NeuroImage: Clinical 20 (2018) 374–379

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Automatic normative quantification of brain tissue volume to support the diagnosis of dementia: A clinical evaluation of diagnostic accuracy

Meike W. Vernooij[a,b,*,1], Bas Jasperse[a,1], Rebecca Steketee[a], Marcel Koek[a,c], Henri Vrooman[a,c], M. Arfan Ikram[a,b,d], Janne Papma[d], Aad van der Lugt[a], Marion Smits[a], Wiro J. Niessen[a,c,e]

[a] Radiology and Nuclear Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[b] Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[c] Medical informatics, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[d] Neurology of Erasmus, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[e] Faculty of Applied Sciences, Delft University of Technology, Delft, the Netherlands

A B S T R A C T

Objectives: To assesses whether automated brain image analysis with quantification of structural brain changes improves diagnostic accuracy in a memory clinic setting.
Methods: In 42 memory clinic patients, we evaluated whether automated quantification of brain tissue volumes, hippocampal volume and white matter lesion volume improves diagnostic accuracy for Alzheimer's disease (AD) and frontotemporal dementia (FTD), compared to visual interpretation. Reference data were derived from a dementia-free aging population (n = 4915, aged > 45 years), and were expressed as age- and sex-specific percentiles. Experienced radiologists determined the most likely imaging-based diagnosis based on structural brain MRI using three strategies (visual assessment of MRI only, quantitative normative information only, or a combination of both). Diagnostic accuracy of each strategy was calculated with the clinical diagnosis as the reference standard.
Results: Providing radiologists with only quantitative data decreased diagnostic accuracy both for AD and FTD compared to conventional visual rating. The combination of quantitative with visual information, however, led to better diagnostic accuracy compared to only visual ratings for AD. This was not the case for FTD.
Conclusion: Quantitative assessment of structural brain MRI combined with a reference standard in addition to standard visual assessment may improve diagnostic accuracy in a memory clinic setting.

1. Introduction

Dementia is a clinical syndrome caused by various brain diseases, of which Alzheimer's disease (AD), vascular dementia and frontotemporal dementia (FTD) are most frequent (Ferri et al., 2005). In early onset dementia AD is the most common cause (approximately 34%) but FTD is also relatively prevalent (12%) (Van der Flier and Scheltens, 2005), which frequently causes a clinical diagnostic dilemma. Dementia is diagnosed clinically as a cognitive disorder interfering with activities of daily life according to core clinical criteria. Imaging biomarkers, cognitive profiling, genetic information, and cerebrospinal fluid (CSF) biochemistry features provide supportive evidence for differential diagnosis (Dubois et al., 2007; McKhann et al., 2011). Yet, the NIA-AA criteria (McKhann et al., 2011) for AD diagnosis only have a sensitivity of 65% for distinguishing probable AD from FTD, with considerable overlap in clinical symptoms, especially in early disease stages (Harris et al., 2015). Early and accurate identification of dementia's underlying causes is important for proper and tailored patient management as well as upcoming disease-modifying treatment options (Ballard et al., 2011; Mattila et al., 2012).

By visualizing structural brain changes associated with specific pathological substrates, Magnetic Resonance imaging (MRI) plays an important role in dementia diagnosis and subtype differentiation (Vernooij and Smits, 2012). MRI interpretation in dementia diagnosis can be challenging, as early brain abnormalities may be difficult to detect visually, especially in early stages of the disease. Additionally, brain changes due to a neurodegenerative disorder may be difficult to distinguish from those related to normal aging.

One way to potentially improve diagnostic accuracy and confidence is to quantify brain structures from an individual patient and compare these to age- and sex-specific reference data from a healthy population (Brewer, 2009; Ross et al., 2015). Although several MR brain quantification methods are now becoming available and gradually finding their way into clinical applications (Ross et al., 2015; Brewer et al., 2009; Ross et al., 2013), there is no clear concept on how they should be implemented in radiology reading or reporting practice. Whether

* Corresponding author at: Department of Radiology and Nuclear Medicine, Room Nd-546, Erasmus MC, Rotterdam, the Netherlands.
E-mail address: m.vernooij@erasmusmc.nl (M.W. Vernooij).
[1] Both authors contributed equally to this manuscript.

https://doi.org/10.1016/j.nicl.2018.08.004
Received 26 February 2018; Received in revised form 18 July 2018; Accepted 3 August 2018
Available online 09 August 2018

quantitative information improves diagnostic accuracy, and, if so, it can be used in isolation or should be considered together with other imaging information is not known.

In this study, we implemented automated quantification of brain tissue volumes, hippocampal volumes and white matter lesion volumes in our memory clinic and compared these volumes to population reference data. Our aim was to compare three different strategies, namely visual rating of brain MR scans only, quantitative normative assessment only, and a combination of both visual rating and quantitative assessment, to the reference standard of multidisciplinary clinical consensus diagnosis, and to assess diagnostic agreement of these strategies between two observers.

