Validation of an automatic tool for the rapid measurement of brain atrophy and white matter hyperintensity: QyScore®

Enrica Cavedo, Philippe Tran, Urielle Thoprakarn, Jean-Baptiste Martini, Antoine Movschin, Christine Delmaire, Florent Gariel, Damien Heidelberg, Nadya Pyatigorskaya, Sébastian Ströer, et al.

European Radiology
https://doi.org/10.1007/s00330-021-08385-9

MAGNETIC RESONANCE

Validation of an automatic tool for the rapid measurement of brain atrophy and white matter hyperintensity: QyScore®

Enrica Cavedo[1] · Philippe Tran[1,2] · Urielle Thoprakarn[1] · Jean-Baptiste Martini[1] · Antoine Movschin[1] · Christine Delmaire[3] · Florent Gariel[4] · Damien Heidelberg[5,6] · Nadya Pyatigorskaya[7] · Sébastian Ströer[7] · Pierre Krolak-Salmon[8,9,10] · Francois Cotton[11,12] · Clarisse Longo dos Santos[1] · Didier Dormont[2,7]

Received: 7 December 2020 / Revised: 15 September 2021 / Accepted: 21 October 2021

Abstract

Objectives QyScore® is an imaging analysis tool certified in Europe (CE marked) and the US (FDA cleared) for the automatic volumetry of grey and white matter (GM and WM respectively), hippocampus (HP), amygdala (AM), and white matter hyperintensity (WMH). Here we compare QyScore® performances with the consensus of expert neuroradiologists.

Methods Dice similarity coefficient (DSC) and the relative volume difference (RVD) for GM, WM volumes were calculated on 50 3DT1 images. DSC and the F1 metrics were calculated for WMH on 130 3DT1 and FLAIR images. For each index, we identified thresholds of reliability based on current literature review results. We hypothesized that DSC/F1 scores obtained using QyScore® markers would be higher than the threshold. In contrast, RVD scores would be lower. Regression analysis and Bland–Altman plots were obtained to evaluate QyScore® performance in comparison to the consensus of three expert neuroradiologists.

Results The lower bound of the DSC/F1 confidence intervals was higher than the threshold for the GM, WM, HP, AM, and WMH, and the higher bounds of the RVD confidence interval were below the threshold for the WM, GM, HP, and AM. QyScore®, compared with the consensus of three expert neuroradiologists, provides reliable performance for the automatic segmentation of the GM and WM volumes, and HP and AM volumes, as well as WMH volumes.

Conclusions QyScore® represents a reliable medical device in comparison with the consensus of expert neuroradiologists. Therefore, QyScore® could be implemented in clinical trials and clinical routine to support the diagnosis and longitudinal monitoring of neurological diseases.

Key Points

- *QyScore® provides reliable automatic segmentation of brain structures in comparison with the consensus of three expert neuroradiologists.*
- *QyScore® automatic segmentation could be performed on MRI images using different vendors and protocols of acquisition. In addition, the fast segmentation process saves time over manual and semi-automatic methods.*
- *QyScore® could be implemented in clinical trials and clinical routine to support the diagnosis and longitudinal monitoring of neurological diseases.*

✉ Enrica Cavedo
ecavedo@qynapse.com

1 Qynapse SAS, 130 rue de Lourmel, 75015 Paris, France

2 Equipe-Projet ARAMIS, ICM, CNRS UMR 7225, Inserm U1117, Sorbonne Université UMR_S 1127, Centre Inria de Paris, Groupe Hospitalier Pitié-Salpêtrière Charles Foix, Faculté de Médecine Sorbonne Université, Paris, France

3 Fondation Adolphe de Rothschild, Paris, France

4 Department of Neuroradiology, University Hospital of Bordeaux, Bordeaux, France

5 Faculty of Medicine, Claude-Bernard Lyon 1 University, 69000 Lyon, France

6 Service de Radiologie and Laboratoire d'anatomie de Rockefeller, centre hospitalier Lyon Sud, hospices civils de Lyon, 69000 Lyon, France

7 Department of Neuroradiology, Groupe Hospitalier Pitié-Salpêtrière, AP-HP, Sorbonne Université UMR_S 1127, Paris, France

8 Clinical and Research Memory Centre of Lyon, Hospices Civils de Lyon, Lyon, France

9 University of Lyon, Lyon, France

10 INSERM, U1028; UMR CNRS 5292, Lyon Neuroscience Research Center, Lyon, France

11 Radiology Department, centre hospitalier Lyon-Sud, hospices civils de Lyon, 69310 Pierre-Bénite, France

12 Inserm U1044, CNRS UMR 5220, CREATIS, Université Lyon-1, 69100 Villeurbanne, France

Published online: 01 January 2022

Keywords Magnetic resonance imaging · White matter · Hippocampus · Automated quantification

Abbreviations

AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AM	Amygdala
AVE	Absolute volume error
DSC	Dice similarity coefficient
FDA	Food and Drug Administration
FLAIR	T2 fluid attenuated inversion recovery
FTD	Frontotemporal dementia
FTLDNI	Frontotemporal Lobar Degeneration Neuroimaging Initiative
GM	Grey matter
HC	Healthy controls
HP	Hippocampus
kNN	K-nearest neighbors
MC	Multiple sclerosis
MCI	Mild cognitive impairment
MRI	Magnetic resonance imaging
NINCDS/ADRD	National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association
OASIS	Open Access Series of Imaging Studies
PD	Parkinson's disease
PPMI	Parkinson Progression Markers Initiative
RVD	Relative volume difference
SPM12	Statistical Parametric Mapping software v12
SVM	Support vector machines
WM	White matter
WMH	White matter hyperintensity

Introduction

Neurological disorders represent a major public health problem in Europe and the rest of the world [1]. A systematic analysis for the Global Burden of Disease Study 2016 showed that neurological disorders were the leading cause of disability-adjusted life-years (worldwide 276 million) and the second leading cause of death (worldwide 90 million) [2].

Magnetic resonance imaging (MRI) technology and the development of MRI markers of neurological diseases have been improved substantially in both research and clinical environments over the last 30 years [3]. Automated MRI segmentation methods have been used in addition to visual analysis and manual segmentation assessments [4–6], improving the early diagnosis of neurological disease and the development of effective drugs [7–9]. In addition, large-scale multi-institutional research studies [10, 11] have worked in synergy for the implementation of standardized imaging acquisition protocols in the research environment and clinical setting [12, 13]. These advancements highlight a need for an MRI volumetric analysis tool suitable for routine clinical use. Few automated segmentation software are currently approved by regulatory agencies (such as the FDA) and therefore included in the clinical routine workflow.

