The impact of automated hippocampal volumetry on diagnostic confidence in patients with suspected Alzheimer's disease: an EADC study

Paolo Bosco[a], Alberto Redolfi[a], Martina Bocchetta[b], Clarissa Ferrari[c], Anna Mega[a], Samantha Galluzzi[a], Mark Austin[d], Andrea Chincarini[e], D. Louis Collins[f,g], Simon Duchesne[g], Bénédicte Maréchal[h,i,l], Alexis Roche[h,i,l], Francesco Sensi[e], Robin Wolz[d], Montserrat Alegret[m], Frederic Assal[n], Mircea Balasa[o], Christine Bastin[p], Anastasia Bougea[q], Derya Durusu Emek-Savaş[r,s], Sebastiaan Engelborghs[t,u], Timo Grimmer[v], Galina Grosu[w], Milica G. Kramberger[x], Brian Lawlor[y], Gorana Mandic Stojmenovic[z], Mihaela Marinescu[aa], Patrizia Mecocci[bb], José Luis Molinuevo[o], Ricardo Morais[cc], Ellis Niemantsverdriet[t], Flavio Nobili[dd], Konstantinos Ntovas[ee], Sarah O'Dwyer[y], George P. Paraskevas[q], Luca Pelini[bb], Agnese Picco[dd,ff], Eric Salmon[p], Isabel Santana[gg], Oscar Sotolongo-Grau[m], Luiza Spiru[hh,ii], Elka Stefanova[z,ll], Katarina Surlan Popovic[mm], Magda Tsolaki[ee], Görsev G. Yener[s,nn,oo], Dina Zekry[pp], Giovanni B. Frisoni[a,ff 1*]

[a]Laboratory of Alzheimer's Neuroimaging and Epidemiology, IRCCS Istituto Centro S. Giovanni di Dio Fatebenefratelli, Brescia, Italy
[b]Dementia Research Centre, Department of Neurodegenerative Disease, UCL Institute of Neurology, Queen Square, London, UK
[c]IRCCS Istituto Centro S. Giovanni di Dio Fatebenefratelli, Brescia, Italy
[d]IXICO Plc, London, UK
[e]INFN, Genova, Italy
[f]McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
[g]True Positive Medical Devices Inc., Quebec City, QC, Canada.
[h]Advanced Clinical Imaging Technology (HC CMEA SUI DI BM PI), Siemens Healthcare AG, Lausanne, Switzerland
[i]Department of Radiology, University Hospital (CHUV), Lausanne, Switzerland.
[l]LTS5, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[m]Alzheimer Research Center and Memory Clinic, Fundació ACE, Institut Català de Neurociències Aplicades, Barcelona, Spain
[n]University Hospitals and University of Geneva, Geneva, Switzerland.
[o]Alzheimer's and other cognitive disorder unit, Hospital Clinic, Barcelona, Spain
[p]GIGA-CRC in vivo imaging and Memory Clinic, University of Liège, Belgium
[q]First Department of Neurology, Eginition Hospital Kapodistrian University, Medical school of Athens, Greece.
[r]Department of Psychology, Dokuz Eylül University, Izmir, Turkey
[s]Department of Neurosciences, Dokuz Eylül University, Izmir, Turkey
[t]Reference Center for Biological Markers of Dementia (BIODEM), University of Antwerp, Antwerp, Belgium
[u]Memory Clinic and Department of Neurology, Hospital Network Antwerp (ZNA) Hoge Beuken and Middelheim, Antwerp, Belgium
[v]Klinikum rechts der Isar, Technische Universität München, Munich, Germany
[w]Radiology and Medical Imagery, Elias University Clinical Hospital, Bucharest, Romania
[x]Centre for Cognitive impairments, Department of Neurology, University Medical Center Ljubljana, Slovenia
[y]Mercer's Institute for Successful Ageing, St. James's Hospital, Dublin , Ireland
[z]Institute of Neurology, CCS, Belgrade, Serbia
[aa]Dpt, Geriatrics-Gerontology and Old Age Psychiatry, Elias University Clinic, Bucharest, Romania
[bb]Istituto di Gerontologia e Geriatria, Università degli Studi di Perugia, Perugia, Italy
[cc]Medical Imaging Department, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal
[dd]Clinical Neurology (DINOGMI), University of Genoa and IRCCS AOU San Martino-IST, Genoa, Italy
[ee]3rd Department of Neurology, Aristotle University of Thessaloniki, Greece
[ff]Memory Clinic and LANVIE - Laboratory of Neuroimaging of Aging, University Hospitals and University of Geneva, Geneva, Switzerland.
[gg]Neurology Department, , Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal
[hh]Carol Davila University of Medicine, Bucharest, Romania
[ii]Ana Aslan Intl Foundation-Memory Clinic, Bucharest, Romania
[ll]Faculty of Medicine, University of Belgrade, Belgrade, Serbia
[mm]Clinical institute of radiology University Medical Center Ljubljana, Slovenia
[nn]Department of Neurology, Dokuz Eylül University, Izmir, Turkey
[oo]Brain Dynamics Multidisciplinary Research Center, Dokuz Eylül University, Izmir, Turkey
[pp]Department of Internal Medicine, Rehabilitation and Geriatrics, University Hospitals and University of Geneva, Geneva, Switzerland.

*Corresponding author Giovanni B. Frisoni
E-mail address gfrisoni@fatebenefratelli.eu
Tel.: +39-030-350136 or +41 22 305 57 60

Abstract

INTRODUCTION: Hippocampal volume is a core biomarker of Alzheimer's Disease (AD). However, its contribution over the standard diagnostic workup is unclear.

METHODS: 356 patients, under clinical evaluation for cognitive impairment, with suspected AD and MMSE≥20, were recruited across 17 European memory clinics. After the traditional diagnostic workup, diagnostic confidence of AD pathology (DCAD) was estimated by the physicians in charge. The latter were provided with the results of automated hippocampal volumetry in standardized format and DCAD was reassessed.

RESULTS: An increment of one interquartile range in hippocampal volume was associated with a mean change of DCAD of -8.0% (95% Credible Interval:[-11.5,-5.0]). Automated hippocampal volumetry showed a statistically significant impact on DCAD beyond the contributions of neuropsychology, FDG-PET/SPECT and CSF markers (-8.5, CrI:[-11.5,-5.6]; -14.1, CrI:[-19.3,-8.8]; -10.6, CrI:[-14.6,-6.1] respectively).

