Radiological classification of dementia from anatomical MRI assisted by machine learning-derived maps

Pierre Chagué, Béatrice Marro, Sarah Fadili, Marion Houot, Alexandre Morin, Jorge Samper-González, Paul Beunon, Lionel Arrivé, Didier Dormont, Bruno Dubois, et al.

Radiological classification of dementia from anatomical MRI assisted by machine learning-derived maps

Pierre Chagué[1,2,3]MD, Béatrice Marro[1], MD; Sarah Fadili[1];MD, Marion Houot[2], MSc; Alexandre Morin, MD[2,5]; Jorge Samper-González[2,3]; Paul Beunon[1]; Lionel Arrivé[1], MD; Didier Dormont, MD,[2,3,4]; Bruno Dubois, MD[2,5], Marc Teichmann, MD, PhD[2,5], Stéphane Epelbaum, MD, PhD[2,3,5] and Olivier Colliot, PhD[2,3,4,5]

[1] AP-HP, Hôpital Saint-Antoine, Department of radiology, Paris, France
[2.] Institut du Cerveau et de la Moelle épinière, ICM, Inserm, U 1127, CNRS, UMR 7225, Sorbonne Université, F-75013 Paris, France
[3] Inria, Aramis-project team, Paris, France
[4]AP-HP, Hôpital de la Pitié-Salpêtrière, Department of Neuroradiology, F-75013, Paris, France
[5] AP-HP, Hôpital de la Pitié-Salpêtrière, Department of Neurology, Institut de la Mémoire et de la Maladie d'Alzheimer (IM2A), F-75013, Paris, France

*Corresponding author:
Olivier Colliot
ICM – Brain and Spinal Cord Institute
ARAMIS team
Pitié-Salpêtrière Hospital
47-83, boulevard de l'Hôpital, 75651 Paris Cedex 13, France
E-mail: olivier.colliot@upmc.fr

Keywords:

Alzheimer's disease; Dementia; diagnosis; anatomical MRI; artificial intelligence

Acknowledgements

All our thoughts are with Anne Bertrand (MD, PhD). She passed away on March 2nd 2018. She was in a touring-skiers group led by a guide and swept by an avalanche in the French Alps. Anne was the initiator of the study, she designed and supervised it until she passed away.

Source of funding
The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), from the European Union H2020 program (project EuroPOND, grant number 666992), and from the Abeona Foundation (project Brain@Scale).

Role of the funding source

The sponsors had no role in study design, data analysis or interpretation, writing or decision to submit the report for publication.

Authors contributions

PC had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concepts and study design: PC, AB, OC
Acquisition, analysis or interpretation of data interpretation: all authors
Manuscript drafting or manuscript revision for important intellectual content: all authors
Approval of final version of submitted manuscript: all authors
Literature research: PC, AM, AB, OC
Statistical analysis: MH
Administrative, technical, or material support: JSG, OC, AB
Study supervision: OC, AB

Disclosure statement

Competing financial interests related to the present article: none to disclose for all authors.

Competing financial interests unrelated to the present article: OC reports having received consulting fees from AskBio (2020), having received fees for writing a lay audience short paper from Expression Santé (2019), having received speaker fees for a lay audience presentation from Palais de la découverte (2017) and that his laboratory has received grants (paid to the institution) from Qynapse (2017-present). Members from his laboratory have co-supervised a PhD thesis with myBrainTechnologies (2016-present). OC spouse is an employee of myBrainTechnologies (2015-present). O.C. has submitted a patent to the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices) (2016).

Abstract

Background and Purpose.

Many artificial intelligence tools are currently being developed to assist diagnosis of dementia from magnetic resonance imaging (MRI). However, these tools have so far been difficult to integrate in the clinical routine workflow. In this work, we propose a new simple way to use them and assess their utility for improving diagnostic accuracy.

Materials and Methods.

We studied 34 patients with early-onset Alzheimer's disease (EOAD), 49 with late-onset AD (LOAD), 39 with frontotemporal dementia (FTD) and 24 with depression from the pre-existing cohort CLIN-AD. Support vector machine (SVM) automatic classifiers using 3D T1 MRI were trained to distinguish: LOAD vs Depression, FTD vs LOAD, EOAD vs Depression, EOAD vs FTD. We extracted SVM weight maps, which are tridimensional representations of discriminant atrophy patterns used by the classifier to take its decisions and we printed posters of these maps. Four radiologists (2 senior neuroradiologists and 2 unspecialized junior radiologists) performed a visual classification of the 4 diagnostic pairs using 3D T1 MRI. Classifications were performed twice: first with standard radiological reading and then using SVM weight maps as a guide.

Results.

Diagnostic performance was significantly improved by the use of the weight maps for the two junior radiologists in the case of FTD vs EOAD. Improvement was over 10 points of diagnostic accuracy.

Conclusion.

This tool can improve the diagnostic accuracy of junior radiologists and could be integrated in the clinical routine workflow.

Keywords:

Alzheimer's disease; Dementia; diagnosis; anatomical MRI; artificial intelligence

Introduction

Dementia linked to neurodegenerative diseases is associated with morphological changes in the central nervous system[1]. Brain magnetic resonance imaging (MRI) is hence systematically recommended in the dementia workup: it helps differentiating a majority of dementias from one another, normal from pathological brain aging and could reveal differential diagnoses (chronic subdural hematoma, slow growing frontal tumors…). MRI can identify areas of atrophy that can suggest a particular cause of dementia, such as atrophy of medial temporal structures in late-onset Alzheimer disease (LOAD)[2] or anterior atrophy in frontotemporal dementia (FTD) [3]. Assessment of regional atrophy using MRI in dementia disorders have been extensively studied using visual, semi-quantitative ratings, computer-based volumetry, and whole-brain gray matter (WBGM) morphometry[3].

Advances in machine learning have led to the development of artificial intelligence (AI) algorithms to assist diagnosis of dementia based on T1-weighted MRI. Many studies showed that automatic support vector machine (SVM)[4] classification based on WBGM maps can differentiate AD patients from healthy controls with high accuracy[5,6]. Fewer studies exist on differential diagnosis of cognitive disorders. Several studies assessed the performance for distinguishing AD from bvFTD patients, with accuracies ranging from 80% to 90%[7,8,910,11]. Koikkalainen et al., [12], Tong et al.[13], and Morin et al.[14] studied various types of dementia and reported high accuracies for some of them (AD, FTD) but lower accuracies for others (Lewy body disease, cortico-basal degeneration).