To be validated and implemented in clinical practice, segmentation algorithms included in medical devices should demonstrate equal or better performance than the assessment performed by expert neuroradiologists. The present study aims to describe the comparison between the performance of the brain segmentation algorithms included in QyScore® and the manual segmentations or manual segmentation correction conducted by three expert neuroradiologists. Here we hypothesize that the segmentation algorithms included in QyScore® show reliable performance in comparison with the consensus of three expert neuroradiologists.

Materials and methods

QyScore® is a CE-marked and FDA-cleared software, developed by Qynapse (https://www.qynapse.com/), that provides segmentation and volumetric measurements of grey matter (GM), white matter (WM), hippocampus (HP), and amygdala (AM) from 3DT1 images, as well as white matter hyperintensities (WMHs) from 3DT1 and FLAIR images. In addition, z-scores and percentiles are obtained from the comparison with a large normative database of healthy controls. The normative database includes cognitively intact individuals between the ages of 20 and 90 years, coming from European and North American databases [10, 11] (https://brain-development.org/ixi-dataset; https://www.humanconnectome.org/study/hcp-young-adult). The full panel of MRI markers described above is quantified in 15 min per patient. It has a user-friendly interface, including 3D navigation of MRI images (Fig. 1). The outputs of the software include an electronic report and color overlays of the regional segmentation on the selected brain image for visualisation.

Populations and cohorts

The experimental validation of QyScore® was performed using data from different cohorts: the Alzheimer's Disease

Fig. 1 QyScore® visual interface, including 3D image navigation and volumetric results, compared with a normative dataset of healthy individuals

Neuroimaging Initiative (ADNI) [10], the Open Access Series of Imaging Studies (OASIS) [14], the KIKI2009/Kirby [15], the Parkinson Progression Markers Initiative (PPMI) [11], the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI), the ANMerge dataset [16], and a public segmentation database of Multiple Sclerosis patients (LIMTS) [17]. Data from three additional cohorts were incorporated to enrich the sample: the Clinically Isolated Syndrome–COGnitive (SCI-COG) cohort, the REACTIV database, and the MEMORA cohort.

We created three different databases: one for the measurement of GM and WM volumes, one for HP and AM volumes, and one for WMH volumes. Subjects in each database were well-balanced according to age (ranging from 20 to 90 years old), sex, scanner field strength (1.5 T/3 T), and type of MRI acquisition (2D/3D for FLAIR images only). To ensure as broad a sample as possible in terms of volumetric measurements, we selected a heterogeneous population composed of healthy controls (HCs), mild cognitive impairment (MCI), Alzheimer's disease (AD), Parkinson's disease (PD), frontotemporal dementia (FTD), and multiple sclerosis (MS) patients.

The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. For up-to-date information, see http://adni.loni.usc.edu/www.adni-info.org.

The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. OASIS is made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN). It provides cross-sectional MRI data in young, middle aged, non-demented, and demented older adults [14].

FTLDNI was funded through the National Institute of Aging and started in 2010. The primary goals of FTLDNI were to identify neuroimaging modalities and methods of analysis for tracking frontotemporal lobar degeneration—and to assess the value of imaging versus other biomarkers in diagnostic roles. The principal investigator of FTLDNI was Dr. Howard Rosen, MD, at the University of California, San Francisco. The data are the result of collaborative efforts at three sites in North America. For up-to-date information on participation and protocol, please visit http://memory.ucsf.edu/research/studies/nifd

Clinical diagnostic criteria for each diagnosis considered were NINCDS/ADRDA and Clinical Dementia Rating Scale

Table 1 Recommended neuroimaging parameters for analyzing QyScore® markers

	T1		T2FLAIR 2D		T2FLAIR 3D	
	Ideal	Tolerated	Ideal	Tolerated	Ideal	Tolerated
Voxel size in plane (mm × mm)	[1, 1]	Min: [0.43, 0.43] Max: [1.5, 1.5]	[1, 1]	Min: [0.43, 0.43] Max: [1.1, 1.1]	[1, 1]	Min: [0.43, 0.43] Max: [1.1, 1.1]
Field of view (mm)	*x*: [152, 340] *y*: [220, 340] *z*: [152, 340]	*x*: [144, 500] *y*: [170, 340] *z*: [132, 340]	*x*: [152, 340] *y*: [220, 340] *z*: [152, 340]	*x*: [144, 500] *y*: [170, 340] *z*: [132, 340]	*x*: [152, 340] *y*: [220, 340] *z*: [152, 340]	*x*: [144, 500] *y*: [170, 340] *z*: [132, 340]
Slice thickness (mm)	1	Min: 0.68 Max: 2	3.3	Min: 2 Max: 5	1	Min: 0.8 Max: 2
Interslice gap (mm)	0.0	Min: − 1.0 Max: 0	0.0	Min: 0.0 Max: 1.0	0.0	Min: − 1.0 Max: 0
Interpolation	No	Yes	No	Yes	No	Yes
Acquisition direction	Sagittal	Axial, Coronal	Axial	Coronal, Sagittal	Sagittal	Axial, Coronal
Acquisition type	3D	2D with isotropic voxel	2D	2D	3D	3D
Field strength	1.5 T, 3 T	1.5 T, 3 T	1.5 T, 3 T	1.5 T, 3 T	1.5 T, 3 T	1.5 T, 3 T

(CDR) = 1 for AD [14, 18]. PD was defined according to the features described by Marek and colleagues [11]. FTD was defined according to frontotemporal dementia consortium criteria [19] and MS, according to Polman and colleagues [20].

MRI data

QyScore® analysis can be performed using MRI sequences with recommended acquisition parameters on 1.5- and 3-Tesla (1.5 T and 3 T) scanners, as reported in Table 1. More specifically, the software retrieves DICOM MRI data (non-contrast 3DT1 series acquired using 1.5 T or 3 T scanners and T2 fluid attenuated inversion recovery (FLAIR) series acquired using 3-Tesla scanners) from a DICOM server and sends them to an analysis server.

From the analysis server, before computing MRI analysis, QyScore® performs a quality check of MRI parameters to verify that the parameters are in line with the ones recommended (Table 1). The range of recommended parameters was selected for their suitability in a clinical routine setting. The analysis is performed if the parameters are within the "ideal" or "tolerated" ranges, as described in Table 1. When the acquisition parameters are not in line with the ones recommended, QyScore® does not perform the analysis. Then, QyScore® performs the automatic segmentation of GM, WM, HP, AM, and WMH.

For the present study, images were acquired using MRI scanners from different manufacturers: General Electric Healthcare (GE), Siemens Medical Solutions, Philips Medical Systems.

3D T1-weighted MRI images, used to quantify atrophy, were acquired either on 1.5 T or 3 T scanners, using exclusively gradient-echo 3D sequences. FLAIR images, used to quantify white matter hyperintensities, were acquired on 3 T scanners either in 2D or 3D.