DISCUSSION: There is a measurable effect of hippocampal volume on DCAD even when used on top of the traditional diagnostic workup.

Key Words Alzheimer's disease, hippocampal volume, biomarkers, diagnostic confidence of AD, Medial Temporal Lobe Atrophy.

1. Background

Over the last decade, many steps have been performed to improve and update the diagnostic criteria of Alzheimer's disease (AD) [[1], [2], [3]]. The International Working Group [[4], [5]] criteria stated that positivity of one or more biomarkers of brain amyloidosis and neuronal injury is associated with a high likelihood of AD.

Specifically, the core AD biomarkers are divided into (1) amyloidosis biomarkers (decreased levels of amyloid beta 42 [Aβ42] in the cerebrospinal fluid [CSF] and increased binding of amyloid brain imaging ligands on positron emission tomography [PET]) and (2) neuronal injury biomarkers such as medial temporal atrophy (MTA), hippocampal volume reduction (both assessed on T1-weighted magnetic resonance images [MRIs] [[6], [7]]), increased total tau or phosphotau CSF levels, cortical, temporoparietal, and posterior cingulate cortex hypometabolism on 18F-fluorodeoxyglucose positron emission tomography (FDG-PET), and hypoperfusion on single-photon emission computed tomography (SPECT).

However, the authors of these revised criteria are extremely cautious on recommending the use of these biomarkers in a clinical setting [2]. Indeed, the clinical use of most, if not all, of the biomarkers mentioned previously is affected by the lack of standard operating procedures for their assessment [[8], [9]]. Furthermore, to show that a proposed biomarker combination can significantly enhance the diagnostic accuracy over the pure clinical workup, a proper validation is needed [3]. Some longitudinal studies have been promoted to assess whether an extended range of biomarkers, added to the traditional clinical assessment, can improve the diagnostic accuracy, and their role is still under discussion [10]. Preliminary results suggest that a combination of imaging (in particular, the assessment of hippocampal volume and regional glucose metabolism by FDG-PET) and CSF biomarkers (in particular, Aβ42 levels) can improve prediction of progression from mild cognitive impairment (MCI) to AD dementia, compared to baseline clinical testing [[11], [12], [13], [14], [15], [16]].

Among these core AD biomarkers, hippocampal volume is one of the most established and validated [17], and it is used in research studies to stage the progression of AD neurodegeneration across the entire spectrum of the disease [18]. Moreover, there is a widespread agreement on its clinical significance, even if its validation process in a clinical framework is still ongoing [19]. Recently, an important effort has been made to improve the accuracy and reproducibility of manual hippocampal volume measurements thanks to the Harmonized Protocol for Hippocampal Segmentation project [[18], [20]]. However, manual segmentation is not feasible in routine clinical practice because it is a time-consuming task that requires highly trained operators. For this reason, fully automated hippocampal volumetry using standardized techniques represents a practical alternative to manual methods.

In this study, we aimed to investigate the perception of diagnostic utility of automated hippocampal volumetry among leading European dementia centers. In the context of Albert et al. [2] that proposed "a probabilistic framework for the way in which biomarkers may be used to provide increasing levels of certainty that AD pathology is the cause of an individual's cognitive decline," we investigated the physician's confidence in the patient's cognitive impairment being due to AD pathology (diagnostic confidence of AD [DCAD]), and we evaluated the impact of automated hippocampal volume assessment on this confidence level within a multicentric clinical setting.

In other words, the study aimed to measure how clinicians are influenced in their clinical diagnostic confidence by different clinical information and by the available biomarkers, hippocampal volumetry in particular. Indeed, because the clinical DCAD is what really determines the management of the patients and because the usage of biomarkers is rapidly increasing in the clinical practice, even if conclusive validation studies are still ongoing, a

direct measure that assesses how the recent diagnostic criteria are being considered and incorporated in a clinical context is certainly beneficial.

2. Methods

2.1 *Data collection*

Three hundred and fifty-six consecutive patients, under clinical evaluation for cognitive impairment and suspected AD etiology, were recruited across 17 centers of the European Alzheimer's Disease Consortium (EADC www.eadc.info).

The inclusion criteria in our study were a Mini-Mental State Examination (MMSE) score ≥ 20, the availability of a volumetric MRI brain scan (T1-weighted volumetric brain scan acquired on a 1.5T or a 3T scanner with MPRAGE or IR-SPGR sequences and maximum linear voxel dimension of 1.5 mm [20]) and suspected AD pathology. On the basis of usual local practices, physicians provided an initial estimate of DCAD on a structured scale ranging from 15% to 85%. We chose to limit the diagnostic confidence to this range in order to exclude cases in which the etiological confidence was close to certainty.

The set of information for each recruited subject, describing physicians' evaluations according to the usual local clinical practices, is the following:

- Age and sex (mandatory);
- MMSE score (mandatory);
- Neuropsychological assessment of long term memory, executive functions, language, visual spatial abilities domains (possible values normal or pathological according to local cutpoints);
- Visual rating on T1 weighted MRI scan (MTA score 0-4) [6]

- Visual assessment of FDG-PET or SPECT brain scan (possible values: normal or pathological with temporal/parietal pattern, pathological with frontal/temporal pattern, pathological with mixed pattern);
- Aβ42, Tau or phospho-tau CSF levels (possible values: normal or pathological according to local cutpoints);

Physicians were asked to provide the MRI scan to the lead investigators, which then proceeded with obtaining automated hippocampal volumetry using one of the following six algorithms: ACM-AdaBoost [21], Freesurfer [22], LEAP [23], GDISeg [24], VolMETRIX [25][26] and MorphoBox [27][28]. The specific algorithm was assigned randomly for each subject. This random assignment was designed to ensure balance in sample size across centers and algorithms.

A standardized volumetric report, including a graph showing left and right hippocampal volumes referenced to the algorithm-specific normative distribution used for the validation of the selected algorithm (Figure 1) was provided to the clinicians. These reports were carefully standardized irrespective of the algorithm used, to keep the clinicians blind to the algorithms and avoid biases. Clinicians then provided a final estimate of DCAD (ranging from 0% to 100%) by taking the volumetric measurements into account.

Finally, they reported for each subject whether the hippocampal volumetry had an impact on diagnostic confidence with a 4-level Likert scale (Possible answers: YES, significantly; YES, somewhat; YES, slightly; NO, not at all). Physicians' evaluations were recorded through *ad hoc* web-based questionnaires.

