Hindawi Publishing Corporation
International Journal of Alzheimer's Disease
Volume 2012, Article ID 979804, 6 pages
doi:10.1155/2012/979804

Research Article

Test-Retest Reliability of a New Medial Temporal Atrophy Morphological Metric

Simon Duchesne,[1,2] Fernando Valdivia,[2] Abderazzak Mouiha,[2] and Nicolas Robitaille[2]

[1] *Département de Radiologie, Faculté de Médecine, Université Laval, Quebec, QC, Canada G1V 0A6*
[2] *Institut Universitaire en Santé Mentale de Québec, Quebec, QC, Canada G1J 2G3*

Correspondence should be addressed to Simon Duchesne, simon.duchesne@crulrg.ulaval.ca

Received 9 May 2012; Revised 11 July 2012; Accepted 20 August 2012

Academic Editor: Michelle M. Mielke

Clinicians and researchers alike are in need of quantitative and robust measurement tools to assess medial temporal lobe atrophy (MTA) due to Alzheimer's disease (AD). We recently proposed a morphological metric, extracted from T1-weighted magnetic resonance images (MRI), to track and estimate MTA in cohorts of controls, AD, and mild cognitive impairment subjects, at high-risk of progression to dementia. In this paper, we investigated its reliability through analysis of within-session scan/repeat images and scan/rescans from large multicenter studies. In total, we used MRI data from 1051 subjects recruited at over 60 centers. We processed the data identically and calculated our metric for each individual, based on the concept of distance in a high-dimensional space of intensity and shape characteristics. Over 759 subjects, the scan/repeat change in the mean was 1.97% (SD: 21.2%). Over three subjects, the scan/rescan change in the mean was 0.89% (SD: 22.1%). At this level, the minimum trial size required to detect this difference is 68 individuals for both samples. Our scan/repeat and scan/rescan results demonstrate that our MTA assessment metric shows high reliability, a necessary component of validity.

1. Introduction

Early detection of Alzheimer's dementia (AD), critical for treatment success, is a high-priority research area. The development of disease-modifying treatment strategies requires objective characterization techniques and quantitative biomarkers able to identify AD with higher accuracy and at a much earlier stage than clinically based assessment [1]. Given that structural magnetic resonance imaging (MRI) (e.g., T1 weighted) on 1 to 3 Tesla clinical scanners allows the *in vivo* assessment of changes such as medial temporal lobe atrophy (MTA) due to AD, it has been proposed to fulfill the role of quantitative biomarkers in recent reports [2, 3].

We have developed a sophisticated automated image processing method for the purpose of evaluating MTA in the context of AD. We recently proposed a single, high-dimensional morphological metric called the disease evaluation factor (DEF) extracted from T1-weighted MRI and able to track and estimate disease state [4]. In our previous report we provided estimates of this metric's efficiency at the discrimination of cognitively normal, control subjects (CTRL) from probable AD patients, as well as the prediction of conversion in mild cognitive impairment (MCI) subjects to probable AD.

Thorough technique verification, validation, and evaluation are necessary, however, in order for imaging biomarkers such as the DEF to be used in clinical trials enrichment, and more importantly, as a diagnostic aid to community physicians. As an essential component of the verification process, comprehensive metrological investigation of MRI-based metrics must include reliability testing.

Reliability is an important component of the precision of a measurement and relates to the consistency of measurements taken by a single person or instrument on the same item and under the same conditions. A less-than-perfect test-retest reliability causes test-retest variability, reducing confidence in the result and decreasing the test's statistical power. Reliability testing is particularly important for MRI-based metrics, which, while acquired with similar protocols, will show dissimilar intensity contrasts for the same tissue

types [5]. These systematic and random variations are machine dependent and can be corrected for the most part via image denoising [6], bias field inhomogeneity estimation [7], and intensity standardization [8].

In this paper we investigated the reliability of our DEF metric through analysis of cross-sectional (i.e., one timepoint) scan/repeat scan and scan/rescan images from two multicentric studies. First, we took advantage of the fact that subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study received two within-session T1-weighted scans at their baseline visit to test for scan/repeat scan analysis. Further, we employed data on three participants in the Pilot European ADNI that had been scanned at seven different sites in a short timeframe to test for Scan/Rescan reliability. We report minimum clinical trial sample size increases at various different levels based on the calculated detection threshold.

Reliability analysis is an important, necessary, and often overlooked step between bench and bedside in the research and clinical contexts.