2. Materials and methods

2.1. Patient population

Between December 2009 and September 2011, all new patients who visited our memory clinic and who (Ferri et al., 2005) underwent MRI as part of clinical work up, and (Van der Flier and Scheltens, 2005) received a clinical diagnosis of AD, FTD or MCI, were eligible for this retrospective study. Our memory clinic is specialized in early onset dementia, hence we see a higher proportion of rare dementias (such as FTD) and patients with early disease onset. A total of 42 patients were eligible, 21 patients with AD, 15 with FTD, and 6 with MCI. The clinical diagnosis was based on expert panel consensus using standard diagnostic criteria (McKhann et al., 2011; Rascovsky et al., 2011) and all available information, including neuropsychological information, brain MRI, CSF (if available) and neurological examination.

Brain MRI scans were acquired at 3.0 T (GE Healthcare, US), according to a standardized protocol, including sagittal 3D T1-weighted (T1w) inversion recovery (IR) fast spoiled gradient recalled echo (FSPGR) scans with axial and coronal reconstructions (perpendicular to the long axis of the hippocampus); fluid attenuated inversion recovery scans (FLAIR); and T2w scans. Supplementary Table 1 provides all relevant MRI parameters.

2.2. Reference population

Reference data were obtained from 4915 non-demented participants (mean age 64 yrs., range 45.7–100.0) from a population-based longitudinal study among community dwelling subjects (Hofman et al., 2015; Ikram et al., 2015)

All scans were acquired on a single 1.5 T MR imaging system (GE Healthcare, US). The imaging protocol (Supplementary Table 1) included a 3D T1w IR-FSPGR, a proton density (PD)–weighted sequence and a FLAIR sequence. The PD sequence was applied with a long TR, resulting in bright CSF as in T2w images.

2.3. Brain tissue, white matter lesion and hippocampal volume quantification

Gray matter (GM), white matter (WM), white matter lesions (WML) and CSF segmentation was performed with a fully automated method (Vrooman et al., 2007) extended with WML segmentation (de Boer et al., 2009). This involved the segmentation of CSF, GM, and WM by an atlas-based k-nearest neighbor classifier on the MRI data. The classifier was trained by registering brain atlases to the subjects (Vrooman et al., 2007). The GM classification was then used to determine a WML intensity threshold value in a FLAIR scan. Applying this threshold to the FLAIR scan yielded the WML segmentation (de Boer et al., 2009). Total brain volume was calculated by summing WM, GM and WML volumes. Intracranial volume was defined as the sum of total brain volume and CSF volumes. T1w scans were processed using FreeSurfer (4.5.0) to obtain hippocampal volumes (Dale et al., 1999; Desikan et al., 2006).

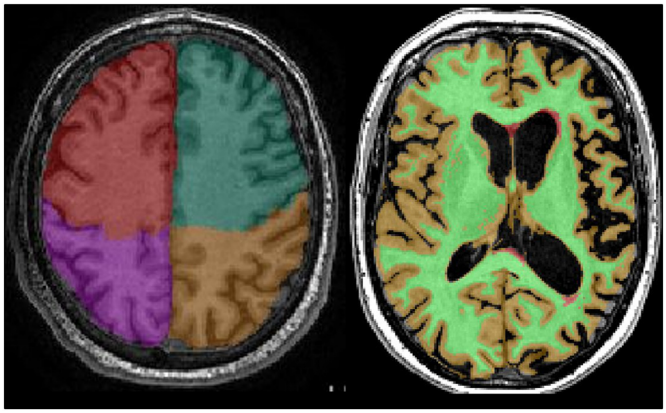

Fig. 1. Brain segmentation.

2.4. Lobar volume quantification

To obtain lobar brain volumes, a multi-atlas approach was used (Vibha et al., 2018). Six template scans (atlases) were created in which the frontal, parietal, temporal and occipital lobes of the left and right hemisphere were outlined (Bokde et al., 2005). These atlases were non-rigidly registered (Klein et al., 2010) to a subject brain MRI and labels were assigned to each voxel using majority voting. By combining this lobar mask with the original tissue segmentation, volumes for each brain lobe were calculated (Ikram et al., 2010). Fig. 1 provides an example of the atlas and segmentation results. Intracranial volume (ICV) was used to correct for inter-individual differences in head size, by dividing each volume by ICV in each subject.

2.5. Visual inspection

All patient tissue and lobar segmentation results were visually checked for segmentation errors, revealing no substantial errors. No manual corrections were performed, as this would ultimately hamper translating the workflow to clinical practice.

For the 4945 reference subjects, outliers (defined as 2.0 standard deviations from the mean) were found for total brain (n = 134), white matter lesion (n = 66) and hippocampal volumes (n = 172). Outliers were visually checked and if caused by segmentation errors, bad scan quality or significant structural abnormalities, scans were excluded (n = 30) resulting in 4915 scans for creating reference curves.

