influence of the measured variability on the definition of, for example, cutoff points for clinical trial enrichment, automated volumetric measurements using the same 1.5-T and 3-T data and framework represent logical extensions to this work. A well-characterized biomarker can then be applied in providing assistance in clinical decision making. Currently, different commercial products are available or tested that provide an automated and fast analysis of a patient's HCV from brain MRI [16,17]. To evaluate the generalizability of our results, preliminary evaluation was also performed on 160 1.5-T test/retest images, from which the hippocampus was segmented using the LEAP algorithm, from the IXI data set [40]. The unsigned variability of $\Delta_{\text{unsigned}} = 1.77 \pm 1.21\%$ is highly consistent with the results reported here. Additional studies are required to assess the test/retest performance of other HCV algorithms. Furthermore, to extend the state of the art beyond the currently "qualified" biomarker of low HCV in a single baseline measurement, the evaluation of longitudinal consistency of automated methods for hippocampal volumetry is another interesting future direction.

One limitation of our study lies in the use of only within-session test/retest images at each field strength. This eliminates any variability resulting from positioning and day-to-day physiological variations. Although these variabilities are included in the interfield strength measurements, they are conflated with imager effects. Analyses on different data sets incorporating such noncontiguous repeat images could help to disentangle these different sources of variability. Our study, furthermore, assumes an exclusion of subjects with reduced HCV for other reasons than AD, such as alcohol use, stress, or head injury. Future work might address how such effects could be accounted for to allow an inclusion of affected subject groups.

5. Conclusions

We addressed the regulatory requirement to characterize carefully the reproducibility of HCV measurements using the LEAP algorithm and a large test/retest cohort of 153 ADNI subjects. This data set allowed the quantification of test/retest performance for 1.5 T and 3 T separately, and a comparison between 1.5 T and 3 T, yielding average unsigned variabilities (absolute difference divided by average) in HCVs of 1.51%, 1.52%, and 2.68%. These equated to absolute average volumetric variabilities of 32.2 mm^3, 33.6 mm^3, and 59.5 mm^3, respectively. ICCs of 0.994, 0.992, and 0.970 were measured for these three comparisons. No bias with the size of the hippocampus or diagnostic category was observed. Only a small bias between field strengths (mean signed difference of 1.17%, compared with a standard deviation across subjects of 3.07%) was observed. Although the evaluation on different data sets precludes a direct comparison, our variability measurements are comparable with what has been reported for manual volumetry and compare favorably with other published results on automated hippocampal volumetry. These reproducibility characteristics and demonstration of robustness to scanner type represent a fundamental characterization that is critical to meet the necessary regulatory requirements for use in clinical trials and health care.

Acknowledgments

This work was funded in part under the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist/). Some of the work was performed as part of the Critical Path Institute Coalition against Major Diseases Biomarker Qualification Project led by Diane Stephenson.

RESEARCH IN CONTEXT

1. Systematic review: We reviewed the literature using traditional (e.g., PubMed) sources and meeting abstracts and presentations. As presented in the article, different publications exist with respect to the variability of automated hippocampal volumetry. Our work fills the gap identified by the Food and Drug Administration with respect to a rigorous analysis of this variability on a large cohort.
2. Interpretation: Our work shows that even on a large data set, current automated algorithms for hippocampal volumetry perform robustly enough to fulfill the purpose that is currently laid out in regulatory processes to use automatically extracted HCV as an enrichment biomarker in clinical trials of Alzheimer's disease.
3. Future work: The article describes a general framework of how to asses the variability of an automated method for the extraction of the volume of brain structures. The same approach can be applied in the qualification process of comparable biomarkers for potentially different purposes.

References

[1] Alzheimer's Disease International. World Alzheimer report 2009. London: Alzheimer's Disease International; 2009.