Automated Qyscore® imaging markers

All QyScore® imaging markers (GM, WM, HP, AM, and WMH volumes) are widely employed in the imaging field as markers of brain atrophy [21–24] and white matter hyperintensities [24]. The algorithms used for volume calculation are based on existing segmentation methods, as described in detail below. Fifteen minutes are needed for the segmentation of all structures (less than 10 min for T1-weighted MRI images only). Whole GM and WM volumes were quantified using the Statistical Parametric Mapping software v12 (SPM12). SPM is a software highly used for GM and WM segmentation both in research and in clinical data [25]. Hippocampus and AM volumes were measured using an improved version of SACHA (Segmentation Automatique Compétitive de l'Hippocampe et de l'Amygdale), a fast and fully automatic hybrid segmentation tool previously described in detail [26, 27].

White matter hyperintensity volumes were measured using a method based on the WHASA (White matter Hyperintensities Automated Segmentation Algorithm) automatic segmentation method [28].

Expert consensus on manual segmentation and semi-automatic assessment of imaging markers

Grey matter, white matter, hippocampus, and amygdala volumes

Three expert neuroradiologists manually edited and corrected the segmentation of GM and WM (rater's initials: C.S., N.P., S.S.). First, GM and WM segmentations were performed using Freesurfer [29]. Then, GM and WM segmentations were corrected using the software ITK-SNAP according to their experience. Since no protocols were found in the literature, we asked clinicians to check the segmentation automatically performed to identify important segmentation errors, mainly at the cortex level.

Three expert neuroradiologists manually delineated HP and AM according to the anatomical constraints described by Chupin and colleagues [26].

For each marker considered (GM, WM, HP, and AM), we established a consensus from the corrections/segmentations obtained from the expert tracers by the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm. The STAPLE algorithm considers a collection of segmentations and computes a probabilistic estimate of the true segmentation by estimating an optimal combination of the segmentations [30]. The consensus was obtained using the Computational Radiology Kit (CRKIT) software that includes the STAPLE algorithm.

Table 2 Thresholds considered for each reliability measure

QyScore® markers	Thresholds			
	DSC	RVD (%)	AVE (mL)	F1
GM	0.8	9.59	–	–
WM	0.8	9.18	–	–
HP	0.7	31.5	–	–
AM	0.7	31.5	–	–
WMH low (< 5 mL)	0.27	–	2	–
WMH medium (5–15 mL)	0.51	–	5	–
WMH high (15–30 mL)	0.64	–	10	–
WMH very high (> 30 mL)	0.67	–	15	–
WMH whole sample	0.47	–	–	0.3

Abbreviations: *GM* grey matter, *WM* white matter, *HP* hippocampus, *AM* amygdala, *WMH* white matter hyperintensity, *DSC* dice similarity coefficient, *RVD* relative volume difference, *AVE* absolute volume error, *SD* standard deviation. Values in parentheses indicate references

White matter hyperintensity volume

Some of the data used for the validation of WHASA are publicly available ($N = 30$) [17]. We used the consensus derived from the WMH manual segmentation of three anonymous expert raters reported in the manuscript of Lesjak and colleagues [16]. On the remaining data ($N = 100$), segmentations were first initiated using the Lesion Segmentation Toolbox (LST) [31], then three expert neuroradiologists (rater's initials: C.D., D.H., N.P.) manually corrected the WMH segmentation. The consensus among raters was obtained using the STAPLE algorithm.

Measures of reliability

The reliability measures considered for the validation of QyScore® imaging markers were the dice similarity coefficient (DSC) and the relative volume difference (RVD) for GM, WM, HP and AM segmentations. The DSC, the absolute volume error (AVE), and the F1 metrics were considered for WMH segmentation. For each measure of reliability related to a QyScore® marker, we identified from the literature the performances of other segmentation methods comparable to QyScore®. As reported in Table 2 [26, 32–40], we averaged their performances to set fair values of reliability. We then considered these values of reliability as thresholds for the subsequent statistical analysis.

Dice similarity coefficient

The DSC [35] is the most frequently used statistical validation metric employed to evaluate the performance of both the reproducibility of manual segmentations/corrections and the spatial overlap accuracy of automated segmentation methods [4]. We identified from the literature DSC thresholds for each QyScore® marker to test our statistical hypothesis as reported in Table 2. The values identified in Table 2 are usually considered an excellent match between two segmentations and are in line with the values reported in previous studies [32]. The DSC threshold for WMH segmentation was defined based on four categories of lesions load: low (< 5 mL), medium (5–15 mL), high (15–30 mL), very high (> 30) mL). The DSC thresholds for these four categories of WMH were obtained by averaging the DSC values reported in the supplementary material of the study by Commowick [34] et al. for each category considered. Categories and DSC thresholds are described in Table 2.

The relative volume difference

We calculated the RVD between the volume of the structure automatically segmented and the volume obtained from the consensus of the three expert neuroradiologists. The RVD threshold (Table 2) was obtained among the different RVD values acquired from freeware packages widely used by experts and described in the manuscript of Mendrik and colleagues [41]. To the best of our knowledge, this study is the only one reporting several RVD values for GM and WM segmentation [41]. RVD thresholds for the HP and AM, reported in Table 2, represent the average of RVDs from a selection of studies published in the literature [36, 37, 42–46].

The absolute volume error and the F1 score

We measured the AVE and F1 score exclusively for WMH. Evaluation of WMH detection relies on determining how many WMH have been correctly or incorrectly detected. The F1-score ranges from 0 to 1 and provides an idea of the detection performance (perfect detection: 1). It is calculated from (i) the number of WMH in each segmentation (expert consensus and QyScore® automatic segmentation), (ii) the number of WMH correctly detected from the experts' consensus, and (iii) the number of WMH in the automatic segmentation for which there is a WMH from the consensus, as suggested by Commowick and colleagues [34].

Thresholds for both metrics (AVE and F1 score) for WMH were defined using the values reported in the supplementary material of the study conducted by Commowick and colleagues [34] (Table 2).

Statistical analysis

The validity of QyScore® imaging markers was tested in comparison to the consensus obtained by three experts. We identified thresholds of reliability measures based on the performances reported in the literature from other similar segmentation methods (Table 2). The null hypothesis was that the measures of comparison DSC/F1 scores (the higher, the better) were equal to/below the chosen threshold—and the alternative hypothesis was that the metrics were higher, indicating improved accuracy when comparing QyScore® with the consensus obtained by three experts. For DSC and F1 scores, the lower bound of the 97.5% confidence interval was compared to the threshold, indicating that the DSC and F1 scores obtained using QyScore® were significantly better to the DSC and F1 scores reported in the literature for similar segmentation methods. For RVD and AVE (the lower, the better), the null hypothesis was that the metrics were equal to/above the threshold, and the alternative hypothesis is that the metrics were lower. The upper bound of the 97.5% confidence interval was compared to the threshold, indicating that the RVD and AVE obtained using QyScore® were significantly inferior to the RVD and AVE reported in the literature for similar segmentation methods.