2.6. Reference curves

Age- and sex-specific percentile curves were generated for each quantitative parameter (total brain, lobar brain, hippocampal and WML volumes) using the LMS method (Cole and Green, 1992). Percentile curves (Fig. 2) were generated using the VGAM (1.0–0) package for R (3.2.3).

For each patient, the age-appropriate percentile value, referred to as “Volume percentile” (Vperc) was calculated for each of the brain volumes, and plotted on the reference curves.

2.7. Rating strategies

Two experienced neuro-radiologists (M.S. and M.W.V., each with more than three years of experience in reading memory clinic scans), blinded to all patient characteristics except age and sex, independently provided an imaging-based diagnosis. To reflect a realistic clinical scenario, the raters selected a diagnosis from three categories: AD, FTD, or alternative diagnosis (including no dementia). They were unaware of the proportion of AD, FTD and MCI in the sample.

We assessed three diagnostic strategies:

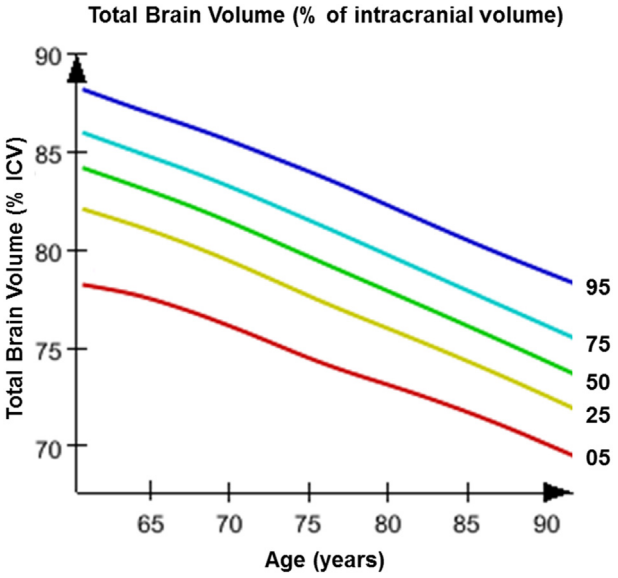

Fig. 2. Percentile curve for total brain volume.

Firstly, a visual interpretation of the brain MR imaging scans was performed. Raters interpreted patterns of atrophy and presence of vascular lesions using the 3D T1w FSPGR (including coronal reformats), the T2w, and the T2w-FLAIR sequences by applying standardized visual rating scales such as the global cortical atrophy scale and Koedam scale for lobar atrophy, the medial temporal atrophy scale for hippocampal atrophy, and the Fazekas scale for WML (Scheltens et al., 1998; Scheltens et al., 1995; Pasquier et al., 1996; Koedam et al., 2011). Each rater independently based their final diagnosis on the combination of these visual ratings.

Secondly, both raters were provided with Vperc only and provided a diagnosis solely based on these. Raters were left free how to interpret the Vperc (no cut-off values prescribed). As quantitative normative assessment is an evolving concept, we made a deliberate choice not to provide the raters with directions or cut-off values, to be able to assess the effect of having quantitative information available for diagnosis. Additionally, assessing relative values, i.e. Vperc of one structure compared to other structures, is equally important as applying absolute cut-offs to separate regions.

Thirdly, the raters reviewed the brain MRI together with the associated Vperc to come to a diagnosis.

The above strategies were each separated by three months, with patient identification numbers altered to ensure that current assessments could not be related to previous assessments.

2.8. Statistical analysis

For all combinations of assessment strategy and rater, diagnostic accuracy for AD and FTD diagnosis was determined as the sum of the true positive and negative cases divided by the total number of cases. Differences in accuracies between strategies for each diagnosis and for each rater were assessed with McNemar tests. Inter-rater agreement per strategy was calculated using Cohen's κ. In addition to the cross-sectional analysis, we also used follow up information (mean follow up 2.8 yrs., range 0–6.1 years) for possible change in clinical diagnosis and recalculated diagnostic accuracies. Finally, to assess the performance of subjective interpretation of the quantitative information by the clinicians (i.e. without specific cut-offs) in comparison to the use of absolute cut-offs, we determined optimal cut-off values to discriminate between diagnoses, based on Vperc of relevant brain regions (MCI versus AD & FTD based on hippocampal Vperc; FTD versus AD & MCI based on frontal and temporal Vperc, and AD versus FTD & MCI based on hippocampal and parietal Vperc). For each cut-off point, we calculated the distance from the maximum sensitivity and specificity as follows: distance = $\sqrt{[(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2]}$, and subsequently located the point where distance was minimal. We compared the diagnostic accuracy at this optimal value with the performance of both raters.

We separately assessed the correlation between visual rating of WML burden (Fazekas score (Scheltens et al., 1998)) and automated quantification. Spearman rank correlation was calculated between the raters' Fazekas scores and total WML volume (% of ICV) and between Fazekas scores and the WML Vperc.

