Chiara Fornari and Giorgio G. Fumagalli*

Visual rating scales of brain atrophy in dementia: a systematic review of reliability and diagnostic accuracy

https://doi.org/10.1515/revneuro-2025-0066
Received May 15, 2025; accepted July 21, 2025;
published online September 17, 2025

Abstract: Visual rating scales of brain atrophy allow to quantify brain regional atrophy and could be used as a diagnostic marker for differentiating neurodegenerative diseases. This paper aims to review the visual rating scales for dementia, focusing on their reliability, the correlation with volumetric imaging measures, and their predictive accuracy. Following the PRISMA guidelines, we systematically searched in Pubmed, Web of Science, Scopus, and MEDLINE databases until November 4, 2024. Of the 441 articles extracted, we included in the review 28 papers. All the scales reached a fair to excellent level of inter and intra rater agreement. Furthermore, negative correlations were found between the rating in each scale and brain volumetric measures. Lastly, the discriminative abilities exhibited variability according to the scale and the population comparisons. Visual rating scales of atrophy provide a reliable method for distinguishing physiological aging from pathological conditions, and among neurodegenerative forms. For an accurate differential diagnosis, it is essential to employ scales with the highest diagnostic precision.

Keywords: dementia; visual rating scales; differential diagnosis; neurodegenerative diseases; brain atrophy

1 Introduction

The incidence of age-related diseases, such as neurodegenerative diseases, is rising due to the life expectancy increase (Jongsiriyanyong and Limpawattana 2018). It has been widely proved that early diagnosis can improve the prognosis, enhancing the quality of life in patients and in their caregivers (de Vugt and Verhey 2013). To provide a faster diagnosis, it is necessary to implement tools that can easily, efficiently and accurately discriminate non-invasively between physiological aging and neurological disorders. Early electrophysiological biomarkers have been found in order to detect mild cognitive impairment (MCI) in older adults (for a meta-analysis: Buzi et al. 2023), along with magnetic resonance imaging (MRI) markers of brain atrophy (for a meta-analysis: Lombardo et al. 2020). Most studies have focused on the early diagnosis of Alzheimer's disease (AD) dementia, primarily in comparison with healthy controls (HC), neglecting other neurodegenerative forms (Koikkalainen et al. 2016). However, in clinical practice the differential diagnosis between neurodegenerative forms is essential (Koikkalainen et al. 2016).

Brain MRI is a routine exam that is required for the clinical evaluation in order to exclude treatable conditions, to evaluate the vascular burden, and to assess the atrophy. It remains one of the most accessible diagnostic exams for the assessment of neurodegeneration. In clinical practice, the interpretation of brain atrophy relies on qualitative visual assessment of MRI (Loreto et al. 2023). However, various methods have been developed to enhance the objectivity and sensitivity of MRI assessment. Among them, the visual rating scales provide a more accessible approach by allowing clinicians to visually assess MRI and assign a semi-quantitative measure of regional brain atrophy using a Likert scale (Loreto et al. 2023), without the requirements of specialist expertise or software (Harper et al. 2015). Unfortunately, the implementation of the visual rating scales in the clinical and in the clinical research practices remains limited. The assessment of the medial temporal lobe atrophy by using the medial temporal atrophy scale (MTA) (Scheltens et al. 1992) is widely and predominantly used, however it may not be the best choice for the diagnostic differentiation among neurodegenerative dementia forms. Furthermore, in order to promote wider use in clinical practice, it is essential that alternative visual rating scales, which have also been demonstrated to be reliable and have been validated through automated methods, are employed as useful and discriminative tools for diagnostic purposes.

In this review we examine the agreement between raters, the correlation between the visual rating scale scores and volumetric measures, and the predictive ability of the

***Corresponding author: Giorgio G. Fumagalli**, Center for Mind/Brain Sciences (CIMeC), University of Trento, Corso Bettini 31, 38068 Rovereto, TN, Italy, E-mail: giorgio.fumagalli@unitn.it
Chiara Fornari, Center for Mind/Brain Sciences (CIMeC), University of Trento, Corso Bettini 31, 38068 Rovereto, TN, Italy, E-mail: chiara.fornari@unitn.it

visual rating scales of brain atrophy in distinguishing different forms of dementia, highlighting their applicability in the clinical and research settings as biomarkers of neurodegeneration.

2 Methods

2.1 Literature search

The literature extraction was conducted on November 4, 2024, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statements for systematic reviews and meta-analyses (Page et al. 2021). PubMed, Web of Science, Scopus, and MEDLINE databases were extracted by the first author using the following keywords string: (MCI OR mild cognitive impairment OR preclinical dementia OR early onset dementia OR prodromal OR early diagnosis OR Alzheimer OR FTD OR frontotemporal dementia OR PCA OR posterior cortical atrophy OR LBD OR Lewy body dementia) AND (visual rating OR visual assessment OR atrophy rating) AND (reproducibility OR rater) AND (MRI OR magnetic resonance imaging OR T1-weighted) NOT stroke (Table S1).

No filters and limitations to the year of publication were applied. Once the papers were screened for abstract, both the authors cross-checked for the eligibility criteria to proceed with the selection process.

2.2 Quality Assessment (QA)

To assess the risk of bias of each study to be included in the analysis, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2, Whiting et al. 2011) was used.

2.3 Study eligibility

Eligible papers were selected according to the following inclusion criteria: (1) the presence of groups comparison between healthy controls and patients with dementia or between patients with different neurodegenerative disorders; (2) studies had to rate slice obtained by MRI; (3) studies had to show the inter-rater agreement (reliability) and either correlations with voxel-based morphometry measures and/or diagnostic accuracy of the visual rating scales.

Papers were discarded according to the following exclusion criteria: (1) studies including psychiatric diseases (i.e., depression); (2) studies that were methodologically inappropriate (e.g., study design, within group, studies with incomplete data, studies that did not consider the analyses of our interest, studies that evaluated the slices obtained by computed tomography, and/or that were no open access); (3) not in English; (4) reviews and meta-analyses.

3 Results

3.1 Study selection

The keywords filled in each one of the databases queried returned a total of 430 articles, of which 241 from PubMed, 49 from Web of Science, 55 from Scopus, and 85 from MEDLINE, while 11 were retrieved by other resources. After removing duplicates, title and abstract of 262 articles were independently screened and then cross-checked by the authors. This first skimming step allowed us to proceed with the full-text screening of 91 articles, 63 of which were excluded for the following reasons: they did not allow open access (n = 1), the study design did not match with our research question (n = 40), they investigated other diseases (n = 5), they were out of topic (n = 2), data were incomplete (n = 3), and the analyses did not include correlations or diagnostic accuracy (n = 12). Where articles showed missing information, authors were contacted. At last, 28 studies were included in the review. The characteristics of the studies meeting our inclusion criteria are shown in Table 1, and the literature search is depicted in the PRISMA flowchart (Figure 1).

Results of the QA are reported in the Supplementary material (Table S2).

3.2 Temporal areas

3.2.1 Medial temporal atrophy (MTA)

The Scheltens scale focuses on the bilateral medial temporal lobe (MTL) atrophy, looking at the width of the choroid fissure, the width of the temporal horn and the height of the hippocampus in the T1-weighted coronal plane (Scheltens et al. 1992). The degree of atrophy in each of these regions is combined to produce a score reflecting overall MTL atrophy. The scale is composed of five increments (Figure S4) ranging from 0 to 4 in which scores ≤1 indicating the absence of AD, and scores ≥2 indicating the presence of AD (Harper et al. 2015). The Scheltens et al. (1992) scale has been included in the research criteria for the diagnosis of AD (Dubois et al. 2007).

The inter-rater agreement measured with Cohen's weighted kappa ($w\kappa$) in the original paper ranged from 0.72

Table 1: Characteristics of the studies included in the review.

First author	Sample size (F/M)	Independent and blind raters	Training	MRI weight (field)	Scales included
Mårtensson et al. (2020)	MCI; SCD	3 (2 neurorad, 1 software)	No	T1 (3 T)	MTA
Fischbach-Boulanger et al. (2018)	100 MCI (66/34); 50 AD (34/16)	4 (2 neurorad, 2 radiologists)	–	T1, T2 (3 T)	MTA
Koedam et al. (2011)	60 AD (30/30); 38 SCD (20/18); others [10 LBD, 10 FTLD]	3 (1 neurorad, 1 MD, 1 PhD)	–	T1 (3 T)	MTA, PA, GCA
Möller et al. (2014)	229 AD; 128 SCD	2 (1 little experience, 1 long experience)	–	T1 (3 T)	MTA, PA, GCA
Harper et al. (2016)	101 AD (39/62); 28 LBD (7/21); 55 FTLD (24/31); 73 HC (35/38)	2 (1PhD, 1 neurologist)	Yes	T1 (1 T, 1.5 T, 3 T)	MTA, PA, AT, OF, AC, FI
Silhan et al. (2021)	26 earlyAD (12/14); 21 yHC (10/11); 32 lateAD (13/19); 36 oHC (22/14)	2 (neurologist, PhD student MD)	–	T1 (1.5)	Hip-hop, PAS
Li et al. (2019)	37 AD (23/14); 29 MCI (20/9); 21 HC (15/6)	2	Standardization of the criteria	T2 (1.5 T)	FA, PA, occipital precuneus, hippocampal, lateral temporal, lateral ventricle enlargement
Chow et al. (2011)	21 FTD (14/7); 14 AD (8/6); 21 HC (15/6)	4 neurologists	References images were provided	T1, T2 (1.5 T)	AC, AT
Sheng et al. (2020)	Cohort A: 73 aMCI (40/33); 48 HC (29/19) Cohort B: 33 aMCI (16/17); 45 HC (27/18)	Cohort A: 3 (2 neurorad, 1 physician) Cohort B: 3 neurorad	Yes	T1 (3 T)	MTA, PA
Fumagalli et al. (2018)	148 HC (58/90); Presymptomatic 66 GRN (25/41); 42 c9orf72 (17/25); 24 MAPT (10/14) Symptomatic: 17 GRN (6/11); 31 c9orf72 (22/9); 15 MAPT (11/4)	2 (2 neurologists)	Yes	T1 (3 T, 1.5 T)	MTA, PA, OF, AC, FI, AT
Molinder et al. (2021)	105 HC (67/38); 184 SCD (114/70); 249 MCI (140/109); 98 AD (61/37); 25 SVD (6/19); 51 mixed (35/16); 40 other (16/24)	2 (2 neurorad)	Rater 1	T1 (0.5 T, 1.5 T)	MTA
Enkirch et al. (2018)	60 SCD; 60 AD	2 (1 radiologist; 1 neurorad)	–	T1 (3 T)	MTA
Tolboom et al. (2010)	21 AD; 20 HC	2 neurorad	Yes	T1	MTA
Min et al. (2017)	30 AD (23/7); 25 HC (16/9)	2 (1 experienced neurorad, 1 radiologist)	Yes	T1 (3 T)	MTA
Vanhoenacker et al. (2017)	25 HC (16/9); 27 MCI (14/13); 27 AD (15/12)	2	References images were provided	T2 (3 T, 1.5 T)	MTA, PA, GCA-F
Jang et al. (2015)	94 AD (57/37); 101 MCI (54/47); 65 HC (47/18)	3	–	T1 (1.5 T, 3 T)	CVRS
Davies et al. (2009)	8 HC; 8 AD; 9 svPPA; 11 bvFTD	3 (2 neurologists, 1 neuropsychologist)	One rater	T1	15 regions
Westman et al. (2011)	81 HC (45/36); 101 MCI (52/49); 75 AD (50/25)	1	–	T1	MTA
Chen et al. (2010)	49 AD; 68 MCI; 149 HC	2 radiologists	Yes	T1 (3 T, 4 T), T2 (3 T)	BALI
Boutet et al. (2012)	30 HC (19/11); 19 sMCI (12/7); 11 cMCI (6/5); 30 AD (16/14)	6 radiologists (2 experts, 4 non-experts)	References images were provided	T1	MTA
Ferreira et al. (2016)	329 HC (164/165); 421 MCI (164/257); 286 AD (157/129); 12 bvFTD (9/3); 9 nfvPPA (6/3); 13 svPPA (8/5)	2 (1 with experience, 1 trained)	Rater 2	T1	GCA-F