Furthermore, we compared volumes obtained from the automated QyScore® imaging markers with the consensus obtained by three experts using regression analysis and Bland-Altmann plots.

We used the following library in Python to perform statistical analysis: Scikit-learn (http://scikit-learn.org/stable/), version 0.19.1; Scipy (https://www.scipy.org/), version 0.17.0; NumPy (http://www.numpy.org/), version 1.14.3; Nipype (https://nipype.readthedocs.io), version 1.1.2.

Results

All 180 MRI images passed the quality control performed by QyScore®.

Table 3 describes the main features of each database's selection criteria and the mean absolute volume of each QyScore® marker. Each database was furthermore constituted of HC and clinical patients as follows: GM and WM database (24 HC, 2 AD, 2 MS, 2 PD), HP and AM database

Table 3 Distribution of selection criteria (sex, age, magnetic field, type) in each database and absolute volumes of QyScore® markers

Database		Volumes (mL) (mean (sd))	Sex (M/W)	Age mean (sd)	Magnetic field (1.5 T/3 T)	Type (2D/3D)
GM and WM ($N=30$)	GM	672.67 (82.02)	16/14	53.51 (20.48)	15/15	0/30
	WM	462.19 (51.85)				
HP and AM ($N=50$)*	HP	5.83 (0.71)	27/23	52.84 (20.40)	24/26	0/50
	AM	2.90 (0.43)				
WMH ($N=130$)		18,76 (17,62)	60/70	62.95 (19.35)	0/130	70/60

Abbreviations: *M* men, *W* women, *GM* grey matter, *WM* white matter, *HP* hippocampus, *AM* amygdala, *WMH* white matter hyperintensity, *sd* standard deviation

*The 30 images included for the GM and WM analysis were also included in the dataset for the HP and AM analysis

Table 4 Average measures and standard deviations of overlap and volumetric agreement between the segmentation performed by each neuroradiologist and their consensus obtained using the STAPLE algorithm for each QyScore® marker

		DICE	RVD	AVE	F1
GM ($N=30$)	Manual tracer 1	0.99 (0.01)	0.18 (0.33)	–	–
	Manual tracer 2	0.99 (0.01)	0.08 (0.06)	–	–
	Manual tracer 3	1.00 (0.01)	0.04 (0.04)	–	–
WM ($N=30$)	Manual tracer 1	0.99 (0.01)	0.15 (0.43)	–	–
	Manual tracer 2	0.99 (0.08)	0.09 (0.08)	–	–
	Manual tracer 3	0.99 (0.01)	0.07 (0.07)	–	–
HP ($N=50$)	Manual tracer 1	0.98 (0.07)	1.38 (2.00)	–	–
	Manual tracer 2	0.97 (0.01)	2.67 (3.24)	–	–
	Manual tracer 3	0.95 (0.01)	3.85 (2.27)	–	–
AM ($N=50$)	Manual tracer 1	0.90 (0.11)	7.91 (10.59)	–	–
	Manual tracer 2	0.86 (0.11)	12.59 (12.78)	–	–
	Manual tracer 3	0.89 (0.11)	5.56 (6.94)	–	–
WMH ($N=100$)	Manual tracer 1	0.88 (0.12)	–	2.70 (4.75)	0.83 (0.16)
	Manual tracer 2	0.85 (0.18)	–	2.51 (3.21)	0.70 (0.23)
	Manual tracer 3	0.77 (0.21)	–	1.50 (1.80)	0.62 (0.27)

Abbreviations: *GM* grey matter, *WM* white matter, *HP* hippocampus, *AM* amygdala, *WMH* white matter hyperintensity, *DSC* dice similarity coefficient, *RVD* relative volume difference, *AVE* absolute volume error; values in parentheses indicate standard deviation

(37 HC, 4 AD, 6 MS, 3 PD), WMH database (20 HC, 49 AD, 6 FTD, 45 MS, 2 PD, 10 with MCI). The type of manufacturers used for the acquisition of the MRI images was distributed among GM and WM database (2 GE, 9 Philips, 19 Siemens), HP and AM database (4 GE, 13 Philips, 33 Siemens), WMH database (30 GE, 26 Philips, 74 Siemens). Consensus results and the contribution of each expert neuroradiologist EW reported in Table 4.

Validation results for GM and WM segmentations

Comparison with the experts' consensus showed that GM and WM segmentations, obtained using QyScore®, displayed the lower bound of the DSC confidence intervals (GM 97.5% CI: 0.848, 0.866; WM 97.5% CI: 0.892, 0.907 respectively) above the DSC threshold (0.8), as well as the higher bounds of the RVD confidence interval (GM 97.5% CI: 5.578, 8.464; WM 97.5% CI: 2.985, 6.425) below the RVD threshold (9.59% for the GM and 9.18% for the WM respectively). We found consistent results after the stratification by field strength; the mean DSC for the GM was equal to 0.87 for 1.5 T and 0.86 for 3 T; the mean DSC for the WM was equal to 0.92 for 1.5 T and 0.90 for 3 T. Coefficient of determination was equal to 0.91 for the GM and 0.92 for the WM (Fig. 2A), whilst the means (95% confidence interval (CI)) of Bland–Altman plots were − 36.51 (95% CI: 35.39, − 108.42) for GM and 13.33 (95% CI: 63.04, − 36.38) for WM (Fig. 2B).