Statistical analyses of diagnostic accuracy (i.e. of assessment strategy; rater; and optimal cut-off values) and of inter-rater agreement were performed with SPSS version 22. To assess differences in accuracies between strategies for each diagnosis and for each rater, Python version 2.7.11+ and the McNemar implementation of the statsmodels python package (version 0.6.1) were used. An α of 0.05 was considered as threshold for statistical significance.

3. Results

Table 1 shows patient characteristics. Mean age of AD patients was 66.1 +/− 8.5 years, and 43% were female; FTD patients were on average younger (60.0 +/− 6.5 years), and 40% were female.

Tables 2 shows regional brain and lesion volumes expressed as percentage of intracranial volume (%ICV), and age appropriate Vperc, respectively. Differences in regional brain volumes between patient groups were more evident when expressed as Vperc than as %ICV. These differences were evident in brain regions that are known to be affected in AD and FTD (i.e. hippocampus, frontal and temporal lobes).

Table 3 shows diagnostic accuracy for AD and FTD of both observers and all three strategies. Table 4 shows the results of the comparisons of accuracy between strategies (McNemar test). Neither for AD nor FTD did the use of quantitative information alone improve diagnostic accuracy compared to visual assessment only. Moreover, for FTD this strategy led to a significantly worse accuracy ($p =/ < 0.01$ for both raters). For AD, diagnostic accuracy improved with the combined visual/quantitative strategy compared to visual assessment only (73.8% for both raters compared to 59.5% in rater A ($p = .03$) and 66.7% in rater B ($p = .45$) for visual only strategy). For FTD, visual assessment only performed slightly better than the combined visual/quantitative strategies, but this difference was not statistically significant ($p = .4$); diagnostic accuracy was high with both strategies.

During a mean follow up of 2.8 years (0.0–6.1 years), 7 of the 42 initial diagnoses had changed. This did not lead to a change in accuracy of any of the strategies (Supplementary Table 2).

Supplementary Table 3 shows the comparison of automated classification accuracy using optimal cut-off values with rater classification. Especially for FTD using these optimal cut-off values improves performance. The correlation between Fazekas score (0–3) and total WML

Table 1
Patient characteristics.

	MCI ($N = 6$)	AD ($N = 21$)	FTD ($N = 15$)	Total ($N = 42$)
Age in y, mean (SD)	63.2 (8.5)	66.1 (8.5)	60.0 (6.5)	63.5 (8.1)
Gender male:female	2:4	12:9	9:6	23:19
Duration in y, mean (SD)	1.1 (1.8–3.6)	2.1 (1.1–3.3)	2.1 (1.5–3.0)	2.1 (1.3–3.2)
MMSE, median (IQR)	25.5 (23.0–26.8)	24.0 (22.0–25.0)	25.5 (22.5–29.0)	24.5 (22.3–27.0)

MCI = mild cognitive impairment, AD = Alzheimer's disease, FTD = frontotemporal dementia, SD = standard deviation, IQR = interquartile range.

Table 2
Median (IQR) of brain, lobar, hippocampal and white matter lesion volumes expressed as percentage of intracranial volume (%ICV) and as volume percentiles (Vperc) for the diagnosis groups.

Volume	%ICV			Vperc		
	MCI	AD	FTD	MCI	AD	FTD
Total brain	84.8 (84.1–86.6)	83.6 (82.5–84.9)	82.0 (80.9–84.3)	86.6 (52.5–98.2)	73.8 (25.6–88.5)	16.2 (4.2–61.0)
WML	1.0 (0.3–1.7)	1.1 (0.3–2.0)	0.4 (0.2–0.9)	78.0 (67.9–95.2)	88.5 (75.4–95.0)	82.4 (54.2–96.3)
Frontal right	14.2 (14.0–14.4)	13.8 (13.4–14.3)	13.2 (12.5–14.3)	30.4 (21.8–33.1)	10.6 (5.1–47.4)	0.9 (0.0–46.9)
Frontal left	14.3 (14.2–14.5)	14.0 (13.3–14.9)	13.2 (12.0–14.9)	29.8 (13.7–42.4)	19.8 (3.8–73.9)	0.6 (0.0–68.8)
Temporal right	8.6 (8.4–9.3)	8.3 (7.9–8.5)	7.9 (7.9–8.9)	63.6 (26.0–96.5)	19.2 (4.8–47.1)	4.9 (2.3–74.2)
Temporal left	8.3 (8.0–8.4)	7.7 (7.5–7.9)	7.4 (6.9–7.9)	89.3 (57.2–94.1)	29.9 (17.6–50.6)	6.4 (0.2–39.2)
Parietal right	8.7 (8.6–9.0)	8.8 (8.4–9.1)	8.9 (8.1–9.4)	78.8 (60.1–95.5)	82.0 (51.3–98.0)	93.1 (25.3–99.8)
Parietal left	9.3 (9.1–9.4)	9.3 (9.1–9.7)	9.7 (9.4–10.1)	81.8 (64.2–90.4)	78.7 (49.1–98.4)	96.8 (83.0–99.7)
Occipital right	5.5 (5.2–5.8)	5.3 (5.2–5.7)	5.4 (5.2–6.0)	69.1 (45.4–91.3)	60.4 (37.6–91.9)	59.2 (32.5–98.3)
Occipital left	5.4 (5.2–5.7)	5.3 (5.0–5.5)	5.5 (5.3–5.8)	85.9 (62.0–98.6)	72.7 (29.0–86.8)	88.5 (72.1–98.6)
Right hippocampus	0.357 (0.310–0.388)	0.282 (0.258–0.322)	0.301 (0.276–0.338)	28.3 (7.9–69.8)	2.28 (0.7–12.4)	2.7 (0.6–14.6)
Left hippocampus	0.363 (0.262–0.402)	0.290 (0.247–0.328)	0.290 (0.263–0.313)	32.3 (0.7–63.5)	3.2 (0.4–14.0)	1.8 (0.2–6.1)