2. Materials and Methods

2.1. Ethics. Institutional review boards of all participating institutions approved the procedures for this study. Written informed consent was obtained from all participants or surrogates. More information about ADNI[1] and Pilot European ADNI investigators are provided in the Acknowledgments.

2.2. Subjects. In this study we used data from three different studies, totaling 1051 subjects from over 60 centers.

(i) The first was the *Mapping group*, consisting of 145 young control subjects from the International Consortium for Brain Mapping database [9].

(ii) The second was the *Classification group*, which consisted in 70 probable AD and 69 CTRL subjects from the LENITEM database [10]. We required those first two groups to build our high-dimensional metric;

(iii) The third was the *Scan/Repeat Test Group*, which consisted in 1518 baseline MRIs (scan + same-session repeat scans) from 759 CTRL, MCI, and probable AD subjects participating in ADNI, acquired on more than 50 different 1.5T scanners using a similar 3D T1-weighted MP-RAGE protocol [11]. Inclusion criteria to the ADNI study were as follows.

(a) CTRL are MMSE scores [12] between 24–30 (inclusive), a CDR [13] of 0, nondepressed, non-MCI, and nondemented. The age range of normal subjects was roughly matched to that of MCI and mild AD subjects.

(b) MCI subjects are MMSE scores between 24–30 (inclusive), a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II [14], a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.

(c) Mild AD is MMSE scores between 20–26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS/ADRDA criteria for probable AD [15].

From the complete ADNI dataset of 822 subjects at baseline, we selected individuals for the *Scan/Repeat Test Group* that had both valid entry images and processed images that passed *automated* quality control [16].

(iv) Finally, the fourth was the *Scan/Rescan Test Group*, which was obtained with permission from the multicentric Pilot European ADNI project [17]. It included data from three healthy volunteers acting as human quality control phantoms for the study.

2.3. MRI Acquisitions. Subjects in the *Mapping group* were scanned in Montreal, QC, Canada on a Philips Healthcare Gyroscan 1.5T scanner (Best, The Netherlands) using a T1-weighted fast gradient echo sequence (sagittal acquisition, TR = 18 ms, TE = 10 ms, $1 \times 1 \times 1\,\text{mm}^3$ voxels, flip angle 30°).

Subjects in the *Classification group* were scanned in Brescia, Italy on a single Philips Healthcare Gyroscan 1.0T scanner (Best, The Netherlands) using a T1-weighted fast field echo sequence (sagittal acquisition, TR = 25 ms, TE = 6.9 ms, $1 \times 1 \times 1,3\,\text{mm}^3$ voxels).

Subjects in the *Scan/Repeat Test Group* were scanned on over 50 different 1.5T scanners (GE Medical Systems; Siemens Healthcare; Philips Healthcare) using a 3D T1-weighted MP-RAGE protocol or its equivalent [11]. In this protocol, within the same scan session, there were two 3D T1-weighted images acquired, allowing us to test reliability on this scan/repeat pair. The subject was not taken out of the scanner between acquisitions.

Subjects in the *Scan/Rescan Test Group* were scanned within the span of few weeks at seven different European centers (Sites 1 to 7), using the ADNI study 3D T1-weighted MP-RAGE protocol [11]. Six centers collected scan/rescan sessions, where the subject was taken out of the scanner between acquisitions. This allows us to estimate scan/rescan reliability on 18 comparison pairs.

2.4. Initial Image Processing. We processed all MRI volumes identically using the MINC image processing toolbox (http://www.bic.mni.mcgill.ca/ServicesSoftware/HomePage) and local software as follows: (a) noise removal [6]; (b) raw scanner intensity inhomogeneity correction [7]; (c) global registration (12 degrees of freedom) [18] to the reference image space defined by the BrainWeb T1-weighted image [19] (1-mm resolution, 0% noise, 0% nonuniformity), maximizing the mutual information between the two volumes [20]; (d) resampling to a 1-mm^3 isotropic grid; (e) linear clamping to (0–100) intensity range; (f) intensity standardization [8]; (g) nonlinear registration of individual standardized subject images to the BrainWeb reference;

(h) computation of determinants of the Jacobian of the deformation field [21].

2.5. High-Dimensional Metric. We generated a low-dimensional feature space with the *Mapping group* using Principal Components Analysis of (a) T1w MRI intensity z-score maps, as a proxy of tissue composition and (2) determinant maps, as a proxy of tissue atrophy. After computing components, data from the *Mapping group* were no longer used in the study.