The study has been performed with the informed consent of each participant and obtained ethics approval from the Ethics Committee of IRCCS San Giovanni di Dio-Fatebenefratelli, Brescia (Italy).

2.2 Data processing

Hippocampal segmentations with ACM-AdaBoost, Freesurfer, LEAP and GDISeg were performed on the neuGRID platform (www.neugrid4you.eu) [29][30]. The MorphoBox prototype and VolMETRIX were run on proprietary resources provided by the algorithm developers.

Importantly, no superimposed specific criteria were required for the automated methods, such as segmenting the hippocampus in a common space, but the hippocampal volumetry providers were free to use their specific settings. Indeed, each algorithm uses different methods to normalize the volumetric information across subjects: some methods (e.g. Freesurfer and LEAP), after providing the segmented hippocampal structures in native space, standardize the volumes by using the total intracranial volume, others (e.g. ACM-AdaBoost, GDISeg, VolMETRIX) primarily register the brain scan with affine registration techniques to a common brain template, thus allowing inflation or deflation of the structures with normalization effects on the volumes across subjects during the following segmentation steps. This implies that the final outcome of each algorithm cannot be directly compared to others in terms of volume. Accordingly, the information provided to the clinicians was the volumetric data of a single subject by contrasting it with an algorithm-specific normative distribution (one algorithm out of six had a gender-specific normative distribution) described in terms of percentiles curves (Figure 1). Therefore, what really could impact the DCAD in our setting was not the hippocampal volume value itself, but its relative position with respect to the represented population of healthy controls. For this reason, we chose to derive a new measure, called Normalized Volume Distance (NVD), defined as the distance between the measured hippocampal volume (HV) and the median hippocampal volume at the corresponding age, divided by the Inter Quartile Range at that age.

$$NVD = \frac{HV - \widetilde{HV}_{age}}{IQR_{age}}$$

The IQR is defined as the difference between the 75th percentile and the 25th at each age. IQR is a well-known measure of statistical dispersion and it can be defined in non-normal distributions. Obviously this definition can be problematic with very skewed distributions, but in our case 5 algorithms out of 6 had a symmetric distribution and the remaining had a very slight asymmetry with maximum absolute value of Bowley's coefficient=0.17 (range (-1,1) where 0 means perfect symmetry) [31].

Thus the NVD represents the relative position of a single volume with respect to a normal distribution taking into account the age and the width of the normative distribution. Hereinafter, we will consider the NVD as the main outcome of the automated methods for hippocampal segmentation.

2.3 Statistical analysis

Considering the violation of the postulates for normal distribution of the target outcome DCAD, all analyses regarding this variable as dependent variable were performed accordingly. In particular, the distribution of diagnostic confidence (i.e. DCAD expressed on the scale 0-100) was heteroskedastic and skewed with a range 0-100 (Figure 2), corresponding (by rescaling for 100 and without loss in generalization) to a Beta distribution. Thus, in order to evaluate the impact of both traditional clinical assessment and hippocampal volume assessment on DCAD, a special case of generalized linear models known as Beta regression models [32] was adopted.

Due to the complexity of the performed models (univariate, multiple and for repeated measures beta regression models), Bayesian inference, based on Markov Chain Monte Carlo (MCMC) method, was adopted as model estimation procedure [33]). The goodness of fit of

the models was evaluated by pseudo-R^2 (ranges 0-1) computed as the square of the correlation between the fitted values (drawn from the posterior predictive distribution) and the observed values [34]. Therefor the greater the magnitude of the correlation between the predicted values and the actual values, the greater the R-squared. Coherently with Bayesian inference, the significance of the parameter estimates was evaluated by 95% posterior credibility interval [95% CrI] (interval not including zero detects significant estimate). Convergence of MCMC chains of Bayesian models was assessed by potential scale reduction factor [35].

All the analyses were performed in R, a language and environment for statistical computing (version 3.2.5, R Core Team, 2015). Univariate Beta regression models and Beta regression models for repeated measures were carried out by 'betareg' and 'zoib' R packages.

3. Results

3.1 Dataset description

Among 356 recruited subjects 45% were male; the mean age was 69.0 (SD 11.1) years and mean MMSE score was 24.9 (SD 3.1). The overall frequencies of the considered clinical measures and biomarkers assessed for the patients are reported in Table 1. Almost all patients underwent a neuropsychological assessment including tests for long-term memory, executive functions, language and visuospatial abilities. About one third of patients were evaluated by FDG-PET (28%) or perfusion SPECT (6%) and 53% of the subjects were assessed for MTA on MRI. About a fourth underwent a lumbar puncture with evaluation of CSF-biomarkers, including Aβ42, tau or phospho-tau analyses.

Table 1. Descriptive statistics of the clinical variables and biomarkers used in the assessment of the 356 enrolled patients.

	n (%)		Mean (SD) or Frequency/n (%)
Sociodemographics			
Sex	356 (100)	F	197 (55)
Age	356 (100)	Years	69.0 (11.1)
Cognitive variables			
MMSE	356 (100)	Score	24.9 (3.1)
Long-Term Memory Test	350 (98)	Abnormal	268 (77)
Executive Functions Test	350 (98)	Abnormal	223 (64)
Language Test	350 (98)	Abnormal	119 (34)
Visuospatial Abilities Test	350 (98)	Abnormal	167 (48)
Biomarkers			
FDG-PET visual assessment	101 (28)	Normal Frontal-Temporal Pattern Mixed pattern Temporal-Parietal Pattern	18 (18) 7 (7) 24 (24) 52 (51)
SPECT visual assessment	21 (6)	Normal Frontal-Temporal Pattern Mixed pattern Temporal-Parietal Pattern	1 (5) 5 (24) 5 (24) 10 (47)
MTA Visual Assessment (MTA score [6])	189 (53)	0 1 2 3 4	18 (10) 47 (25) 72 (38) 42 (22) 10 (5)
CSF tau or CSF phospho-tau	93 (26)	Abnormal	65 (70)
CSF Aβ42	94 (26)	Abnormal	55 (58)