IQR = interquartile range
%ICV = percentage of intracranial volume.
Vperc = volume percentile (age appropriate percentile value calculated for each of the volumes, plotted on the reference curves).
MCI = mild cognitive impairment, AD = Alzheimer's disease, FTD = frontotemporal dementia.
WML = white matter lesion.

Table 3
Accuracy of the different rating scenarios per observer.

Observer	Disease	Scenario	TP + TN/all	Accuracy
A	AD	Visual only	25/42	59.5
		Quantitative only	22/42	52.4
		Combined	31/42	73.8
	FTD	Visual only	38/42	90.5
		Quantitative only	28/42	66.7
		Combined	35/42	83.3
B	AD	Visual only	28/42	66.7
		Percentiles only	26/42	61.9
		Combined	31/42	73.8
	FTD	Visual only	36/42	85.7
		Combined	33/42	78.6

AD = Alzheimer's disease; FTD = frontotemporal dementia; TP = true positive; TN = true negative.
Accuracy was calculated as (TP + TN/all subjects)*100%.

Table 4
Comparison of the three different strategies.

		AD		FTD	
Observer	Comparison	Higher accuracy for	P*	Higher accuracy for	P*
A	visual vs Vperc	Visual	0.55	Visual	0.01
A	visual vs combined	Combined	0.03	Visual	0.4
A	Vperc vs combined	Combined	0.02	Combined	0.04
B	visual vs Vperc	Visual	0.8	Visual	0.002
B	visual vs combined	Combined	0.45	Visual	0.4
B	Vperc vs combined	Combined	0.27	Combined	0.02

Abbreviations: Vperc = Volume percentile; * = *p*-value for comparison of diagnostic accuracy between two strategies (Mc Nemar test).

volume expressed as %ICV was high, with Spearman correlation coefficient of 0.75 (for both raters, $p < .01$). This correlation dropped to 0.57 ($p < .01$) for Fazekas and Vperc WML, with most variation in Vperc for the Fazekas scores of 0 and 1 (Supplementary Fig. 1).

Interrater agreement between both observers was 69% (kappa 0.55, $p < .01$) when using visual assessment only, 62% (kappa 0.42, $p < .01$) when solely using quantitative information, and 67% (kappa 0.5, $p < .01$) when combining visual and quantitative assessment for diagnosis.

4. Discussion

In the setting of a memory clinic, we evaluated how adding quantitative volumetric brain data and population reference data affect the accuracy of radiologists' MR imaging-based dementia diagnosis. Providing experienced radiologists with only quantitative data significantly decreased diagnostic accuracy compared to conventional visual rating methods. Yet, the combination of quantitative data with visual rating of brain MR imaging suggested better diagnostic accuracy of AD, but not that of FTD.

Strengths of our study are the large dataset of reference subjects from the general population, enabling us to compare patient data to age and sex-specific normative volumetric data. The automated algorithms that were used can be easily implemented in a general clinical setting. We normalized for intracranial volume, as differences in head size would otherwise preclude a fair comparison between individuals. This is illustrated by our findings in WML: we found a high correlation between the visual Fazekas score and automated absolute WML volume, which decreased when WML were age- and sex-adjusted as volume

percentile. Lower visual WML scores in particular showed a wide range of variation in percentile WML load. Although this needs further investigation, it suggests that quantifying WML relative to normal aging may be more sensitive for identifying subjects with a 'higher than normal' relative WML load.