Table 1: (continued)

First author	Sample size (F/M)	Independent and blind raters	Training	MRI weight (field)	Scales included
Wittens et al. (2024)	16 HC (10/6); 33 SCD (19/14); 35 MCI (13/22); 27 DEM (17/10)	3 radiologists	–	T1 (3 T)	MTA
Falgàs et al. (2024)	44 lvPPA (25/19); 19 nfvPPA (10/9); svPPA (11/20); 11 uPPA (8/3); 45 HC (23/22)	2 neurologists	–	T1 (3 T)	MTA, PA, OF, AC, AT, FI
Fumagalli et al. (2020)	15 HC (7/8); 30 AD (22/8); 15 PCA (8/7)	2 neurologists	Yes	T1 (3 T)	MTA, PA, OF, AT
Ferreira et al. (2015)	345 HC; 385 sMCI; 95 cMCI; 322 AD	1	–	T1	MTA, PA, GCA
Falgàs et al. (2020)	42 HC (33/9); 48 aAD (30/18); 22 naAD (10/12); 25 sMCI (12/13); 11 bvFTD (3/8); 9 svPPA (3/6); 5 nfvPPA (2/5); 7 genetic FTD (6/1); 17 yHC (11/6); 14 genetic AD (6/8)	2 experts	–	T1 (3 T)	MTA, PA, OF, AC, AT, FI
Benussi et al. (2024)	117 prodromal FTD genetic carriers (70/47); 281 HC (158/123)	2 neurologists	Yes	T1	AC
Yuan et al. (2019)	100 AD (69/31) [43 mild AD (32/11); 57 severe AD (37/20)]; 100 HC (65/35)	2 radiologists	Yes	T1 (3 T)	MTA, PA, OF, AC, AT, FI

Female (F); male (M); mild cognitive impairment (MCI); stable mild cognitive impairment (sMCI); converted mild cognitive impairment (cMCI); amnestic mild cognitive impairment (aMCI); subjective cognitive decline (SCD); Alzheimer's disease (AD); frontotemporal lobar degeneration (FTLD); Lewy body dementia (LBD); subcortical vascular dementia (SVD); demented patients (DEM); young healthy controls (yHC); older healthy controls (oHC); healthy controls (HC); behavioral variant of frontotemporal dementia (bvFTD); progranulin (GRN); microtubule-associated protein tau (MAPT); chromosome 9 open reading frame 72 (c9orf72); semantic variant of primary progressive aphasia (svPPA); non-fluent variant of primary progressive aphasia (nfvPPA); logopenic variant of primary progressive aphasia (lvPPA); undetermined of primary progressive aphasia (uPPA); posterior cortical atrophy (PCA); amnestic AD (aAD); non-amnestic AD (naAD); doctor of medicine (MD); neuroradiologist (neurorad); medial temporal atrophy (MTA); posterior atrophy (PA); global cortical atrophy (GCA); anterior temporal (AT); orbitofrontal (OF); anterior cingulate (AC); fronto insula (FI); frontal atrophy (FA); hippocampo-horn percentage (hip-hop); parietal atrophy score (PAS); global cortical atrophy-frontal areas (GCA-F); comprehensive visual rating scale (CVRS); brain atrophy and lesion index (BALI); information not included (–).

to 0.84 (Scheltens et al. 1992). The level of expertise in the visual inspection of the MR images can affect ability to find the landmarks and then on the assessment of the level of atrophy; most of the studies in the review included neuroradiologists and radiologists. In general, the inter-rater agreement was fair to excellent (ranging from 0.3 to 0.91), generally higher for the rating of total MTA score than for the right and left sides. Furthermore, even in non-expert readers the agreement was moderate (Boutet et al. 2012), suggesting that the MTA scale can be easily and reliably used by non-experts. Moreover, the implementation of reference images or a training set protocol can affect the performance of the scale (Harper et al. 2015). Most of the studies included reference images or training protocols for the readers and, indeed, the final agreement was high. Lastly, to assess the consistency between measures, the studies have also included the intra-rater agreement for all the images processed or for a pull of them. In our sample, the intra-rater agreement ranged between 0.45 and 0.96, suggesting a moderate to near perfect agreement within the same rater.

The MRI parameters, as the weight, may also have an impact on the readability of the MR images: Fischbach-Boulanger et al. (2018) made a comparison between T1-weighted and T2-weighted images. In the study the inter-rater agreement was fair for both the two weights, suggesting similar abilities of the raters in the detection of the hippocampus atrophy through the MTA scale (Fischbach-Boulanger et al. 2018). Interestingly, Fischbach-Boulanger et al. (2018) also found that in the T2-weighted images, raters overestimated the atrophy compared to the T1-weighted images. Lastly, Molinder et al. (2021) pointed out the differences in the agreement considering the strength of the magnetic field. Considering both the 0.5 T and the 1.5 T, the inter-rater agreement was fair and the intra-rater agreement was substantial, showing that the magnetic field strength does not affect the rater's score.

Moderate negative correlations have been detected between MTA scores volumetric measurements of the hippocampus, such as gray matter density or volume (Boutet et al. 2012; Fischbach-Boulanger et al. 2018; Fumagalli et al. 2018; Harper et al. 2016; Mårtensson et al. 2020;

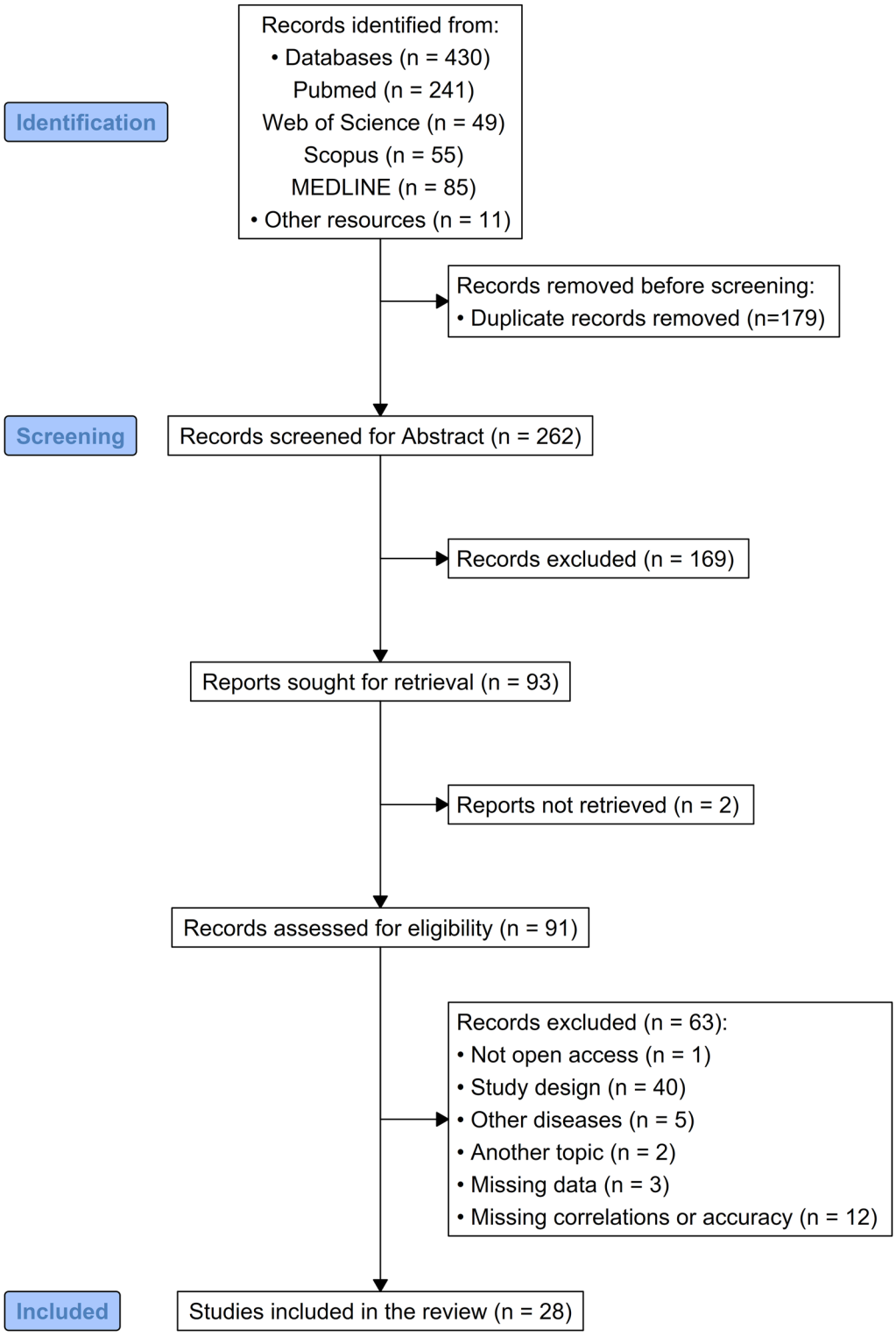

Figure 1: The flowchart illustrates the study selection process following the PRISMA guidelines for meta-analyses and systematic reviews.

Molinder et al. 2021; Sheng et al. 2020; Wittens et al. 2024; Yuan et al. 2019). Furthermore, Enkirch et al. (2018) also found correlations between the MTA score and the cerebrospinal fluid (CSF) amyloid markers. Interestingly, the authors also assessed the entorhinal cortex atrophy by using the ERICA scale, which seemed to have higher diagnostic accuracy than the MTA scale in AD (Enkirch et al. 2018).

In the original paper of Scheltens et al. (1992) a sensitivity of 81 % and specificity of 67 % for AD versus age-matched controls was reported. The implementation of higher MRI field strength had further increased the reliability of the scale (Harper et al. 2015). For example, Koedam et al. (2011) using a magnetic field of 3 T found that the specificity for AD versus age-matched controls reached 92 %, and in the study of Wittens et al. (2024) the specificity was even higher up to 94 %. Surprisingly, the visual evaluation of T2-weighted images had higher discriminating power than the T1-weighted (Fischbach-Boulanger et al. 2018).