[2] Jack CR Jr, Petersen RC, Xu Y, O'Brien PC, Smith GE, Ivnik RJ, et al. Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. Neurology 2000;55:484–90.
[3] Fox NC, Cousens S, Scahill R, Harvey RJ, Rossor MN. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects. Arch Neurol 2000;57:339–44.
[4] Van Petten C. Relationship between HCV and memory ability in healthy individuals across the lifespan: review and meta-analysis. Neuropsychologia 2004;42:1394–413.
[5] Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, et al. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 2009;132:1067–77.
[6] Soininen HS, Partanen K, Pitkanen A, Vainio P, Hänninen T, Hallikainen M, et al. Volumetric MRI analysis of the amygdala and the hippocampus in subjects with age-associated memory impairment: correlation to visual and verbal memory. Neurology 1994;44:1660–8.
[7] Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D. LEAP: Learning Embeddings for Atlas Propagation. Neuroimage 2010; 49:1316–25.
[8] Braak H, Braak E. Neuropathological staging of Alzheimer-related changes. Acta Neuropathol 1999;82:239–59.
[9] Bobinski M, de Leon MJ, Wegiel J, DeSanti S, Convit A, Saint Louis LA, et al. The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. Neuroscience 1999;95:721–5.
[10] Jack C Jr, Knopman D, Jagust W, Shaw L, Aisen P, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 2010;27:685–91.
[11] Jack CR, Vemuri P, Wiste HJ, Weigand SD, Aisen PS, Trojanowski JQ, et al. Evidence for ordering of Alzheimer disease biomarkers. Arch Neurol 2011;68:1526–35.
[12] Convit A, De Leon MJ, Tarshish C, De Santi S, Tsui W, Rusinek H, et al. Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. Neurobiol Aging 1997; 18:131–8.
[13] Gosche KM, Mortimer JA, Smith CD, Markesbery WR, Snowdon DA. Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study. Neurology 2002;58:1476–82.
[14] Jack CR Jr, Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, et al. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 1999;52:1397–403.
[15] Hampel H, Frank R, Broich K, Teipel S, Katz J, Hardy RG, et al. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. Nat Rev Drug Discovery 2010;9:560–74.
[16] NeuroQuant. Available at: http://www.cortechs.net/products/neuroquant.php. Accessed January 17, 2014.
[17] Assessa. Available at: http://www.myassessa.com/. Accessed January 17, 2014.
[18] Hill DLG, Schwarz AJ, Isaac M, Pani L, Vamvakas S, Hemmings R, et al. CAMD/EMA biomarker qualification of hippocampal volume for enrichment of clinical trials in pre-dementia stages of Alzheimer's disease. Alzheimers Dement 2013 [in press].
[19] Qualification process for drug development tools: draft guidance. October 2010. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf. Accessed January 17, 2014.
[20] Woodcok J, Woosley R. The FDA Critical Path Initiative and its influence on new drug development. Annu Rev Med 2008;59:1–12.
[21] Frisoni GB, Jack CR. Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. Alzheimer Dement 2011;7:171–4.
[22] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–55.
[23] Carmichael OT, Aizenstein HA, Davis SW, Becker JT, Thompson PT, Meltzer CC, et al. Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. Neuroimage 2005; 27:979–90.
[24] Chupin M, Mukuna-Bantumbakulu AR, Hasboun D, Bardinet E, Baillet S, Kinkingnéhun S. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on control subjects and patients with Alzheimer's disease. Neuroimage 2007;34:996–1019.
[25] Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 2010;51:1345–59.
[26] Lötjönen J, Wolz R, Koikkalainen J, Julkunen V, Thurfjell L, Lundqvist R. Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. Neuroimage 2011; 56:185–96.
[27] Vos SJB, van Rossum IA, Verhey F, Knol DL, Soininen H, Wahlund LO, et al. Prediction of Alzheimer disease in subjects with amnestic and nonamnestic MCI. Neurology 2013;80:1124–32.
[28] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's Disease Neuroimaging Initiative. Neuroimag Clin North Am 2005;15:869–77.
[29] Jack CR Jr, Petersen RC, Xu Y, O'Brien PC, Smith GE, Ivnik RJ, et al. Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. Neurology 1998;51:993–9.
[30] Jovicich J, Czanner S, Han X, Salat D, van der Kouwe A, Quinn B, et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of image sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage 2009;46:177–92.
[31] Pfefferbaum A, Rohlfing T, Rosenbloom MJ, Sullivan EV. Combining Atlas-based parcellation of regional brain data acquired across scanners at 1.5T and 3.0T field strengths. Neuroimage 2012;60:940–51.
[32] Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imag 2008;27:685–91.
[33] Alzheimer's Disease Neuroimaging Initiative. Available at: http://adni.loni.usc.edu/. Accessed on January 17, 2014.
[34] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan E, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imag 2010;29:1310–20.
[35] Wolz R, Heckemann RA, Aljabar P, Hajnal JV, Hammers A, Lötjönen J, et al. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. Neuroimage 2010; 52:109–18.
[36] Rajaratnam N. Reliability formulas for independent decision data when reliability data are matched. Psychometrika 1960;25:261–71.
[37] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–48.
[38] Hogg RV, Ledolter J. Engineering statistics. New York: MacMillan; 1987.
[39] Gibbons JD. Nonparametric statistical inference. New York: Marcel Dekker; 1985.
[40] IXI Dataset. Available at: http://biomedic.doc.ic.ac.uk/brain-development/index.php?n=Main.Datasets. Accessed on January 17, 2014.
[41] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
[42] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986:307–10.
[43] Gunter JL, Bernstein MA, Borowski BJ, Ward CP, Britson PJ, Felmlee JP, et al. Measurement of MRI scanner performance with the ADNI phantom. Med Phys 2009;36:2193–205.
[44] Wyman BT, Harvey DJ, Crawford K, Bernstein MA, Carmichael O, Cole PE, et al. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimer Dement 2013;9:332–7.