There are also limitations that need to be considered. Firstly, the sample size was modest. This is inherent to the nature of the sample, as it included only patients with early onset (< 65 years) dementia, which is less prevalent than late onset dementia. We specifically selected this sample because diagnosis in early onset dementia is much more challenging than in late onset dementia. A second limitation is that subjects in the reference population were all scanned on a single scanner using the same scan parameters. Patients were scanned using a different scanner, with different field strength and scan parameters. These differences may hamper comparison between subjects and application of absolute volume cut-off values. However, relative comparisons of volumes between regions within one patient will still be valid. Inter-scanner effects have been studied (Cover et al., 2011; Wolz et al., 2014; Opfer et al., 2016; Abdulkadir et al., 2011; Kruggel et al., 2010) and future studies should focus on developing quantitative markers that are robust to inter-scanner differences (Puonti et al., 2013; van Opbroek et al., 2015). Finally, as reference standard we used the clinical diagnosis based on established criteria and including the full clinical picture. Although this is the most optimal diagnosis in the setting of lack of pathological confirmation, the clinical diagnosis may still be wrong, especially in the early disease stages. We specifically investigated this issue by repeating analyses with available follow up data and found that diagnostic accuracy did not change substantially.

Another issue is that patients may have mixed or multiple pathologies, which is difficult to detect clinically, but may be detected better by volumetric quantification. Our current study design would not be able to show this potential advantage. Having a group of MCI patients as a control condition instead of cognitively normal subjects or 'healthy controls' may also have attenuated our ability to distinguish the three groups, since MCI is a heterogeneous group among which subjects may have brain changes that are in the spectrum of AD abnormalities. Yet, this composition of the patient group optimally reflects the clinical setting, as MCI is a very common alternative diagnosis in a memory clinic population.

Of greater importance than the absolute diagnostic performance is the comparison between the three strategies. For AD the best diagnostic strategy appears to be using quantitative information combined with visual inspection, providing the highest accuracy and highest interrater agreement. The addition of quantitative information to visual inspection may provide added value to the experienced rater, either by providing clues for interpreting the quantitative information or by directing attention to brain regions that may only show subtle changes on visual inspection. The added value of quantitative information was solely present for diagnosis of AD and not for FTD. This was rather unexpected, but may be due to patterns of atrophy being more visually obvious in FTD than in AD, even in the early stage of disease, as evidenced by the high diagnostic accuracy of visual inspection alone. At present, accuracy is not yet sufficient for clinical implementation (ranging from 52.4–66.7% when clinicians subjectively interpreted quantitative information only and from 73.8–83.3% when they combined quantitative and visual information). Longitudinal imaging may further improve performance of quantitative assessment, as the accelerated rate of atrophy associated with progression of the disease will probably be more evident in the quantitative information. Relative regional decreases in volume in particular will facilitate (differential) diagnosis (Mak et al., 2015; Scahill and Fox, 2007).

Still, the discrepancy between the value of quantitative information and visual information for diagnosis seems to be in contrast with other studies investigating the relationship between qualitative and quantitative assessment of MRI for dementia diagnosis. For example Harper et al. (Harper et al., 2016) found a high correlation between regional visual scales and voxel-based morphometry (VBM). Our study was however not limited to (disease specific) regions. Moreover, in the Harper study, VBM results were not corrected for age, which may have resulted in more exaggerated measures of volume loss, due to both aging and neurodegeneration, than in our study.

Although vascular dementia patients were not included, we evaluated the agreement between a qualitative and quantitative assessment of WML, which could be used in the context of diagnosing vascular dementia. The correlation between the visual Fazekas score and automated absolute WML volume was very high, but decreased when WML were age- and sex-adjusted (as Vperc). In particular the lower visual WML scores showed a wide range of variation in percentile WML load. Although this needs further investigation, this may indicate that quantitative evaluation of WML against the background of normal aging may be more sensitive to identifying subjects with a 'higher than normal' relative WML load.

Interestingly, dementia diagnosis based solely on quantitative information had poor accuracy as well as low concordance between raters, lower than based on visual inspection alone. We therefore also evaluated automated classification when using optimal thresholds of the quantitative image features, which did improve on rater accuracy. It should be noted that selecting optimal cut-points on the data causes overestimation of the performance. Nevertheless, it suggests that interpretation of percentile curves warrants new guidelines for interpretation, and more experience, to improve diagnostic accuracy. At present, our results suggest that quantitative image information should not be used as stand-alone information, without visual inspection of scans. Future studies should focus on providing cut-off values to determine 'significant atrophy' or guidelines on how to interpret the quantitative information, also to rule out training effects that may arise due to the novelty of the method. In the current study, we aimed to simulate the current clinical process, which uses hippocampal volume and lobar volume as the most important diagnostic imaging markers in dementia. Our objective was to investigate whether normative values for these structures improved or at least resembled the accuracy of the visual assessment. However, research literature has put forward several potentially specifically affected structures in neurodegeneration (e.g. entorhinal cortex and subcortical structures such as caudate and putamen), and providing percentiles of these structures could potentially be informative to the raters. This would however introduce additional hurdles for interpretation and an important learning curve. Nevertheless, our implementation does provide the opportunity to add more elaborate and refined features of brain volume loss, which may ultimately exceed visual rating performance. This process may be extended by exploiting many more image features or image information by employing e.g. machine learning or deep learning approaches, which have received increased interest in recent years and have the potential to improve subject classification. Future efforts could therefore be directed towards training diagnostic classifiers based on multiple imaging markers extracted from the reference data, or directly investigating which information in the reference imaging data is most informative for differential diagnosis. In a challenge comparing performance of computer-aided diagnosis algorithms to classify subjects into normal controls, MCI and AD it was shown that methods including more imaging biomarkers (e.g. hippocampal volume, shape and texture) performed best (Bron et al., 2015). Future research should focus on determining which (combination of) quantitative imaging biomarkers is most informative in computer-aided diagnosis of dementia. In view of our results it is to be expected that providing reference curves of such imaging biomarkers, in combination with visual assessment, will provide the most accurate diagnosis of AD.