Furthermore, automated tools are now available to assess hippocampal atrophy and differentiate between physiological and pathological aging. Considering AD versus HC, comparable predictive and discriminatory abilities have been found between automated tools (e.g., Westman et al. 2011; Min et al. 2017), volumetric measures (e.g., Boutet et al. 2012; Molinder et al. 2021), and the MTA scale.

The implementation of the MTA scale was also extended with the purpose of differentiating between the prodromal stage of the neurodegenerative disease (i.e., MCI) or the risk factor for the onset of the AD, such as in people with subjective cognitive decline (SCD). Sheng et al. (2020) found a fair discriminative power between amnestic MCI (aMCI) and HC at the MTA scale which was comparable to the discriminative power obtained by the hippocampal GM density measured by volumetric measures. Good discriminative power between MCI and HC has also been detected by Wittens et al. (2024), and by Westman et al. (2011). Furthermore, MTA was also implemented for the differentiation of typical AD from atypical AD (i.e., posterior cortical atrophy (PCA) in Fumagalli et al. (2020)).

Moreover, the MTA scale has been also used for the longitudinal analyses, assessing the conversion to AD (Boutet et al. 2012; Ferreira et al. 2015). The authors concluded that the accuracy with an automatic tool was significantly better than the visual assessment for all non-expert readers regarding the differentiation between converted MCI (cMCI) and HC. While, for the stable MCI (sMCI) compared to HC, the difference between the accuracy of automatic and visual assessments was not significant (Boutet et al. 2012). However, the MTA scale is useful for determining MCI prognosis (Ferreira et al. 2015). Lastly, MTA scale has also been implemented for the differentiation of non-AD neurodegenerative diseases (e.g., Harper et al. 2016; Falgàs et al. 2020; Molinder et al. 2021; Falgàs et al. 2024), genetic forms (Falgàs et al. 2020; Fumagalli et al. 2018), and disease severity (Yuan et al. 2019), obtaining a fair to good discriminatory ability.

Software-based analyses have also been conducted to distinguish the early stages of neurodegenerative diseases from physiological brain changes in the brain. For example, Mårtensson et al. (2020) found that the software analysis had similar reliability MTA for differentiating MCI and SCD (Mårtensson et al. 2020).

To summarize, MTA scale was originally developed to assess AD-related atrophy but has recently been extended even to the prodromal stages and to other neurodegenerative diseases. The visual assessment of brain atrophy is reliable, consistent with volumetric hippocampal measures, and discriminates between neurodegenerative diseases and physiological aging. The results are reported in Tables 1 and 2.

3.2.2 Anterior temporal (AT)

The scale implemented by Davies et al. (2006) and Kipps et al. (2007) focuses on the bilateral anterior temporal lobe atrophy in the T1-weighted coronal plane by using five increments (Harper et al. 2015; Figure S4). The scale was implemented for the frontotemporal dementia (FTD) population to assess the prognosis (Davies et al. 2006). Moreover, the implementation of the AT scale has been extended to genetic variants (Falgàs et al. 2020; Fumagalli et al. 2018), to the assessment of other neurodegenerative forms related to the frontotemporal lobar degeneration (FTLD) (Harper et al. 2016), such as primary progressive aphasia (PPA) (Falgàs et al. 2020, 2024), and to other neurodegenerative diseases, such as AD (Fumagalli et al. 2020; Yuan et al. 2019).

The inter-rater and intra-rater agreement measured with Cohen's kappa (κ) was 0.71 and 0.83, respectively (Davies et al. 2006). In our sample, the inter-rater agreement ranged from 0.57 to 0.93, meaning a moderate to near perfect agreement. The intra-rater agreement ranged from 0.63 to 0.95, meaning substantial to near perfect agreement. Interestingly, the highest agreement scores were related to the assessment of PPA variants (Falgàs et al. 2024). Furthermore, most of the studies found negative correlations between the AT score and GM atrophy, suggesting that it is a reliable measure of anterior temporal lobe atrophy (Falgàs et al. 2024; Fumagalli et al. 2018, 2020; Yuan et al. 2019).

The AT scale accurately discriminates between FTLD and both AD and Lewy body dementia (LBD) (Harper et al. 2016); between PPA variants from HC (Falgàs et al. 2020), except for the non-fluent PPA (nfvPPA) (Falgàs et al. 2024); among the PPA variants, mainly in differentiating the semantic PPA (svPPA) from the other variants (Falgàs et al. 2024); and between AD from HC (Fumagalli et al. 2020; Yuan et al. 2019). In our sample, no study has assessed the longitudinal changes in the AT brain atrophy to understand the predictive diagnostic accuracy of the prognosis.

To summarize, AT as a measure of the brain atrophy in the anterior temporal areas is reliable, consistent with

Table 2: The table summarizes the results regarding the agreement and the accuracy of the scale measuring temporal areas atrophy.

First author	Agreement	Correlation visual vs volumetric	Accuracy	Conclusion
MTA				
Mårtensson et al. (2020)	INTER: $w\kappa$ between the raters (0.3, 0.4), and raters vs software (0.3, 0.61)	Negative correlations: hip volume and rating (−0.50 to −0.58); hip volume and software (−0.58 to −0.61)	–	MTA is a reliable alternative to automatic image segmentation
Fischbach-Boulanger et al. (2018)	INTER: left hip (T1: κ = 0.35; T2: κ = 0.31). Right hip (T1: κ = 0.35; T2: κ = 0.30) INTRA: T2 higher atrophy ratings than in the T1	Negative correlations for visual rating in left (T1: r = −0.68; T2: r = −0.69) and right atrophy (T1: r = −0.716; T2: r = −0.701)	T2 visual rating had higher discrimination power than T1. MTA score provides slightly better discrimination power than volumetric measurements	T2 more adequate rating for AD and with a better distinction for MCI vs AD
Koedam et al. (2011)	INTER: $w\kappa$ (0.82 0.9) INTRA: $w\kappa$ (0.91 0.95)	–	Sens = 0.45, Spec = 0.92	MTA is useful for distinguishing AD and other dementias
Harper et al. (2016)	INTER: ICC≥0.83	Negative partial correlation between rating scores and GM density	AD vs HC: AUC = 0.82 AD vs LBD: AUC = 0.67 FTLD vs HC: AUC = 0.92 FTLD vs LBD: AUC = 0.81	MTA for distinguishing each pathological group from HC
Sheng et al. (2020)	INTER: cohort B ICC (0.8 0.9) INTRA: cohort B ICC (0.7 0.9)	Negative correlations between GM and MTA scores (r = −0.58) in both left (r = −0.53) and right (r = −0.56)	Cohort A: MTA: AUC = 0.78; GM volume: AUC = 0.84 GM density: AUC = 0.78 Cohort B: MTA: AUC = 0.82	MTA and volumetric measures have similar discriminatory power
Fumagalli et al. (2018)	INTER: ICC = 0.88 INTRA: ICC (0.9 0.96)	Negative correlations between MTA score and GM atrophy	–	Differences in MTA in genetic mutations
Molinder et al. (2021)	INTER: right (0.5 T: $w\kappa$ = 0.59; 1.5 T: $w\kappa$ = 0.53). Left (0.5 T: $w\kappa$ = 0.65; 1.5 T: $w\kappa$ = 0.67) INTRA: right (0.5 T: $w\kappa$ = 0.78; 1.5 T: $w\kappa$ = 0.71). Left (0.5 T: $w\kappa$ = 0.78; 1.5 T: $w\kappa$ = 0.8)	Negative correlations between volumetric estimates and MTA score on the right (r = −0.64) and the left (r = −0.68) side	MTA and volumetric measure had good discriminatory ability for AD vs HC, for mixed vs HC. SVD vs HC was fair. MTA: AD (AUC = 0.87), SVD (AUC = 0.79), Mixed (AUC = 0.86) Volumetric: AD (AUC = 0.91), SVD (AUC = 0.75), Mixed (AUC = 0.94)	MTA is a reliable and valid marker of medial temporal lobe atrophy
Enkirch et al. (2018)	MTA INTER: $c\kappa$ = 0.91 (0.87 0.94) ERICA INTER: $c\kappa$ = 0.88 (0.8 0.97)	Both ERICA and MTA had significant correlations with CSF parameters	ERICA (AUC = 0.93) was higher than the MTA (AUC = 0.82) Similar discrimination ability for MCI vs SCD with the ERICA (AUC = 0.66) and the MTA (AUC = 0.65)	ERICA has high diagnostic accuracy but similar discriminatory power to MTA
Tolboom et al. (2010)	INTER: $c\kappa$ = 0.90	–	AUC = 0.8; Sens = 0.71; Spec = 0.9	Accuracy of visual assessment of MTA is lower than PET but still good
Min et al. (2017)	INTER: $w\kappa$ (0.69 0.78) INTRA: $w\kappa$ (0.86 0.9)	–	MTA AUC (0.813 0.831). No significant differences in AUC between the automated tool and the MTA	Automated assessment and visual inspection accurately differentiate AD from HC
Vanhoenacker et al. (2017)	INTER: r = 0.86 (0.79 0.91)	–	AD vs HC and AD vs MCI: AUC = 0.92 (cut-off of 1.25), AUC = 0.70 (cut off of 0.75), AUC = 0.89 (cut off of 2.25), AUC = 0.8 (cut off of 1.75). AUC = 0.77 with cut-off of 1 and age<70; AUC = 0.74 with a cut-off of 1.75 and age>70	MTA high diagnostic performance

Table 2: (continued)