We then projected intensity and determinant data from the *Classification group* into the space defined by the principal components and used a system of supervised linear classifiers with forward stepwise regression (*p-to-enter* 0.05) to identify a restricted set of eigenvectors $\{\lambda_f\}$ forming a hyperplane that best separated the two classes under study (CTRL versus probable AD). After computing the classification function, data from the *Classification group* were no longer used in the study.

Finally, we projected *Test Group* data in the $\{\lambda_f\}$ eigenvector space. The morphological DEF metric is based on the concept of distance within the space defined by eigenvectors $\{\lambda_f\}$ [4]. Specifically, in this embodiment it consists in the calculated Mahalanobis distance (1) for each subject's image between the position p of a subject's image in the *Mapping group* feature space, along the restricted set of principal components, and the centroids of coordinates formed by the CTRL subjects of the *Classification group*.

The Mahalanobis distance between p and a group G is given by

$$\text{mahal}(p, G) = \sqrt{(p - \mu_G) S_G^{-1} (p - \mu_G)}, \quad (1)$$

where μ_G and S_G are respectively the mean and covariance matrix of group G.

2.6. Experimental Design. We first tested reliability in the ADNI *Scan/Repeat Test Group*, that is, between within-session scan/repeat scan pairs, at a single study timepoint (namely, baseline scans). Secondly, we tested reliability in the Pilot European ADNI *Scan/Rescan Test Group*, that is, within-scanner scan/rescan pairs. For each reliability estimate, we calculated the change in the mean, standard deviation, and Pearson retest correlation. Finally, we estimated the impact of the reliability thresholds on the minimum trial size required to discriminate probable AD versus CTRL subjects, using conservative power assumptions, for cross-sectional evaluations.

3. Results

3.1. Scan/Repeat Scan Reliability. Over the 759 subjects of the ADNI dataset, the scan/repeat change in the mean was 1.97% (95% CI: 0.46%–3.48%), with standard deviation 21.2% (*cf.* Figure 1), and Pearson retest correlation $r = 0.9381$.

We ensured there were no statistical differences in reliability between scan/repeat scans in either CTRL or probable AD groups using the diagnostic provided by ADNI (*cf.* Figure 2).

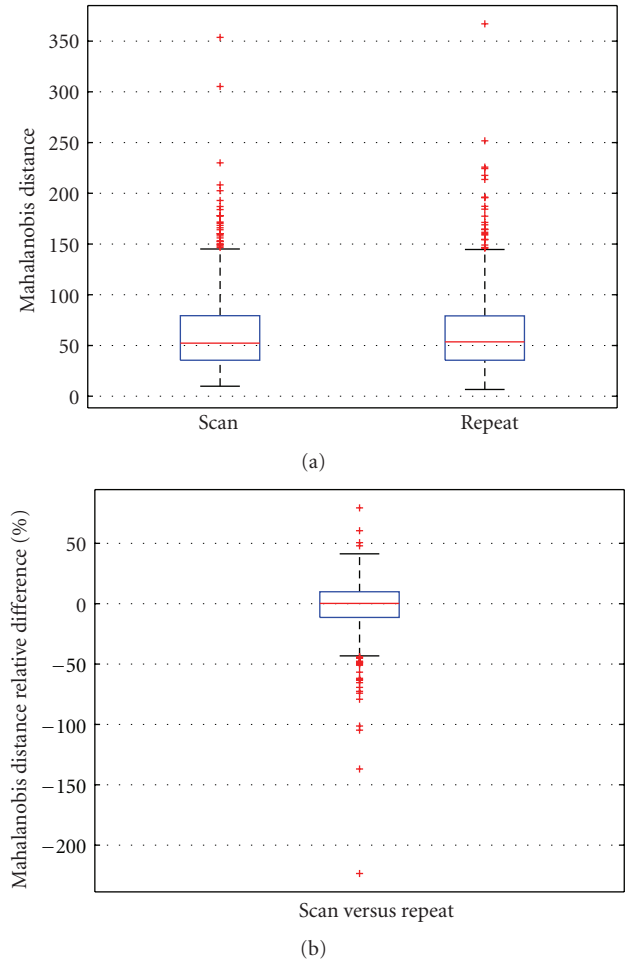

Figure 1: Absolute distances (a) and relative difference in % (b) for the DEF factor between within-session scan and repeat T1-weighted MR scans for 759 baseline ADNI subjects. The change in mean was 1.97%, with 95% confidence intervals 0.46%–3.48%.