[45] Jack CR Jr, Barkhof F, Bernstein MA, Cantillon M, Cole PE, DeCarli C, et al. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheimers Dement 2011;7:474–85.

[46] Leung KK, Barnes J, Modat M, Ridgway GR, Bartlett JW, Fox NC, et al. Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. Neuroimage 2011; 55:1091–108.

Appendix A

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations as a 60 million$, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance (MR) imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and Alzheimer's disease. 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 principle investigator of this initiative is Michael W. Weiner, MD (VA Medical Center and University of California, San Francisco, CA). The 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 more than 50 sites across the United States and Canada. The initial goal of the ADNI was to recruit 800 adults, ages 55 to 90 years, to participate in the research—approximately 200 cognitively normal older individuals to be monitored for 3 years, 400 people with amnestic mild cognitive impairment to be monitored for 3 years, and 200 people with early Alzheimer's disease to be monitored for 2 years. For up-to-date information, see [34]. For all subjects, T1-weighted MP-RAGE volumetric MR images were acquired at 1.5 T at baseline and during regular follow-up intervals. For every subject and at every time point, two back-to-back (test/retest) repeat images were acquired during a single imaging session. For a subset of subjects, images were also acquired at 3 T [25]. The one image per field strength and time point that is identified as "best" by the ADNI MR core is preprocessed with a pipeline using GradWarp, B1 nonuniformity correction and bias field correction using N3 [32].

In the ADNI study, subjects were imaged at more than 50 sites in North America, resulting in a diversity of imaging vendors and models. Images of 1.5 T and 3 T were acquired on 10 and 9 different imaging models, respectively, at 33 different sites. Images at both field strengths were acquired, in general, with a slice thickness of 1.20 mm. The average in-plane resolution was slightly higher and less variable at 3 T, at 1.00 ± 0.01 mm (range, 1–1.02 mm) compared with 1.01 ± 0.13 mm (range, 0.94–1.30 mm) at 1.5 T.

The quality of the 3-T images in ADNI-1 is heterogeneous because some of the imagers used were early 3-T models that did not have array head coils and are therefore unrepresentative of modern 3-T imagers. The analysis was therefore performed independently for individual imagers to assess a potential influence on test/retest accuracy on imagers that have been identified previously to be problematic [43].

The 3-T baseline images were acquired 27 ± 18 days (mean ± standard deviation; minimum, 3 days; maximum, 103days) after the 1.5-T baseline image. We therefore analyzed independently the groups of subjects that had (1) less than or equal to 30 days (n = 109) or (2) less than or equal to 45 days (n = 136) between images at both field strengths. For 17 subjects, the 3-T image was acquired more than 45 days after the 1.5-T image, with a maximum of 103 days.