Validation results for HP and AM segmentations

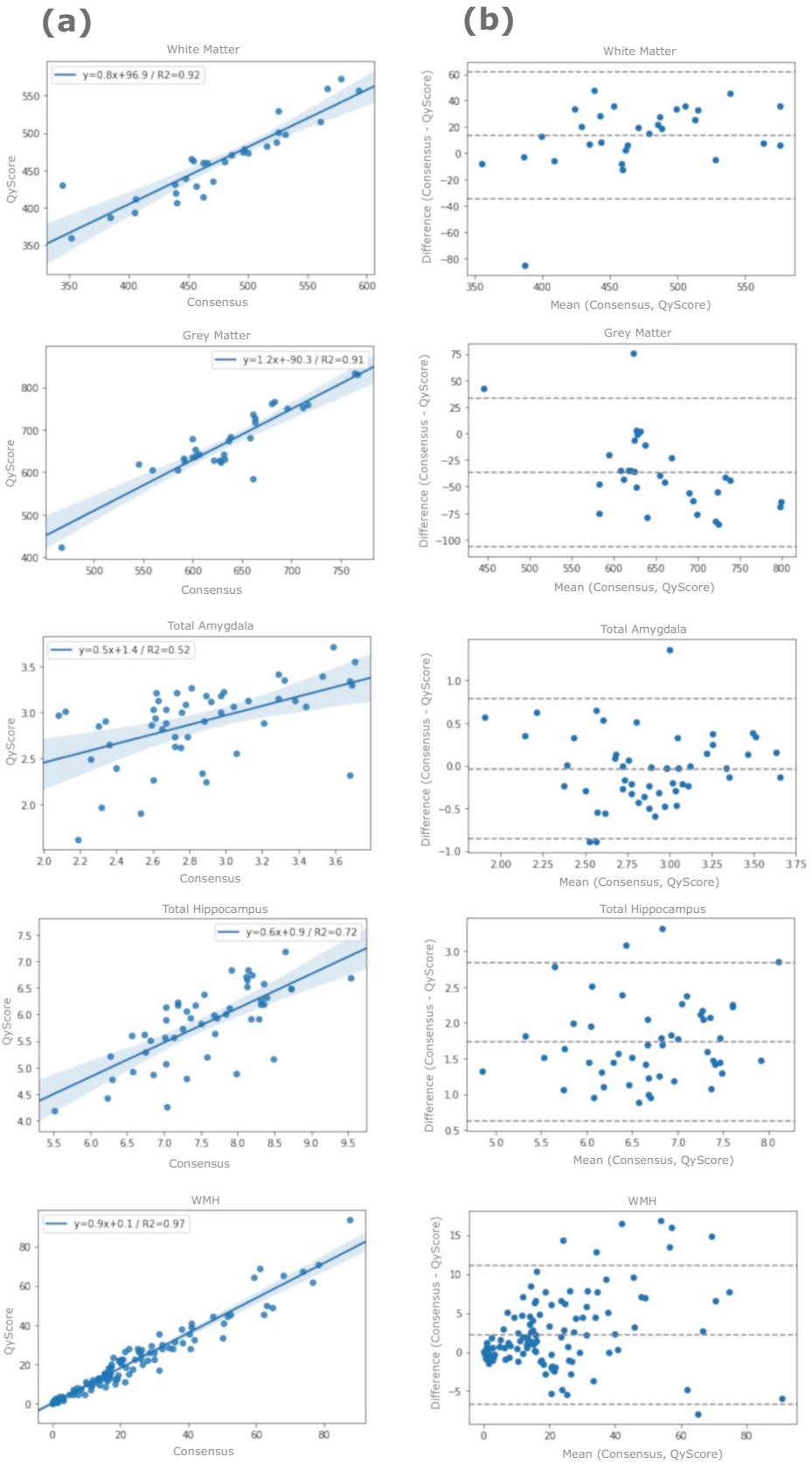
HP and AM segmentations showed the lower bound of the DSC confidence intervals (HP 97.5% CI: 0.801, 0.818; AM 97.5% CI: 0.763, 0.786 respectively) above the DSC threshold (0.7), as well as the higher bounds of the RVD confidence interval (HP 97.5% CI: 20.886, 24.648; AM 97.5% CI: 9.203, 15.023) below the RVD threshold (31.5%). We found consistent results after the stratification by field strength; the mean DSC for the HP was equal to 0.81 for 1.5 T and 0.80 for 3 T; the mean DSC for the AM was equal to 0.75 for 1.5 T and 0.76 for 3 T.

Coefficient of determination was equal to 0.72 for the HP and 0.52 for the AM (Fig. 2A), while the means (95% CI) of Bland–Altman plots were − 1.63 (95% CI: − 0.93, − 3.23) for HP and 0.02 (95% CI: 0.86, − 0.81) for AM (Fig. 2B).

Validation results for WMH segmentations

As with the previous markers, WMH segmentations also satisfied the alternative hypothesis for each comparison considered. As reported in Table 5, for all the WMH load considered, the lower bound of the DSC confidence intervals was above the DSC threshold, and the higher bound of the AVE confidence interval was below the AVE threshold.

Coefficient of determination was equal to 0.97 (Fig. 2A), while the mean (95% CI) of Bland–Altman plot was − 2.16 (95% CI: 11.17, − 6.84) (Fig. 2B).


(a)
(b)
White Matter
y=0.8x+96.9 / R2=0.92
QyScore
Consensus
Difference (Consensus - QyScore)
Mean (Consensus, QyScore)
Grey Matter
y=1.2x+-90.3 / R2=0.91
Total Amygdala
y=0.5x+1.4 / R2=0.52
Total Hippocampus
y=0.6x+0.9 / R2=0.72
WMH
y=0.9x+0.1 / R2=0.97

◂**Fig. 2** Scatterplots from regression analysis (**a**) and Bland–Altman plots (**b**) showing good concordance between QyScore® markers and the consensus from the expert raters

Discussion

Our results showed that the fully automatic tool, QyScore®, accurately measures brain imaging markers such as GM, WM, HP, AM, and WMH volumes. In comparison with the consensus of three expert neuroradiologists, QyScore® showed good concordance for all its imaging markers: GM, WM, HP, AM, and WMH—as depicted by the Bland–Altman plots and similarity measures. These results support the reliability of QyScore® compared with the experts' consensus in segmenting brain structure essentials for detecting brain atrophy in several brain diseases such as Alzheimer's, Parkinson's, and multiple sclerosis. The value of reliability indices (DSC, AVE, RVD, F1), obtained from the comparison between QyScore® and the consensus from three expert neuroradiologists, satisfied the alternative hypothesis for each comparison considered. Thresholds were set based on data present in the literature using other automatic segmentation methods, suggesting that QyScore® performs equally to or better than other automated methods currently used in the research context. In addition, results obtained by QyScore® are consistent using T1-weighted and FLAIR sequences from common clinical platforms and regardless of field strength for T1-weighted images.

Grey matter and white matter

The performance of markers quantified using QyScore® is in accordance with the currently available methods for the automatic segmentation of brain MRI described in the literature.

In particular, results from the MRBrainS challenge—comparing automatic and semi-automatic methods to segment GM and WM to the gold standard—concluded that SPM12 was the most robust, accurate, and fastest algorithm among the freeware packages evaluated [41]. A further study assessing whole brain and GM atrophy in multiple sclerosis showed for GM a DSC equal to 0.90 for SPM12 that was significantly higher to the one found for MSmetrix (DSC = 0.59) [47]. The quantification of GM atrophy using QyScore® in a heterogeneous sample of individuals with different pathophysiology showed an intermediate DSC (GM 97.5% CI: 0.848–0.866).