In conclusion, this study indicates that age-appropriate percentile values of automatically quantified regional brain volumes may improve accuracy and inter-rater agreement of the radiological diagnosis of AD. Further studies should focus on overcoming the present technical limitations and on developing guidelines on the interpretation of such

quantitative biomarkers.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2018.08.004.

Funding

Wiro Niessen and Meike Vernooij were partially funded by the EU FP7 framework project VPH-Dare-IT (601055) and the Horizon 2020 project EuroPOND (666992).

Funding for this project was further provided by the Coolsingel Foundation ('Stichting Coolsingel') under project nr. 2012–86 ('Automatic digital assessment of brain scans').

Typical segmentation result for automated lobar segmentation (left panel, showing frontal and parietal lobes) and automated tissue segmentation (right panel, with white matter lesions indicated in red).

Example of percentile curve showing total brain volume for male subjects derived from the reference population. The curve shows the decrease of total brain volume (y-axis; expressed as percentage of intracranial volume) in relation to age (x-axis). The lines indicate different percentile lines (range from 5th to 95th percentile; with the green line indicating the 50th percentile line).

References

Abdulkadir, A., Mortamet, B., Vemuri, P., Jack Jr., C.R., Krueger, G., Klöppel, et al., 2011. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. NeuroImage 58, 785–792.

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., Jones, E., 2011. Alzheimer's disease. Lancet 377, 1019–1031.

Bokde, A.L., Teipel, S.J., Schwarz, R., Leinsinger, G., Buerger, K., Moeller, T., et al., 2005. Reliable manual segmentation of the frontal, parietal, temporal, and occipital lobes on magnetic resonance images of healthy subjects. Brain Res Brain Res Protoc. 14, 135–145.

Brewer, J.B., 2009. Fully-automated volumetric mri with normative ranges: translation to clinical practice. Behav. Neurol. 21, 21–28.

Brewer, J.B., Magda, S., Airriess, C., Smith, M.E., 2009. Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in alzheimer disease. AJNR Am. J. Neuroradiol. 30, 578–580.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., et al., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: the caddementia challenge. NeuroImage 111, 562–579.

Cole, T.J., Green, P.J., 1992. Smoothing reference centile curves: the lms method and penalized likelihood. Stat. Med. 11, 1305–1319.

Cover, K.S., van Schijndel, R.A., van Dijk, R.W., Redolfi, A., Knol, D.L., Firsoni, G.B., et al., 2011. Assessing the reproducibility of the SienaX and Siena brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. Psychiatry Res. 193, 182–190.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage. 9, 179–194.

de Boer, R., Vrooman, H.A., van der Lijn, F., Vernooij, M.W., Ikram, M.A., van der Lugt, A., et al., 2009. White matter lesion extension to automatic brain tissue segmentation on mri. NeuroImage 45, 1151–1161.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. NeuroImage 31, 968–980.

Dubois, B., Feldman, H.H., Jacova, C., Dekosky, S.T., Barberger-Gateau, P., Cummings, J., et al., 2007. Research criteria for the diagnosis of alzheimer's disease: revising the nincds-adrda criteria. Lancet Neurol. 6, 734–746.

Ferri, C.P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., et al., 2005. Global prevalence of dementia: a delphi consensus study. Lancet 366, 2112–2117.

Harper, L., Fumagalli, G.G., Barkhof, F., Scheltens, P., O'Brien, J.T., Bouwman, F., et al., 2016. MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases. Brain 139, 1211–1225.

Harris, J.M., Thompson, J.C., Gall, C., Richardson, A.M., Neary, D., du Plessis, D., et al., 2015. Do nia-aa criteria distinguish alzheimer's disease from frontotemporal dementia? Alzheimers Dement. 11, 207–215.

Hofman, A., Brusselle, G.G., Darwish Murad, S., van Duijn, C.M., Franco, O.H., Goedegebure, A., et al., 2015. The Rotterdam study: 2016 objectives and design update. Eur. J. Epidemiol. 30, 661–708.