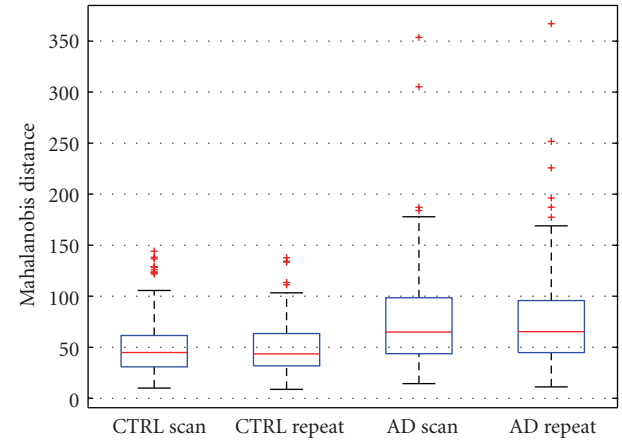

Figure 2: Absolute Mahalanobis distances (DEF factor) between scan/repeat scans for ADNI 203 CTRL subjects (left) and 169 probable AD subjects (right). While the between-group difference was significant, there were no statistical differences in reliability within each diagnostic group.

As reported previously [4], the difference in DEF averages between probable AD and CTRL was 55%. At this level, the minimum trial size required to detect this difference is 62 individuals for both samples ($\alpha = 0.05$; $\beta = 0.50$) (*cf.* Figure 3). Due to the 1.97% minimum precision threshold of the technique, to reach identical power the trial size must increase to 68 individuals.

To evaluate whether the scan/repeat scan distance was smaller than the distance to any one image's nearest neighbor (scan or repeat), we proceeded by calculating all pairwise distances between subjects in the scan/repeat dataset. The comparison shows that the nearest neighbor in nearly all cases was the scan/repeat pair, as opposed to one of the possible neighbor (*cf.* Figure 4).

3.2. Scan/Rescan Reliability. Over the three subjects of the Pilot European ADNI dataset, the scan/rescan change in the mean was 0.89% (95% CI: −14.34%, +12.56%) (*cf.* Figure 5), standard deviation 22.1%, and retest Pearson correlation $r = 0.8609$. Based on similar assumptions, the 0.89% precision threshold of the technique implies an increase in trial size from 62 to 64 individuals.

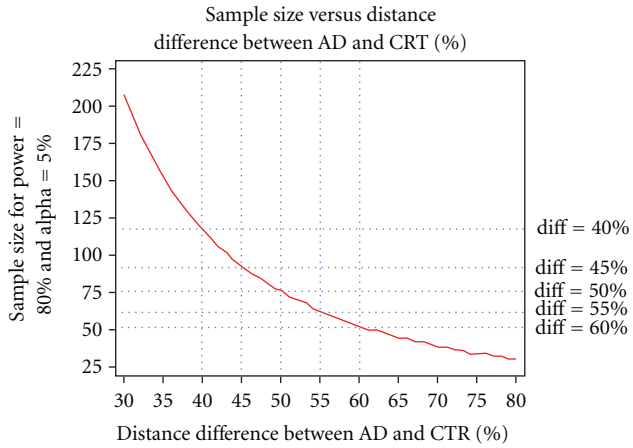

Figure 3: Sample sizes necessary to detect a given DEF difference (in %) between groups at 80% power and alpha level of 0.05.

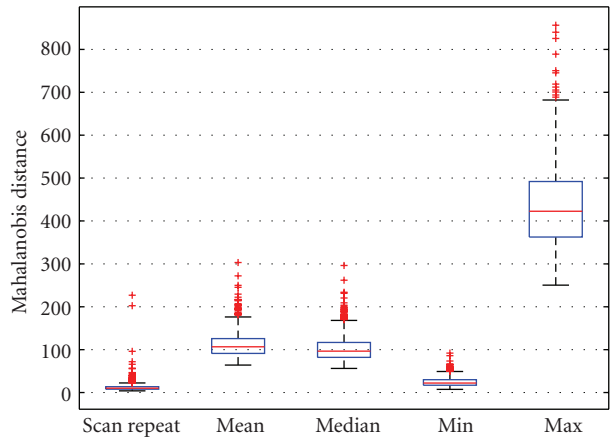

Figure 4: Comparison of scan/repeat scan distance versus all pairwise distances in the *Scan/Repeat Test Group* of 759 ADNI subjects shows that the closest image in the high-dimensional feature space remains its own repeat.