Hippocampus and amygdala

Regarding the HP and AM segmentation, a previous version of QyScore®'s algorithm using SACHA has shown better segmentation results compared to semi-automatic methods and other segmentation methods based on atlases [48]. The overlap between SACHA segmentation and the manual segmentation (detected by the DSC) was higher compared to the ones obtained for Freesurfer and FSL/FIRST [33]. HP measurements with Freesurfer were superior to FIRST [33]; however, several studies reported an overestimation of HP volume measured using Freesurfer [33, 46, 49]. Moreover, SACHA showed good accuracy in detecting mild cognitive impairment (MCI) and AD patients [50].

White matter hyperintensities

A previous version of QyScore®'s algorithm WHASA was validated on clinical routine MRI images showing a high intra-class coefficient of correlation with manual segmentation [28], which supports our results. Furthermore, WHASA was previously compared to other methods such as Freesurfer, a thresholding approach, and other methods based on k-nearest neighbors (kNN) and support vector machine (SVM) algorithms [28]. WHASA showed a better performance than Freesurfer and the thresholding approach as well as a comparable performance to the one obtained from kNN and SVM methods [28]. The spatial overlap between

Table 5 Thresholds of the reliability measures (DSC, AVE, F1) for each category of WMH as well as confidence intervals of WMH volumes measured using QyScore® in comparison with experts' consensus

QyScore® markers	Thresholds			QyScore® 97.5% confidence intervals		
	DSC	AVE (mL)	F1	DSC	AVE (mL)	F1
WMH low (< 5 mL)	0.27	2	–	**0.284**, 0.414	0.329, **0.666**	–
WMH medium (5–15 mL)	0.51	5	–	**0.647**, 0.716	1.152, **2.518**	–
WMH high (15–30 mL)	0.64	10	–	**0.728**, 0.771	2.648, **4.120** Ok	–
WMH very high (> 30 mL)	0.67	15	–	**0.761**, 0.811	6.186, **10.417**	–
WMH whole sample	0.47	–	0.3	**0.616**, 0.674	–	**0.352**, 0.3911

Bold numbers highlight the fact that the values were above or below the defined tresholds

Abbreviations: *WMH* white matter hyperintensity, *DSC* dice similarity coefficient, *AVE* absolute volume error, *mL* milliliter

WHASA segmentation and manual segmentation was higher compared to the spatial overlap that was found between established automatic methods and manual segmentation [51].

Implications of the use of automatic algorithms and medical devices in clinical routine

Several medical devices measuring brain imaging markers for neurological diseases are currently available [51–55]. However, multiple elements prevent the comparison between QyScore® performances and other devices due to (i) different indices of reliability used for their validation [54], (ii) the use of different validation methods [53, 56], and (iii) heterogeneous target populations among studies (exclusively young healthy controls, or MS or AD patients only) [51, 53, 54].

Volumetric MRI measures (such as the HP or GM volume) have been proposed as surrogate markers of in vivo brain atrophy and neurodegeneration [57, 58] and might be used for the early diagnosis, monitoring, and secondary outcome in clinical trials for several neurological diseases. An additional advantage of measuring brain atrophy in clinical trials is that fewer subjects need to be included [59, 60].

In line with our evidence, a recent research study has also demonstrated that quantitative reports, alongside routine visual MRI assessment, improves sensitivity and accuracy for detecting volume loss in AD compared to visual assessment alone [61].

Gold standard methods, such as manual tracing, are difficult, time-consuming and prone to inter- and intra-rater variability. In contrast, automatic methods have the obvious advantage of being consistent and fast compared to manual or semi-automatic methods [62]. For this reason, the validation of automated measures of MRI brain segmentation is of substantial importance to support practitioners in clinical settings and pharmacological companies in the context of new drug development [46].

Thanks to the reliability of the main results presented in the present manuscript, QyScore® has been considered suitable by regulatory bodies worldwide (CE and FDA). The panel of MRI markers (GM, WM, HP, AM, and WMH volumes) measured by QyScore® can support clinicians in the diagnosis and monitoring of clinical progression of neurological diseases such as MCI and AD dementia, MS, Parkinson's, and other neurodegenerative disorders. However, the clinicians make the final clinical decision based on their expert review of the QyScore® results.

The automated measure of MRI markers allowing fast segmentation of brain volumes and WMH (15 min) overcomes the time-consuming and subjective nature of the manual approach. QyScore® has reliable markers that can be implemented in clinical trials as primary/secondary outcomes to investigate the disease-modifying effect of treatments. In this regard, the HP atrophy measured using SACHA was already employed as the primary endpoint in a clinical trial aimed at investigating the efficacy of donepezil treatment in suspected prodromal AD patients [9].

Furthermore, QyScore® results can be incorporated into the clinical reports—providing additional neuroimaging information to clinicians that may be employed during their clinical and radiological assessments. QyScore® outputs can assist clinicians in the clinical diagnosis and monitoring, as well as in the choice of the most appropriate treatment in clinical practice.

Study limitations

Some limitations of the present study should be considered. The current version of the QyScore® algorithm for GM segmentation is not optimized for the exclusion of white matter lesions. In this regard, review of results is required before considering their use for clinical reports. This functionality will be included in a future update of the software. QyScore® algorithms have been mainly validated on open-source research cohorts, and further studies are ongoing in a clinical routine setting. A direct comparison with other medical devices, test–retest, and longitudinal data is also needed.

Conclusions

QyScore® provides reliable automatic segmentation of brain structures compared to the experts' consensus and other semi-automatic and automatic software described in the literature. Our results support the implementation of medical devices, such as QyScore® using neuroimaging methods, in clinical routine for supporting the diagnosis and monitoring of brain disorders.

QyScore® markers could also be implemented in clinical trials to test the efficacy of new drugs for neurological diseases. Reliable measures of brain atrophy and white matter hyperintensities, such as the ones provided by QyScore®, will help us move into more personalized and evidence-based medicine for neurological diseases. Further retrospective and prospective validation studies, as well as test–retest reliability studies, are currently ongoing on real-world data to demonstrate the diagnostic performance and the clinical impact of using QyScore® in neurological disorders.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08385-9.

Acknowledgements Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Part of the data used in the present manuscript were provided in part by OASIS: cross-sectional—principal investigators D. Marcus, R. Buckner, J. Csernansky, J. Morris, P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI—a public-private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including (list the full names of all of the PPMI funding partners found at www.ppmi-info.org/fundingpartners). Data collection and sharing for this project was funded by the Frontotemporal Lobar Degeneration Neuroimaging Initiative (National Institutes of Health Grant R01 AG032306). The study is coordinated through the University of California, San Francisco, Memory and Aging Center. FTLDNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. The ANMerge update of the AddNeroMed data has received partial support from the Innovative Medicines Initiative Joint Undertaking "AETIONOMY" under grant agreement no. 115568, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. The update of the original AddNeuroMed data was conducted by Colin Birkenbihl, Sarah Westwood, Liu Shi, Alejo Nevado-Holgado, EricWestman, Simon Lovestone, and Martin Hofmann-Apitius. Part of the data used in the present manuscript come from the REACTIV Study; we acknowledge Dr. Delphine Lamargue-Hamel, principal investigator of the REACTIV study, and Dr. Mathilde Deloire, clinical research coordinator and the professor Aurélie Ruet team leader.