Ikram, M.A., Vrooman, H.A., Vernooij, M.W., den Heijer, T., Hofman, A., Niessen, W.J., et al., 2010. Brain tissue volumes in relation to cognitive function and risk of dementia. Neurobiol. Aging 31, 378–386.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Koudstaal, P.J., Krestin, G.P., Hofman, A., et al., 2015. The Rotterdam scan study: design update 2016 and main findings. Eur. J. Epidemiol. 30, 1299–1315.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205.

Koedam, E.L., Lehmann, M., van der Flier, W.M., Scheltens, P., Pijnenburg, Y.A., Fox, N., et al., 2011. Visual assessment of posterior atrophy development of a mri rating scale. Eur. Radiol. 21, 2618–2625.

Kruggel, F., Turner, J., Muftuler, L.T., 2010. Alzheimer's disease neuroimaging initiative. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. NeuroImage 49, 2123–2133.

Mak, E., Su, L., Williams, G.B., Watson, R., Firbank, M., et al., 2015. Longitudinal assessment of global and regional atrophy rates in Alzheimer's disease and dementia with Lewy bodies. Neuroimage Clin. 7, 456–462.

Mattila, J., Soininen, H., Koikkalainen, J., Rueckert, D., Wolz, R., Waldemar, G., et al., 2012. Optimizing the diagnosis of early alzheimer's disease in mild cognitive impairment subjects. J. Alzheimers Dis. 32, 969–979.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr., C.R., Kawas, C.H., et al., 2011. 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. 7, 263–269.

Opfer, R., Suppa, P., Kepp, T., Spies, L., Schippling, S., Huppertz, H.J., et al., 2016. Atlas based brain volumetry: how to distinguish regional volume changes due to biological or physiological effects from inherent noise of the methodology. Magn. Reson. Imaging 34, 455–461.

Pasquier, F., Leys, D., Weerts, J.G., Mounier-Vehier, F., Barkhof, F., Scheltens, P., 1996. Inter- and intraobserver reproducibility of cerebral atrophy assessment on mri scans with hemispheric infarcts. Eur. Neurol. 36, 268–272.

Puonti, O., Iglesias, J.E., Van Leemput, K., 2013. Fast, sequence adaptive parcellation of brain mr using parametric models. Med Image Comput Comput Assist Interv. 16, 727–734.

Rascovsky, K., Hodges, J.R., Knopman, D., Mendez, M.F., Kramer, J.H., Neuhaus, J., et al., 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 134, 2456–2477.

Ross, D.E., Ochs, A.L., Seabaugh, J.M., Shrader, C.R., 2013. Alzheimer's disease neuroimaging I. Man versus machine: comparison of radiologists' interpretations and neuroquant(r) volumetric analyses of brain mris in patients with traumatic brain injury. J Neuropsychiatry Clin Neurosci. 25, 32–39.

Ross, D.E., Ochs, A.L., Desmit, M.E., Seabaugh, J.M., Havranek, M.D., 2015. Alzheimer's disease neuroimaging I. Man versus machine part 2: comparison of radiologists' interpretations and neuroquant measures of brain asymmetry and progressive atrophy in patients with traumatic brain injury. J Neuropsychiatry Clin Neurosci. 27, 147–152.

Scahill, R.I., Fox, N.C., 2007 Dec. Longitudinal imaging in dementia. Br J Radiol. 80, S92–S98. https://doi.org/10.1259/bjr/78981552. Spec No 2.

Scheltens, P., Launer, L.J., Barkhof, F., Weinstein, H.C., van Gool, W.A., 1995. Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: Interobserver reliability. J. Neurol. 242, 557–560.

Scheltens, P., Erkinjunti, T., Leys, D., Wahlund, L.O., Inzitari, D., del Ser, T., et al., 1998. White matter changes on ct and mri: an overview of visual rating scales. European task force on age-related white matter changes. Eur. Neurol. 39, 80–89.

Van der Flier, W.M., Scheltens, P., 2005. Epidemiology and risk factors of dementia. J Neurol Neurosurg Pyschiatry. 76 (Suppl. 5), v2–v7.

van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M., 2015. Transfer learning improves supervised image segmentation across imaging protocols. IEEE Trans. Med. Imaging 34, 1018–1030.

Vernooij, M.W., Smits, M., 2012. Structural neuroimaging in aging and alzheimer's disease. Neuroimaging Clin. N. Am. 22, 33–55 (vii-viii).

Vibha, D., Tiemeier, H., Mirza, S.S., Adams, H.H.H., Niessen, W.J., et al., 2018. Brain volumes and longitudinal cognitive change: a population-based study. Alzheimer Dis. Assoc. Disord. 32, 43–49.

Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., et al., 2007. Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. NeuroImage 37, 71–81.

Wolz, R., Schwarz, A.J., Yu, P., Cole, P.E., Ruckert, D., Jack Jr., C.R., et al., 2014. Robustness of automated hippocampal volumetry across magnetic resonance field strengths and repeat images. Alzheimers Dement. 10 430–438.e2.