Artificial intelligence tools are thus potentially useful for the diagnosis of dementia. However, these software tools are very rarely, if not never, used in clinical routine. A major reason is that their operation is burdensome.

In this paper, we propose a simple way to integrate results from an AI tool into the clinical routine workflow and assess its potential utility for improving diagnostic accuracy of radiologists. Specifically, we extracted weight maps indicating which brain areas are used by the algorithm to take its decision. We studied whether these weight maps improved the diagnostic accuracy for four diagnostic pairs: late-onset AD (LOAD) vs depression, LOAD vs FTD, FTD vs early-onset AD (EOAD), and depression vs EOAD. These four pairs were chosen for their clinical relevance and because the automatic classifier achieved good performances. Four radiologists performed classifications for each diagnostic pairs. Classifications were performed twice: first with standard radiological reading and then using the weight maps as a guide.

Material and Methods

Participants

The subjects were retrospectively recruited from the ClinAd cohort[15], constituted in a tertiary academic expert memory center (Institute for Memory and Alzheimer's disease, Pitié-Salpêtrière University Hospital). 992 patients were included in this cohort from 2005 to 2014. All patients had neurological, biological and neuropsychological evaluations. Cerebrospinal fluid (CSF) ABeta1-42, tau and phosphorylated tau was available for all participants. At inclusion, patients and their relatives were informed that anonymized data could be used in subsequent research studies. No explicit consent was needed according to French legislation for this study because all clinical and biological data were generated during a routine clinical work-up and were retrospectively extracted. However, regulations concerning electronic filing, as defined by CNIL (Commission Nationale de l'Informatique et des Libertés), were followed during all the investigations.

Morin et al.[14] studied 239 patients (corresponding to nine different diagnoses) from this cohort to evaluate the diagnostic accuracy of an automatic classifier based on WBGM segmentation maps extracted from T1-weighted MRI. From all the tested diagnostic pairs in Morin et al[14], we retained LOAD vs depression, LOAD vs FTD, EOAD vs depression and EOAD vs FTD because of their clinical relevance, and because of the good performance of the classifier. This resulted in the inclusion of 146 patients in our study corresponding to four diagnostic groups: EOAD (34 patients), LOAD (49 patients), FTD (39 patients) and depression (24 patients). For each patient, the diagnosis was assessed by a group of 3 neurologists based on clinical, biological and imaging data, following international consensus criteria for AD (IWG-2)[16], fronto-temporal dementia of the behavioral type (FTD)[17], and depression[18]. This consensus diagnosis formed the reference standard. The automatic classifier results and the two neuroradiological classifications (index tests) results were not available to assessors of the reference standard. As clinical presentations and atrophy patterns depend on the age of onset of AD[17, 18], the AD group was separated into Early-onset AD (EOAD) and Late-Onset-AD (LOAD), with age of onset respectively before and after 65 years.

MRI Acquisition

All patients had a brain MRI performed in clinical routine in the Department of Neuroradiology at Pitié-Salpêtrière Hospital on one out of four machines (3T MRI, Sigma HD; GE Healthcare or 1.5T MRI Optima 450; GE Healthcare or 1.5T MRI Horizon; GE Healthcare or 1T MRI Panorama; Philips). All MRI included a 3D T1-weighted sequence with a spatial resolution ranging from $0.5 \times 0.5 \times 1.2 mm^3$ to $1 \times 1 \times 1.2 mm^3$ that was used for SVM classification and for neuroradiological classification. Since imaging was performed as part of clinical routine, MRI acquisition parameters were not homogenized.

Construction of the computer-aided diagnosis tool

All T1-weighted MRI images were segmented into Gray Matter (GM), White Matter (WM) and CSF tissues maps using the Statistical Parametric Mapping unified segmentation routine with the default parameters (SPM12)[19]. A population template was calculated from GM and WM tissue maps using the DARTEL[20] diffeomorphic registration algorithm. All GM tissue maps were normalized to MNI space and spatially smoothed with a 12mm isotropic kernel. This kernel size is larger than what is often used in voxel-based morphometry study. It was chosen based on previous experiments.[14] where we found that the performances were slightly higher with this kernel size compared to lower or no smoothing.

Classifiers were created using Support vector machine (SVM) for each pair of diagnostic groups (LOAD vs depression, LOAD vs FTD, EOAD vs depression and EOAD vs FTD). 10-fold cross validations (CV) were performed to evaluate the performance of the classifiers with another nested 10-fold CV for unbiased search of the optimal value of the regularization hyperparameter.

SVM classifiers give a weight to each feature (here, each voxel of the image). Therefore, they provide a weight map for each pair of diagnoses. The higher the absolute value of the weight, the higher the importance of the feature in the classification. Weights can be positive or negative, indicating that the feature is taken into account when predicting a diagnosis or the other. In the weight maps, the values were normalized between 1 and -1. The weight maps describe the most discriminant atrophy areas for the SVM classifiers. From the 3D weight maps, we selected and printed on paper a set of slices, in the three planes of space.

First and second radiological classifications

To assess whether weight maps can assist radiological evaluation, two sets of radiological classifications were performed: one without and one with weight maps. Each classification and each diagnostic pair was assessed by four radiologists (two senior neuro-radiologists with more than two years of neuroradiology experience and two residents junior with one to four years of radiological experience). Only T1-weighted MRI was available to the radiologists who were blind to all other patient data. For each of the four diagnostic pairs, the patients were presented to the radiologist in randomized order (LOAD vs Depression: 73 patients, LOAD vs FTD: 88 patients, EOAD vs FTD: 73 patients and EOAD vs depression: 58 patients). No time limit was set. All radiologists viewed scans on their own and were asked to avoid any discussion of the cases with the other radiologists. In the first classification, radiologists used only their own knowledge. In the second classification, performed within 12 weeks after the first, the radiologists had printed weight maps to help spot the regions highlighted by the classifiers. Note that the performances were not revealed to the readers until the end of the study.