First author	Agreement	Correlation visual vs volumetric	Accuracy	Conclusion
Westman et al. (2011)	INTRA: right *wκ* = 0.81 and left *wκ* = 0.78. *wκ* = 0.93 on both sides.	–	MTA: AUC = 0.81, Sens = 0.787, Spec = 0.827 Multivariate method: AUC = 0.827, Sens = 0.773, Spec = 0.877	Visual assessment gives similar prediction accuracy to multivariate classification
Boutet et al. (2012)	INTRA: kappa for expert (0.58 0.63). Kappa for non-expert readers (0.45 0.74)	Negative correlations for MTA score and volumetric analysis (experts). Negative correlations in left side for the non-experts; significant also on the right but only for 3 non-experts	AD vs HC: accuracy differences for one non-expert (left side only) between the automatic and MTA cMCI vs HC: accuracy with automatic was better than the MTA for all non-experts. sMCI vs HC: no significant difference between automatic and MTA measures	Visual and volumetric measures are equally good for the classification of AD, sMCI and HC and less good for cMCI vs HC.
Wittens et al. (2024)	INTER: total ICC = 0.84; left ICC = 0.82; right ICC = 0.79	Negative correlation for total hip volumes and MTA rating	MTA DEM vs HC: AUC = 0.953 (sens = 0.778, spec = 0.938) MCI vs HC: AUC = 0.812 (sens = 0.8, spec = 0.75) Volumetric DEM vs SCD: AUC = 0.887 (sens = 0.741, spec = 0.879)	MTA is good at differentiating DEM vs HC. Volumetric measures are good at differentiating DEM vs SCD
Falgàs et al. (2024)	INTER: right *wκ* = 0.7; left *wκ* = 0.86 INTRA: rater 1, left and right *wκ* = 0.91; rater2 right *wκ* = 0.91 and left *wκ* = 0.95	Negative correlations between the scale and GM atrophy	MTA was not the best	MTA is a reliable and valid marker of MTL atrophy but not the best predictor for PPA
Fumagalli et al. (2020)	INTER: *wκ* = 0.85 INTRA: rater 1 *wκ* = 0.92; rater 2 *wκ* = 0.84	Negative correlations between the scale and GM atrophy	PCA vs HC: AUC = 0.85 AD vs HC: AUC = 0.86	MTA is a reliable and valid marker of MTL atrophy
Ferreira et al. (2015)	INTRA: left *wκ* = 0.93 and right *wκ* = 0.94	–	AD vs HC: AUC = 0.838 cMCI vs sMCI: AUC = 0.624	MTA is useful for assessing regional brain atrophy and aiding AD diagnosis, and for determining MCI prognosis
Falgàs et al. (2020)	INTER: left ICC = 0.87, right ICC = 0.83 INTRA: rater 1 left ICC = 0.82 and right ICC = 0.76. Rater 2 left ICC = 0.73 and right ICC = 0.72	–	aAD vs HC: AUC = 0.77 Genetic AD vs young HC: AUC = 0.78 FTD vs HC: AUC = 0.84 svPPA vs HC: AUC = 0.96 Genetic FTD vs HC: AUC = 0.94 FTD vs aAD: AUC = 0.75 FTD vs naAD: AUC = 0.82	Little utility in differential diagnosis of early onset dementia. None of the scale met the requirements for being a valid diagnostic biomarker.
Yuan et al. (2019)	INTER: right ICC = 0.82; left ICC = 0.83	Negative correlations between the scale and GM volume ($r = -0.5$)	Mild AD vs HC: AUC = 0.79 Severe/moderate AD vs HC: AUC = 0.9	MTA is a reliable and valid marker of hip atrophy, and discriminates between AD and HC.
AT				
Harper et al. (2016)	INTER: ICC≥0.57	A small region in the left superior parietal lobule/supramarginal gyrus was positively correlated with GM density	FTLD vs AD + LBD: AUC = 0.63	AT is associated with FTLD pathologies
Fumagalli et al. (2018)	INTER: ICC = 0.77 INTRA: ICC (0.95 0.96)	Negative correlations between the scale and GM atrophy	–	Differences in AT in genetic mutations

Table 2: (continued)

First author	Agreement	Correlation visual vs volumetric	Accuracy	Conclusion
Falgàs et al. (2024)	INTER: right $w\kappa$ = 0.81; left $w\kappa$ = 0.93 INTRA: rater 1, right $w\kappa$ = 0.95 and left $w\kappa$ = 0.94; rater 2 right $w\kappa$ = 0.94 and left $w\kappa$ = 0.95	Negative correlations between the scale and GM atrophy	Left AT lvPPA vs HC: AUC = 0.921 svPPA vs HC: AUC = 0.999 uPPA vs HC: AUC = 0.953 svPPA vs lvPPA: AUC = 0.94 svPPA vs nfvPPA: AUC = 0.925 svPPA vs uPPA: AUC = 0.868	Unstructured expert review is sufficient to confirm or exclude svPPA from other PPAs
Fumagalli et al. (2020)	INTER: $w\kappa$ = 0.64 INTRA: rater 1 $w\kappa$ = 0.79; rater 2 $w\kappa$ = 0.77	Negative correlations between the scale and GM atrophy	PCA vs HC: AUC = 0.82 AD vs HC: AUC = 0.86	AT is a reliable and valid marker of anterior temporal atrophy
Falgàs et al. (2020)	INTER: left ICC = 0.86, right ICC = 0.89 INTRA: rater 1 left ICC = 0.67 and right ICC = 0.68. Rater 2 left ICC = 0.63 and right ICC = 0.75	–	naAD vs HC: AUC = 0.83 FTD vs HC: AUC = 0.90 svPPA vs HC: AUC = 0.98 Genetic FTD vs HC: AUC = 0.94 FTD vs aAD: AUC = 0.82 FTD vs naAD: AUC = 0.84	Little utility in differential diagnosis of early onset dementia. None of the scale met the requirements for being a valid diagnostic biomarker.
Yuan et al. (2019)	INTER: right ICC = 0.75; left ICC = 0.74	Negative correlations between the scale and GM volume (r = −0.22)	Mild AD vs HC: AUC = 0.77 Severe/moderate AD vs HC: AUC = 0.89	MTA is a reliable and valid marker of anterior temporal atrophy, and discriminates between AD and HC.

Healhty controls (HC); mild cognitive impairment (MCI); stable mild cognitive impairment (sMCI); converted mild cognitive impairment (cMCI); amnestic mild cognitive impairment (aMCI); subjective cognitive decline (SCD); Alzheimer's disease (AD); frontotemporal lobar degeneration (FTLD); Lewy body dementia (LBD); subcortical vascular dementia (SVD); demented patients (DEM); posterior cortical atrophy (PCA); primary progressive aphasia (PPA); non-fluent variant of primary progressive aphasia (nfvPPA); logopenic variant of primary progressive aphasia (lvPPA); semantic variant of primary progressive aphasia (svPPA); undetermined variant of primary progressive aphasia (uPPA); amnestic AD (aAD); non amnestic AD (naAD); doctor of medicine (MD); Cohen's weighted kappa ($w\kappa$); Fleiss's kappa (κ); Cohen's kappa ($c\kappa$) Interclass correlation coefficient (ICC); Spearman's rho (r); gray matter (GM); area under the curve (AUC); hippocampus (hip); versus (vs); inter-rater agreement (INTER); intra-rater agreement (INTRA); cerebrospinal fluid (CSF); medial temporal lobe (MTL); medial temporal atrophy (MTA); information not included (–); sensitivity (sens); specificity (spec).

volumetric measures and discriminates between neurodegenerative diseases, between different forms of PPA, and physiological aging (for the studies details see Tables 1 and 2).

3.3 Frontal areas

3.3.1 Fronto-insula (FI)

The fronto-insula (FI) scale (Davies et al. 2009; Fumagalli et al. 2014) assesses the atrophy of the circular sulcus of the insula in the coronal plane on the slice where the anterior commissure becomes visible and the two following posterior (Falgàs et al. 2024) by using four points rating (from 0: no atrophy to 3: severe; Figure S4). The inter-rater and the intra-rater agreement ranged between substantial and near to perfect. Negative correlations were found between the FI scale and the GM density (Harper et al. 2016) and atrophy (Falgàs et al. 2024; Fumagalli et al. 2018; Yuan et al. 2019), suggesting that the scale is a reliable measure of insula atrophy. The scale had good discriminatory ability in distinguishing between the early onset of AD and HC (Falgàs et al. 2020; Harper et al. 2016), and AD severity (Yuan et al. 2019). While the discriminatory abilities between genetic mutations [poor: Fumagalli et al. (2018); excellent: Falgàs et al. (2020)], and PPA variants [poor: Falgàs et al. (2024); good: Falgàs et al. (2020)] are still entangled. The results are reported in Tables 1 and 3.

3.3.2 Orbitofrontal (OF)

The orbitofrontal (OF) scale (Davies et al. 2009; Fumagalli et al. 2014) evaluates the olfactory sulcus in the coronal plane by using four grade system (Figure S4) ranging from 0 (no atrophy) to 3 (severe atrophy) on the most anterior slice where the corpus callosum becomes visible (Falgàs et al. 2024). The inter-rater and the intra-rater agreement ranged between substantial and near to perfect. The negative correlations found between the OF scale and the GM density (Harper et al. 2016) and atrophy (Falgàs et al. 2024; Fumagalli et al. 2018, 2020; Yuan et al. 2019) suggest

Table 3: The table summarizes the results regarding the agreement and the accuracy of the scale measuring frontal areas atrophy.

First author	Agreement	Correlation visual vs volumetric	Accuracy	Conclusion
FI				
Harper et al. (2016)	INTER: ICC≥0.72	Negative correlations between the scale and GM density	FI best scale for distinguish between early onset AD and young HC (AUC = 0.89)	Age effects in the FI suggests the need to account for age in the visual assessment
Fumagalli et al. (2018)	INTER: ICC = 0.75 INTRA: ICC (0.82 0.91)	Negative correlations between the scale and GM atrophy	–	FI is a reliable and valid marker of insula atrophy among groups of genetic mutations
Falgàs et al. (2024)	INTER: right *wκ* = 0.66; left *wκ* = 0.78 INTRA: rater 1, right *wκ* = 0.82 and left *wκ* = 0.90; rater2 right *wκ* = 0.87 and left *wκ* = 0.92	Negative correlations between the scale and GM atrophy	FI was not the best	FI is a reliable and valid marker of insula atrophy but not the best predictor for PPA
Falgàs et al. (2020)	INTER: left ICC = 0.90, right ICC = 0.89 INTRA: rater 1 left ICC = 0.75 and right ICC = 0.63. Rater 2 left ICC = 0.81 and right ICC = 0.74	–	early onset AD vs HC: AUC = 0.76 aAD vs HC: AUC = 0.77 naAD vs HC: ACU = 0.78 FTD vs HC: AUC = 0.86 svPPA vs HC: AUC = 0.89 Genetic FTD vs HC: AUC = 0.94	Little utility in differential diagnosis of early onset dementia. None of the scale met the requirements for being a valid diagnostic biomarker.
Yuan et al. (2019)	INTER: right ICC = 0.75; left ICC = 0.77	Negative correlations between the scale and GM volume (r = −0.42)	Mild AD vs HC: AUC = 0.70 Severe/moderate AD vs HC: AUC = 0.88	FI is a reliable and valid marker of insula atrophy, and discriminates between AD and HC.
OF				
Harper et al. (2016)	INTER: ICC≥0.72	Negative correlations between the scale and GM density	LBD vs HC: AUC = 0.7 LBD vs AD: AUC = 0.52 FTLD vs AD: AUC = 0.68	OF is valid, reliable, but had poor discriminative value
Fumagalli et al. (2018)	INTER: ICC = 0.82 INTRA: ICC (0.89 0.97)	Negative correlations between the scale and GM atrophy	–	OF is a reliable and valid marker of insula trophy among groups of genetic mutations
Falgàs et al. (2024)	INTER: right *wκ* = 0.76; left *wκ* = 0.8 INTRA: rater 1, right *wκ* = 0.8 and left *wκ* = 0.83; rater2 right *wκ* = 0.84 and left *wκ* = 0.84	Negative correlations between the scale and GM atrophy	OF was not the best	OF is a reliable and valid marker of orbital atrophy but not the best predictor for PPA
Fumagalli et al. (2020)	INTER: *wκ* = 0.75 INTRA: rater 1 *wκ* = 0.89; rater 2 *wκ* = 0.8	Negative correlations between the scale and GM atrophy	AD vs HC: AUC = 0.73	OF is a reliable and valid marker of orbital atrophy and fairly distinguish between AD and HC
Falgàs et al. (2020)	INTER: left ICC = 0.90, right ICC = 0.91 INTRA: rater 1 left ICC = 0.65 and right ICC = 0.75. Rater 2 left ICC = 0.75 and right ICC = 0.72	–	early onset AD vs HC: AUC = 0.75 naAD vs HC: ACU = 0.81 FTD vs HC: AUC = 0.8 svPPA vs HC: AUC = 0.79 Genetic FTD vs HC: AUC = 0.92	Little utility in differential diagnosis of early onset dementia. None of the scale met the requirements for being a valid diagnostic biomarker.
Yuan et al. (2019)	INTER: right ICC = 0.7; left ICC = 0.7	Negative correlations between the scale and GM volume (r = −0.3)	Mild AD vs HC: AUC = 0.79 Severe/moderate AD vs HC: AUC = 0.9	OF is a reliable and valid marker, and discriminates between AD and HC.
AC				
Harper et al. (2016)	INTER: ICC≥0.61	Negative correlations between the scale and GM density	LBD vs AD + FTLD: AUC = 0.52	AC is a reliable and valid marker of cingulate atrophy but with poor discriminatory abilities