3.2 Diagnostic confidence of AD (DCAD) before hippocampal volumetry

Table 2 reports the parameters of the univariate models that describe the DCAD expressed by clinicians before disclosing automated hippocampal volumetry. When considered one by one, the variables included in the clinical assessment have a statistically significant impact on the

diagnostic confidence in the vast majority of the cases. However, if we consider the explained variance, we observe that neuropsychological test scores in executive, language and visuospatial domains provided a poor explanation of DCAD variability (pseudo R^2 less than 0.05 for all). Among neuropsychological tests, MMSE and Long Term Memory domain tests had the biggest effect on variance: an increment of one point of MMSE caused an average decrease of 4% (β = -4.2) of confidence that the reported cognitive impairments were actually due to AD, while an abnormal score in a long-term memory test increased AD diagnostic confidence by 24% (β = 24.2) on average. The variables, which were more explanatory for the DCAD variability were the CSF Aβ42 analysis (pseudo-R^2=0.46, β=29.4) and the visual assessment of FDG-PET brain scans (pseudo-R^2=0.45 and β>30 for all the categories and with higher β for temporal-parietal and mixed patterns). Similarly, MTA scores and CSF tau levels showed a remarkable impact (pseudo-R^2=0.32 and β>33 for MTA scores > 2; pseudo-R^2=0.35 and β=27.6 for CSF tau levels).

Table 2. Univariate Beta regression models for initial diagnostic confidence of AD (DCAD). The variables explaining the DCAD variability better (higher pseudo-R2) are CSF Aβ42 analysis, visual assessment of FDG-PET brain scans, CSF tau analysis and MTA scores.

Predictors	**Explained variance** ***pseudo-R^2***	**Impact on initial DCAD** ***Estimates# [95% CrI]***
Age (+1)	0.05	β = 0.4 [0.2, 0.6]*
Gender (M vs F)	0.005	β = -3.2 [-7.8, 1.5]
MMSE (+1)	0.28	β = -4.2 [-4.1, -3.3]*
NPSY Long-Term Memory (abnormal vs normal)	0.23	β = 24.3 [20.0, 27.6]*
NPSY Executive Functions (abnormal vs normal)	0.05	β = 10.2 [4.9, 14.7]*
NPSY Language (abnormal vs normal)	0.02	β = 6.9 [1.7, 12.4]*
NPSY Visuospatial Abilities (abnormal vs normal)	0.04	β = 8.9 [4.3, 13.2] *
FDG-PET	0.45	
Temporal-Parietal Pattern vs normal		β = 35.9 [29.6, 40.9]*

Frontal-Temporal Pattern vs normal		β = 30.7 [18.8, 39.4]*
Mixed Pattern vs normal		β = 34.9 [28.2, 40.3]*
SPECT	0.28	
Temporal-Parietal Pattern vs normal		β = 36.6 [-3.2, 48.6]
Frontal-Temporal Pattern vs normal		β = 31.6 [-11.7, 48.2]
Mixed Pattern vs normal		β = 41.1 [-0.1, 49.1]
MTA Visual Assessment	0.32	
1 vs 0		β = 24.2 [15.2, 32.8]*
2 vs 0		β = 32.6 [25.6, 38.0]*
3 vs 0		β = 35.8 [29.9, 40.5]*
4 vs 0		β = 33.7 [23.3, 40.9]*
CSF tau (abnormal vs normal)	0.35	β = 27.6 [20.1, 33.7]*
CSF Aβ42 (abnormal vs normal)	0.46	β = 29.4 [23.8, 34.5]*

The β values indicate the mean variation in confidence (range 0-100) for one unit increase of the independent continuous variables or for change of category in categorical predictors. *β significantly differs from 0 with a posterior probability > 95%. Pseudo-R2 estimates the goodness of fit (ranges 0-1)

3.3 Impact of hippocampal volumetry on DCAD

3.3.1 Perceived impact on DCAD

When directly asked about the impact of hippocampal volumetry on the DCAD, in 24.4% of cases (87/356) clinicians reported that the additional information "significantly" changed their initial diagnostic confidence. In 27.0% (96/356) and in 28.4% (101/356) of cases they felt "somewhat" or "slightly" impacted, respectively. In 20.2% (72/356) of cases they didn't change their initial belief "at all".

3.3.2 Measured impact on DCAD

In Figure 3, the DCAD before and after the disclosure of hippocampal volumetric information (NVD) is described in a scatter plot. Points drawn in cooler colors, which denote a negative distance from the median of the normative distribution, are in most of the cases

located below the bisector of the plane, thus showing that lower NVD values generally increased clinicians' DCAD. On the contrary, points drawn in warmer colors are in the majority of the cases above the bisector or in any cases close to it. Therefore, non-atrophic NVD in general diminished or at least did not increase clinicians' DCAD. The impact of NVD looks more evident when the initial confidence of AD was low (≤50%) and weaker when the initial confidence of AD was >50%. In particular, we had 12 cases out of 356 for which the confidence of AD changed from ≥50% to <50% and 35 cases for which the confidence of AD changed from ≤50% to >50%.

3.3.3 Impact evaluation of hippocampal volumetry alongside other clinical variables on DCAD: repeated measures Beta regression models

The evaluation of the impact of automated hippocampal volume information was carried out by modelling the DCAD variable, gathered pre and post hippocampal volume information, in a repeated measures Beta regression model framework. The regression coefficient of hippocampal NVD alone on DCAD is β=-8.0 (95% CrI:[-11.5, -5.0]), i.e. a decrease of 1 Interquartile Range (IQR) in the NVD induced a mean increase of 8.0% in the confidence level of AD in the clinicians' opinion.

The quantification of the impact of the automated hippocampal volume information together with the other clinical variables is reported in Table 3.

Table 3. Repeated measures Beta regression models for diagnostic confidence of AD (DCAD). Hippocampal volumetry, expressed in terms of Normalized Volume Distance (NVD), is perceived as a significant biomarker for AD in combination with neuropsychological assessment (model a), visual assessment of brain FDG-PET/SPECT

(model c), CSF markers (model d) and visual medio-temporal atrophy (MTA) score (model b).

