MEDIAL TEMPORAL ATROPHY IN PRECLINICAL DEMENTIA: VISUAL AND AUTOMATED ASSESSMENT DURING SIX YEAR FOLLOW-UP

Gustav Mårtensson[1,*], Claes Håkansson[2], Joana B. Pereira[1], Sebastian Palmqvist[3,4], Oskar Hansson[3,4], Danielle van Westen[2,5,†], and Eric Westman[1,6,†]

[*]Corresponding author: gustav.martensson@ki.se
[†]Shared last author.
[1]Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden.
[2]Diagnostic Radiology, Institution for Clinical Sciences, Lund University, Lund, Sweden.
[3]Clinical Memory Research Unit, Department of Clinical Sciences in Malmö, Lund University, Lund, Sweden.
[4]Memory Clinic, Skåne University Hospital, Malmö, Sweden.
[5]Image and Function, Skåne University Hospital, Lund, Sweden.
[6]Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.

ABSTRACT

Medial temporal lobe (MTL) atrophy is an important morphological marker of many dementias and is closely related to cognitive decline. In this study we aimed to characterize longitudinal progression of MTL atrophy in 93 individuals with subjective cognitive decline and mild cognitive impairment followed up over six years, and to assess if clinical rating scales are able to detect these changes. All MRI images were visually rated according to Scheltens' scale of medial temporal atrophy (MTA) by two neuroradiologists and AVRA, a software for automated MTA ratings. The images were also segmented using FreeSurfer's longitudinal pipeline in order to compare the MTA ratings to volumes of the hippocampi and inferior lateral ventricles. We found that MTL atrophy rates increased with CSF biomarker abnormality, used to define preclinical stages of Alzheimer's Disease. Both AVRA's and the radiologists' MTA ratings showed a similar longitudinal trajectory as the subcortical volumes, suggesting that visual rating scales provide a valid alternative to automatic segmentations. While the MTA scores from each radiologist showed strong correlations to subcortical volumes, the inter-rater agreement was low. We conclude that the main limitation of quantifying MTL atrophy with visual ratings in clinics is the subjectiveness of the assessment.

1 Introduction

Atrophy of the medial temporal lobe (MTL) is an important diagnostic biomarker in many different dementias, including Alzheimer's Disease (AD). In research we quantify atrophy using automatic softwares that compute volume or thickness measures of regions of interests, specified by a neuroanatomical atlas. These softwares are either not sufficiently reliable for clinical usage, and the ones that have been approved are not widely implemented. To quantify atrophy in clinics, radiologists visually assess the degree of atrophy in a brain region according to established rating scales.

The most widely used rating scale in clinical practice is Scheltens' scale of Medial Temporal Atrophy (MTA) (Scheltens et al., 1992; Vernooij et al., 2019). The MTA scale quantifies the level of atrophy in hippocampus (HC) and its surrounding structures, the choroid fissure and inferior lateral ventricle (ILV). The MTA scale has been shown to reliably distinguish individuals with AD from healthy elderly (Scheltens et al., 1992; Wahlund et al., 1999; Westman et al., 2011). It is an ordinal scale ranging from 0 (no atrophy) to 4 (end-stage atrophy) where an integer score is given for each hemisphere. In Fig. 1 we provide examples of each score. Several studies have reported on the diagnostic ability and relevant clinical cut-offs of the MTA scale (Westman et al., 2011; Scheltens et al., 1992; Ferreira et al., 2015), and others have argued for the importance of reporting MTA in the clinical routing (Torisson et al., 2015; Håkansson et al., 2019; Wahlund et al., 2017).

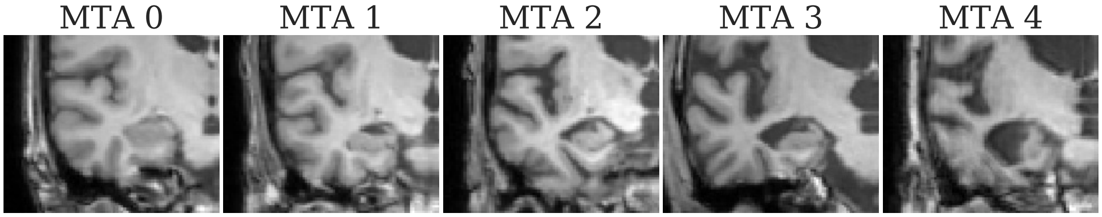

Figure 1: Example of the Scheltens' MTA scale, with progressive atrophy of the hippocampus, the choroid fissure and the inferior lateral ventricle. The image selected for each score was given the same rating by both radiologists in this study. Each hemisphere is rated individually.

Longitudinal progression of medial temporal atrophy, quantified through e.g. hippocampal volumes, has been studied in cognitively normal subjects as well as in preclinical, prodromal and probable AD (Rusinek et al., 2003; Ridha et al., 2006; Henneman et al., 2009a; Pettigrew et al., 2017). The reported annual decrease in HC volume varies between the studies. Rusinek et al. (2003) found a 0.36% volume loss/year for cognitively stable subjects, and a greater loss (1%/year) in individuals with cognitive decline. Henneman et al. (2009a) reported 2.2% annual HC volume loss for healthy controls, with greater atrophy rates in patients with MCI (-3.8%/year) and AD (-4.0%/year). Another study reported an up to 8% decrease in HC volume per year in asymptomatic individuals at risk of familial AD (Fox et al., 1996). Despite the large interest in longitudinal MTL atrophy, no studies have investigated how these corresponds to clinical MTA ratings.

The aim of this paper is to investigate longitudinal changes in the MTL in individuals with subjective cognitive decline (SCD) or mild cognitive impairment (MCI), and whether clinical rating scales can detect these changes. Two neuroradiologists and AVRA (Automatic Visual Ratings of Atrophy)—our recently developed software providing automated continuous MTA scores—rated 93 individuals scanned four times over six years using Scheltens' MTA scale. Further, all images were segmented using the longitudinal FreeSurfer pipeline to extract hippocampal and inferior lateral ventricle volumes. We calculated atrophy rates of the MTL for visual and automated measures to understand what progression to expect in different stages of preclinical dementia.

2 Methods

2.1 Study population

The study population was part of the prospective and longitudinal Swedish BioFINDER (Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably) study (www.biofinder.se) and comprised non-demented people with subjective and objective cognitive decline. All patients were consecutively enrolled at three outpatient memory clinics and were assessed by physicians specialized in dementia disorders. Inclusion criteria were: 1) referred to the memory clinics because of cognitive symptoms, 2) not fulfilling dementia criteria, 3) MMSE score of 24–30 points, 4) age 60–80 years and, 5) fluent in Swedish. Exclusion criteria were: 1) cognitive impairment that without doubt could be explained by a condition other than prodromal dementia, 2) severe somatic disease, and 3) refusing lumbar puncture or neuropsychological investigation. A neuropsychological battery assessing four broad cognitive domains including verbal ability, visuospatial construction, episodic memory, and executive functions was performed and a senior neuropsychologist then stratified all patients into those with SCD (no measurable cognitive deficits) or MCI according to the consensus criteria for MCI (Petersen, 2004). From this larger cohort we selected all individuals who had been followed up three times over the course of six years.

As in the work by Pettigrew et al. (2017), we stratified the cohort based on abnormality in CSF amyloid-β (A) and phosphorylated tau (T) levels analyzed with Euroimmun essays (EUROIMMUN AG, Lübeck, Germany). We applied the cut-off Aβ_{42}/Aβ_{40} < 0.10 (Janelidze et al., 2016) to define A^+ and p-tau > 72 pg/ml (Mattsson et al., 2018) for T^+. This yielded the subgroups A^-T^- (i.e. denoting normal amyloid-β and p-tau levels), A^+T^-, and A^+T^+. No individuals displayed the CSF combination A^-T^+. Demographics and clinical characteristics of these groups are summarized in Table 1.