Table 3: (continued)

First author	Agreement	Correlation visual vs volumetric	Accuracy	Conclusion
Fumagalli et al. (2018)	INTER: ICC = 0.74 INTRA: ICC (0.82 0.9)	Negative correlations between the scale and GM atrophy	–	AC is a reliable and valid marker of cingulate atrophy among groups of genetic mutations
Falgàs et al. (2024)	INTER: right $w\kappa$ = 0.71; left $w\kappa$ = 0.93 INTRA: rater 1, right $w\kappa$ = 0.88 and left $w\kappa$ = 0.94; rater2 right $w\kappa$ = 0.86 and left $w\kappa$ = 0.95	Negative correlations between the scale and GM atrophy	AC was not the best	AC is a reliable and valid marker of cingulate atrophy but not the best predictor for PPA
Falgàs et al. (2020)	INTER: left ICC = 0.88, right ICC = 0.91 INTRA: rater 1 left ICC = 0.88 and right ICC = 0.65. Rater 2 left ICC = 0.86 and right ICC = 0.82	–	early onset AD vs HC: AUC = 0.77 aAD vs HC: ACU = 0.8 FTD vs HC: ACU = 0.83 nfvPPA vs HC: ACU = 0.78 Genetic FTD vs HC: AUC = 0.97	Little utility in differential diagnosis of early onset dementia. None of the scale met the requirements for being a valid diagnostic biomarker.
Benussi et al. (2024)	INTER: ICC = 0.78	–	Prodromal FTD vs HC: AUC = 0.69	The inclusion of AC and plasma neurofilament light increase the accuracy to AUC = 0.9
Yuan et al. (2019)	INTER: right ICC = 0.7; left ICC = 0.71	Negative correlations between the scale and GM volume (r = −0.31)	Mild AD vs HC: AUC = 0.76 Severe/moderate AD vs HC: AUC = 0.9	AC is a reliable and valid marker, and discriminates between AD and HC.

Fronto-insula (FI); orbitofrontal (OF); anterior cingulate (AC); healthy controls (HC); Alzheimer's disease (AD); frontotemporal lobar degeneration (FTLD); Lewy body dementia (LBD); posterior cortical atrophy (PCA); primary progressive aphasia (PPA); frontotemporal dementia (FTD); amnestic AD (aAD); non amnestic AD (naAD); semantic variant of primary progressive aphasia (svPPA); Cohen's weighted kappa ($w\kappa$); interclass correlation coefficient (ICC); correlation value (r); gray matter (GM); area under the curve (AUC); versus (vs); inter-rater agreement (INTER); intra-rater agreement (INTRA); information not included (–).

that the scale is a reliable proxy of olfactory sulcus enlargement. The scale had poor discriminatory values in distinguishing LBD and AD (Harper et al. 2016), LBD and HC (Harper et al. 2016), and FTD and AD (Harper et al. 2016). Furthermore, OF scale is not the best for discriminating PPA variants (Falgàs et al. 2024), but the discriminatory ability is fair between svPPA and HC (Falgàs et al. 2020). However, it fairly distinguishes AD and HC (Fumagalli et al. 2020; Yuan et al. 2019), even in early-onset AD (Falgàs et al. 2020), FTD and HC (Falgàs et al. 2020), and is excellent in the genetic forms of FTD (Falgàs et al. 2020). The results are reported in Tables 1 and 3.

3.3.3 Anterior cingulate (AC)

The anterior cingulate (AC) scale (Davies et al. 2009; Fumagalli et al. 2014) assesses the cingulate sulcus in the coronal plane with four grade system (Figure S4) ranging from 0 (no atrophy) to 3 (severe atrophy) on the most anterior slice where the corpus callosum becomes visible (Falgàs et al. 2024). Both the inter-rater and the intra-rater agreement ranged between substantial and near to perfect. The negative correlations found between the AC scale and the GM density (Harper et al. 2016) and atrophy (Falgàs et al. 2024; Fumagalli et al. 2018; Yuan et al. 2019) suggest that the scale is a reliable measure of anterior cingulate sulcus enlargement. The scale had poor discriminatory abilities in distinguishing LBD and AD + FTD (Harper et al. 2016), prodromal FTD and HC (Benussi et al. 2024), and was not the best for distinguishing PPA variants (Falgàs et al. 2024). However, it showed fair ability in the distinction between AD and HC (Falgàs et al. 2020; Yuan et al. 2019), between nfvPPA and HC (Falgàs et al. 2024), good between FTD and HC (Falgàs et al. 2020), and excellent abilities in distinguishing FTD genetic variants from HC (Falgàs et al. 2020). Interestingly, the combination of plasma neurofilament light and AC scale showed an excellent accuracy in distinguishing prodromal FTD and HC (Benussi et al. 2024). The results are reported in Tables 1 and 3.

3.3.4 Global cortical atrophy-frontal areas (GCA-F)

Two studies evaluated the global atrophy score (GCA) only the in frontal lobe (GCA-F) (Pasquier et al. 1996; Figure S4), finding an inter-rater agreement of $w\kappa$ = 0.59 (Ferreira et al.

2016) and Spearman's rho of 0.257 (Vanhoenacker et al. 2017), and a substantial ($w\kappa$ = 0.7) intra-rater agreement (Ferreira et al. 2016). Negative correlations were found between CGA-F scores and both GM volume (Ferreira et al. 2016) and GM thickness (Ferreira et al. 2016), suggesting that GCA-F is a reliable measure of frontal lobe atrophy. Among the articles included, no study has assessed the diagnostic accuracy of the scale for neurodegenerative diseases. General information about the studies is included in Table 1.

To summarize, the visual rating scale assessing the frontal areas showed high reliability and correlations with GM volumetric measures. The discriminative abilities are excellent in distinguishing FTD genetic forms from controls, and the FI scale seems the most trustable in differentiating other neurodegenerative forms from controls.

3.4 Parietal areas

3.4.1 Posterior atrophy (PA)

The Koedam scale focuses on the parietal lobe atrophy, looking at the posterior cingulate sulcus, precuneus, and the parieto–temporal sulcus in the T1-weighted sagittal and coronal plane (Harper et al. 2015; Koedam et al. 2011). The degree of atrophy in each of these regions is combined to produce a score reflecting overall PA atrophy. The scale is composed of four increments (Figure S4) ranging from 0 to 3 in which scores <1 indicating the absence of PA (Harper et al. 2015).

Both the inter-rater agreement and the intra-rater agreement of the scale are substantial to near perfect (ranging between 0.69–0.85 and 0.71–0.95), respectively. In the original study by Koedam et al. (2011), raters did not undergo a training session; while in other studies (Fumagalli et al. 2018; Harper et al. 2016; Sheng et al. 2020) raters received a training protocol. The presence of training for the raters in this case seems not to have an impact on the raters' consistency and reliability.

The degree of atrophy measured by the PA scale negatively correlated with GM volume in the same posterior areas in all of the studies (Falgàs et al. 2024; Fumagalli et al. 2018, 2020; Harper et al. 2016; Möller et al. 2014; Sheng et al. 2020; Yuan et al. 2019).

The PA scale accurately discriminates between pathological and physiological aging (Falgàs et al. 2020; Ferreira et al. 2015; Möller et al. 2014; Yuan et al. 2019), between different forms of neurodegeneration (Koedam et al. 2011), between those MCI that will convert to AD from the stable MCI (Ferreira et al. 2015), and even in the genetic variants (Falgàs et al. 2020; Fumagalli et al. 2018). Furthermore, the PA scale can fairly distinguish between HC and MCI patients (Sheng et al. 2020). Furthermore, the discriminative power of the scale was similar to the automated classification method (Harper et al. 2016). Interestingly, the combination of MTA and PA scales increased the discriminative ability than the single scale scores both in the AD (Falgàs et al. 2020; Koedam et al. 2011) and in the MCI (Sheng et al. 2020) population.

To summarize, PA as a measure of the brain atrophy in the parietal areas is reliable, consistent with volumetric measures, and discriminates between neurodegenerative diseases and physiological aging. The combination of PA with other visual rating scales, such as the MTA, increases the discriminative power of each scale. Results are reported in Tables 1 and 4.

3.5 Global atrophy

3.5.1 Global cortical atrophy (GCA)

The global cortical atrophy (GCA) evaluates brain atrophy in 13 regions in each hemisphere and the final score is given by the sum of each 13 regions scores (Pasquier et al. 1996). This scale was designed to assess atrophy following stroke and was not specifically developed to evaluate neurodegenerative dementias. However, it has also been applied in the assessment of neurodegenerative diseases. The inter-rater and intra-rater agreement of the GCA scale was evaluated by Koedam et al. (2011), reporting values of $w\kappa > 0.7$ and $w\kappa = 0.85$, respectively, while Ferreira et al. (2015) reported a lower value of intra-rater agreement ($w\kappa$ = 0.7) but still substantial. However, poor diagnostic (AD vs HC, AUC = 0.653) and predictive performances (sMCI vs cMCI, AUC = 0.581) were found (Ferreira et al. 2015). General information about the studies is included in Table 1.

3.6 Other scales of atrophy

Assessing the frontal brain atrophy is crucial for the differential diagnosis between neurodegenerative diseases or psychiatric disorders. For this reason, several visual rating scales have been developed. Among them Davies et al. (2009) developed a scale that includes 15 frontotemporal brain regions, rated in the T1-weighted images by using a scale ranging between 0 (no atrophy) to 4 (most abnormal). Both the inter-rater and intra-rater agreement were substantial and the scales correlated with the volumetric estimation method. Discriminative analysis pointed out disease-specific impairment in several sets of brain regions,

Table 4: The table summarizes the results regarding the agreement and the accuracy of the PA scale.