	Variables in the model	n subjects	pseudo-R2	Clinical variable *Predictors*	Impact on initial DCAD *Estimates# [95% CrI]*
a	Neuropsychology Hippocampal volumetry (NVD)	350	0.29	NVD (+1IQR)	β = -8.5 [-11.0, -5.6]*
				MMSE (+1)	β = -2.3 [-3.1, -1.5]*
				NPSY Long-Term Memory (abn vs n)	β = 16.8 [12.2, 21.1]*
				NPSY Executive Functions (abn vs n)	β = -2.8 [-7.4,1.5]
				NPSY Language (abn vs n)	β = 2.5 [-1.7, 6.4]
				NPSY Visuospatial Abilities (abn vs n)	β = 5.4 [1.4, 9.5] *
b	Neuropsychology Hippocampal volumetry (NVD) MTA Visual Assessment	184	0.32	NVD (+1IQR)	β = -5.7 [-9.9,-1.8] *
				MMSE (+1)	β = -1.0 [-2.1, 0.5]
				NPSY Long-Term Memory (abn vs n)	β = 8.5 [1.8, 14.8]*
				NPSY Visuospatial Abilities (abn vs n)	β = 6.9 [1.4, 12.0]*
				MTA Visual Assessment	
				1 vs 0	β = 28.2 [20.3, 34.3]*
				2 vs 0	β = 36.4 [30.1, 41.0]*
				3 vs 0	β = 36.6 [29.7, 41.0]*
				4 vs 0	β = 32.8 [21.2,40.1]*
c	Neuropsychology Hippocampal volumetry (NVD) FDG-PET/SPECT	116	0.50	NVD (+1IQR)	β = -14.1 [-19.3, -8.8]*
				MMSE (+1)	β = -2.0 [-3.4, -0.8]*
				NPSY Long-Term Memory (abn vs n)	β = 13.9 [6.7, 20.7]*
				NPSY Executive Functions (abn vs n)	β = -4.7 [-11.8, 2.9]
				NPSY Language (abn vs n)	β = -3.7 [-9.9, 2.6]
				NPSY Visuospatial Abilities (abn vs n)	β = 2.7 [-3.9, 9.2]
				FDG-PET/SPECT (abn vs n)	β = 27.9 [21.1, 32.9]*
d	Neuropsychology Hippocampal volumetry (NVD) CSF measures	92	0.59	NVD (+1IQR)	β = -10.6 [-14.6, -6.1]*
				MMSE (+1)	β = -1.7 [-2.9, -0.5]*
				NPSY Long-Term Memory (abn vs n)	β = 15.9 [6.8, 23.6]*
				CSF tau (abn vs n)	β = 18.1 [11.5, 24.3]*
				CSF Aβ42 (abn vs n)	β = 21.7 [16.2, 27.0]*
e	Neuropsychology Hippocampal volumetry (NVD) (subjects not evaluated for MTA)	166	0.52	NVD (+1IQR)	β = -12.5 [-15.7, -9.3]*
				MMSE (+1)	β = -3.1 [-3.9, -2.2]*
				NPSY Long-Term Memory (abn vs n)	β = 24.2 [18.3, 29.2]*
				NPSY Executive Functions (abn vs n)	β = -8.2 [-13.8, -2.2]*
				NPSY Language (abn vs n)	β = -0.1 [-4.5, 5.0]
				NPSY Visuospatial Abilities (abn vs n)	β = 0.5 [-4.7, 5.6]
f	Neuropsychology Hippocampal volumetry (NVD) (subjects evaluated for MTA)	184	0.22	NVD (+1IQR)	β = -6.0 [-9.9, -1.6]*
				MMSE (+1)	β = -2.22 [-3.3, -1.2]*
				NPSY Long-Term Memory (abn vs n)	β = 14.3 [7.1, 20.2]*
				NPSY Executive Functions (abn vs n)	β = -0.7 [-6.4, 5.0]
				NPSY Language (abn vs n)	β = 1.9 [-4.4, 9.3]
				NPSY Visuospatial Abilities (abn vs n)	β = 9.1 [3.0, 15.3]*

The β values indicate the mean variation in confidence (range 0-100) for one unit increase of the independent continuous variables or for change in categories in categorical predictors. The sample size can vary in the construction of the different models depending on the availability of the included predictors (e.g. the model which includes neuropsychological assessment and CSF measurements can rely on 92 subjects that have both information available). *β significantly differs from 0 with a posterior probability > 95%. Pseudo-R2 estimates the goodness of fit (ranges 0-1)

The impact of NVD on DCAD remained significant, also if added to other clinical variables. The best models (in terms of explained DCAD variability) were those including CSF biomarkers (model d), pseudo-R^2=0.59) and visual assessment of FDG/SPECT (model c), pseudo-R^2=0.50) and the mean confidence variations, due to an increase of 1 IQR of NVD measured on the respective age-matched normative population, are β=-10.6% (95% CrI:[-14.6, -6.1]) and β=-14.1% (95% CrI:[-19.3, -8.8]), respectively. Likewise, the impact was remarkable when the hippocampal volumetric information was added to neuropsychological assessment only (β=-8.5%, 95% CrI:[-11.0, -5.6]), although the model fit decreases to pseudo-R^2=0.29 (model a).

The impact was weaker when automated hippocampal volumetry was added as predictor to MTA scores (model b): β=-5.7%, 95% CrI:[-9.9,-1.8] and pseudo-R^2=0.32). More precisely, the impact of NVD on DCAD was significantly different when hippocampal volumetry information was added to neuropsychological assessment of subjects which were already visually assessed for MTA (model f) with respect to subjects without MTA assessment (model e). Indeed, the pseudo-R^2 increases from 0.22 to 0.52 when the NVD was available as the sole atrophy measure of the medial temporal lobe. The correspondent β increases from -6.0% to -12.5%.

Among neuropsychological variables, MMSE and Long Term Memory domain remained the variables that affected diagnostic confidence of AD the most, while language domain did not show any significant impact on it. Differently, visuospatial and executive function domains showed a less robust influence on clinicians' confidence, depending on the combination of variables included in the model and on the underlying sample: regression coefficients for visual domain were significant in models a, b, f and not significant in c and e.

4. Discussion

Results show that physicians perceive automated hippocampal volumetry, in combination with neuropsychological assessment, visual assessment of brain FDG-PET/SPECT, and CSF biomarkers as a valuable biomarker for AD.