Table 1: Demographics of the included participants at baseline. p-values were computed using Kruskal-Wallis H-test, testing the null hypothesis that medians are equal in all subgroups.

	All	A^-T^-	A^+T^-	A^+T^+	p-value
N	93	54	18	21	—
SCD/MCI	61/32	42/12	8/10	11/10	—
Age at bl.	70.06 ± 5.41	69.71 ± 5.57	70.18 ± 4.55	70.86 ± 5.58	0.208
Sex, F (%)	57.0	64.8	50.0	42.9	0.001
ApoE4 carriers (%)	38.7	14.8	66.7	76.2	<.001
Education (years)	12.01 ± 3.30	11.91 ± 3.34	11.67 ± 3.42	12.57 ± 3.02	0.108
MMSE at bl.	28.26 ± 1.72	28.57 ± 1.46	28.06 ± 1.75	27.62 ± 2.06	<.001
ADAS-DWR at bl	4.24 ± 2.73	3.41 ± 2.39	5.17 ± 2.41	5.38 ± 3.05	<.001
CSF $A\beta_{42/40}$ ratio	0.12 ± 0.04	0.15 ± 0.03	0.09 ± 0.02	0.07 ± 0.02	—
CSF $A\beta_{42}$ (pg/ml)	611.8 ± 259.4	788.0 ± 182.7	370.6 ± 135.5	399.3 ± 145.3	—
CSF p-tau (pg/ml)	56.7 ± 36.5	35.7 ± 10.9	47.9 ± 14.4	113.4 ± 27.6	—
N conv. to dementia (to AD)	19 (13)	5 (0)	5 (5)	9 (8)	—

2.2 MRI protocol

All T_1-weighted MRI scans were acquired with an MPRAGE protocol on a 3T Siemens TrioTim with the following parameters: 1.2 mm slice thickness, 0.98 mm inplane resolution, 3.37 ms Echo Time, 1950 ms Repetition Time, 900 ms Inversion Time, and 9° Flip angle.

2.3 Visual assessments

Two neuroradiologists (*Rad. 1* and *Rad. 2*) rated all available images according to Scheltens' MTA scale (Scheltens et al., 1992), see Fig. 1. The raters were blinded to sex, age, diagnosis, amyloid-β and tau status, subject ID and timepoint to not bias the ratings. Both radiologists assess MTA on a regular basis as part of their clinical work but have not trained together to facilitate ratings consistency.

2.4 Automated methods

For automated MTA ratings we used our recently proposed software AVRA[1] v0.8 (Mårtensson et al., 2019a). Briefly, AVRA is a deep learning model that was trained on more than 3000 MRI images from multiple cohorts rated by a single radiologist (none of the raters in the current study). It is based on convolutional neural networks and predicts MTA from features extracted from the raw images (i.e., not volumetric data), similar to how a radiologist would perform the assessment. The model has previously demonstrated substantial inter-rater agreement levels in multiple imaging cohorts from various memory clinics (Mårtensson et al., 2019a,b). The first step of the processing pipeline of AVRA is to align the anterior and posterior commissures (AC-PC) using FSL FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002). We visually inspected the rigid registrations to ensure that the AC-PC alignment had not failed, but no AVRA ratings were discarded based on this. Contrary to the radiologist ratings, AVRA outputs continuous MTA scores. This allows for capturing more subtle longitudinal changes in the MTA scores that is not possible with a discretized scale.

All MRI scans were processed through TheHiveDB system (Muehlboeck et al., 2014) with FreeSurfer[2] 6.0.0 for automatic segmentation of cortical and subcortical structures, such as hippocampi and inferior lateral ventricles (Dale et al., 1999; Fischl et al., 2004). First, all images were processed cross-sectionally, and each output was visually inspected to detect images with inaccurate hippocampal or ventrical segmentation. Images that passed quality control were re-processed with FreeSurfer's longitudinal pipeline for more consistent segmentation (Reuter et al., 2012). The longitudinal output was once again visually inspected and cases with poor hippocampal or ventrical segmentations excluded. In total, 339 (out of 372) images from 87 (out of 93) participants were included in the study for further analyses.

2.5 Analyses

The analyses revolve around two central themes: to study the sensitivity and reliability of MTA ratings in a longitudinal setting, and to characterize medial temporal atrophy in preclinical dementia.

[1] *Automatic visual Ratings of Atrophy*, freely available at `https://github.com/gsmartensson/avra_public`

[2] Freely available at `http://surfer.nmr.mgh.harvard.edu/`

Table 2: Inter-rater agreements (κ_w) and Spearman correlations (r_s) between radiologists' ratings, hippocampal (HC) and inferior lateral ventricle (ILV) volumes.

		Rad. 1		Rad. 2		AVRA	
Measure	Metric	Left	Right	Left	Right	Left	Right
Rad. 1	κ_w			0.30	0.36	0.58	0.61
Rad. 2	κ_w	0.30	0.36			0.30	0.35
AVRA	κ_w	0.58	0.61	0.30	0.35		
Rad. 1	r_s			0.77	0.75	0.74	0.76
Rad. 2	r_s	0.77	0.75			0.84	0.84
AVRA	r_s	0.74	0.76	0.84	0.84		
HC vol.	r_s	-0.57	-0.49	-0.53	-0.47	-0.54	-0.55
ILV vol.	r_s	0.78	0.79	0.82	0.84	0.87	0.88
MMSE	r_s	-0.39	-0.36	-0.34	-0.33	-0.30	-0.31
ADAS-DWR	r_s	0.45	0.39	0.47	0.41	0.41	0.39

For the first theme we used Cohen's weighted kappa $\kappa_w \in [-1, 1]$ to assess the agreement between two sets of ratings. As there is no ground truth available, κ_w is a common metric to report in studies using visual ratings, where a high inter- and intra-rater agreement suggests that the rater is reliable and consistent. We further compared the manual and automated ratings to hippocampal and inferior lateral ventricle volumes. Although MTA is rated in a single slice—and does not assess volumes—we assume that reliable ratings should (anti-)correlate strongly with HC and ILV volumes. We studied visual rating sensitivity, i,e. their ability to capture MTL atrophy, by comparing within-subject changes in MTA ratings ("ΔMTA") to changes in HC and ILV volume ("ΔHC" and "ΔILV").

Characterizing MTL atrophy in preclinical dementia was done by studying the cross-sectional and longitudinal progression of MTA scores, HC volumes and ILV volumes as a function of age. We approximated the average annual change in MTA scores ("ΔMTA/year"), Mini Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-delayed word recall (ADAS-DWR), HC and ILV volumes by fitting a least-squares regression line for each individual and measure. To study the effect of clinical status (i.e. SCD or MCI), we performed additional analyses on SCD and MCI subjects separately within each CSF group. The analyses including volumetric data were performed on the subset of images that passed the visual quality control.

3 Results

The rating agreements between radiologists and AVRA, and their correlations to HC and ILV volumes, are shown in Table 2. The weighted kappa agreements between the raters ranged from *fair* ($\kappa_w \in$ [0.2,0.4)) to *substantial* ($\kappa_w \in$ [0.6,0.8)) (Landis and Koch, 1977). All sets of ratings demonstrated similar Spearman correlations strengths to HC ($r_s \in$ [-0.57,-0.47]) and ILV ($r_s \in$ [0.78,0.88]) volumes. Violinplots illustrating the distribution of HC and ILV volumes per MTA score and rater are shown in Fig. 2, where we note that Rad. 1 systematically gave higher MTA scores than Rad. 2. We include confusion matrices between rating sets for left and right hemispheres as Supplementary data, Tables S1-S3.

In Table 3 the baseline characteristics and average annual progression rates among study participants for all sets of ratings, MTL volumes, MMSE and ADAS-DWR are shown. No clear pattern was found between CSF groups in the cross-sectional baseline measures. However, all MTL measures showed that the atrophy rates increased with progressing AD CSF pathology, with the exception of the ratings from Rad. 1 that showed a milder progression in the A$^+$T$^-$ group. By assessing the SCD and MCI patients separately, we observed that the CSF group differences in atrophy rates were larger in the MCI subset. We further noted that the atrophy rates were greater in the SCD subjects in A$^+$T$^+$ than in the MCI patients in the A$^-$T$^-$ group.