First author	Agreement	Correlation visual vs volumetric	Accuracy	Conclusion
Koedam et al. (2011)	INTER: $wκ$ = 0.73 INTRA: $wκ$ = 0.93 (rater 1) $wκ$ = 0.95 (rater 2)	–	AD: Sens = 0.58, Spec = 0.95 Combining MTA and PA: Sens = 0.73, Spec = 0.87	PA discriminates between AD and other dementias. The combination of PA and MTA is better.
Möller et al. (2014)	INTER: $wκ$ = 0.85	Patients with a degree of PA had lower total brain volume than patients without PA. Increasing PA more atrophy of the whole brain was detected	PA discriminates well between atrophy and no atrophy	PA scale is quantitatively validated and reliably reflects GM atrophy in parietal regions
Harper et al. (2016)	INTER: ICC≥0.71	Negative partial correlation of GM with visual rating score	AD vs FTD: AUC = 0.6 AD vs DLB + FTD: AUC = 0.54 DLB vs FTD: AUC = 0.54	Automated classification as good as expert reads
Sheng et al. (2020)	INTER: ICC = 0.845 INTRA: ICC (0.709 0.832)	Negative correlations of GM measures with visual rating score	AUC = 0.725 The combination of the MTA and PA showed higher classification accuracy compared with single visual rating scale (cohortA: AUC = 0.818; cohortB: AUC = 0.824)	The combined MTA and PA rating scales have similar discriminative power than the GM assessment. Diagnostic value for distinguishing aMCI vs HC
Fumagalli et al. (2018)	INTER: ICC>0.73 INTRA: rater 1 ICC>0.82, rater 2 ICC>0.89	Negative correlations of PA score with GM measures	–	Differences in AT in genetic mutations
Falgàs et al. (2024)	INTER: right $wκ$ = 0.73; left $wκ$ = 0.7 INTRA: rater 1, right $wκ$ = 0.81 and left $wκ$ = 0.82; rater2 right $wκ$ = 0.88 and left $wκ$ = 0.82	Negative correlations between the scale and GM atrophy	nfvPPA vs HC: left (AUC = 0.817)	PA can accurately discriminate nfvPPA from HC
Fumagalli et al. (2020)	INTER: $wκ$ = 0.69 INTRA: rater 1, $wκ$ = 0.92; rater2, $wκ$ = 0.84	Negative correlations of PA score with GM measures	PCA vs HC: AUC = 0.83 AD vs HC: AUC = 0.73	Visual rating scales have been validated also in PCA
Ferreira et al. (2015)	INTRA: $wκ$ = 0.72	–	AD vs HC: AUC = 0.567 cMCI vs sMCI: AUC = 0.547	PA is useful for assessing regional brain atrophy and aiding AD diagnosis, and for determining MCI prognosis
Falgàs et al. (2020)	INTER: left ICC = 0.85, right ICC = 0.84 INTRA: rater 1 left ICC = 0.75 and right ICC = 0.84. Rater 2 left ICC = 0.74 and right ICC = 0.81	–	naAD vs HC: AUC = 0.84 Genetic AD vs young HC: AUC = 0.81 Genetic FTD vs HC: AUC = 0.92	Little utility in differential diagnosis of early onset dementia. None of the scale met the requirements for being a valid diagnostic biomarker.
Yuan et al. (2019)	INTER: right ICC = 0.74; left ICC = 0.74	Negative correlations between the scale and GM volume (r = −0.45)	Mild AD vs HC: AUC = 0.69 Severe/moderate AD vs HC: AUC = 0.85	PA is a reliable and valid marker, and discriminates between AD and HC.

Parietal atrophy (PA); medial temporal atrophy (MTA); healthy controls (HC); mild cognitive impairment (MCI); converted mild cognitive impairment (cMCI); stable mild cognitive impairment (sMCI); amnestic mild cognitive impairment (aMCI); subjective cognitive decline (SCD); Alzheimer's disease (AD); frontotemporal lobar degeneration (FTLD); frontotemporal dementia (FTD); Lewy body dementia (LBD); posterior cortical atrophy (PCA); non-fluent variant of primary progressive aphasia (nfvPPA); non amnestic AD (naAD); Cohen's weighted kappa ($wκ$); interclass correlation coefficient (ICC); gray matter (GM); area under the curve (AUC); correlation value (r); versus (vs); inter-rater agreement (INTER); intra-rater agreement (INTRA); information not included (–); sensitivity (sens); specificity (spec).

suggesting that the scale can differentiate between neuropathological and physiological aging (Davies et al. 2009).

Chow and colleagues (2011) adapted the scale proposed by Davies et al. (2009) to evaluate the atrophy in the left anterior cingulate and left anterior temporal regions (Harper et al. 2015). The aim of the authors was to investigate the ability of clinicians in detecting localized brain atrophy in MRI patient's scans. The authors found that the raters agreed about the presence of atrophy in those areas. Furthermore, negative correlations between visual rating scores and volumetric measures were detected (Chow et al. 2011). The scale differentiates fairly accurately between FTD patients, and between AD and HC.

Other in-house scales for the rating of brain atrophy and lesions have been developed. Among them, the brain atrophy and lesion index (BALI) was developed by adapting existing visual rating scales (Chen et al. 2010). The BALI categories and grades the gray matter lesions and small vessels, the periventricular lesions, the deep white matter lesions, the basal ganglia and surrounding lesions, the infratentorial region lesions, and the global atrophy, and other lesions (Chen et al. 2010). High inter-rater reliability was found assessing both the T1 and the T2-weighted images, with a fair discriminative ability in differentiating AD patients (Chen et al. 2010). Moreover, the comprehensive visual rating scale (CVRS) has been developed to screen patients with cognitive decline (Jang et al. 2015). The CVRS is made of four parts which scales hippocampal atrophy, cortical atrophy, ventricular enlargement, and small vessel disease. High inter-rater and intra-rater agreement were found. Furthermore, the CVRS subscales correlated with volumetric reduction in the brain regions that the visual scales intended to rate (Jang et al. 2015). Regarding the discriminability of the scale, the AUC of the CVRS was greater than the one of any other single subscale and volumetric measurement (Jang et al. 2015), differentiating fairly between HC and MCI (AUC = 0.7), and satisfactorily between MCI and AD (AUC = 0.68). In the study of Li et al. (2019) the evaluation of T2-weighted images on several brain areas was performed (e.g., frontal, parietal, occipital, precuneus, hippocampus, lateral temporal, and lateral ventricle enlargement), showing a moderate to high consistency in the rater agreements, and correlations with volumetric scores (Li et al. 2019). The discriminative abilities were comparable with the voxel-based specific regional analysis system, but the visual rating scores were not able to differentiate MCI from HC (Li et al. 2019).

Lastly, Silhan et al. (2021) developed the hippocam po-horn percentage (Hip-hop) scale for the assessment of the hippocampus and the parietal atrophy score (PAS) for the evaluation of the atrophy degree in the parietal lobe structures. Both inter-raters and intra-rater agreement were almost near to perfect. Moreover, the hip-hop scale perfectly discriminated between older HC and late AD, and it differentiated fair to good young HC and early AD. On the other hand, the PAS had poor discrimination ability between older HC and late AD, but it discriminated fairly between young HC and early AD (Silhan et al. 2021). Results are reported in Table 1 and Table S3.

4 Discussion and conclusions

The review aimed to assess the reliability, validity, and accuracy of the visual rating scales of atrophy in neurodegenerative diseases as a tool for the clinical differentiation between different forms. We extensively evaluated raters' ability, the correlation between the scales scores and the gray matter brain volume, and the diagnostic accuracy of the scales. Our results showed that: (1) the inter-raters agreement varied from fair to nearly perfect, (2) the intra-rater agreement was at least good across all the studies (see Figure 2 for a summary). Moreover, we showed that (3) visual rating scores were negatively correlated with volumetric measures of atrophy assessed with automated tools. Lastly, we showed that (4) different scales can accurately distinguish between various neurodegenerative diseases (see Figure 3 for a summary). Therefore,

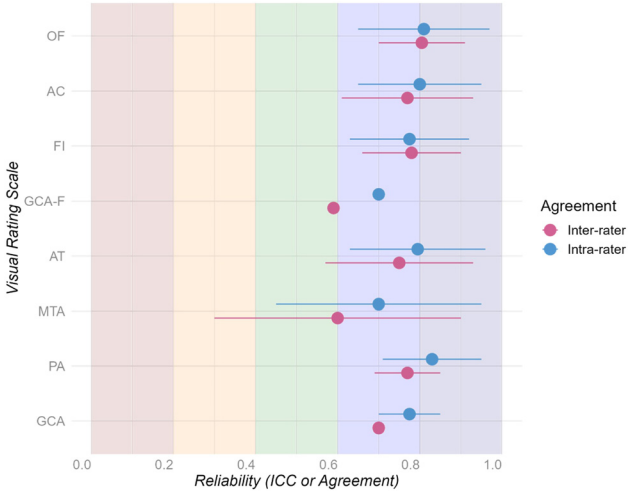

Figure 2: The plot summarizes the inter-rater (purple) and intra-rater (light blue) agreements of the visual rating scales. Poor (red): from 0 to 0.2; fair (orange): from 0.21 to 0.4; moderate (green): from 0.41 to 0.6; substantial (blue): from 0.61 to 0.8; near perfect to perfect (dark blue): from 0.81 to 1. The dot represents the midpoint measured between the lower and the upper agreement level retrieved in the studies. Abbreviations: orbito-frontal (OF); anterior temporal (AT); fronto-insula (FI); anterior cingulate (AC); medial temporal atrophy (MTA); posterior atrophy (PA); global cortical atrophy (GCA); global cortical atrophy-frontal areas (GCA-F). The plot was built in Rstudio (R version 4.3.1) using the ggplot2 (Wickham 2016) and dplyr (Wickham et al. 2023) libraries.