Data used in preparation of this article were obtained from the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI) database. The investigators at NIFD/FTLDNI contributed to the design and implementation of FTLDNI and/or provided data, but did not participate in analysis or writing of this report (unless otherwise listed). The FTLDNI investigators included the following individuals:

Howard Rosen; University of California, San Francisco (PI)

Bradford C. Dickerson; Harvard Medical School and Massachusetts General Hospital

Kimoko Domoto-Reilly; University of Washington School of Medicine

David Knopman; Mayo Clinic, Rochester

Bradley F. Boeve; Mayo Clinic Rochester

Adam L. Boxer; University of California, San Francisco

John Kornak; University of California, San Francisco

Bruce L. Miller; University of California, San Francisco

William W. Seeley; University of California, San Francisco

Maria-Luisa Gorno-Tempini; University of California, San Francisco

Scott McGinnis; University of California, San Francisco

Maria Luisa Mandelli; University of California, San Francisco

Funding The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Enrica Cavedo.

Conflict of interest EC, PT, and CLS are employees of Qynapse. UT, JBM, and AM are not currently employees of Qynapse. This work has been performed during their previous position at Qynapse. FG, PKS, and FC served as consultants/advisors for Qynapse. CD, FG, DH, NP, SS, and DD have nothing to disclose.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional review board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in previous studies for the all the public available cohorts cited in the Manuscript.

Methodology
- retrospective
- cross-sectional study
- multicenter study

References

1. DiLuca M, Olesen J (2014) The Cost of brain diseases: a burden or a challenge? Neuron 82:1205–1208. https://doi.org/10.1016/j.neuron.2014.05.044
2. Feigin VL, Nichols E, Alam T et al (2019) Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet

Neurol 18:459–480. https://doi.org/10.1016/S1474-4422(18)30499-X
3. Lockhart SN, DeCarli C (2014) Structural imaging measures of brain aging. Neuropsychol Rev 24:271–289. https://doi.org/10.1007/s11065-014-9268-3
4. Dill V, Franco AR, Pinho MS (2015) Automated methods for hippocampus segmentation: the evolution and a review of the state of the art. Neuroinformatics 13:133–150. https://doi.org/10.1007/s12021-014-9243-4
5. Teipel SJ, Grothe M, Lista S, Toschi N, Garaci FG,Hampel H (2013) Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease. Med Clin North Am 97:399–424. https://doi.org/10.1016/j.mcna.2012.12.013
6. Egger C, Opfer R, Wang C et al (2017) MRI FLAIR lesion segmentation in multiple sclerosis: does automated segmentation hold up with manual annotation? Neuroimage Clin 13:264–270. https://doi.org/10.1016/j.nicl.2016.11.020
7. Jessen F, Hampel H (2009) MRI as a surrogate marker in clinical trials in Alzheimer's disease. Oxford University Press
8. Albert MS, DeKosky ST, Dickson D et al (2011) 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 7:270–279. https://doi.org/10.1016/j.jalz.2011.03.008
9. Dubois B, Chupin M, Hampel H et al (2015) Donepezil decreases annual rate of hippocampal atrophy in suspected prodromal Alzheimer's disease. Alzheimers Dement 11:1041–1049. https://doi.org/10.1016/j.jalz.2014.10.003
10. Petersen RC, Aisen PS, Beckett LA et al (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI). Neurology 74:201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25
11. Marek K, Chowdhury S, Siderowf A et al (2018) The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. Ann Clin Transl Neurol 5:1460–1477. https://doi.org/10.1002/acn3.644
12. Frisoni GB, Henneman WJP, Weiner MW et al (2008) The pilot European Alzheimer's Disease Neuroimaging Initiative of the European Alzheimer's Disease Consortium. Alzheimers Dement 4:255–264. https://doi.org/10.1016/j.jalz.2008.04.009
13. Cavedo E, Redolfi A, Angeloni F et al (2014) The Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI): validation of structural MR imaging. J Alzheimers Dis JAD 40:941–952. https://doi.org/10.3233/JAD-132666
14. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL (2007) Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 19:1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498
15. Landman BA, Huang AJ, Gifford A et al (2011) Multi-parametric neuroimaging reproducibility: a 3-T resource study. Neuroimage 54:2854–2866. https://doi.org/10.1016/j.neuroimage.2010.11.047
16. Birkenbihl C, Westwood S, Shi L et al (2020) ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. J Alzheimers Dis 79(1):423–431
17. Lesjak Ž, Galimzianova A, Koren A et al (2018) A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics 16:51–63. https://doi.org/10.1007/s12021-017-9348-7
18. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) 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 34:939–944. https://doi.org/10.1212/wnl.34.7.939
19. Rascovsky K, Hodges JR, Knopman D et al (2011) Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain J Neurol 134:2456–2477. https://doi.org/10.1093/brain/awr179
20. Polman CH, Reingold SC, Banwell B et al (2011) Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 69:292–302. https://doi.org/10.1002/ana.22366
21. Pini L, Pievani M, Bocchetta M et al (2016) Brain atrophy in Alzheimer's disease and aging. Ageing Res Rev 30:25–48. https://doi.org/10.1016/j.arr.2016.01.002
22. Teipel S, Kilimann I, Thyrian JR, Kloppel S, Hoffmann W (2018) Potential role of neuroimaging markers for early diagnosis of dementia in primary care. Curr Alzheimer Res 15:18–27. https://doi.org/10.2174/1567205014666170908093846
23. Delgado-Alvarado M, Gago B, Navalpotro-Gomez I, Jiménez-Urbieta H, Rodriguez-Oroz MC (2016) Biomarkers for dementia and mild cognitive impairment in Parkinson's disease. Mov Disord 31:861–881. https://doi.org/10.1002/mds.26662
24. Louapre C, Bodini B, Lubetzki C, Léoraha F, Brunoa S (2017) Imaging markers of multiple sclerosis prognosis. Curr Opin Neurol 30:231–236. https://doi.org/10.1097/WCO.0000000000000456
25. Guo C, Ferreira D, Fink K, Westman E, Granberg T (2019) Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. Eur Radiol 29:1355–1364. https://doi.org/10.1007/s00330-018-5710-x
26. Chupin M, Mukuna-Bantumbakulu AR, Hasboun D et al (2007) Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. Neuroimage 34:996–1019. https://doi.org/10.1016/j.neuroimage.2006.10.035
27. Chupin M, Hammers A, Liu RSN et al (2009) Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. Neuroimage 46:749–761. https://doi.org/10.1016/j.neuroimage.2009.02.013
28. Samaille T, Fillon L, Cuingnet R et al (2012) Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation. PLoS One 7:e48953. https://doi.org/10.1371/journal.pone.0048953
29. Fischl B (2012) FreeSurfer. Neuroimage 62:774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021
30. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23:903–921. https://doi.org/10.1109/TMI.2004.828354
31. Schmidt P, Gaser C, Arsic M et al (2012) An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. Neuroimage 59:3774–3783. https://doi.org/10.1016/j.neuroimage.2011.11.032
32. Guindon B, Zhang Y (2017) Application of the dice coefficient to accuracy assessment of object-based image classification. Can J Remote Sens 43:48–61. https://doi.org/10.1080/07038992.2017.1259557
33. Morey RA, Petty CM, Xu Y et al (2009) A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45:855–866. https://doi.org/10.1016/j.neuroimage.2008.12.033
34. Commowick O, Istace A, Kain M et al (2018) Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci Rep 8:13650. https://doi.org/10.1038/s41598-018-31911-7
35. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302. https://doi.org/10.2307/1932409