Statistical analysis

Demographic, clinical and cognitive measures were compared between the 4 groups using χ^2 test for categorical variables and one-way ANOVA for continuous variables. When the global test was significant, post hoc Tukey test was performed for continuous variables and pairwise χ^2 test with Benjamini Hochberg correction for categorical variables.

Classification performances were assessed using the balanced accuracy defined as: (sensitivity+specificity)/2. To statistically compare the peformances between the two sets of radiological classifications (without and with weight maps) and between the classifier and the radiologists (first reading), we used Generalized Linear Mixed Models (GLMM) with logit link and binomial distribution. Details about the GLMMs are provided in Supplementary Material S1.

Statistical analyses were performed by M.H. using R 3.5.0. The package lme4 (version 1.1-17) was used to perform GLMM.

Results

Population

The population's description is presented in Table 1. As expected, age was significantly different among the four groups. In the Depression group, there were significantly more women than in the FTD and LOAD groups. MMSE scores were significantly different between groups. This was expected since these neurodegenerative conditions do not have the same cognitive profile and since depression has a low impact on this score. There was no significant difference between groups regarding the MRI magnetic field.

Table 1. Demographic and clinical characteristics of the population.

	Depression n = 24 (16.44%)	EOAD n = 34 (23.29%)	FTD n = 39 (26.71%)	LOAD n = 49 (33.56%)	pvalue [ʃ]
Age	64.33 ± 7.36[#£]	59.29 ± 4.58[&$£]	66.26 ± 9.30[#£]	73.04 ± 5.92[&#$]	<0.001*
Gender					0.044*
Female	18 (75.00%)[$£]	21 (63.64%)	18 (46.15%)[&]	22 (44.90%)[&]	
Male	6 (25.00%)	12 (36.36%)	21 (53.85%)	27 (55.10%)	
Evolution (years)	5.91 ± 8.50[#]	2.67 ± 1.54[&]	3.43 ± 1.95	3.39 ± 2.20	0.022*
MMS	25.24 ± 3.18[#]	19.31 ± 6.43[&$]	23.00 ± 4.32[#]	22.38 ± 5.38	0.001*
Magnetic fields					0.293
1T	18 (75.00%)	21 (61.76%)	19 (48.72%)	24 (48.98%)	
1.5T	3 (12.50%)	7 (20.59%)	7 (17.95%)	9 (18.37%)	
3T	3 (12.50%)	6 (17.65%)	13 (33.33%)	16 (32.65%)	

Note. Counts, percentages, means and standard deviations are shown four groups, as well as p-values, to indicate statistically significant group differences. Values are expressed as Mean values ± Standard Deviation

[ʃ] p-values for the comparison between the four groups using one-way ANOVA for continuous variables and chi-square test for qualitative variables

Following signs indicate which groups significantly differ: [&] group differs from Depression; [#] group differs from EOAD; [$] group differs from FTD; [£] group differs from LOAD.

* Statistically significant at $p < 0.05$

Weight maps

The printed weight maps that were presented to the radiologists are shown in Figure 1. For LOAD vs Depression, areas contributing to the classification as LOAD are almost exclusively localized in hippocampus, while for Depression areas are of small size, more diffuse, and without lobar predominance except in insular areas. For LOAD vs FTD, LOAD classification is associated with bilateral hippocampal and medial parietal atrophy, while areas associated with FTD predominate prominently in the frontal lobe and temporal lobes and next to the ventricles. For EOAD vs FTD, EOAD classification is associated with bilateral atrophy in medial temporal, retrosplenial cortex and medial parietal cortex (precuneus), while FTD involves mainly the frontal lobes and in particular the cingulate gyrus with some involvement of the head of caudate nucleus and bilateral cerebellar atrophy. For EOAD vs Depression, areas contributing to EOAD classification involve bilateral atrophy of the medial parietal cortex and of the hippocampus atrophy as well as areas next to the ventricles, while for depression areas are more diffuse without lobar predominance except in insular areas. However, there is bilateral cerebellar involvement that is absent in EOAD.

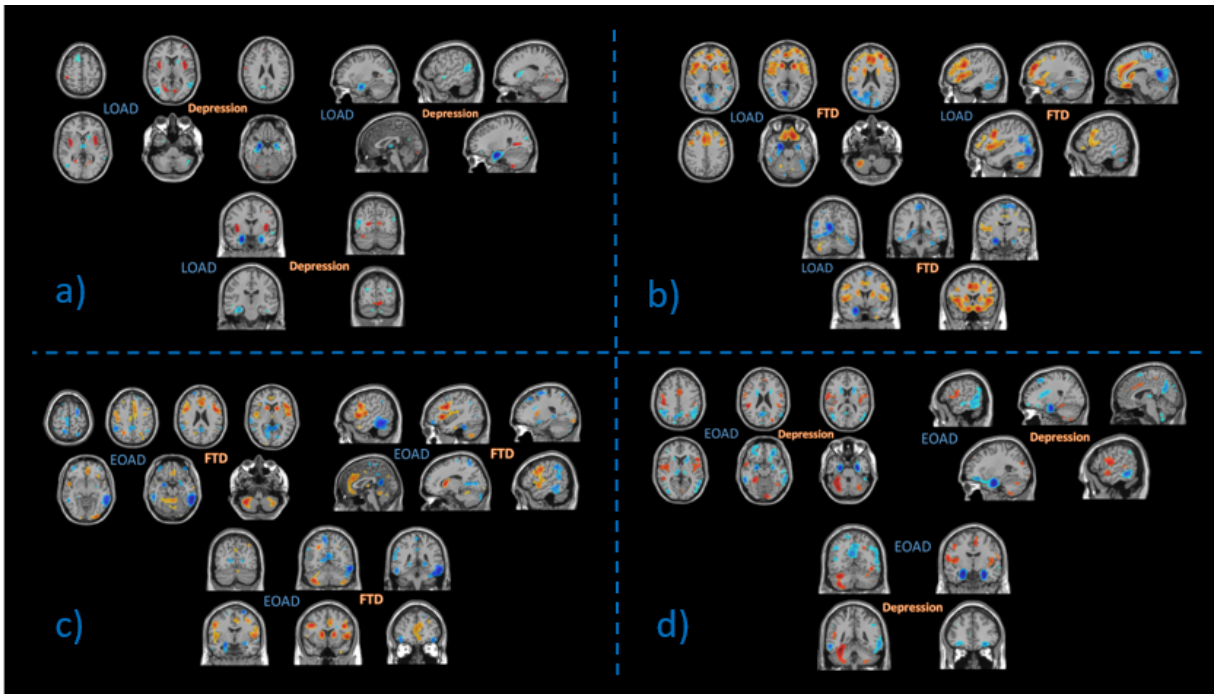

Figure 1: Printed set of slices from the 3D weight map corresponding to:
a) LOAD vs Depression. Blue (resp. orange) areas correspond to regions in which atrophy increases the likelihood of classification as LOAD (resp. Depression), b) LOAD vs FTD, c) EOAD vs FTD and d) EOAD vs Depression

Classifications

Balanced accuracies for the two radiological classifications and for the SVM classifier are reported in Table 2.