To assess the contribution of automated hippocampal volume, we started by evaluating which were the usual local practices in EADC centers for cognitive impairment evaluation in suspected AD patients. In a recent work, Bocchetta et al. [37] investigated through a survey the use of AD biomarkers in the EADC centers and assessed their perceived usefulness for the etiologic diagnosis of MCI. In this work, we performed instead a direct measure of the frequency of use and the perceived usefulness in terms of diagnostic confidence. As expected, we found that almost all patients underwent a neuropsychological assessment inclusive of evaluation on different domains. With respect to the 16% reported in [37], in our sample, 28% (FDG-PET) plus 6% (SPECT) of subjects were evaluated for the detection of hypometabolism areas, and 53% of the subjects were rated for atrophy of the medial temporal lobe (75% were estimated by the survey in [37]). The frequency of usage of CSF markers instead (26%) is very similar to that reported in [37] (22%).

In Bocchetta et al., the 45% of participants of the survey perceived MTA scores as "moderately" contributing to DCAD. Moreover, 79% of the responders felt "very/extremely" comfortable delivering a diagnosis of MCI due to AD when both amyloid and neuronal injury biomarkers were abnormal, results that are in line with the criteria developed by the International Working Group. In our work, we measured how those beliefs corresponded to the real variability of DCAD with respect to the clinical variables and available biomarkers. In particular, the best models explaining the DCAD were those where CSF markers (both amyloid and tau), FDG-PET/SPECT, and MTA scores were included. This confirms that, when both amyloid and neuronal injury biomarkers were abnormal, the clinicians were very confident that cognitive impairment was due to AD.

In this framework, we assessed the added contribution of automated hippocampal volumetry on DCAD. The results show that hippocampal volumetry in combination with neuropsychological assessment, visual assessment of FDG/SPECT, CSF biomarkers, and MTA scores had a statistically significant impact. The results were confirmed by the perception of clinicians who declared to be influenced by hippocampal volume with an impact rated from "slight" to "significant" in 80% of the cases. However, our models suggest that clinicians considered hippocampal volume information in a similar way as MTA score. Indeed, when MTA score was not available, clinicians were more confident on AD diagnosis when the hippocampal volumetric report showed a low volume in comparison to the normative population. On the contrary, when MTA score was available, clinicians rarely modified their DCAD even when the automated hippocampal volumetry was added. Moreover, when considered together (model b), the MTA score appeared to have more impact on DCAD than hippocampal NVD. In addition, the models including MTA showed a very similar model fit (pseudo-R2 = 0.32) both before and after disclosure of NVD. Given the fact that automated hippocampal volumetry is by definition a simple, quantitative,

reproducible measure that does not require a specific training, and because the estimated impact on DCAD is similar to that provided by MTA score, it can provide a significant added value in the diagnostic process of AD, especially in centers which do not include neuronal injury biomarkers in their clinical routine. This also points to the potential need for volumetry measurements on other structures, such as cortices and ventricles.

This study has some limitations. First, our sample represents a fraction (17/66) of the EADC centers. Therefore, even if the participating centers are well distributed across Europe, we cannot generalize our findings as representative of the real-life clinical activities across Europe in the assessment of cognitive impairment and suspected AD. Moreover, the fact itself to be part of the EADC entails that the participating centers are part of a selected group more likely to use biomarkers for diagnosis.

Second, except for the MMSE score, our study aggregated outcomes coming from neuropsychological tests that were acquired and dichotomized in normal/abnormal according to local clinical practices and protocols. The same caveat is valid for CSF biomarker. Even if it seemed reasonable to aggregate such information in the scope of this work, we cannot rule out that different tests or protocols may have different impact on DCAD.

In our study, we included only one biomarker of amyloidosis (CSF Aβ42 levels), and we did not consider amyloid-PET imaging. We chose not to include amyloid imaging because this examination was not available in many of our participating centers, and even if available, it was prescribed for a specific subtype of patients [38]. Once the use of amyloid imaging will be more widespread, future similar works will need to integrate amyloid imaging, as already done by Grundman et al. [39] that reported a significant alteration in physicians' diagnostic thinking due to amyloid imaging results.

In this work, we used six different methods for hippocampal segmentation (randomly assigned and balanced on the involved centers and patients), but we did not separate the results on the diagnostic confidence by algorithm. Our aim was to verify whether the automated volumetry in general had an impact on DCAD and how the information on the hippocampal volume was treated by clinicians during the diagnostic process. For this reason, the design of the study prescribed a standard and uniform template to present hippocampal volumetry to clinicians, even if extracted by different methods. However, in principle, we cannot exclude that different methods for hippocampus segmentation, once known by the clinicians in terms of accuracy, could have a different impact on the diagnostic process. Comparisons between outcomes provided by different algorithms were out of the scope of this work and will be the subject of a different study.

5. Acknowledgements

The study has been funded in part by the project “Development of operational research diagnostic criteria for diagnosis of Alzheimer's disease in the preclinical/predementia phase and implementation of SOPs for imaging and CSF biomarkers in Memory Clinics. An integrated care pathway for early diagnosis and best management in the National Health Service of five Italian regions,” Italian Ministry of Health, grant code NET-2011-02346784.

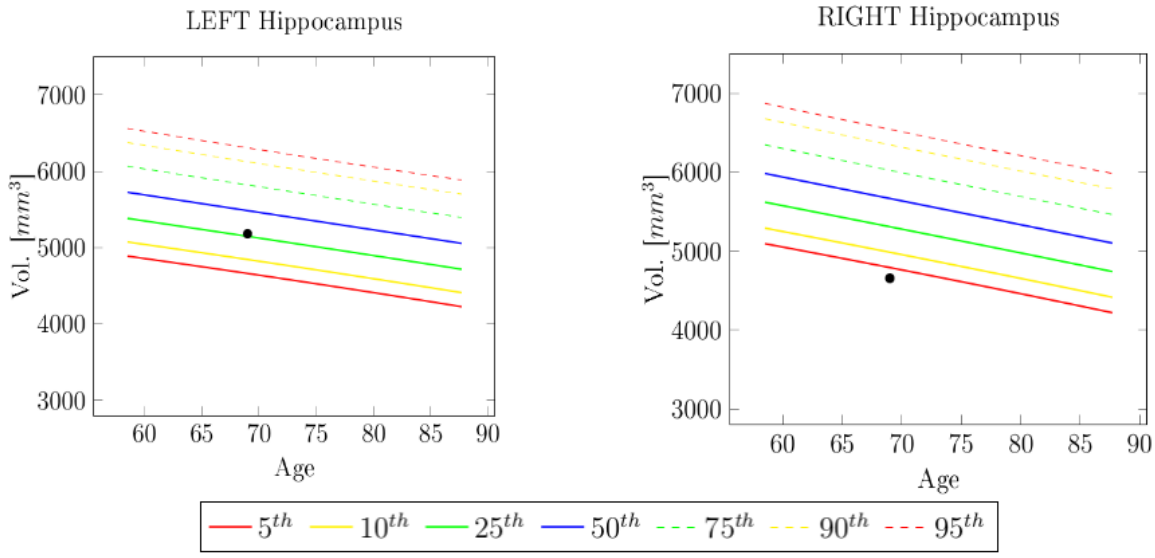

Figure 1: Example of hippocampal volumetric report provided to the clinicians: the left and right hippocampal volumes of the subject are shown compared to algorithm-specific normative populations (for more details see methods)