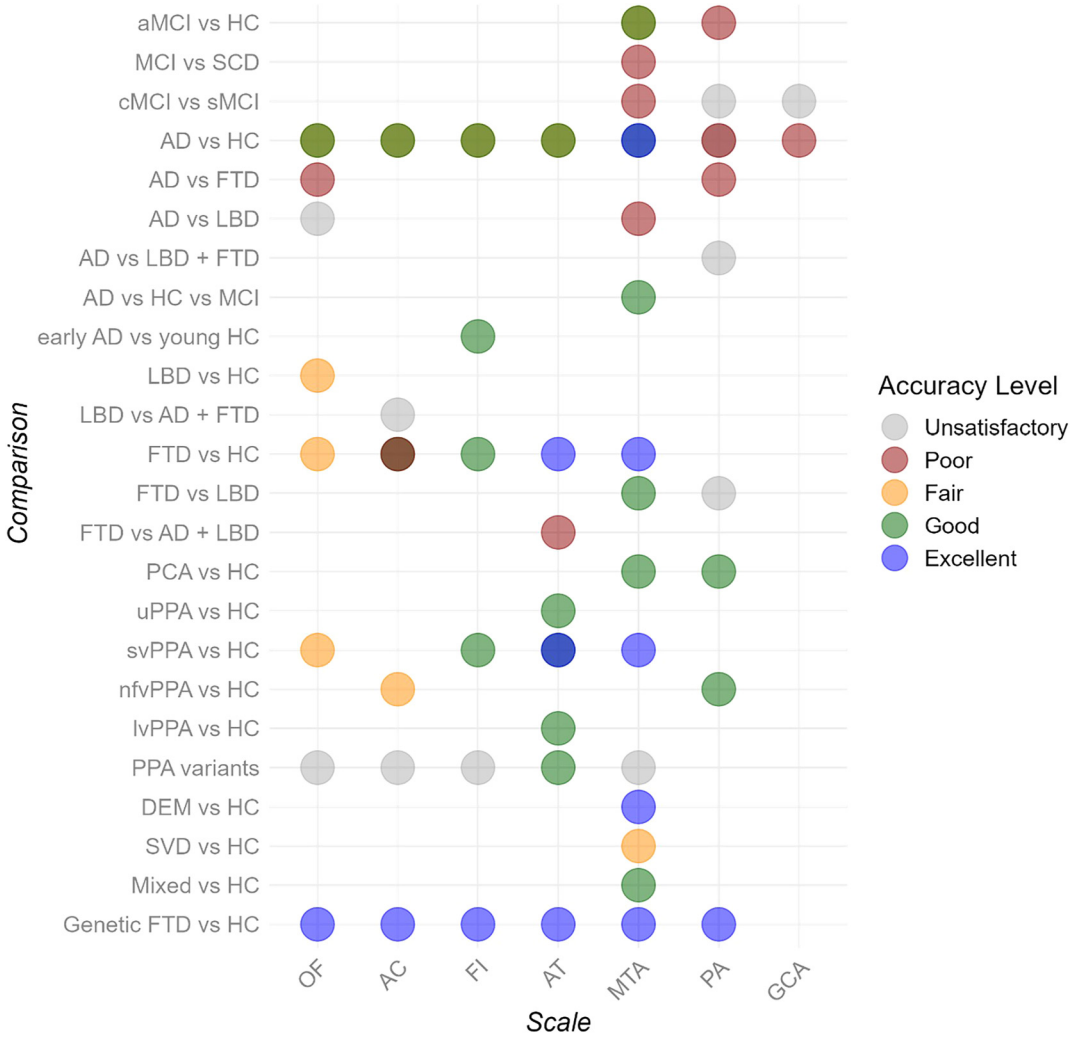

Figure 3: The plot summarizes the accuracy of the visual rating scales. Unsatisfactory (gray): AUC < 60; poor (red): AUC < 70; fair (orange): AUC < 80; good (green): AUC < 90; excellent (blue): AUC ≥ 90. Darker blue represents from good to excellent (MTA in AD vs HC; AT in svPPA vs HC), darker red represents from unsatisfactory to poor (PA in AD vs HC), darker green represents from fair to good (MTA in aMCI vs HC; AT in AD vs HC; FI in AD vs HC; OF in AD vs HC; AC in AD vs HC), and brown represents from poor to fair (AC in FTD vs HC). Abbreviations: amnestic mild cognitive impairment (aMCI); healthy controls (HC); subjective cognitive decline (SCD); Alzheimer's disease (AD); frontotemporal dementia (FTD); Lewy body dementia (LBD); posterior cortical atrophy (PCA); semantic variant of primary progressive aphasia (svPPA); non-fluent variant of primary progressive aphasia (nfvPPA); logopenic variant of primary progressive aphasia (lvPPA); undetermined of primary progressive aphasia (uPPA); primary progressive aphasia (PPA); dementia (DEM); subcortical vascular dementia (SVD); orbito-frontal (OF); anterior temporal (AT); fronto-insula (FI); anterior cingulate (AC); medial temporal atrophy (MTA); posterior atrophy (PA); global cortical atrophy (GCA). The plot was built in Rstudio (R version 4.3.1) using the ggplot2 (Wickham 2016) and dplyr (Wickham et al. 2023) libraries.

the literature review highlights which scales may be most appropriate for specific pathologies, enhancing the accuracy of differential diagnosis.

Harper and colleagues (2015) concluded their review by stating that it was surprising that visual rating scales had not been more widely adopted in routine clinical assessment. After 10 years, this remains an open question. This is particularly striking given that the assessment of hippocampal atrophy using the MTA scale is already justified in clinical routine for AD (Boutet et al. 2012), as MTA is a reliable diagnostic marker distinguishing AD from healthy controls (Park et al. 2021). Interestingly, an incremental diagnostic value for AD could be reached through the combination of different scales, such as the MTA and the PA (e.g., Falgàs et al. 2020; Sheng et al. 2020). However, visual rating scales are still underutilized for the differential diagnosis of other neurodegenerative conditions, despite their potential to complement automated tools and increase diagnostic accuracy. Further research is needed to determine whether integrating the evaluation of other brain areas could improve the differentiation of atypical AD (i.e., PCA) and non-AD dementias (i.e., LBD), ultimately contributing to the

development of standardized criteria and consensus guidelines (Loreto et al. 2023).

In this regard, the implementation of the visual rating scales, such as MTA and/or PA, may not be the best option for the differentiation in the early stages of the neurodegenerative diseases, as their accuracy is generally low, and only fair when differentiating MCI from controls. This result aligns with a recent paper assessing the preclinical state by using the visual atrophy scales (Socher et al. 2025). This could be due to the reduced sensitivity of these scales in distinguishing physiological brain aging from early pathological changes, where age-related atrophy can mask neurodegenerative processes. In contrast, automated measures of gray matter volume appear to be more precise than MTA alone in this population (Sheng et al. 2020), and further research is needed to explore the potential benefits of implementing specialized and automated software for atrophy detection. However, visual rating scales remain crucial tools for clinical assessment, as they allow for a rapid, cost-effective, and accessible evaluation of atrophy patterns. Importantly, different scales might be more suitable for different neurodegenerative diseases. For instance, while MTA is particularly useful for detecting AD-related atrophy, PA may be more relevant for identifying atypical AD. Similarly, frontal atrophy scales such as the FI scale could play a critical role in differentiating FTD from HC, and the AT scale may be implemented for the PPAs spectrum. On the contrary, the GCA scale appeared less optimal for isolating disease-specific atrophy patterns. While our review places greater emphasis on AD, reflecting the predominance of AD-focused studies in literature, we incorporated findings from other neurodegenerative syndromes whenever available. Nevertheless, the number of studies examining visual rating scales in conditions such as FTD, PPA, and other non-AD dementias remains limited. This underrepresentation highlights a critical gap in literature and underscores the need for future research to validate and adapt these scales for a wider range of neurodegenerative diseases.

In addition, although negative correlations between MTA scores and CSF biomarkers have been reported in AD (Enkirch et al. 2018), evidence on the relationship between visual rating scales and PET or CSF biomarkers in other neurodegenerative conditions, such as FTD, remains scarce and needs further investigation. Future studies should further explore the potential of integrating multiple scales to improve diagnostic accuracy in different forms of dementia.

One limitation of our study is the absence of measures of cognitive status and disease severity, such as neuropsychological test scores about the global cognition (e.g., the Mini-Mental State Examination) or domain specific assessments (e.g., Trail making test, and Rey–Osterrieth complex figure). Other limitations include the small number of studies investigating certain visual rating scales (e.g., FI and AC), and the lack of stratification based on age specific cut-off values, as this information was not consistently reported across studies. Additionally, we did not apply exclusion criteria related to the minimum sample size, MRI acquisition protocols, or scanner field strength, which may introduce variability in imaging quality and interpretation. However, these methodological characteristics (e.g., sample size and sex distribution, MRI field strength) were extracted from the included studies, thereby providing context for interpreting potential sources of heterogeneity. Furthermore, demographic heterogeneity, such as differences in age, and ethnicity, was not systematically accounted for, potentially affecting both the generalizability and diagnostic accuracy of the findings. Finally, we acknowledge the potential for publication bias, as studies reporting significant or positive findings may be more likely to be published and included in our review. This may have influenced the overall representation of diagnostic performance and study characteristics in the included literature.

In conclusion, this review provides an overview of the most widely used visual rating scales in both clinical and research settings, emphasizing their reliability and clinical utility. Our findings suggest that visual rating scales remain valuable tools for the assessment of brain atrophy. While some scales, such as MTA, are already well-established for AD diagnosis, the differential diagnostic potential of other scales should be further explored. By highlighting the most useful and robust visual rating scales for differentiating between physiological and pathological aging, as well as among different neurodegenerative diseases, our review underscores the importance of a more systematic and widespread implementation of these tools in clinical practice.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: During the preparation of this work the authors used ChatGPT and DeepL Write in order to check for grammar mistakes and to improve the readability of the work. After using this tool/service, the authors reviewed and edited the content as needed.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

Benussi, A., Premi, E., Grassi, M., Alberici, A., Cantoni, V., Gazzina, S., Archetti, S., Gasparotti, R., Fumagalli, G.G., Bouzigues, A., et al. (2024). Diagnostic accuracy of research criteria for prodromal frontotemporal dementia. *Alzheimers Res. Ther.* 16: 10.

Boutet, C., Chupin, M., Colliot, O., Sarazin, M., Mutlu, G., Drier, A., Pellot, A., Dormont, D., Lehéricy, S., and Alzheimer's Disease Neuroimaging Initiative (2012). Is radiological evaluation as good as computer-based volumetry to assess hippocampal atrophy in Alzheimer's disease? *Neuroradiology* 54: 1321–1330.

Buzi, G., Fornari, C., Perinelli, A., and Mazza, V. (2023). Functional connectivity changes in mild cognitive impairment: a meta-analysis of M/EEG studies. *Clin. Neurophysiol.* 156: 183–195.

Chen, W., Song, X., Zhang, Y., Darvesh, S., Zhang, N., D'Arcy, R.C., Black, S., and Rockwood, K. (2010). An MRI-based semiquantitative index for the evaluation of brain atrophy and lesions in Alzheimer's disease, mild cognitive impairment and normal aging. *Dement Geriatr. Cogn. Disord.* 30: 121–130.

Chow, T.W., Gao, F., Links, K.A., Ween, J.E., Tang-Wai, D.F., Ramirez, J., Scott, C.J., Freedman, M., Stuss, D.T., and Black, S.E. (2011). Visual rating versus volumetry to detect frontotemporal dementia. *Dement Geriatr. Cogn. Disord.* 31: 371–378.

Davies, R.R., Kipps, C.M., Mitchell, J., Kril, J.J., Halliday, G.M., and Hodges, J.R. (2006). Progression in frontotemporal dementia: identifying a benign behavioral variant by magnetic resonance imaging. *Arch. Neurol.* 63: 1627–1631.

Davies, R.R., Scahill, V.L., Graham, A., Williams, G.B., Graham, K.S., and Hodges, J.R. (2009). Development of an MRI rating scale for multiple brain regions: comparison with volumetrics and with voxel-based morphometry. *Neuroradiology* 51: 491–503.

de Vugt, M.E. and Verhey, F.R. (2013). The impact of early dementia diagnosis and intervention on informal caregivers. *Prog. Neurobiol.* 110: 54–62.

Dubois, B., Feldman, H.H., Jacova, C., Dekosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.* 6: 734–746.