Table 2. Balanced accuracies to distinguish between diagnostic pairs for the two radiological classifications of the four radiologists and for the SVM classifier.

	LOAD vs Depression (alone/with maps)	LOAD vs FTD (alone/with maps)	FTD vs EOAD (alone/with maps)	EOAD vs Depression (alone/with maps)
Radiologist				
Senior1	0.70/0.68	0.77/0.77	0.75/0.76	0.70/0.65
Senior2	0.69/0.70	0.62/0.69	0.59/0.72	0.63/0.67
Junior1	0.74/0.84	0.77/0.80	0.68/0.84	0.63/0.68
Junior2	0.71/0.68	0.69/0.72	0.67/0.80	0.69/0.76
SVM Classifier	0.72	0.72	0.80	0.82

Statistical comparisons of the performances between the two radiological classifications and between the classifier and radiologists are summarized in Table 3. Full results of the GLMM are presented in Supplementary Tables S1 and S2. Across the four radiologists, the use

of weight maps significantly improved the diagnostic performances for FTD vs EOAD ($p<0.001$) but not for other diagnostic pairs. Looking at each radiologist separately, the use of weight maps significantly improved the performance for the two junior radiologists for FTD vs EOAD (OR=3.83±1.89 p=0.007 and OR=2.81±1.31 p=0.027, respectively). Although large in magnitude (respectively, 10 and 13 percent points of balanced accuracies), two other improvements did not reach statistical significance (FTD vs EOAD for Senior 2: OR=2.08±0.90 p=0.090; LOAD vs Depression for Junior 1: OR=3.03±1.7 p= 0.053). The performances of the SVM classifier were significantly higher than that of all four radiologists (first classification) for EOAD vs Depression (Junior 1: p=0.002; Junior 2: p=0.026; Senior 1: p=0.044; Senior 2: p=0.001) and that of two of the four radiologists for FTD vs EOAD (Junior 1: p=0.066; Junior 2: p=0.040; Senior 1: p=0.47; Senior 2: p=0.002) but not for other diagnostic pairs.

Table 3. Summary of the results of statistical comparison using GLMM between the first and the second radiological classification (with and without weight maps), indicated in the column "WMs", and between the automatic classifier and the first radiological classification, indicated in column "C".

	LOAD vs Depression		LOAD vs FTD		FTD vs EOAD		EOAD vs Depression	
	WMs	C	WMs	C	WMs	C	WMs	C
Overall effect	0.81		0.43		**0.0008***		0.62	
Junior1	*0.054*[#]	0.81	0.57	0.44	**0.007***	*0.066*[#]	0.53	**0.002***
Junior2	0.98	0.10	0.35	0.46	**0.027***	**0.040***	0.37	**0.026***
Senior1	0.33	0.24	0.81	0.33	0.89	0.47	0.26	**0.044***
Senior2	0.66	0.10	0.20	0.10	*0.090*[#]	**0.002***	0.28	**0.001***

Results are presented as p values.
WM: Results of the comparison between the first and the second radiological classification, assessing the impact of using weight maps (WMs) on diagnostic performance. The line 'Overall effect' refers to the test on the overall improvement across all four radiologists. The lines 'Junior1' to 'Senior2' refer to the posthoc tests on the improvement of each radiologist.
C: Results of the comparison between the automatic classifier and the first radiological classification. The lines 'Junior1' to 'Senior2' refer to the posthoc tests comparing the automatic classifier to each radiologist.
* $p<0.05$; [#] $0.05<p<0.1$

Discussion

In this paper, we proposed a simple way to introduce results from an AI tool in the clinical routine workflow and we assessed its value for improving radiological diagnosis of neurodegenerative dementias. We showed that this tool can significantly improve diagnostic performance for some clinically difficult diagnoses, in particular in the case of junior radiologists.

We found a significant improvement for two of the readings. Moreover, two other readings, although not statistically significant, were improved by more than 10 points of balanced accuracy. Significant improvements concerned FTD vs EOAD diagnosis for the two junior radiologists (16 points and 13 points respectively). Importantly, there was no statistically significant decrease in accuracy for any of the readings (decreases were observed in only 3 of the 16 cases and none was significant). The effect of the weight maps was mostly seen for junior radiologists who are not yet specialized. It is likely because they had less knowledge regarding atrophy patterns in different types of dementia. Among the information provided by the WMs, it was interesting that, for the contrast EOAD vs FTD, it pushed the readers to focus on the medial frontal cortex (next to the falx cerebri) instead of looking for global anterior atrophy in FTD. Moreover, each junior radiologist received a quick briefing about atrophy in dementia before the first reading. A brief oral reminder was given to them and both read about the main atrophy patterns. It could thus be that the improvement with the weight maps would have been even higher without this briefing. This could explain the relatively good performance of junior radiologists in the first reading. In Klöppel et al.[9], the mean accuracy for AD vs FTD, for radiologists with at least six years of practice (including four neuroradiologists) was 68.6%. In our study, the average accuracy of junior radiologists was 73% for FTD vs LOAD pair and 67.5% for FTD vs EOAD. Nevertheless, it is true that the fact that senior radiologists' performance was not improved limits the impact of the study. Indeed, the resident readings are usually checked by senior radiologists. However, we would like to point out that detecting moderate increases in classification performance requires very large samples. For instance, a calculation based on binomial low shows that 100 patients par groups are required to estimate accuracy with a standard deviation of 5 points. In our study, the performance of the senior 2 was increased by 13 points for one reading, even though it was not statistically significant.