Appendix B

Our study aims to analyze all Alzheimer's Disease Neuroimaging Initiative (ADNI) subjects for which a baseline and 12-month image is available at both field strengths (1.5 T and 3 T). A total of 161 subjects were imaged at both field strengths and at both time points. For eight subjects, less than two unprocessed MP-RAGE images were available that passed quality control by the ADNI magnetic resonance imaging (MRI) core [43]. Table B1 gives the ADNI subject identification numbers for all included and excluded subjects. The ADNI MRI core has very recently proposed standardized lists of ADNI subjects, and recommended their use wherever possible to improve comparability of analyses across publications based on ADNI data [44]. However, the current list that includes subjects imaged at 1.5 T and 3 T excludes subjects that were imaged on magnetic resonance images that are not considered typical of modern 3-T images (e.g., those with single-channel head coils). Of 118 ADNI subjects that are suggested by the ADNI core to be used in an analysis of 3-T images at month 12, 115 are included in our study population. The three subjects that are not used are part of the eight subjects excluded because their back-to-back images are not available, as mentioned earlier. Subject identification numbers highlighted in Table B1 are also part of the official ADNI 3 T list [44] for month 12 and therefore form the subset that is included in the ADNI list but not in our study. No significant difference in unsigned and signed measurement variability was observed between the subset of our study population that is also part of the ADNI standardized list and the subset of our study population that is not part of this list. No significant difference in performance was observed between different image vendors or models. The Siemens Allegra model, which has been identified previously as problematic [44], did not perform significantly differently than other 3-T images in this study in terms of reproducibility.

Table B1
ADNI subjects included and excluded in our analysis

1.1.1 Included subjects					
002_S_0413	018_S_0450	027_S_0417	082_S_1256	130_S_1337	033_S_1279
002_S_0954	018_S_0633	027_S_0835	094_S_1241	136_S_0086	033_S_1309
002_S_1018	021_S_0332	027_S_1081	094_S_1267	136_S_0194	068_S_1075
002_S_1070	023_S_0030	027_S_1082	094_S_1293	136_S_0195	126_S_0605
002_S_1261	023_S_0031	027_S_1277	100_S_0015	136_S_0196	128_S_1242
002_S_1268	023_S_0058	027_S_1385	100_S_0190	136_S_0300	131_S_0384
005_S_0324	023_S_0061	027_S_1387	100_S_1286	136_S_0426	131_S_0441
005_S_0448	023_S_0078	031_S_0830	116_S_0382	136_S_0429	131_S_0457
005_S_0553	023_S_0139	031_S_1066	116_S_0392	136_S_0579	131_S_1301
005_S_0572	023_S_0331	031_S_1209	116_S_0487	136_S_1227	131_S_1389
005_S_0602	023_S_0376	032_S_0677	116_S_0649	002_S_0729	133_S_0433
005_S_0814	023_S_0388	032_S_1101	116_S_0752	002_S_1280	133_S_0488
007_S_1206	023_S_0604	032_S_1169	116_S_1232	013_S_1035	133_S_0525
007_S_1222	023_S_0625	037_S_0303	116_S_1249	020_S_1288	133_S_0629
012_S_0689	023_S_0855	037_S_0501	126_S_0606	032_S_0187	133_S_0638
012_S_1009	023_S_0916	037_S_1225	127_S_0260	032_S_0479	133_S_0727
012_S_1292	023_S_0926	051_S_1072	127_S_0393	033_S_0514	133_S_0771
012_S_1321	023_S_0963	051_S_1123	127_S_0622	033_S_0724	133_S_0792
016_S_0769	023_S_1046	051_S_1131	127_S_0844	033_S_0725	133_S_0912
016_S_1117	023_S_1126	051_S_1331	128_S_1088	033_S_0733	133_S_0913
016_S_1121	023_S_1190	052_S_1250	128_S_1148	033_S_0920	133_S_1031
016_S_1326	023_S_1247	052_S_1251	130_S_0423	033_S_0922	133_S_1055
018_S_0335	023_S_1262	053_S_0507	130_S_0505	033_S_1016	133_S_1170
018_S_0369	027_S_0307	067_S_0290	130_S_0886	033_S_1086	
018_S_0406	027_S_0403	067_S_0607	130_S_0956	033_S_1098	
018_S_0425	027_S_0404	082_S_0928	130_S_0969	033_S_1116	
Excluded subjects					
013_S_0996	**067_S_1253**	100_S_0035	133_S_0493		
037_S_0588	082_S_1079	**130_S_0449**	**136_S_0184**		

The top part of the table lists the 153 subjects. The bottom part shows Alzheimer's Disease Neuroimaging Initiative subjects excluded because of at least one missing unprocessed image at one time point/field strength combination. Highlighted subjects are part of the official Alzheimer's Disease Neuroimaging Initiative list for 3-T magnetic resonance images acquired at M12 but not part of our data set.