36. Schmidt MF, Storrs JM, Freeman KB et al (2018) A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. Hum Brain Mapp 39:2500–2513. https://doi.org/10.1002/hbm.24017
37. Akudjedu TN, Nabulsi L, Makelyte M et al (2018) A comparative study of segmentation techniques for the quantification of brain subcortical volume. Brain Imaging Behav 12:1678–1695. https://doi.org/10.1007/s11682-018-9835-y
38. Valverde S, Oliver A, Cabezas M, Roura E, Lladó X (2015) Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. J Magn Reson Imaging 41:93–101. https://doi.org/10.1002/jmri.24517
39. Tsang O, Gholipour A, Kehtarnavaz N, Gopinath K, Briggs R, Panahi I (2008) Comparison of tissue segmentation algorithms in neuroimage analysis software tools. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2008:3924–3928. https://doi.org/10.1109/IEMBS.2008.4650068
40. Kazemi K, Noorizadeh N (2014) Quantitative comparison of SPM, FSL, and Brainsuite for brain MR image segmentation. J Biomed Phys Eng 4:13–26
41. Mendrik AM, Vincken KL, Kuijf HJ et al (2015) MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput Intell Neurosci 2015:813696. https://doi.org/10.1155/2015/813696
42. Hosseini M-P, Nazem-Zadeh M-R, Pompili D, Jafari-Khouzani K, Elisevich K, Soltanian-Zadeh H (2016) Comparative performance evaluation of automated segmentation methods of hippocampus from magnetic resonance images of temporal lobe epilepsy patients: comparative performance evaluation of automated segmentation of hippocampus. Med Phys 43:538–553. https://doi.org/10.1118/1.4938411
43. Doring TM, Kubo TTA, Cruz LCH et al (2011) Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. J Magn Reson Imaging 33:565–572. https://doi.org/10.1002/jmri.22473
44. Cherbuin N, Anstey KJ, Réglade-Meslin C, Sachdev PS (2009) In vivo hippocampal measurement and memory: a comparison of manual tracing and automated segmentation in a large community-based sample. PLoS One 4:e5265. https://doi.org/10.1371/journal.pone.0005265
45. Tae WS, Kim SS, Lee KU, Nam E-C, Kim KW (2008) Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. Neuroradiology 50:569–581. https://doi.org/10.1007/s00234-008-0383-9
46. Sánchez-Benavides G, Gómez-Ansón B, Sainz A, Vives Y, Delfino M, Peña-Casanova J (2010) Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. Psychiatry Res Neuroimaging 181:219–225. https://doi.org/10.1016/j.pscychresns.2009.10.011
47. Storelli L, Rocca MA, Pagani E et al (2018) Measurement of whole-brain and gray matter atrophy in multiple sclerosis: assessment with MR imaging. Radiology 288:554–564. https://doi.org/10.1148/radiol.2018172468
48. Chupin M, Gérardin E, Cuingnet R et al (2009) Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19:579–587. https://doi.org/10.1002/hipo.20626
49. Wenger E, Mårtensson J, Noack H et al (2014) Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. Hum Brain Mapp 35:4236–4248. https://doi.org/10.1002/hbm.22473
50. Colliot O, Chételat G, Chupin M et al (2008) Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248:194–201. https://doi.org/10.1148/radiol.2481070876
51. Jain S, Sima DM, Ribbens A et al (2015) Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. Neuroimage Clin 8:367–375. https://doi.org/10.1016/j.nicl.2015.05.003
52. Lötjönen JMP, Wolz R, Koikkalainen JR et al (2010) Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49:2352–2365. https://doi.org/10.1016/j.neuroimage.2009.10.026
53. Lötjönen J, Wolz R, Koikkalainen J et al (2011) Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. Neuroimage 56:185–196. https://doi.org/10.1016/j.neuroimage.2011.01.062
54. Brewer JB, Magda S, Airriess C, Smith ME (2009) Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer disease. AJNR Am J Neuroradiol 30:578–580. https://doi.org/10.3174/ajnr.A1402
55. Pemberton HG, Zaki LAM, Goodkin O et al (2021) Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis—a systematic review. Neuroradiology. https://doi.org/10.1007/s00234-021-02746-3
56. Smeets D, Ribbens A, Sima DM et al (2016) Reliable measurements of brain atrophy in individual patients with multiple sclerosis. Brain Behav 6:e00518. https://doi.org/10.1002/brb3.518
57. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6:67–77. https://doi.org/10.1038/nrneurol.2009.215
58. Marciniewicz E, Bladowska J, Podgórski P, Sąsiadek M (2019) The role of MR volumetry in brain atrophy assessmentin multiple sclerosis: a review of the literature. Adv Clin Exp Med 28:989–999. https://doi.org/10.17219/acem/94137
59. Schott JM, Bartlett JW, Barnes J et al (2010) Reduced sample sizes for atrophy outcomes in Alzheimer's disease trials: baseline adjustment. Neurobiol Aging 31:1452-1462.e2. https://doi.org/10.1016/j.neurobiolaging.2010.04.011
60. De Stefano N, Giorgio A, Battaglini M et al (2010) Assessing brain atrophy rates in a large population of untreated multiple sclerosis subtypes. Neurology 74:1868–1876. https://doi.org/10.1212/WNL.0b013e3181e24136
61. Pemberton HG, Goodkin O, Prados F et al (2021) Automated quantitative MRI volumetry reports support diagnostic interpretation in dementia: a multi-rater, clinical accuracy study. Eur Radiol. https://doi.org/10.1007/s00330-020-07455-8
62. MAGNIMS Study Group, Vrenken H, Jenkinson M et al (2013) Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis. J Neurol 260:2458–2471. https://doi.org/10.1007/s00415-012-6762-5