For some of the pairwise classification, there was no improvement with the use of the weight maps. There are several possible explanations to this phenomenon. For the LOAD vs depression and LOAD vs FTD pairs, it could be that the lower performance of the classifier may make WMs less relevant and in turn lead to a lack of increase in radiological classification performance. Another possibility is that the radiologists had difficulty integrating information from the weight maps. Furthermore, some areas of atrophy spotted by the WMs are difficult to take into account for the human eye. For example, atrophy of the head of the caudate nucleus may be difficult to assess because of the absence of a corresponding widening of a sulcus.

One may wonder if similar improvements could have been obtained by presenting maps of significant groups differences resulting from a standard mass-univariate voxel-based morphometry analysis. This is nevertheless beyond the scope of the present study and was left for future work.

Although not the main objective of our study, we also compared the performance of the radiologists to that of the automatic classifier. We showed that the classifier was significantly more accurate in several of the cases and was never significantly less accurate. In particular, the classifier was significantly better than every radiologist for EOAD vs Depression and better than two radiologists for FTD vs EOAD. Klöppel et al.[9] also report higher or similar accuracies for the automatic classification. Three studies have focused on automatic classification of AD vs FTD [8,21,22], and obtained slightly higher classification accuracies: 72% to 90%, as compared to 72%-80% in our study (for FTD vs. EOAD-LOAD). However, these studies were based on

research datasets while ours used clinical routine data. For the four tested pairs, our classifier provided accuracies up from 72% to 82%. One study evaluated classifiers on a clinical routine dataset[12]. The reported accuracy for FTD vs AD (80%) is consistent with ours. In our opinion, the superiority of the classifier for some tasks by no means implies that AI tools should be used in place of radiological reading for such tasks. Nevertheless, such tools will most likely be used to assist radiological evaluation in the future. This requires technical developments to integrate them in the radiological workflow as well as clinical studies demonstrating their added value.

Structural MRI depicts characteristic patterns of brain atrophy in Alzheimer's disease and FTD[3,23,24]. Nevertheless, radiological diagnosis can be difficult. First, atrophy patterns can vary within the same pathology. AD patterns will change substantially depending on their age of onset[25,26]. Similarly, FTDs represent a family of diseases that vary both clinically and in their imaging presentation, even within the subcategory of behavioral-variant FTD[27,28]. The second difficulty is the overlap of atrophy patterns[29], particularly at an early stage. Discriminative areas displayed by the weight maps are mostly consistent with the literature on atrophy patterns, but specifically highlight the areas that allow for discriminating between groups. Comparing the maps for LOAD vs Depression and EOAD vs Depression, we retrieved the prominent hippocampal atrophy in LOAD while there is also lobar involvement in EOAD. Among the atrophic areas described in depressed patients in previous group studies (hippocampus, central grey nuclei, frontal and insular lobes)[30,31], only the insula appears to discriminate against AD patients. The weight maps comparing FTD to EOAD and LOAD highlight that lobar atrophy is more pronounced in EOAD patients compared to LOAD patients. We notice that the temporal poles do not appear as discriminating atrophy areas. Surprisingly, the cerebellum was also part of the discriminating regions. The role of the cerebellum in cognitive function has been broadly investigated in the last decades but the severity of cerebellar changes in FTD, AD, and cognitive disorders remains unclear[32–34]. For our classifier, cerebellar atrophy was associated to classification as FTD rather than as AD.

The gold standard diagnoses were made in a standardized and multidisciplinary way in line with the latest research guidelines in the field[16,17], thereby providing a solid reference to which radiological diagnosis can be compared. At the same time, the cohort is representative of clinical routine, making the results more generalizable to clinical practice. AD and FTD are sometimes difficult to differentiate clinically because of overlapping symptoms[35]. MRI is systematically recommended for diagnosis of dementia and thus represents no extra examination and cost. Automatic classifiers are effective tools to assist diagnosis from MRI but are still not usable in current practice. The simple tool that we propose leverages recent advances in AI but is still applicable to clinical practice without requiring any specific software or change to the clinical workflow. It can be effective to improve the performance of radiologists and particularly junior radiologists in some differential diagnoses.

One limitation of our study is the binary classification which does not correspond to the clinical practice. However, the aim of WMs is to spot relevant regions when there is a doubt in a differential diagnosis. In order to perform the classification, we only considered the core diagnosis and disregarded mixed pathologies. The use of depression as a control group could be another limitation given that depressive patients present atrophy. However, this situation is representative of the clinical routine: patients presenting with cognitive disorders are usually diagnosed with a neurological or a psychiatric condition, or present with subjective cognitive impairment, and are thus not "pure" control subjects. The imaging data used in this study was from a memory clinical cohort acquired over a period of nine years. Consequently, image quality substantially varied (1.5 T and 3.0 T). Using a more homogenous dataset could potentially improve results. Still, the use of routine imaging data shows that the proposed method can be used in clinical practice. Another limitation is that we did not study all possible contrasts, compared to those reported in Morin et al.[14]. For some of these contrasts, the classifier

accuracy was high (for instance 82% for FTD vs depression). The reason for limiting the number of studied contrasts was simply the limited available time of the radiologists. Furthermore, one can note that each patient was seen twice (for different diagnostic pairs). This is thus in principle possible that the reader could infer the diagnosis based on the pair in which the patient was previously present (for instance, if an MRI is seen in EOAD vs FTD and then in EOAD vs depression, one could conclude that the patient is EOAD). While this is theoretically possible, we believe that the risk is limited, in particular given that not all diagnostic pairs were done on the same day. Finally, a limitation is that the weight maps were built using the same datasets that were used for radiological diagnosis, which is a form of double-dipping[36]. Ideally, the weight maps should have been built using a different dataset, which was unfortunately not available.