In Fig. 3 the trajectories of each study participant are displayed for left MTA (predicted with AVRA), HC volume and ILV volume respectively. (The measures of the right hemisphere showed similar characteristics, and are provided as Supplementary data). The longitudinal trajectories of the FreeSurfer measures were generally smoother than AVRA's MTA scores, which were not monotonically increasing for individuals, suggesting some degree of rating variability. From Fig. 3 we see that the MTL measures of the MCI patients (orange lines) were generally more pathological than the SCD subjects (blue lines), which is confirmed in Table 3. We include examples of MRI scans for all timepoints for randomly selected participants as Supplementary data, Fig. S1.

Table 3: Average baseline (bl) MTA ratings and volumes, and average annual change for individuals with different CSF statuses. Rows in bold denotes entries where the whole CSF group was considered (i.e. SCD's and MCI's), and 'SCD/MCI only' refers to the subset of SCD/MCI subjects within the CSF group. ΔMTA/year refers to the average annual change in MTA score of the study participants. The reported p-values were computed using Kruskal-Wallis H-test to test the null-hypothesis that the population medians of all groups were equal. Applying a Bonferroni correction to a significance level of $\alpha = 0.01$ would render all p-values <.0001 significant.

	A$^-$T$^-$		A$^+$T$^-$		A$^+$T$^+$		p-value	
Measure	Left	Right	Left	Right	Left	Right	Left	Right
Rad. 1: MTA at bl.	**1.17 ± 0.66**	**1.17 ± 0.63**	**1.56 ± 0.68**	**1.28 ± 0.45**	**1.67 ± 0.78**	**1.43 ± 0.58**	**0.0282**	**0.2783**
SCD only	1.07 ± 0.63	1.10 ± 0.61	1.38 ± 0.48	1.38 ± 0.48	1.27 ± 0.45	1.18 ± 0.39	0.2492	0.3542
MCI only	1.50 ± 0.65	1.42 ± 0.64	1.70 ± 0.78	1.20 ± 0.40	2.10 ± 0.83	1.70 ± 0.64	0.2835	0.1948
Rad. 1: ΔMTA/year	**0.05 ± 0.09**	**0.05 ± 0.08**	**0.04 ± 0.07**	**0.07 ± 0.09**	**0.09 ± 0.11**	**0.12 ± 0.10**	**0.0255**	**<.0001**
SCD only	0.05 ± 0.09	0.04 ± 0.08	0.04 ± 0.06	0.03 ± 0.05	0.07 ± 0.10	0.08 ± 0.11	0.8090	0.3825
MCI only	0.05 ± 0.09	0.06 ± 0.08	0.04 ± 0.07	0.11 ± 0.10	0.11 ± 0.11	0.16 ± 0.07	0.0016	<.0001
Rad. 2: MTA at bl.	**0.50 ± 0.71**	**0.56 ± 0.79**	**0.61 ± 0.76**	**0.50 ± 0.69**	**0.62 ± 0.79**	**0.81 ± 0.85**	**0.7581**	**0.3620**
SCD only	0.36 ± 0.65	0.52 ± 0.79	0.50 ± 0.71	0.62 ± 0.70	0.27 ± 0.45	0.45 ± 0.66	0.8081	0.8203
MCI only	1.00 ± 0.71	0.67 ± 0.75	0.70 ± 0.78	0.40 ± 0.66	1.00 ± 0.89	1.20 ± 0.87	0.6051	0.0902
Rad. 2: ΔMTA/year	**0.05 ± 0.08**	**0.06 ± 0.09**	**0.11 ± 0.10**	**0.13 ± 0.07**	**0.15 ± 0.12**	**0.13 ± 0.12**	**<.0001**	**<.0001**
SCD only	0.06 ± 0.09	0.06 ± 0.09	0.04 ± 0.07	0.10 ± 0.06	0.11 ± 0.12	0.10 ± 0.10	0.0155	0.0039
MCI only	0.03 ± 0.07	0.07 ± 0.09	0.16 ± 0.10	0.14 ± 0.07	0.19 ± 0.10	0.17 ± 0.13	<.0001	0.0033
AVRA: MTA at bl.	**1.26 ± 0.58**	**1.26 ± 0.56**	**1.39 ± 0.71**	**1.40 ± 0.64**	**1.20 ± 0.58**	**1.28 ± 0.64**	**0.6503**	**0.5989**
SCD only	1.18 ± 0.55	1.24 ± 0.55	1.10 ± 0.50	1.33 ± 0.69	1.02 ± 0.43	1.01 ± 0.50	0.7771	0.3942
MCI only	1.54 ± 0.60	1.34 ± 0.60	1.62 ± 0.77	1.46 ± 0.58	1.39 ± 0.65	1.57 ± 0.65	0.8216	0.6186
AVRA: ΔMTA/year	**0.04 ± 0.04**	**0.04 ± 0.04**	**0.07 ± 0.05**	**0.08 ± 0.05**	**0.13 ± 0.08**	**0.11 ± 0.08**	**<.0001**	**<.0001**
SCD only	0.04 ± 0.05	0.04 ± 0.04	0.07 ± 0.05	0.07 ± 0.05	0.11 ± 0.09	0.09 ± 0.08	<.0001	<.0001
MCI only	0.04 ± 0.04	0.06 ± 0.04	0.07 ± 0.05	0.09 ± 0.05	0.15 ± 0.07	0.14 ± 0.07	<.0001	<.0001
HC vol at bl. (mm^3)	**3629 ± 432**	**3753 ± 506**	**3698 ± 586**	**3834 ± 567**	**3331 ± 487**	**3433 ± 494**	**0.0858**	**0.1054**
SCD only	3697 ± 414	3773 ± 479	3999 ± 571	4023 ± 547	3530 ± 325	3659 ± 309	0.1740	0.5019
MCI only	3409 ± 415	3686 ± 579	3431 ± 456	3666 ± 531	3151 ± 536	3229 ± 538	0.4706	0.1605
ΔHC/year (mm^3/year)	**-36.3 ± 26.9**	**-39.3 ± 25.5**	**-53.4 ± 29.7**	**-55.4 ± 31.3**	**-93.4 ± 33.2**	**-99.3 ± 42.0**	**<.0001**	**<.0001**
SCD only	-34.7 ± 27.4	-35.7 ± 25.1	-36.8 ± 20.2	-46.0 ± 23.8	-79.4 ± 21.3	-87.8 ± 27.1	<.0001	<.0001
MCI only	-41.4 ± 24.5	-50.9 ± 23.2	-68.1 ± 29.0	-63.8 ± 34.6	-106.0 ± 36.7	-109.7 ± 49.7	<.0001	<.0001
ΔHC/year (%/year)	**-1.0 ± 0.9**	**-1.1 ± 0.8**	**-1.6 ± 1.0**	**-1.5 ± 0.9**	**-2.9 ± 1.0**	**-2.9 ± 1.2**	—	—
ILV vol at bl. (mm^3)	**777 ± 529**	**724 ± 523**	**1053 ± 739**	**860 ± 629**	**858 ± 507**	**817 ± 412**	**0.2098**	**0.3466**
SCD only	700 ± 481	683 ± 472	804 ± 545	839 ± 772	698 ± 241	652 ± 329	0.7090	0.9822
MCI only	1029 ± 596	856 ± 644	1274 ± 813	879 ± 465	1001 ± 627	966 ± 423	0.6443	0.4588
ΔILV/year (mm^3/year)	**38.1 ± 41.3**	**38.9 ± 44.6**	**83.1 ± 74.2**	**84.1 ± 98.5**	**117.1 ± 76.5**	**107.8 ± 94.9**	**<.0001**	**<.0001**
SCD only	37.7 ± 43.5	36.3 ± 45.5	69.8 ± 63.6	98.8 ± 123.3	86.7 ± 83.8	49.3 ± 44.9	<.0001	0.0002
MCI only	39.6 ± 33.3	47.1 ± 40.6	95.0 ± 80.7	71.0 ± 66.6	144.4 ± 56.8	160.4 ± 97.2	<.0001	<.0001
ΔILV/year (%/year)	**4.8 ± 4.0**	**4.5 ± 3.9**	**7.9 ± 4.3**	**9.9 ± 7.0**	**14.7 ± 9.0**	**13.9 ± 10.0**	—	—
ΔMMSE/year	**-0.15 ± 0.47**		**-0.49 ± 0.70**		**-1.13 ± 1.02**		**<.0001**	
SCD only	-0.05 ± 0.30		-0.19 ± 0.34		-0.87 ± 1.05		<.0001	
MCI only	-0.53 ± 0.71		-0.74 ± 0.82		-1.41 ± 0.92		<.0001	
ΔADAS-DWR/year	**-0.04 ± 0.38**		**0.14 ± 0.40**		**0.39 ± 0.50**		**<.0001**	
SCD only	-0.03 ± 0.31		0.01 ± 0.40		0.49 ± 0.62		<.0001	
MCI only	-0.07 ± 0.57		0.25 ± 0.36		0.28 ± 0.26		0.0003	