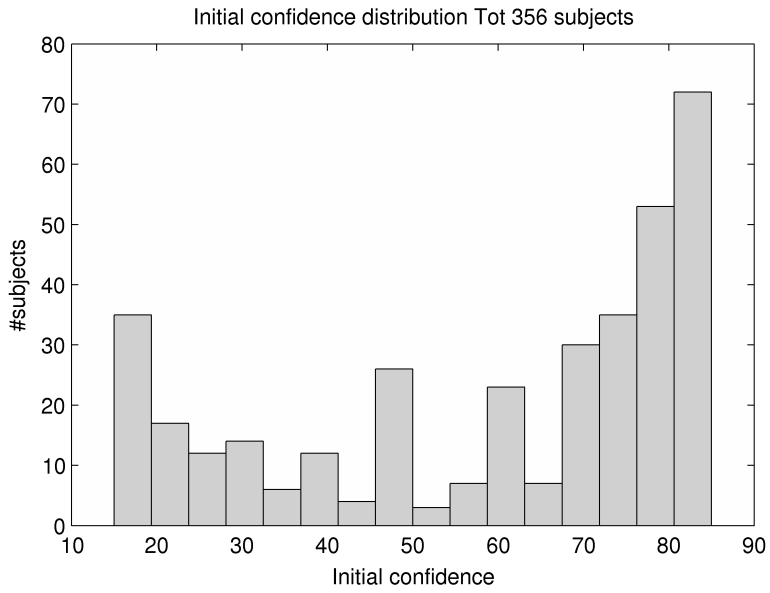

Figure 2: Distribution of confidence for the diagnosis of AD (DCAD) before automated hippocampal volumetry assessment.

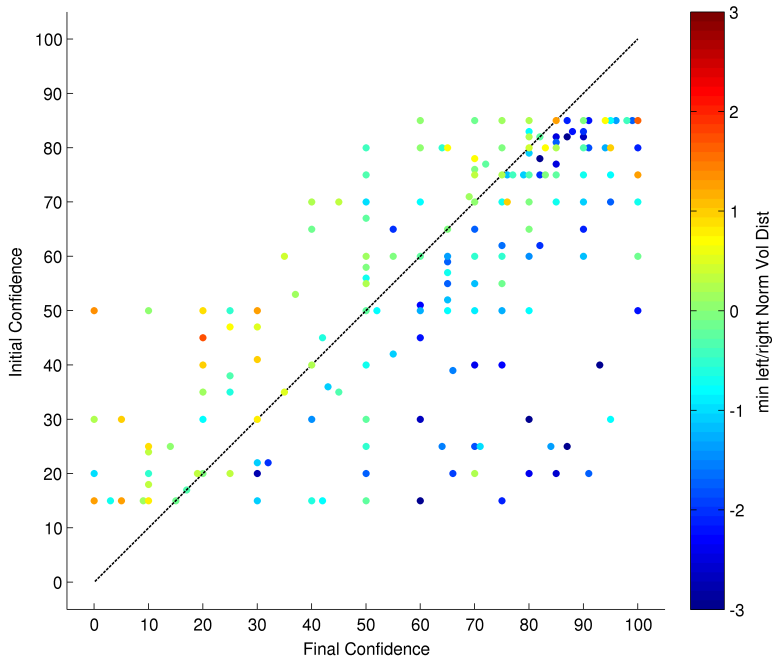

Figure 3: Scatter plot describing the diagnostic confidence of AD (DCAD) before and after the disclosure of hippocampal volumetric information (NVD). Points that lay on the bisector of the plane are those for which there was no change in DCAD, whereas points that are

further from the bisector are those for which the change in DCAD was larger. Warmer colors denote a positive distance from the median of the age-matched normative population (i.e. no hippocampal atrophy). Cooler colors denote a negative distance from the median of the age-matched normative population (i.e. hippocampal atrophy)

References

[1] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280–92. doi:10.1016/j.jalz.2011.03.003.

[2] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:270–9. doi:10.1016/j.jalz.2011.03.008.

[3] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, 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. Alzheimers Dement 2011;7:263–9. doi:10.1016/j.jalz.2011.03.005.

[4] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 2007;6:734–46. doi:10.1016/S1474-4422(07)70178-3.

[5] Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, et al. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol 2010;9:1118–27. doi:10.1016/S1474-4422(10)70223-4.

[6] Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, et al. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry 1992;55:967–72. doi:10.1136/jnnp.55.10.967.

[7] Jack CR. Alliance for aging research AD biomarkers work group: structural MRI. Neurobiol Aging 2011;32 Suppl 1:S48-57. doi:10.1016/j.neurobiolaging.2011.09.011.

[8] Frisoni GB, Hampel H, O'Brien JT, Ritchie K, Winblad B. Revised criteria for Alzheimer's disease: what are the lessons for clinicians? Lancet Neurol 2011;10:598–601. doi:10.1016/S1474-4422(11)70126-0.

[9] Frisoni GB, Bocchetta M, Chételat G, Rabinovici GD, de Leon MJ, Kaye J, et al. Imaging markers for Alzheimer disease: which vs how. Neurology 2013;81:487–500. doi:10.1212/WNL.0b013e31829d86e8.

[10] Hampel H, Lista S, Teipel SJ, Garaci F, Nisticò R, Blennow K, et al. Perspective on future role of biological markers in clinical therapy trials of Alzheimer's disease: A long-range point of view beyond 2020. Biochem Pharmacol 2014;88:426–49. doi:10.1016/j.bcp.2013.11.009.

[11] Shaffer JL, Petrella JR, Sheldon FC, Choudhury KR, Calhoun VD, Coleman RE, et al.

Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. Radiology 2013;266:583–91. doi:10.1148/radiol.12120010.
[12] Prestia A, Caroli A, Herholz K, Reiman E, Chen K, Jagust WJ, et al. Diagnostic accuracy of markers for prodromal Alzheimer's disease in independent clinical series. Alzheimer's Dement 2013;9:677–86. doi:10.1016/j.jalz.2012.09.016.
[13] Choo IH, Ni R, Schöll M, Wall A, Almkvist O, Nordberg A. Combination of 18F-FDG PET and cerebrospinal fluid biomarkers as a better predictor of the progression to Alzheimer's disease in mild cognitive impairment patients. J Alzheimers Dis 2013;33:929–39. doi:10.3233/JAD-2012-121489.
[14] Prestia A, Caroli A, van der Flier WM, Ossenkoppele R, Van Berckel B, Barkhof F, et al. Prediction of dementia in MCI patients based on core diagnostic markers for Alzheimer disease. Neurology 2013;80:1048–56. doi:10.1212/WNL.0b013e3182872830.
[15] Prestia A, Caroli A, Wade SK, van der Flier WM, Ossenkoppele R, Van Berckel B, et al. Prediction of AD dementia by biomarkers following the NIA-AA and IWG diagnostic criteria in MCI patients from three European memory clinics. Alzheimer's Dement 2015;11:1191–201. doi:10.1016/j.jalz.2014.12.001.
[16] Engelborghs S, De Vreese K, Van de Casteele T, Vanderstichele H, Van Everbroeck B, Cras P, et al. Diagnostic performance of a CSF-biomarker panel in autopsy-confirmed dementia. Neurobiol Aging 2008;29:1143–59. doi:10.1016/j.neurobiolaging.2007.02.016.
[17] Apostolova LG, Zarow C, Biado K, Hurtz S, Boccardi M, Somme J, et al. Relationship between hippocampal atrophy and neuropathology markers: a 7T MRI validation study of the EADC-ADNI Harmonized Hippocampal Segmentation Protocol. Alzheimers Dement 2015;11:139–50. doi:10.1016/j.jalz.2015.01.001.
[18] Bocchetta M, Boccardi M, Ganzola R, Apostolova LG, Preboske G, Wolf D, et al. Harmonized benchmark labels of the hippocampus on magnetic resonance: the EADC-ADNI project. Alzheimers Dement 2015;11:151–60.e5. doi:10.1016/j.jalz.2013.12.019.
[19] Frisoni GB, Jack CR, Bocchetta M, Bauer C, Frederiksen KS, Liu Y, et al. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimers Dement 2015;11:111–25. doi:10.1016/j.jalz.2014.05.1756.
[20] Jack CR. Alliance for aging research AD biomarkers work group: structural MRI. Neurobiol Aging 2011;32 Suppl 1:S48-57. doi:10.1016/j.neurobiolaging.2011.09.011.
[21] Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, et al. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. Neuroimage 2008;43:59–68. doi:10.1016/j.neuroimage.2008.07.003.
[22] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–55.
[23] Wolz R, Aljabar P, Hajnal J V, Hammers A, Rueckert D. LEAP: learning embeddings for atlas propagation. Neuroimage 2010;49:1316–25. doi:10.1016/j.neuroimage.2009.09.069.
[24] Chincarini A, Sensi F, Rei L, Gemme G, Squarcia S, Longo R, et al. Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease. Neuroimage 2015;125:834–47. doi:10.1016/j.neuroimage.2015.10.065.

[25] Collins DL, Pruessner JC. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage 2010;52:1355–66. doi:10.1016/j.neuroimage.2010.04.193.
[26] Hu S, Coupé P, Pruessner JC, Collins DL. Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation. Hum Brain Mapp 2014;35:377–95. doi:10.1002/hbm.22183.
[27] Roche A, Ribes D, Bach-Cuadra M, Krüger G. On the convergence of EM-like algorithms for image segmentation using Markov random fields. Med Image Anal 2011;15:830–9. doi:10.1016/j.media.2011.05.002.
[28] Schmitter D, Roche A, Maréchal B, Ribes D, Abdulkadir A, Bach-Cuadra M, et al. An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease. NeuroImage Clin 2015;7:7–17. doi:10.1016/j.nicl.2014.11.001.
[29] Redolfi A, McClatchey R, Anjum A, Zijdenbos A, Manset D, Barkhof F, et al. Grid infrastructures for computational neuroscience: the neuGRID example 2009.
[30] Frisoni GB, Redolfi A, Manset D, Rousseau M-É, Toga A, Evans AC. Virtual imaging laboratories for marker discovery in neurodegenerative diseases. Nat Rev Neurol 2011;7:429–38. doi:10.1038/nrneurol.2011.99.
[31] Groeneveld RA, Meeden G. Measuring skewness and kurtosis. J R Stat Soc Ser D (The Stat 1984;33:391–9. doi:10.2307/2987742.
[32] Ferrari S, Cribari-Neto F. Beta Regression for Modelling Rates and Proportions. J Appl Stat 2004;31:799–815. doi:10.1080/0266476042000214501.
[33] Cepeda-Cuervo E. Beta Regression Models: Joint Mean and Variance Modeling. J Stat Theory Pract 2014;9:134–45. doi:10.1080/15598608.2014.890983.
[34] Gelman A, Pardoe I. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics 2006;48:241–51. doi:10.1198/004017005000000517.
[35] Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci 1992;7:457–511. doi:10.1214/ss/1177011136.
[36] Bocchetta M, Galluzzi S, Kehoe PG, Aguera E, Bernabei R, Bullock R, et al. The use of biomarkers for the etiologic diagnosis of MCI in Europe: An EADC survey. Alzheimer's Dement 2015;11:195–206.e1. doi:10.1016/j.jalz.2014.06.006.
[37] Johnson KA, Minoshima S, Bohnen NI, Donohoe KJ, Foster NL, Herscovitch P, et al. Appropriate use criteria for amyloid PET: a report of the Amyloid Imaging Task Force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association. Alzheimers Dement 2013;9:e-1-16. doi:10.1016/j.jalz.2013.01.002.
[38] Grundman M, Pontecorvo MJ, Salloway SP, Doraiswamy PM, Fleisher AS, Sadowsky CH, et al. Potential impact of amyloid imaging on diagnosis and intended management in patients with progressive cognitive decline. Alzheimer Dis Assoc Disord n.d.;27:4–15. doi:10.1097/WAD.0b013e318279d02a.