The present work shows the potential of the approach to assist radiological diagnosis of dementia. Nevertheless, future studies should be performed to fully demonstrate the value of the approach. First, it would be necessary to use a different dataset (different from the one used to build the weight maps) to assess external validity. A larger sample would be necessary to assess whether the approach can improve the performance of senior radiologists. In our study, the improvement was not significant but this may be due to insufficient statistical power. Moreover, we could assess if similar improvements could be obtained with maps representing group differences rather than derived from a machine learning algorithm. This future study would also include more contrasts. Finally, it would be interesting to extend the work to multi-class, rather than binary, classification.

Conclusion

Although AI represents the future for radiological diagnosis of dementias, it has not yet entered clinical practice. In this paper, we provide a simple way to make use of AI results and demonstrate that it can improve the diagnosis.

ABBREVIATIONS:

EOAD: Early Onset of Alzheimer disease
LOAD: Late-Onset of Alzheimer disease
FTD: Fronto-Temporal Dementia
SWM: Support vector machine
GLMM: Generalized Linear Mixed Model
WMs: Weight maps
WBGM: whole-brain gray matter
AI: artificial intelligence
IWG: international working group
CV: Cross Validation

References

1. Park M, Moon W-J. Structural MR Imaging in the Diagnosis of Alzheimer's Disease and Other Neurodegenerative Dementia: Current Imaging Approach and Future Perspectives. *Korean J Radiol*. 2016;17(6):827-845. doi:10.3348/kjr.2016.17.6.827

2. Scheltens P, van de Pol L. Impact commentaries. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *J Neurol Neurosurg Psychiatry*. 2012;83(11):1038-1040. doi:10.1136/jnnp-2012-302562

3. Yang J, Pan P, Song W, et al. Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation. *J Neurol Sci*. 2012;316(1-2):21-29. doi:10.1016/j.jns.2012.02.010

4. Vapnik VN. An overview of statistical learning theory. *IEEE Trans Neural Netw*. 1999;10(5):988-999. doi:10.1109/72.788640

5. Cuingnet R, Gerardin E, Tessieras J, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage*. 2011;56(2):766-781. doi:10.1016/j.neuroimage.2010.06.013

6. Salvatore C, Battista P, Castiglioni I. Frontiers for the Early Diagnosis of AD by Means of MRI Brain Imaging and Support Vector Machines. *Curr Alzheimer Res*. 2016;13(5):509-533.

7. Möller C, Pijnenburg YAL, van der Flier WM, et al. Alzheimer Disease and Behavioral Variant Frontotemporal Dementia: Automatic Classification Based on Cortical Atrophy for Single-Subject Diagnosis. *Radiology*. 2016;279(3):838-848. doi:10.1148/radiol.2015150220

8. Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark CM. Individual Patient Diagnosis of AD and FTD via High-Dimensional Pattern Classification of MRI. *NeuroImage*. 2008;41(4):1220-1227. doi:10.1016/j.neuroimage.2008.03.050

9. Klöppel S, Stonnington CM, Barnes J, et al. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. *Brain*. 2008;131(11):2969-2974. doi:10.1093/brain/awn239

10. Bouts MJRJ, Möller C, Hafkemeijer A, et al. Single Subject Classification of Alzheimer's Disease and Behavioral Variant Frontotemporal Dementia Using Anatomical, Diffusion Tensor, and Resting-State Functional Magnetic Resonance Imaging. *J Alzheimers Dis JAD*. 2018;62(4):1827-1839. doi:10.3233/JAD-170893

11. Kim JP, Kim J, Park YH, et al. Machine learning based hierarchical classification of frontotemporal dementia and Alzheimer's disease. *NeuroImage Clin*. 2019;23:101811. doi:10.1016/j.nicl.2019.101811

12. Koikkalainen J, Rhodius-Meester H, Tolonen A, et al. Differential diagnosis of neurodegenerative diseases using structural MRI data. *NeuroImage Clin*. 2016;11:435-449. doi:10.1016/j.nicl.2016.02.019

13. Tong T, Ledig C, Guerrero R, et al. Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting. *NeuroImage Clin*. 2017;15:613-624. doi:10.1016/j.nicl.2017.06.012

14. Morin A, Samper-Gonzalez J, Bertrand A, et al. Accuracy of MRI Classification Algorithms in a Tertiary Memory Center Clinical Routine Cohort. *J Alzheimers Dis JAD*. March 2020. doi:10.3233/JAD-190594

15. Teichmann M, Epelbaum S, Samri D, et al. Free and Cued Selective Reminding Test - accuracy for the differential diagnosis of Alzheimer's and neurodegenerative diseases: A large-scale biomarker-characterized monocenter cohort study (ClinAD). *Alzheimers Dement J Alzheimers Assoc*. 2017;13(8):913-923. doi:10.1016/j.jalz.2016.12.014

16. Dubois B, Feldman HH, Jacova C, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol*. 2007;6(8):734-746. doi:10.1016/S1474-4422(07)70178-3

17. Rascovsky K, Hodges JR, Knopman D, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*. 2011;134(9):2456-2477. doi:10.1093/brain/awr179

18. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Fifth Edition. American Psychiatric Association; 2013. doi:10.1176/appi.books.9780890425596

19. Ashburner J, Friston KJ. Voxel-Based Morphometry—The Methods. *NeuroImage*. 2000;11(6):805-821. doi:10.1006/nimg.2000.0582

20. Ashburner J. A fast diffeomorphic image registration algorithm. *NeuroImage*. 2007;38(1):95-113. doi:10.1016/j.neuroimage.2007.07.007

21. Klöppel S, Stonnington CM, Chu C, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain J Neurol*. 2008;131(Pt 3):681-689. doi:10.1093/brain/awm319

22. Bron EE, Smits M, Papma JM, et al. Multiparametric computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural and advanced MRI. *Eur Radiol*. 2017;27(8):3372-3382. doi:10.1007/s00330-016-4691-x

23. Karas GB, Burton EJ, Rombouts SARB, et al. A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *NeuroImage*. 2003;18(4):895-907. doi:10.1016/S1053-8119(03)00041-7

24. Du A-T, Schuff N, Kramer JH, et al. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. *Brain*. 2007;130(4):1159-1166. doi:10.1093/brain/awm016

25. Frisoni GB, Pievani M, Testa C, et al. The topography of grey matter involvement in early and late onset Alzheimer's disease. *Brain J Neurol*. 2007;130(Pt 3):720-730. doi:10.1093/brain/awl377

26. Fiford CM, Ridgway GR, Cash DM, et al. Patterns of progressive atrophy vary with age in Alzheimer's disease patients. *Neurobiol Aging*. 2018;63:22-32. doi:10.1016/j.neurobiolaging.2017.11.002

27. Kobayashi Z, Arai T, Kawakami I, et al. Clinical features of the behavioural variant of frontotemporal dementia that are useful for predicting underlying pathological subtypes of frontotemporal lobar degeneration. *Psychogeriatr Off J Jpn Psychogeriatr Soc*. 2018;18(4):307-312. doi:10.1111/psyg.12334

28. Zhang Y, Tartaglia MC, Schuff N, et al. MRI signatures of brain macrostructural atrophy and microstructural degradation in frontotemporal lobar degeneration subtypes. J Alzheimers Dis. 2013;33:431–444.