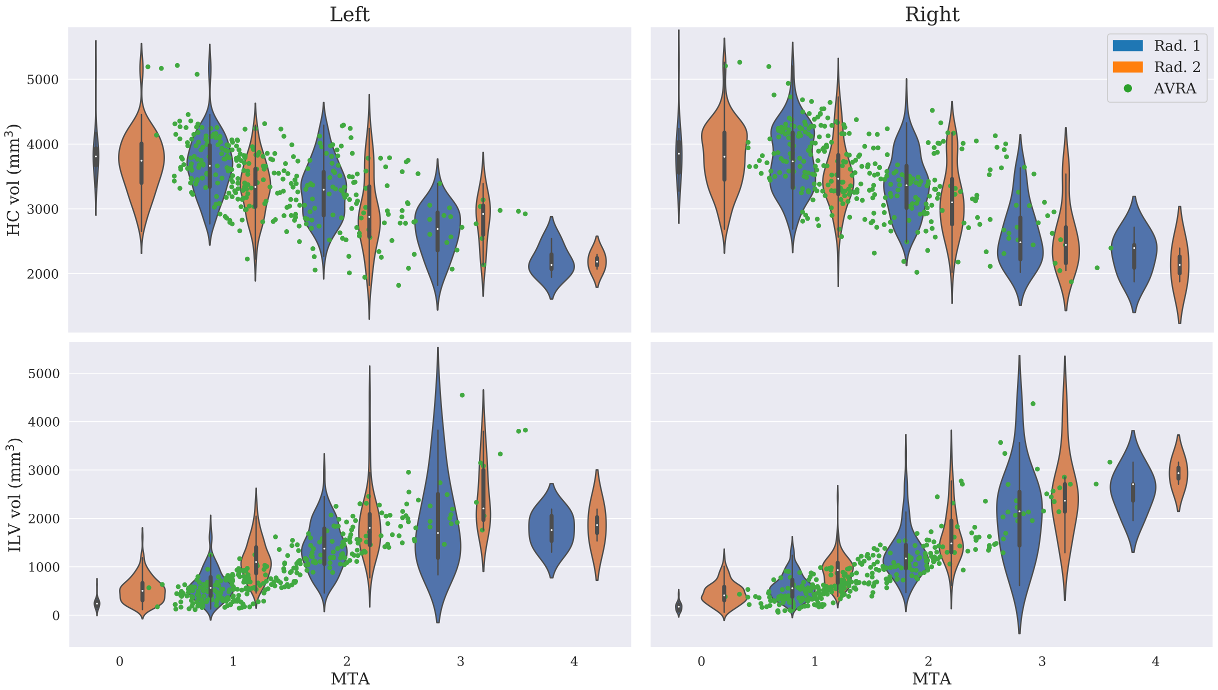

Figure 2: Violinplots of the radiologists' MTA ratings and corresponding hippocampal volume (*top*) and inferior lateral ventricle volume (*bottom*). The width of the violins shows the distribution over volumes for each rating and rater, and the area indicates the number of images given a specific rating. The green dots show AVRA's MTA rating for each image.

To study the sensitivity of the discrete radiologist ratings, we investigated the changes in MTA scores and MTL volumes compared to baseline. In Fig. 4 we show kernel density plots that estimate the distribution of ΔHC and ΔILV for follow-up images given the same MTA score (ΔMTA=0), +1 MTA (ΔMTA=1) and +2 MTA (ΔMTA=2). Both radiologists show similar distributions for ΔMTA=0 and the ΔMTA=1 entries, with a larger shift in means for ΔMTA=2. From these results it was possible to estimate that when ΔHC equaled -238 mm^3 (-8%) and -235 mm^3 (-7%) mm^3 it became more likely that the image was being rated with a higher MTA score, for Rad. 1 and Rad. 2 respectively. Corresponding values for ΔILV were 225 mm^3 (27%) and 254 mm^3 (33%).

4 Discussion

In this study we investigated longitudinal medial temporal atrophy in preclinical dementia, and to what extent it is possible to capture these changes with the Scheltens' MTA scale. We found that both radiologists provided reliable ratings, capable of capturing longitudinal progression, despite low inter-rater agreement. This was due to systematic rating differences between the radiologists, which highlights the issue of using subjective methods to quantify atrophy. Further, we observed increased MTL atrophy rates with worsening cognition and CSF AD pathology. This is the first study to investigate longitudinal MTL atrophy using MTA ratings, which helps bridge the gap between neuroimaging research and clinical radiology.

The rating agreement was only moderate between Rad. 2 and Rad. 1, as well as between Rad. 2 and AVRA. This is slightly lower than inter-rater agreements reported in studies using MTA, normally in the range $\kappa_w \in (0.6, 0.9)$ (Koedam et al., 2011; Cavallin et al., 2012b; Velickaite et al., 2017; Ferreira et al., 2017). All sets of ratings showed strong correlation to both HC and ILV volumes. This was reasonable, given that another recently proposed model estimating MTA was based on a linear combination of HC and ILV volumes (Koikkalainen et al., 2019). Our reported Spearman correlations between MTA and HC volume were stronger than previously reported, with $r_s \in$ [-0.26,-0.37] (Wahlund et al., 1999; Cavallin et al., 2012a). This shows that both radiologists are reliable, but that their rating styles differ—with one being more conservative—leading to low agreements. Since none of the radiologists trained together prior to rating the images, the low κ_w is not surprising. These results demonstrate the issue of using subjective measures to quantify atrophy, where pathological status (normal/abnormal MTA) of a patient may differ depending on which