4. Discussion

Imaging biomarkers such as DEF should be thoroughly verified, validated, and evaluated (following ISO9000:2008) before they can be used to enrich populations in clinical trials and aid community physicians to diagnose prodromal AD clinically. *Verification* consists in assessing that the system is built according to its specifications (i.e., assessing that the system is built correctly) and that test data is accurate. *Validation* consists in assessing that the system actually fulfills the purpose for which it was intended (i.e., assessing that the correct system was built). *Evaluation* consists in assessing that the system is accepted by the end-users and performs well for a specific purpose (i.e., assessing that the system is valuable). These are important, necessary, and often overlooked steps between bench and bedside in the research and clinical contexts.

In this study, we proposed a reliability analysis of our high-dimensional morphological metric in a large-scale multicenter setting. Reliability is a *necessary, but not sufficient, component of validity*. Our scan/repeat and scan/rescan results demonstrate that DEF is a reliable metric for medial temporal lobe atrophy estimations.

We further estimated minimum precision threshold that must be added to the effect size to obtain true cohort sizes in the case of clinical trials. While this resulted in increased number of subjects, this increase is somewhat negligible, especially when comparing trial sizes using DEF to those obtained with other metrics, for example, ADAS-Cog [22] or MMSE [12], as mentioned in Schuff et al. [23].

While large datasets represent one of the strengths of the current study, it is not without its limitations. First is the lack of systematic pathological evaluation in both the *Classification group* and the ADNI data. The former implies that the classification function is not optimal for the task of discriminating CTRL from AD; the latter relates to the stability of the DEF. Further, while the mean and confidence intervals are relatively tight, standard deviations tend to be elevated. While it makes the DEF metric suitable for group studies, more work would be required for individual predictions. However, by design, we refrained from using techniques (e.g., within-subject registration, within-subject intensity normalization) that are specifically aimed at removing random and/or systematic errors in individual subject scanning that are not relevant to the pathology. For example, it is expected that within-subject registration would increase spatial concordance, and hence positional variability in the projected intensity and deformation spaces. Such techniques should be considered when continuing our investigations regarding the longitudinal reliability and overall validity of the DEF.

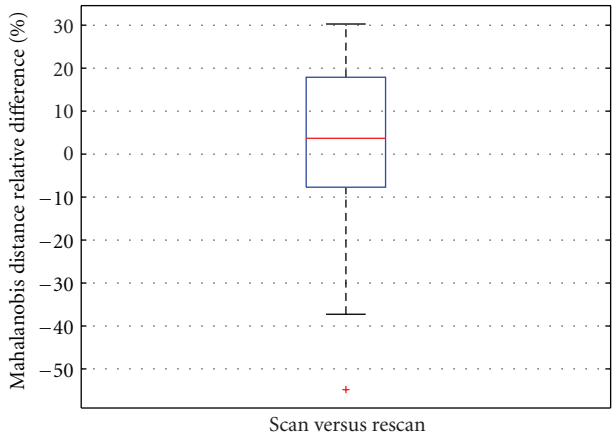

Figure 5: Relative difference in % for the DEF factor between different scan/rescan image pairs for Pilot European ADNI subjects (3 subjects at 6 sites; 18 scan/rescan pairs). The change in mean was 0.89%, with 95% confidence intervals (−14.34%, +12.56%).

Abbreviations

AD: Alzheimer's disease
ADNI: Alzheimer's disease neuroimaging initiative
CTRL: Control subjects
DEF: disease evaluation factor
MCI: mild cognitive impairment
MRI: magnetic resonance imaging
MTA: medial temporal lobe atrophy.

Disclosures

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). 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 paper. A complete listing of ADNI investigators can be found at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Authors' Contribution

All the authors were guarantors of integrity of the entire study. They conducted the study concepts and design. S. Duchesne did the Literature research. Clinical studies and data acquisition included ADNI, Pilot European ADNI, and ICBM. Methods, analysis and interpretation were conducted by S. Duchesne, F. Valdivia, and N. Robitaille. Statistical analysis was conducted by S. Duchesne, A. Mouiha, and N. Robitaille. The first author helped in the preparation of the paper. All the authors were involved in the revision and review of the paper, paper definition of intellectual content, editing, and final version approval.