29. Chan D, Fox NC, Jenkins R, Scahill RI, Crum WR, Rossor MN. Rates of global and regional cerebral atrophy in AD and frontotemporal dementia. *Neurology*. 2001;57(10):1756-1763.

30. Malykhin NV, Coupland NJ. Hippocampal neuroplasticity in major depressive disorder. *Neuroscience*. 2015;309:200-213. doi:10.1016/j.neuroscience.2015.04.047

31. Igata N, Kakeda S, Watanabe K, et al. Voxel-based morphometric brain comparison between healthy subjects and major depressive disorder patients in Japanese with the s/s genotype of 5-HTTLPR. *Sci Rep*. 2017;7(1):3931. doi:10.1038/s41598-017-04347-8

32. Mormina E, Petracca M, Bommarito G, Piaggio N, Cocozza S, Inglese M. Cerebellum and neurodegenerative diseases: Beyond conventional magnetic resonance imaging. *World J Radiol*. 2017;9(10):371-388. doi:10.4329/wjr.v9.i10.371

33. Chen Y, Kumfor F, Landin-Romero R, Irish M, Hodges JR, Piguet O. Cerebellar atrophy and its contribution to cognition in frontotemporal dementias. *Ann Neurol*. 2018;84(1):98-109. doi:10.1002/ana.25271

34. Toniolo S, Serra L, Olivito G, Marra C, Bozzali M, Cercignani M. Patterns of Cerebellar Gray Matter Atrophy Across Alzheimer's Disease Progression. *Front Cell Neurosci*. 2018;12:430. doi:10.3389/fncel.2018.00430

35. Siri S, Benaglio I, Frigerio A, Binetti G, Cappa SF. A brief neuropsychological assessment for the differential diagnosis between frontotemporal dementia and Alzheimer's disease. *Eur J Neurol*. 2001;8(2):125-132.

36. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*. 2009;12(5):535-540. doi:10.1038/nn.2303

Supplementary Material S1.

Details regarding the Generalized Linear Mixed Models (GLMM).

To statistically compare the peformances between the two sets radiological classifications (without and with weight maps) and between the classifier and the radiologists (first reading), we used Generalized Linear Mixed Models (GLMM) with logit link and binomial distribution. We used one model for each diagnostic pair. The binary dependent variable was correct or wrong diagnosis for each patient, each radiologist and each classification. Fixed effects were age of the patient, radiologist, diagnostic class, classification (i.e. radiologist alone or radiologist with weight maps) and random effect was the patient ID. Interactions of order two and three between radiologist, diagnostic class and classification were tested. Post hoc tests were performed for the interaction radiologist:diagnostic_class:classification and radiologist:classification to test wether each radiologist improved its diagnostic performances.

We then compared diagnostic performances of the automatic classifier to those of the radiologists (first classification without weight maps). To that purpose, we also used a GLMM with logit link and binomial distribution. The binary dependent variable was correct or wrong diagnosis for each patient, each radiologist and each classification. Fixed effects were age of the patient, assessor (i.e. classifier and each one of the four radiologists), diagnostic class, assessor:diagnostic_class and random effect was the patient ID. Post hoc tests were performed for the interaction assessor:diagnostic_class and assessor to test whether the classifier had better diagnostic performances than each of the radiologist.

For both GLMM, type II likelihood ratio tests were performed to test fixed effects. Cohen's f2 were calculated, using the marginal R2, for each effect to estimate their size. Normality of residuals and random effects as well as heteroskedasticity were checked visually. Influencers and outliers were checked computing hat values and Cook distances.

Supplementary Material S2

Oral Briefing reminder given to junior before the first classification.

Neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease, FTD are chronic and slow progressing diseases. They generally cause deterioration in the functioning of nerve cells, particularly neurons, which can lead to neurodegeneration. These pathologies are thought to be caused by the abnormal build-up of proteins in and around brain cells. During the evolution of these diseases, areas of cerebral atrophy will appear.
Brain imaging is used to look for a pattern of atrophy suggestive of the suspected neurodegenerative disease.
In Alzheimer's disease, areas of atrophy are mainly found in the hippocampus and in the medial parietal cortex. There is a posterior-anterior gradient. The atrophy is rather symmetrical. In EOAD the lobar atrophy is more diffuse whereas in LOAD atrophy is more confined to the hippocampus.
In FTD, areas of atrophy are mainly found in the frontal and temporal cortex. There is an anterior-posterior gradient of the atrophy and possibly asymmetrical presentation.