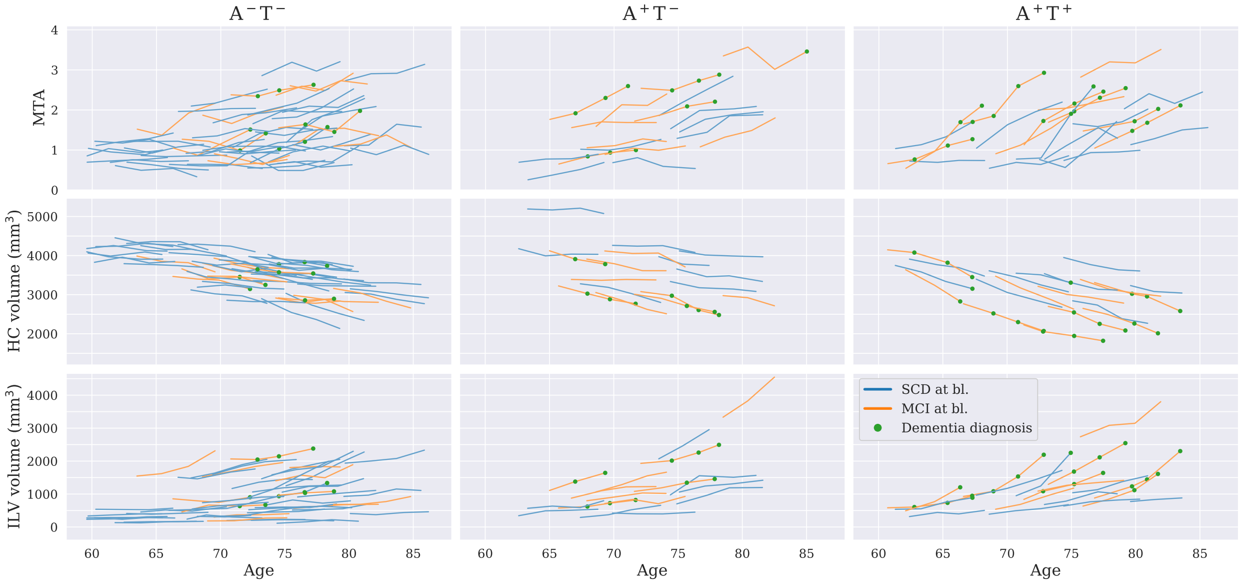

Figure 3: *Top*: AVRA's left MTA ratings plotted against age at scan time for different combinations of Aβ and p-tau abnormality. *Middle/bottom*: corresponding plots for left hippocampal (HC)/inferior lateral ventricles (ILV) volumes. Orange and blue lines show individual trajectories for SCD and MCI patients, respectively. The green dots show if a patient was diagnosed with dementia at the given timepoint.

radiologist performs the rating. On the other hand, 33 images failed the FreeSurfer segmentation upon visual QC. Having to discard almost 10% of the MRI scans due to software issues is not acceptable in a clinical setting. While other segmentation tools may be more reliable than FreeSurfer, inter-scanner variability, scanner software updates and image artifacts will always be obstacles that can influence performance (Guo et al., 2019; Mårtensson et al., 2019b). This does not seem to be an issue for visual ratings, where excellent intra-rater agreement has been demonstrated even across modalities (Wattjes et al., 2009). The benefits of using objective measures will outweigh the disadvantages—particularly as softwares become more robust—but it is important to understand that a software may fail in other ways than humans.

In accordance with previous studies we found increased HC (and ILV) atrophy rates with progressed CSF AD pathology and in MCI patients compared to cognitively normal (CN) subjects (Rusinek et al., 2003; Ridha et al., 2006; Henneman et al., 2009a). Pettigrew et al. (2017) specifically investigated the progression of MTL atrophy in preclinical AD, defined by abnormality in amyloid-β and tau. They also found an increased atrophy rate in individuals with A^+T^+ biomarker profile. They did not find any differences between A^-T^- and A^+T^-. We observed differences in our automatic measures, although these differences where smaller when studying SCD subjects only. Pettigrew and colleagues investigated only CN subjects that were 10-15 years younger (on average) than in our study. Further, we defined CSF abnormality based on established cut-offs, and not by percentiles of the sample distribution. We expect our study sample to be in a more advanced pathological stage, which may explain why our data showed a difference between A^-T^- and A^+T^-. Henneman et al. (2009a) reported differences in both HC volume at baseline and HC atrophy rate between healthy controls and MCI patients, which is consistent with our observed differences between SCD and MCI subjects within each CSF group. However, SCD individuals in the A^+T^+ group displayed greater atrophy rates than A^-T^- and A^+T^- MCI patients. This is in line with another study from Henneman et al. (2009b) which suggested that greater CSF p-tau levels were associated with greater HC atrophy rate.

The same trends as for HC were captured by AVRA's MTA ratings, but not as clear in the radiologist ratings. Most subjects, when using discrete ratings, had the same or +1 MTA score at six-year follow-up compared to baseline. This led to that the computed ΔMTA/year values for Rad. 1 and Rad. 2 merely reflect the ratio of subjects given a higher MTA score within six years. That is, the Δ MTA/year for a subject can "only" assume three values $\{0, 0.15, 0.2\}$ depending on if, or at what timepoint, a higher MTA score is assigned. Thus, we argue that it is not possible to obtain reliable measure of atrophy rates from the integer radiologist ratings in our small study samples. Focusing on the ratings from AVRA only, we found that the average changes in MTA scores were small: between 0.04 and 0.15 per year. This corresponds to roughly 25 years for A^-T^- subjects to progress a "full" MTA score (e.g. "1.0 → 2.0"). For the A^+T^- group the time is 13.3 years, and 8.3 years for A^+T^+. By combining the ΔHC/year entries from Table 3 with the ΔHC value at which it becomes more likely for the radiologists to give a higher MTA score (Fig. 4), we can estimate

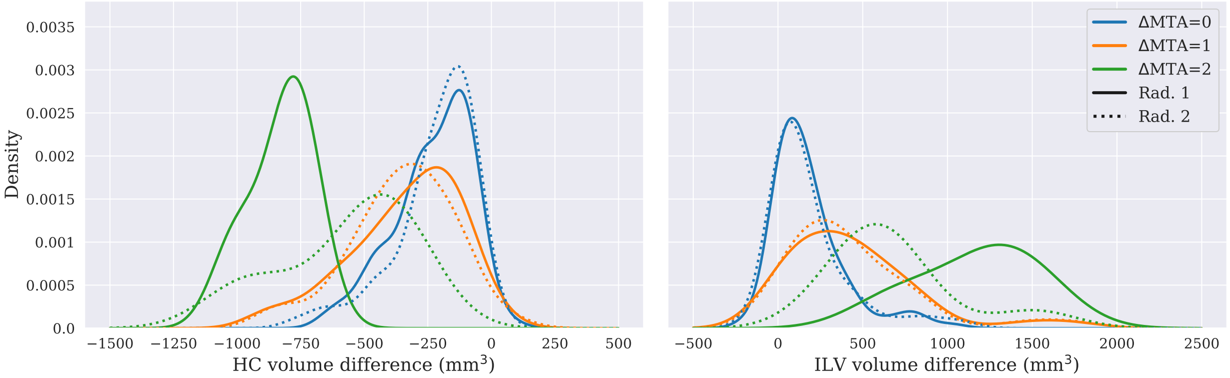

Figure 4: Shows distribution (kernel density plots) of the change in HC (left) and ILV (right) volumes between baseline and follow-up scan. A follow-up image rated the same as the baseline scan are in blue ("0 MTA"), 1 MTA score higher ("+1 MTA") in orange, and 2 MTA scores higher ("+2 MTA") in green. Solid lines are ratings from Rad. 1 and dotted lines from Rad. 2.

how many years it takes for individuals in each CSF group to be more likely to get a higher MTA score at follow-up. Subjects with A^-T^- at baseline are more likely to get a higher score at roughly 6.2 years, A^+T^- at 4.3 years, and A^+T^+ at 2.5 years. The difference in the two methods is that in the latter measure we are estimating the time to reach the next discrete MTA step. That is, borderline cases (e.g. MTA=2.9) are more likely to get a higher score at the next follow-up than individuals with MTA=2.0 at baseline. Assuming that patients being rated MTA=2 by a radiologist have an underlying continuous MTA score, and that these are uniformly distributed on the interval [2,3), a patient in this group would on average have MTA=2.5. The first method (based on AVRA ratings) should thus give roughly twice the conversion time to the second (based on radiologist ratings), which is too short but fairly close. The remaining differences can have multiple explanations. 1) The estimates are crude and based on relatively few subjects with large within-group variability in MTA rates. 2) The calculations are based on atrophy rates being constant over 20 years. This seems unlikely, given that individuals' CSF status and cognition may worsen, which should yield increased atrophy rates according to Table 3. 3) The MTA scale assesses three structures and not just HC atrophy. Further, it has been suggested that atrophy mainly occurs in posterior HC in preclinical AD (Lindberg et al., 2017), leading the HC volume change to occur mainly "outside" the MTA rating slice.