Acknowledgments

This work was supported by operating grants from the Fonds de Recherche Québec-Santé, the Ministère du Développement Économique, de l'Innovation et de l'Exportation du Québec, and the National Science and Engineering Research Council of Canada. S. Duchesne is a Junior 1 Research Scholar from the Fonds de Recherche Québec-Santé. Data collection and sharing was funded by the Alzheimer's Disease Neuroimaging Initiative (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., and Wyeth, as well as nonprofit 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 (http://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, CA, USA. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles, CA, USA. This research was also supported by NIH Grants P30 AG010129, K01 AG030514, and the Dana Foundation. The Pilot European ADNI study was funded thanks to an unrestricted grant by the Alzheimer's Association. Principal Investigator is Giovanni B. Frisoni (Italy), PIs of the Clinical program Bruno Vellas (France), of the MR Imaging program Frederik Barkof (The Netherlands), and of the Biological Marker program Harald Hampel (Ireland/Germany) and Kaj Blennow (Sweden). The authors finally thank the International Consortium for Brain Mapping for access to data.

Endnotes

1. Data used in the preparation of this paper for the *Scan/Repeat Test Group* were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.ucla.edu). The 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 MRI, positron emission tomography, 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 CA, USA. 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 USA 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. http://www.adni-info.org/.

References

[1] B. Vellas, S. Andrieu, C. Sampaio, and G. Wilcock, "Disease-modifying trials in Alzheimer's disease: a European task force consensus," *Lancet Neurology*, vol. 6, no. 1, pp. 56–62, 2007.

[2] G. M. McKhann, D. S. Knopman, H. Chertkow et al., "The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's and Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[3] B. Dubois, H. H. Feldman, C. Jacova et al., "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria," *Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

[4] S. Duchesne and A. Mouiha, "Morphological factor estimation via high-dimensional reduction: prediction of MCI conversion to probable AD," *International Journal of Alzheimer's Disease*, vol. 2011, Article ID 914085, 8 pages, 2011.

[5] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[6] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 27, no. 4, pp. 425–441, 2008.

[7] J. G. Sied, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.

[8] N. Robitaille, A. Mouiha, B. Crépeault et al., "Tissue-based MRI intensity standardization: application to multicentric datasets," *International Journal of Biomedical Imaging*, vol. 2012, Article ID 347120, 11 pages, 2012.

[9] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster, "A probabilistic atlas of the human brain: theory and rationale for its development," *NeuroImage*, vol. 2, no. 2 I, pp. 89–101, 1995.

[10] S. Galluzzi, C. Testa, M. Boccardi et al., "The Italian Brain Normative Archive of structural MR scans: norms for medial temporal atrophy and white matter lesions," *Aging*, vol. 21, no. 4-5, pp. 266–276, 2009.

[11] C. R. Jack Jr., M. A. Bernstein, N. C. Fox et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[12] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini mental state". A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[13] J. C. Morris, "Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type," *International Psychogeriatrics*, vol. 9, no. 1, supplement, pp. 173–176, 1997.

[14] D. Wechsler, *WMS-R Wechsler Memory Scale—Revised Manual*, The Psychological Corporation, Harcourt Brace Jovanovich, Inc, New York, NY, USA, 1987.

[15] G. McKhann, D. Drachman, and M. Folstein, "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease," *Neurology*, vol. 34, no. 7, pp. 939–944, 1984.

[16] S. Duchesne and F. Valdivia, "Quality control of large-scale MRI processing by means of outlier detection," in *Human Brain Mapping*, Quebec City, QC, Canada, 2011.

[17] G. B. Frisoni, W. J. P. Henneman, M. W. Weiner et al., "The pilot European Alzheimer's Disease Neuroimaging Initiative of the European Alzheimer's Disease Consortium," *Alzheimer's and Dementia*, vol. 4, no. 4, pp. 255–264, 2008.

[18] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space," *Journal of Computer Assisted Tomography*, vol. 18, no. 2, pp. 192–205, 1994.

[19] B. Aubert-Broche, A. C. Evans, and L. Collins, "A new improved version of the realistic digital brain phantom," *NeuroImage*, vol. 32, no. 1, pp. 138–145, 2006.

[20] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[21] M. K. Chung, K. J. Worsley, T. Paus et al., "A unified statistical approach to deformation-based morphometry," *NeuroImage*, vol. 14, no. 3, pp. 595–606, 2001.

[22] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease," *American Journal of Psychiatry*, vol. 141, no. 11, pp. 1356–1364, 1984.

[23] N. Schuff, N. Woerner, L. Boreta et al., "MRI of hippocampal volume loss in early Alzheimers disease in relation to ApoE genotype and biomarkers," *Brain*, vol. 132, no. 4, pp. 1067–1077, 2009.