Enkirch, S.J., Traschütz, A., Müller, A., Widmann, C.N., Gielen, G.H., Heneka, M.T., Jurcoane, A., Schild, H.H., and Hattingen, E. (2018). The ERICA Score: an MR imaging-based visual scoring system for the assessment of entorhinal Cortex atrophy in Alzheimer disease. *Radiology* 288: 226–333.

Falgàs, N., Balasa, M., Bargalló, N., Borrego-Écija, S., Ramos-Campoy, O., Fernández-Villullas, G., Bosch, B., Olives, J., Tort-Merino, A., Antonell, A., et al. (2020). Diagnostic accuracy of MRI visual rating scales in the diagnosis of early onset cognitive impairment. *J. Alzheimers Dis.* 73: 1575–1583.

Falgàs, N., Sacchi, L., Carandini, T., Montagut, N., Conte, G., Triulzi, F., Galimberti, D., Arighi, A., Sanchez-Valle, R., and Fumagalli, G.G. (2024). Utility of visual rating scales in primary progressive aphasia. *Alzheimers Res. Ther.* 16: 73.

Ferreira, D., Cavallin, L., Granberg, T., Lindberg, O., Aguilar, C., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., et al. (2016). Quantitative validation of a visual rating scale for frontal atrophy: associations with clinical status, APOE e4, CSF biomarkers and cognition. *Eur. Radiol.* 26: 2597–2610.

Ferreira, D., Cavallin, L., Larsson, E.M., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., et al. (2015). Practical cut-offs for visual rating scales of medial temporal, frontal and posterior atrophy in Alzheimer's disease and mild cognitive impairment. *J. Intern. Med.* 278: 277–290.

Fischbach-Boulanger, C., Fitsiori, A., Noblet, V., Baloglu, S., Oesterle, H., Draghici, S., Philippi, N., Duron, E., Hanon, O., Dietemann, J.L., et al. (2018). T1- or T2-weighted magnetic resonance imaging: what is the best choice to evaluate atrophy of the hippocampus? *Eur. J. Neurol.* 25: 775–781.

Fumagalli, G.G., Basilico, P., Arighi, A., Bocchetta, M., Dick, K.M., Cash, D.M., Harding, S., Mercurio, M., Fenoglio, C., Pietroboni, A.M., et al. (2018). Distinct patterns of brain atrophy in genetic frontotemporal dementia initiative (GENFI) cohort revealed by visual rating scales. *Alzheimers Res. Ther.* 10: 46.

Fumagalli, G.G., Basilico, P., Arighi, A., Mercurio, M., Scarioni, M., Carandini, T., Colombi, A., Pietroboni, A.M., Sacchi, L., Conte, G., et al. (2020). Parieto-occipital sulcus widening differentiates posterior cortical atrophy from typical Alzheimer disease. *Neuroimage Clin.* 28: 102453.

Fumagalli, G.G., Gordon, E., Harper, L., Lehmann, M., Hyare, H., Warren, J.D., Schott, J.M., and Rohrer, J.D. (2014). 9th International conference on frontotemporal dementias P.252 development of a visual rating scale for atrophy of the anterior cingulate, insula and frontal lobes. *Am. J. Neurodegener. Dis.* 3: 1–375.

Harper, L., Barkhof, F., Fox, N.C., and Schott, J.M. (2015). Using visual rating to diagnose dementia: a critical evaluation of MRI atrophy scales. *J. Neurol. Neurosurg. Psychiatry* 86: 1225–1233.

Harper, L., Fumagalli, G.G., Barkhof, F., Scheltens, P., O'Brien, J.T., Bouwman, F., Burton, E.J., Rohrer, J.D., Fox, N.C., Ridgway, G.R., et al. (2016). MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases. *Brain* 139: 1211–1225.

Jang, J.W., Park, S.Y., Park, Y.H., Baek, M.J., Lim, J.S., Youn, Y.C., and Kim, S. (2015). A comprehensive visual rating scale of brain magnetic resonance imaging: application in elderly subjects with Alzheimer's disease, mild cognitive impairment, and normal cognition. *J. Alzheimers Dis.* 44: 1023–1034.

Jongsiriyanyong, S. and Limpawattana, P. (2018). Mild cognitive impairment in clinical practice: a review article. *Am. J. Alzheimers Dis. Other Demen.* 33: 500–507.

Kipps, C.M., Davies, R.R., Mitchell, J., Kril, J.J., Halliday, G.M., and Hodges, J.R. (2007). Clinical significance of lobar atrophy in frontotemporal dementia: application of an MRI visual rating scale. *Dement Geriatr. Cogn. Disord.* 23: 334–342.

Koedam, E.L., Lehmann, M., van der Flier, W.M., Scheltens, P., Pijnenburg, Y.A., Fox, N., Barkhof, F., and Wattjes, M.P. (2011). Visual assessment of posterior atrophy development of a MRI rating scale. *Eur. Radiol.* 21: 2618–2625.

Koikkalainen, J., Rhodius-Meester, H., Tolonen, A., Barkhof, F., Tijms, B., Lemstra, A.W., Tong, T., Guerrero, R., Schuh, A., Ledig, C., et al. (2016). Differential diagnosis of neurodegenerative diseases using structural MRI data. *Neuroimage Clin.* 11: 435–449.

Li, F., Takechi, H., Saito, R., Ayaki, T., Kokuryu, A., Kuzuya, A., and Takahashi, R. (2019). A comparative study: visual rating scores and the voxel-based specific regional analysis system for Alzheimer's disease on magnetic resonance imaging among subjects with Alzheimer's disease, mild cognitive impairment, and normal cognition. *Psychogeriatrics* 19: 95–104.

Lombardi, G., Crescioli, G., Cavedo, E., Lucenteforte, E., Casazza, G., Bellatorre, A.G., Lista, C., Costantino, G., Frisoni, G., Virgili, G., et al. (2020). Structural magnetic resonance imaging for the early diagnosis

of dementia due to Alzheimer's disease in people with mild cognitive impairment. *Cochrane Database Syst. Rev.* 3: CD009628.

Loreto, F., Gontsarova, A., Scott, G., Patel, N., Win, Z., Carswell, C., Perry, R., and Malhotra, P. (2023). Visual atrophy rating scales and amyloid PET status in an Alzheimer's disease clinical cohort. *Ann. Clin. Transl. Neurol.* 10: 619–631.

Mårtensson, G., Håkansson, C., Pereira, J.B., Palmqvist, S., Hansson, O., van Westen, D., and Westman, E. (2020). Medial temporal atrophy in preclinical dementia: visual and automated assessment during six year follow-up. *Neuroimage Clin.* 27: 102310.

Min, J., Moon, W.J., Jeon, J.Y., Choi, J.W., Moon, Y.S., and Han, S.H. (2017). Diagnostic efficacy of structural MRI in patients with mild-to-moderate Alzheimer disease: automated volumetric assessment versus visual assessment. *AJR Am. J. Roentgenol.* 208: 617–623.

Molinder, A., Ziegelitz, D., Maier, S.E., and Eckerström, C. (2021). Validity and reliability of the medial temporal lobe atrophy scale in a memory clinic population. *BMC Neurol.* 21: 289.

Möller, C., van der Flier, W.M., Versteeg, A., Benedictus, M.R., Wattjes, M.P., Koedam, E.L., Scheltens, P., Barkhof, F., and Vrenken, H. (2014). Quantitative regional validation of the visual rating scale for posterior cortical atrophy. *Eur. Radiol.* 24: 397–404.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372: n71.

Park, H.Y., Park, C.R., Suh, C.H., Shim, W.H., and Kim, S.J. (2021). Diagnostic performance of the medial temporal lobe atrophy scale in patients with Alzheimer's disease: a systematic review and meta-analysis. *Eur. Radiol.* 31: 9060–9072.

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

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H.C., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E.C., and Valk, J. (1992). Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *J. Neurol. Neurosurg. Psychiatry* 55: 967–972.

Sheng, C., Sun, Y., Wang, M., Wang, X., Liu, Y., Pang, D., Liu, J., Bi, X., Du, W., Zhao, M., et al. (2020). Combining visual rating scales for medial temporal lobe atrophy and posterior atrophy to identify amnestic mild cognitive impairment from cognitively normal older adults: evidence based on two cohorts. *J. Alzheimers Dis.* 77: 323–337.

Silhan, D., Pashkovska, O., and Bartos, A. (2021). Alzheimer's disease neuroimaging initiative. Hippocampo-Horn percentage and parietal atrophy Score for easy visual assessment of brain atrophy on magnetic resonance imaging in Early- and late-onset Alzheimer's Disease. *J. Alzheimers* 84: 1259–1266.

Socher, K.L.R., Nunes, D.M., Lopes, D.C.P., Coutinho, A.M.N., Faria, D.P., Squarzoni, P., Busatto Filho, G., Buchpiguel, C.A., Nitrini, R., and Brucki, S.M.D. (2025). Diagnosing preclinical and clinical Alzheimer's disease with visual atrophy scales in the clinical practice. *Arq. Neuropsiquiatr.* 83: 1–7.

Tolboom, N., van Der Flier, W.M., Boverhoff, J., Yaqub, M., Wattjes, M.P., Raijmakers, P.G., Barkhof, F., Scheltens, P., Herholz, K., Lammertsma, A.A., et al. (2010). Molecular imaging in the diagnosis of Alzheimer's disease: visual assessment of [11C] PIB and [18F] FDDNP PET images. *JNNP* 81: 882–884.

Vanhoenacker, A.S., Sneyers, B., De Keyzer, F., Heye, S., and Demaerel, P. (2017). Evaluation and clinical correlation of practical cut-offs for visual rating scales of atrophy: normal aging versus mild cognitive impairment and Alzheimer's disease. *Acta Neurol. Belg.* 117: 661–669.

Westman, E., Cavallin, L., Muehlboeck, J.S., Zhang, Y., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Spenger, C., et al. (2011). Sensitivity and specificity of medial temporal lobe visual ratings and multivariate regional MRI classification in Alzheimer's disease. *PLoS One* 6: e22506.

Whiting, P.F., Rutjes, A.W., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M., Sterne, J.A., Bossuyt, P.M., and QUADAS-2 Group (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med*. 155: 529–536.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). dplyr: a grammar of data manipulation. R package version 1.1.4, Available at: https://CRAN.R-project.org/package=dplyr.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York, Available at: https://ggplot2.tidyverse.org.

Wittens, M.M.J., Allemeersch, G.J., Sima, D.M., Vanderhasselt, T., Raeymaeckers, S., Fransen, E., Smeets, D., de Mey, J., Bjerke, M., and Engelborghs, S. (2024). Towards validation in clinical routine: a comparative analysis of visual MTA ratings versus the automated ratio between inferior lateral ventricle and hippocampal volumes in Alzheimer's disease diagnosis. *Neuroradiology* 66: 487–506.

Yuan, Z., Pan, C., Xiao, T., Liu, M., Zhang, W., Jiao, B., Yan, X., Tang, B., and Shen, L. (2019). Multiple visual rating scales based on structural MRI and a novel prediction model combining visual rating scales and age stratification in the diagnosis of Alzheimer's Disease in the Chinese population. *Front. Neurol.* 10: 93.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/revneuro-2025-0066).