Longitudinal MTA scores have, to our knowledge, only previously been reported by Ferreira et al. (2017) in AD patients and CN subjects over a two-year follow-up. This study reported an an MTA change of 0.25/year in CN participants, and 0.4/year in AD patients (estimated from figure). The annual change in MTA scores in CN individuals was higher than those observed in SCD subjects in the current study. However, we believe that our data, comprising four scans per participant and continuous ratings, allows for an accurate estimation of the MTA rate.

A limitation of the current study is that many of the analyses assume a linear relationship between variables. From Fig. 3 the individual slopes for all MTL measures look linear with respect to age, or at least like a reasonable approximation for six years. However, if one was to model ILV as a function of MTA (see Fig. 2), the relationship is clearly not linear. This means that the change in ILV volume between MTA 0-1 is smaller than between MTA 3-4. This may confound the interpretations of Fig. 4, but our study sample was not large enough to consider non-linear relationships. Further, we emphasize that the study sample is not fully representative of A^-T^-, A^+T^- and A^+T^+ groups given that the inclusion criteria excluded (subjective) cognitively normal and dementia patient. The former would likely affect mainly the A^- group results, and the latter the CSF pathological groups.

5 Conclusion

In this study we investigated the sensitivity and reliability of visual assessment of MTL atrophy according to Scheltens' MTA scale in a longitudinal cohort of subjects with subjective cognitive decline and mild cognitive impairment. Our data showed that MTA ratings display the same cross-sectional and longitudinal trends as the volumes of hippocampus and the inferior lateral ventricle, suggesting that the MTA scale is a sensitive alternative to automatic image segmentations. The MTA ratings from two experienced radiologists, and an automated software, were strongly associated to the subcortical volumes as well as cognitive tests, showing that all raters were reliable. However, the inter-rater agreement was low due to systematic rating differences, which highlights the issue of using subjective assessments.

6 Acknowledgements

The authors wish to thank Konstantinos Poulakis for helpful discussions regarding the statistics of this paper. Further, we thank the Swedish Foundation for Strategic Research (SSF), The Swedish Research Council (VR), the Strategic Research Programme in Neuroscience at Karolinska Institutet (StratNeuro), Swedish Brain Power, Center for Innovative Medicine (CIMED), the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, Swedish Brain Foundation, Swedish Alzheimer Foundation, Åke Wiberg Foundation, Stiftelsen Olle Engkvist Byggmästare, the joint research funds of KTH Royal Institute of Technology and Stockholm County Council (HMT), and Birgitta och Sten Westerberg for financial support. The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

References

L. Cavallin, L. Bronge, Y. Zhang, A. R. Øksengard, L. O. Wahlund, L. Fratiglioni, and R. Axelsson. Comparison between visual assessment of MTA and hippocampal volumes in an elderly, non-demented population. *Acta Radiologica*, 53(5):573–579, 2012a. ISSN 02841851. doi: 10.1258/ar.2012.110664.

L. Cavallin, K. Løken, K. Engedal, A. R. Øksengård, L. O. Wahlund, L. Bronge, and R. Axelsson. Overtime reliability of medial temporal lobe atrophy rating in a clinical setting. *Acta Radiologica*, 53(3):318–323, 2012b. ISSN 02841851. doi: 10.1258/ar.2012.110552.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0395.

D. Ferreira, L. Cavallin, E.-M. Larsson, J.-S. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, L.-O. Wahlund, and E. Westman. Practical cut-offs for visual rating scales of medial temporal, frontal and posterior atrophy in Alzheimer's disease and mild cognitive impairment. *Journal of Internal Medicine*, 278(3):277–290, 2015. ISSN 13652796. doi: 10.1111/joim.12358.

D. Ferreira, C. Verhagen, J. A. Hernández-Cabrera, L. Cavallin, C. J. Guo, U. Ekman, J. S. Muehlboeck, A. Simmons, J. Barroso, L. O. Wahlund, and E. Westman. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: Longitudinal trajectories and clinical applications. *Scientific Reports*, 7(April):1–13, 2017. ISSN 20452322. doi: 10.1038/srep46263. URL `http://dx.doi.org/10.1038/srep46263`.

B. Fischl, D. H. Salat, A. J. W. Van Der Kouwe, N. Makris, F. Ségonne, B. T. Quinn, and A. M. Dale. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23(SUPPL. 1):69–84, 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.07.016.

N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor. Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study. *Brain : a journal of neurology*, 119 (Pt 6(1996):2001–7, 1996. ISSN 0006-8950. doi: 10.1093/brain/119.6.2001. URL `http://www.ncbi.nlm.nih.gov/pubmed/9010004`.

C. Guo, D. Ferreira, K. Fink, E. Westman, and T. Granberg. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. *European Radiology*, 29(3):1355–1364, 2019. ISSN 14321084. doi: 10.1007/s00330-018-5710-x.

C. Håkansson, G. Torisson, E. Londos, O. Hansson, and D. van Westen. Structural imaging findings on non-enhanced computed tomography are severely underreported in the primary care diagnostic work-up of subjective cognitive decline. *Neuroradiology*, 61(4):397–404, 2019. ISSN 14321920. doi: 10.1007/s00234-019-02156-6.

W. J. Henneman, J. D. Sluimer, J. Barnes, W. M. Van Der Flier, I. C. Sluimer, N. C. Fox, P. Scheltens, H. Vrenken, and F. Barkhof. Hippocampal atrophy rates in Alzheimer disease: Added value over whole brain volume measures. *Neurology*, 72(11):999–1007, 2009a. ISSN 1526632X. doi: 10.1212/01.wnl.0000344568.09360.31.

W. J. Henneman, H. Vrenken, J. Barnes, I. C. Sluimer, N. A. Verwey, M. A. Blankenstein, M. Klein, N. C. Fox, P. Scheltens, F. Barkhof, and W. M. Van Der Flier. Baseline CSF p-tau levels independently predict progression of hippocampal atrophy in Alzheimer disease. *Neurology*, 73(12):935–940, 2009b. ISSN 1526632X. doi: 10.1212/WNL.0b013e3181b879ac.

S. Janelidze, H. Zetterberg, N. Mattsson, S. Palmqvist, H. Vanderstichele, O. Lindberg, D. van Westen, E. Stomrud, L. Minthon, K. Blennow, and O. Hansson. CSF Aβ42/Aβ40 and Aβ42/Aβ38 ratios: Better diagnostic markers of Alzheimer disease. *Annals of Clinical and Translational Neurology*, 3(3):154–165, 2016. ISSN 23289503. doi: 10.1002/acn3.274.

M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001. ISSN 13618415. doi: 10.1016/S1361-8415(01)00036-6.

M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002. ISSN 10538119. doi: 10.1016/S1053-8119(02)91132-8.

E. L. Koedam, M. Lehmann, W. M. Van Der Flier, P. Scheltens, Y. A. Pijnenburg, N. Fox, F. Barkhof, and M. P. Wattjes. Visual assessment of posterior atrophy development of a MRI rating scale. *European Radiology*, 21(12):2618–2625, 2011. ISSN 09387994. doi: 10.1007/s00330-011-2205-4.

J. R. Koikkalainen, H. F. M. Rhodius-Meester, K. S. Frederiksen, M. Bruun, S. G. Hasselbalch, M. Baroni, P. Mecocci, R. Vanninen, A. Remes, H. Soininen, M. van Gils, W. M. van der Flier, P. Scheltens, F. Barkhof, T. Erkinjuntti, and J. M. P. Lötjönen. Automatically computed rating scales from MRI for patients with cognitive disorders. *European Radiology*, 13(7):P1108, feb 2019. ISSN 0938-7994. doi: 10.1007/s00330-019-06067-1. URL `http://ovidsp.ovid.com/ovidweb.cgi?T=JS{&}PAGE=reference{&}D=emexa{&}NEWS=N{&}AN=620612139http://link.springer.com/10.1007/s00330-019-06067-1`.