This briefing was mainly a summary of a PowerPoint course available on the following website: ***http://www.sfrnet.org/rc/org/sfrnet/htm/Article/2014/20140515-080519-834/src/htm_fullText/fr/BERTRAND_A.pdf***

Supplementary Results

Table S1. Results of the GLMM on having a good classification for each pair comparisons

	LOAD vs Depression (R2m = 0.09 ; R2c = 0.62)			LOAD vs FTD (R2m = 0.06 ; R2c = 0.36)			FTD vs EOAD (R2m = 0.06 ; R2c = 0.43)			EOAD vs Depression (R2m = 0.05 ; R2c = 0.48		
	coefficient ± se ¥	Cohen'f2	pvalue	coefficient ± se ‡	Cohen'f2	pvalue	coefficient ± se ∫	Cohen'f2	pvalue	coefficient ± se ¥	Cohen'f2	pvalue
Diagnosis	-1.301 ± 1.236	0.0803	0.0016*	1.102 ± 0.696	0.0029	0.6255	1.074 ± 0.820	0.0021	0.1030	-1.119 ± 0.948	0.0212	0.0941
Age	0.013 ± 0.059	0.0026	0.8038	-0.077 ± 0.027	0.0237	0.0031*	-0.063 ± 0.037	0.0026	0.0762	-0.019 ± 0.052	0.0004	0.6970
Radiologist		0.0261	0.0004*		0.0127	0.0030*		0.0130	0.0563		0.0034	0.3433
Junior2	1.884 ± 1.184			-0.002 ± 0.605			0.214 ± 0.624			0.015 ± 0.790		
Senior1	0.004 ± 0.995			-0.380 ± 0.595			0.846 ± 0.652			-0.290 ± 0.780		
Senior2	0.555 ± 1.036			-0.697 ± 0.590			-1.448 ± 0.625			0.692 ± 0.830		
Reading	1.162 ± 1.093	0.0009	0.8051	0.808 ± 0.648	0.0015	0.4294	0.856 ± 0.653	0.0148	0.0008*	0.699 ± 0.831	0.0002	0.6201
Diagnosis*Radiologist		0.0041	0.0309*		0.0014	0.4352		0.0066	0.0513		0.0001	0.8616
Junior2	-3.461 ± 1.337			-1.142 ± 0.815			-0.597 ± 0.876			0.820 ± 1.026		
Senior1	-0.722 ± 1.163			0.965 ± 0.872			-0.649 ± 0.907			1.589 ± 1.033		
Senior2	-1.789 ± 1.200			-0.565 ± 0.802			1.849 ± 0.896			-1.295 ± 1.047		
Diagnosis*Reading	-0.154 ± 1.268	0.0001	0.4787	-1.139 ± 0.864	0.0102	0.0014*	1.296 ± 1.043	0.0029	0.1460	-0.705 ± 1.045	0.0008	0.3644
Radiologist*Reading		0.0059	0.1589		0.0005	0.7518		0.0032	0.2665		0.0071	0.3766
Junior2	-2.530 ± 1.642			-0.435 ± 0.897			-0.451 ± 0.915			0.842 ± 1.234		
Senior1	0.744 ± 1.590			0.630 ± 0.922			-0.072 ± 0.981			0.276 ± 1.164		
Senior2	-1.185 ± 1.528			0.270 ± 0.886			1.673 ± 0.935			-1.979 ± 1.172		
Diagnosis*Radiologist*Reading		0.0036	0.0576		0.0054	0.1682		0.0216	0.0003*		0.0164	0.0009*
Junior2	1.693 ± 1.857			0.887 ± 1.173			0.096 ± 1.405			-1.053 ± 1.537		
Senior1	-3.137 ± 1.824			-1.776 ± 1.241			-2.657 ± 1.424			-1.974 ± 1.491		
Senior2	0.346 ± 1.758			-0.064 ± 1.161			-4.789 ± 1.406			3.654 ± 1.497		

Notes. ¥ reference category for the Diagnosis effect = Depression

‡ reference category for the Diagnosis effect = FTD

∫ reference category for the Diagnosis effect = EOAD

For the radiologist effect, Junior1 is the reference category; for the method effect, radiologist alone is the reference category

R2m: coefficient of determination estimated on the fixed part of the GLMMs; R2c : coefficient of determination estimated on the fixed and random parts of the GLMMs

Table S2. Results of the GLMM on having a good classification to compare classifier and radiologist alone for each pair comparisons

	LOAD vs Depression (R2m = 0.06 ; R2c = 0.59)			LOAD vs FTD (R2m = 0.09 ; R2c = 0.32)			FTD vs EOAD (R2m = 0.05 ; R2c = 0.51)			EOAD vs Depression (R2m = 0.08 ; R2c = 0.51)		
	coefficient ± se ¥	Cohen'f2	pvalue	coefficient ± se ‡	Cohen'f2	pvalue	coefficient ± se ∫	Cohen'f2	pvalue	coefficient ± se ¥	Cohen'f2	pvalue
Diagnosis	0.946 ± 1.065	0.0368	0.0679	1.754 ± 0.634	0.0394	0.0049*	0.324 ± 0.962	0.0163	0.0451*	3.688 ± 1.354	0.0051	0.6799
Age	0.005 ± 0.053	0.0004	0.9456	-0.082 ± 0.025	0.0531	0.0006*	-0.085 ± 0.044	0.0051	0.0388*	-0.014 ± 0.050	-0.0012	0.7893
Assessor		0.0192	0.3040		0.0136	0.0499*		0.0244	0.0171*		0.0382	0.0064*
Junior1	1.061 ± 0.867			0.609 ± 0.561			-1.403 ± 0.709			0.298 ± 0.760		
Junior2	2.688 ± 1.100			0.609 ± 0.561			-1.195 ± 0.710			0.295 ± 0.760		
Senior1	1.061 ± 0.867			0.290 ± 0.549			-0.522 ± 0.725			0.008 ± 0.748		
Senior2	1.507 ± 0.912			-0.008 ± 0.542			-2.996 ± 0.753			0.942 ± 0.804		
Diagnosis*Assessor		0.0130	0.0004*		0.0150	0.0778		0.0133	0.0108*		0.0430	<0.0001*
Junior1	-1.830 ± 1.073			-0.616 ± 0.798	1.670 ± 0.77	.256 ± 0.82	0.962 ± 0.972			-4.739 ± 1.442		
Junior2	-4.925 ± 1.297			-1.142 ± 0.815			0.343 ± 0.967			-3.941 ± 1.426		
Senior1	-2.506 ± 1.076			0.965 ± 0.872			0.299 ± 0.990			-3.212 ± 1.412		
Senior2	-3.432 ± 1.123			-0.565 ± 0.802			3.003 ± 1.018			-5.957 ± 1.494		

Notes. ¥ reference category for the Diagnosis effect = Depression

‡ reference category for the Diagnosis effect = FTD

∫ reference category for the Diagnosis effect = EOAD

For the Assessor effect, classifier is the reference category

R2m: coefficient of determination estimated on the fixed part of the GLMMs; R2c: coefficient of determination estimated on the fixed and random parts of the GLMMs