J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159, 1977. ISSN 0006341X. doi: 10.2307/2529310. URL `http://www.jstor.org/stable/2529310?origin=crossref`.

O. Lindberg, G. Mårtensson, E. Stomrud, S. Palmqvist, L. O. Wahlund, E. Westman, and O. Hansson. Atrophy of the posterior subiculum is associated with memory impairment, Tau- and Aβ pathology in non-demented individuals. *Frontiers in Aging Neuroscience*, 9(SEP):1–12, 2017. ISSN 16634365. doi: 10.3389/fnagi.2017.00306.

G. Mårtensson, D. Ferreira, L. Cavallin, J.-S. Muehlboeck, L.-O. Wahlund, C. Wang, and E. Westman. AVRA: Automatic Visual Ratings of Atrophy from MRI images using Recurrent Convolutional Neural Networks. *NeuroImage: Clinical*, 23(March):101872, 2019a. ISSN 2213-1582. doi: 10.1016/j.nicl.2019.101872. URL `http://arxiv.org/abs/1901.00418`.

G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. Kramberger, J.-P. Taylor, J. Hort, J. Snædal, J. Kulisevsky, F. Blanc, A. Antonini, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, D. Aarsland, and E. Westman. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *arXiv preprint*, pages 1–18, 2019b. URL `http://arxiv.org/abs/1911.00515`.

N. Mattsson, R. Smith, O. Strandberg, S. Palmqvist, M. Schöll, P. S. Insel, D. Hägerström, T. Ohlsson, H. Zetterberg, K. Blennow, J. Jögi, and O. Hansson. Comparing 18 F-AV-1451 with CSF t-tau and p-tau for diagnosis of Alzheimer disease. *Neurology*, 90(5):e388–e395, 2018. ISSN 1526632X. doi: 10.1212/WNL.0000000000004887.

J.-S. Muehlboeck, E. Westman, and A. Simmons. TheHiveDB image data management and analysis framework. *Frontiers in Neuroinformatics*, 7(January):49, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2013.00049. URL `http://journal.frontiersin.org/article/10.3389/fninf.2013.00049/abstract`.

R. C. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194, 2004. ISSN 09546820. doi: 10.1111/j.1365-2796.2004.01388.x.

C. Pettigrew, A. Soldan, K. Sloane, Q. Cai, J. Wang, M. C. Wang, A. Moghekar, M. I. Miller, and M. Albert. Progressive medial temporal lobe atrophy during preclinical Alzheimer's disease. *NeuroImage: Clinical*, 16(August):439–446, 2017. ISSN 22131582. doi: 10.1016/j.nicl.2017.08.022. URL `https://doi.org/10.1016/j.nicl.2017.08.022`.

M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, jul 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.02.084. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdfhttps://linkinghub.elsevier.com/retrieve/pii/S1053811912002765`.

B. H. Ridha, J. Barnes, J. W. Bartlett, A. Godbolt, T. Pepple, M. N. Rossor, and N. C. Fox. Tracking atrophy progression in familial Alzheimer's disease: a serial MRI study. *Lancet Neurology*, 5(10):828–834, 2006. ISSN 14744422. doi: 10.1016/S1474-4422(06)70550-6.

H. Rusinek, S. De Santi, D. Frid, W. H. Tsui, C. Y. Tarshish, A. Convit, and M. J. De Leon. Regional Brain Atrophy Rate Predicts Future Cognitive Decline: 6-Year Longitudinal MR Imaging Study of Normal Aging. *Radiology*, 229 (3):691–696, 2003. ISSN 00338419. doi: 10.1148/radiol.2293021299.

P. Scheltens, D. Leys, F. Barkhof, D. Huglo, H. C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters, and J. Valk. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology Neurosurgery, and Psychiatry*, 55:967–972, 1992. ISSN 0022-3050. doi: 10.1136/jnnp.55.10.967.

G. Torisson, D. Van Westen, L. Stavenow, L. Minthon, and E. Londos. Medial temporal lobe atrophy is underreported and may have important clinical correlates in medical inpatients. *BMC Geriatrics*, 15(1):1–8, 2015. ISSN 14712318. doi: 10.1186/s12877-015-0066-4.

V. Velickaite, D. Ferreira, L. Cavallin, L. Lind, H. Ahlström, L. Kilander, E. Westman, and E. M. Larsson. Medial temporal lobe atrophy ratings in a large 75-year-old population-based cohort: gender-corrected and education-corrected normative data. *European Radiology*, pages 1–9, 2017. ISSN 14321084. doi: 10.1007/s00330-017-5103-6.

M. W. Vernooij, F. B. Pizzini, R. Schmidt, M. Smits, T. A. Yousry, N. Bargallo, G. B. Frisoni, S. Haller, and F. Barkhof. Dementia imaging in clinical practice: a European-wide survey of 193 centres and conclusions by the ESNR working group. *Neuroradiology*, 61(6):633–642, 2019. ISSN 14321920. doi: 10.1007/s00234-019-02188-y.

L.-O. Wahlund, P. Julin, J. Lindqvist, and P. Scheltens. Visual assessment of medial temporal lobe atrophy in demented and healthy control subjects: correlation with volumetry. *Psychiatry Research: Neuroimaging*, 90(3):193–199, 1999. ISSN 09254927. doi: 10.1016/S0925-4927(99)00016-5. URL `https://ac.els-cdn.com/S0925492799000165/1-s2.0-S0925492799000165-main.pdf?{_}tid=33ac94fc-d111-4137-88d9-2f6ca702583e{&}acdnat=1526123326{_}dfd495990f9ade84488074d8bdf84427{%}0Ahttp://linkinghub.elsevier.com/retrieve/pii/S0925492799000165`.

L. O. Wahlund, E. Westman, D. van Westen, A. Wallin, S. Shams, L. Cavallin, and E. M. Larsson. Imaging biomarkers of dementia: recommended visual rating scales with teaching cases. *Insights into Imaging*, 8(1):79–90, 2017. ISSN 18694101. doi: 10.1007/s13244-016-0521-6. URL `http://dx.doi.org/10.1007/s13244-016-0521-6`.

M. P. Wattjes, W. J. P. Henneman, W. M. van der Flier, O. de Vries, F. Träber, J. J. G. Geurts, P. Scheltens, H. Vrenken, and F. Barkhof. Diagnostic Imaging of Patients in a Memory Clinic: Comparison of MR Imaging and 64–Detector Row CT. *Radiology*, 253(1):174–183, 2009. ISSN 0033-8419. doi: 10.1148/radiol.2531082262. URL `http://pubs.rsna.org/doi/10.1148/radiol.2531082262`.

E. Westman, L. Cavallin, J. S. Muehlboeck, Y. Zhang, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, C. Spenger, S. Lovestone, A. Simmons, and L. O. Wahlund. Sensitivity and specificity of medial temporal lobe visual ratings and multivariate regional MRI classification in Alzheimer's disease. *PLoS ONE*, 6(7), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0022506.

A Supplementary data

As additional information we provide visual examples of the MTA rating slice for four subjects in Fig. S1. Figure S2 is the same plot as Fig. 3 but for the right hemisphere. Confusion matrices for all rating sets are shown in Table S1-S3.

Table S1: Confusion matrices for left and right MTA ratings between Rad. 1 and Rad. 2.

Left		Rad. 1				
		0	1	2	3	4
Rad. 2	0	19	145	7	4	0
	1	0	36	73	9	0
	2	0	2	41	19	1
	3	0	0	2	7	2
	4	0	0	0	0	3

Right		Rad. 1				
		0	1	2	3	4
Rad. 2	0	19	128	10	0	0
	1	0	56	71	3	0
	2	0	4	48	12	1
	3	0	0	1	12	3
	4	0	0	0	0	2

Table S2: Confusion matrices for left and right MTA ratings between Rad. 1 and AVRA.

Left		Rad. 1				
		0	1	2	3	4
AVRA	0	1	5	0	0	0
	1	18	158	26	5	0
	2	0	19	83	19	1
	3	0	1	14	13	5
	4	0	0	0	2	0

Right		Rad. 1				
		0	1	2	3	4
AVRA	0	1	6	1	0	0
	1	18	157	28	0	0
	2	0	24	85	10	0
	3	0	1	16	17	5
	4	0	0	0	0	1

Table S3: Confusion matrices for left and right MTA ratings between Rad. 2 and AVRA.

Left		Rad. 2				
		0	1	2	3	4
AVRA	0	6	0	0	0	0
	1	162	44	1	0	0
	2	7	74	41	0	0
	3	0	0	20	10	3
	4	0	0	1	1	0

Right		Rad. 2				
		0	1	2	3	4
AVRA	0	7	1	0	0	0
	1	143	60	0	0	0
	2	7	66	45	1	0
	3	0	3	20	15	1
	4	0	0	0	0	1

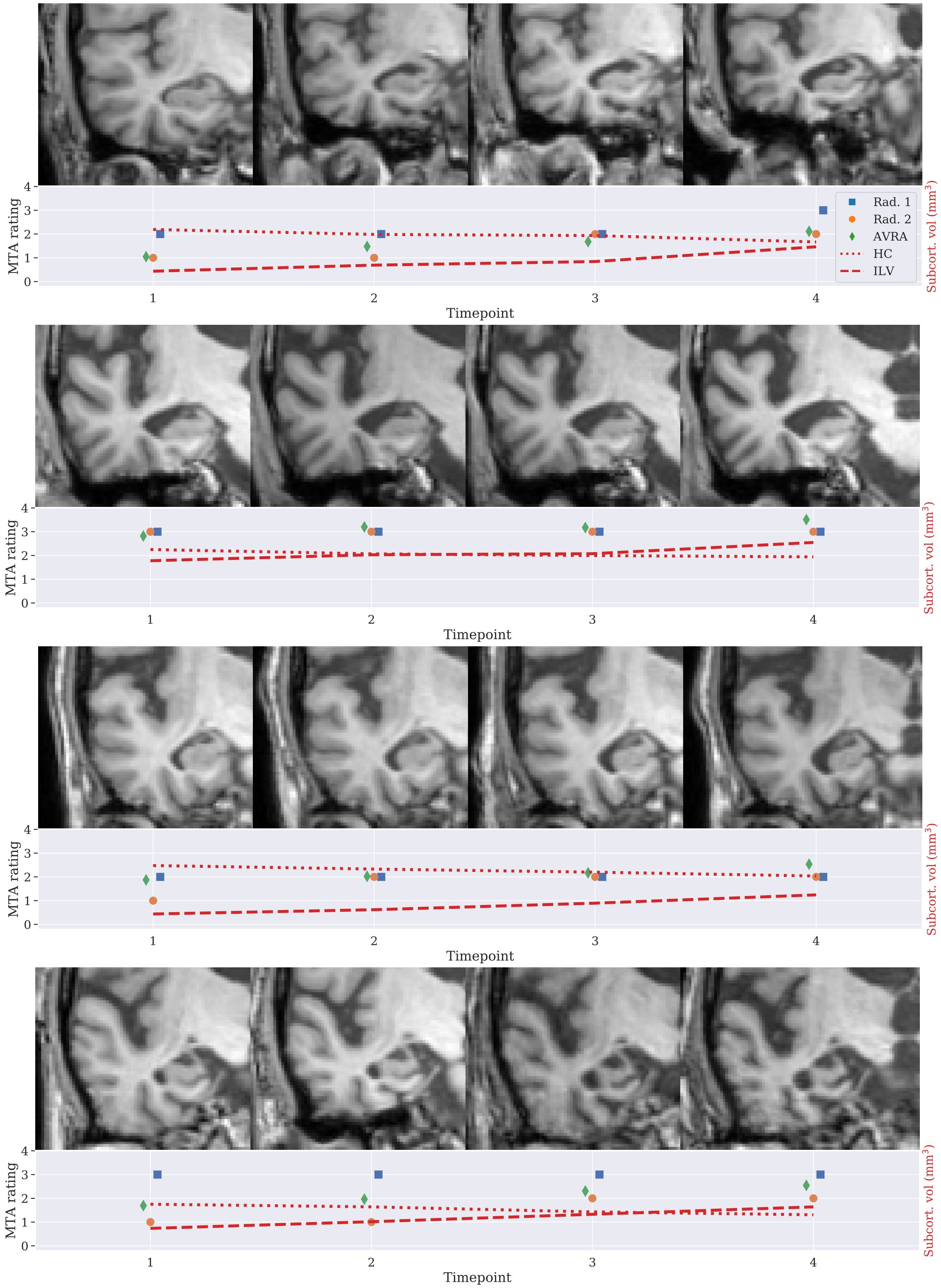

Figure S1: Rating slices at each timepoint for study four participants and corresponding MTA ratings and MTL volumes.)

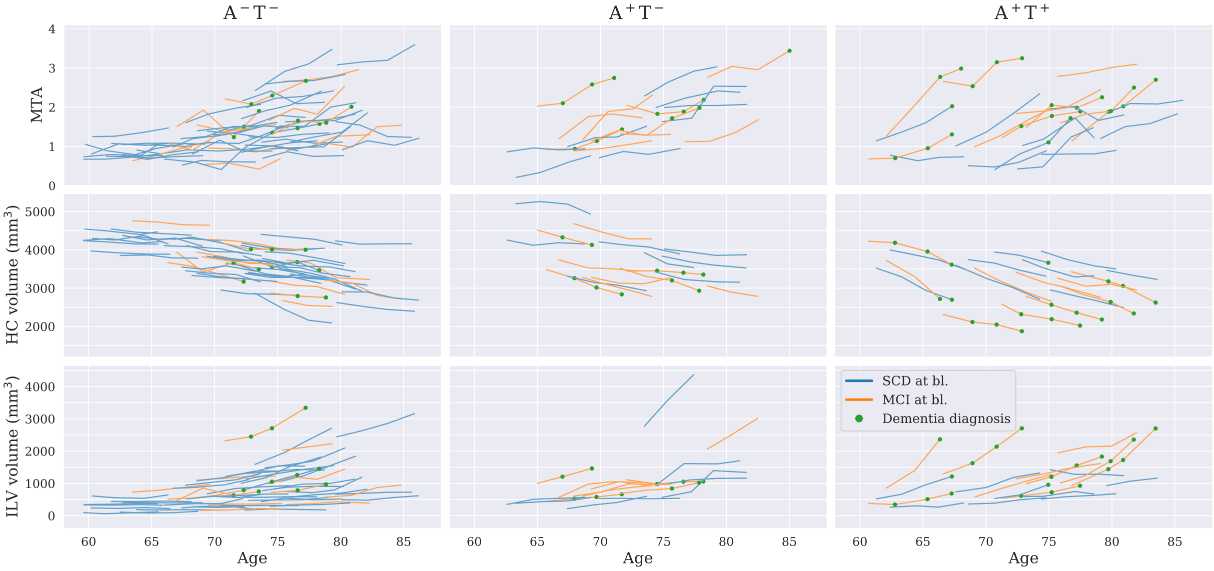

Figure S2: *Top*: AVRA's **right** MTA ratings plotted against age at scan time for different combinations of Aβ and p-tau abnormality. *Middle/bottom*: corresponding plots for right hippocampal (HC)/inferior lateral ventricles (ILV) volumes. Orange and blue lines show individual trajectories for SCD and MCI patients, respectively. The green dots show if a patient was diagnosed with dementia at the given timepoint.