# MAIN BODY

## Title

**Automatically computed rating scales from MRI for patients with cognitive disorders**


## Abstract

**Objectives.** To study whether visual MRI rating scales used in diagnostics of cognitive disorders can be estimated computationally and to compare the visual rating scales with their computed counterparts in differential diagnostics.

**Methods.** A set of volumetry and voxel-based morphometry imaging biomarkers was extracted from T1-weighted and FLAIR images. A regression model was developed for estimating visual rating scale values from a combination of imaging biomarkers. We studied three visual rating scales: medial temporal lobe atrophy (MTA), global cortical atrophy (GCA) and white matter hyperintensities (WMHs) measured by the Fazekas scale. Images and visual ratings from the Amsterdam Dementia Cohort (ADC) (N=513) were used to develop the models and cross validate them. The PredictND (N=672) and ADNI (N=752) cohorts were used for independent validation to test generalizability.

**Results.** The correlation coefficients between visual and computed rating scale values were 0.83/0.78 (MTA-Left), 0.83/0.79 (MTA-Right), 0.64/0.64 (GCA) and 0.76/0.75 (Fazekas) in ADC/PredictND cohorts. When performance in differential diagnostics was studied for the main types of dementia, the highest balanced accuracy, 0.75-0.86, was observed for separating different dementias from cognitively normal subjects using computed GCA. The lowest accuracy of about 0.5 for all the visual and computed scales was observed for the differentiation between Alzheimer's disease and frontotemporal lobar degeneration. Computed scales produced higher balanced accuracies than visual scales for MTA and GCA (statistically significant).

**Conclusions.** MTA, GCA and WMHs can be reliably estimated automatically helping to provide consistent imaging biomarkers for diagnosing cognitive disorders, even among less-experienced readers.

## Keywords

magnetic resonance imaging, cognition disorders, atrophy

## Key points

Visual rating scales used in diagnostics of cognitive disorders can be estimated computationally from MRI images with intra-class correlations ranging from 0.64 (GCA) to 0.84 (MTA).

Computed scales provided high diagnostic accuracy with single-subject data (area under the receiver operating curve range, 0.84-0.94).

## Abbreviations

| | |
|---|---|
| AD | Alzheimer's disease |
| ADC | Amsterdam dementia cohort |
| ADNI | Alzheimer's disease neuroimaging initiative |
| BACC | balanced accuracy |
| CN | cognitively normal |
| DLB | dementia with Lewy bodies |
| FLAIR | fluid-attenuated inversion recovery |
| FTLD | frontotemporal lobar degeneration |
| GCA | global cortical atrophy |
| ICC | intra class correlations |
| MTA | medial temporal lobe atrophy on the left (MTA-L) and right (MTA-R) |
| OTH | other dementias but AD, VaD, FTD and, DLB |

| | |
|---|---|
| VaD | vascular dementia |
| VBM | voxel-based morphometry |
| WMH | white matter hyperintensities |

## Introduction

Clinical differential diagnosis of cognitive disorders is challenging. The most common underlying diseases include Alzheimer's disease (AD), vascular dementia (VaD), dementia with Lewy bodies (DLB) and frontotemporal lobar degeneration (FTLD). Early and precise diagnosis is important both for therapeutical and research purposes (1-6).

Magnetic resonance imaging (MRI) is a standard tool in clinical diagnostics of cognitive disorders, historically to rule out other pathologies, while current guidelines advise the use of MRI to find evidence for underlying patterns of neurodegeneration (1,4,5,6). Mediotemporal atrophy is often seen in typical AD, while young onset patients with an atypical presentation show more frequent parietal atrophy (7,8). In FTLD, atrophy is focused on frontal and temporal regions, but overall global atrophy is also present with increasing age. In VaD, white matter hyperintensities (WMHs) are essential (9-11), however WMHs become more abundant with increasing age (12). DLB patients typically show little atrophy on MRI. These patterns of neurodegeneration are typically visually assessed. To make visual reads more uniform, visual rating scales are commonly used in the clinical and research settings, especially in Europe. A recent survey shows that about 75 % of radiologists use visual scales in Europe (13). Medial temporal lobe atrophy can be evaluated using a 5-point rating scale (MTA, range 0-4) (14) and global cortical atrophy (GCA) using a 4-point rating scale (range 0-3) (15). There is also a specific visual rating scale, Koedam score (range 0-3), for assessing posterior atrophy (8), useful for the atypical form of AD. WMHs can be rated using the Fazekas scale (range 0-3) (9-11). Table 1 provides details on these rating scales.

Visual rating scales produce semi-quantitative information about the underlying pathology and consider more than just the volume of a specific region. However, they are coarse and biased by subjective visual interpretation. Computational imaging biomarkers, such as the hippocampal volume, aim to measure this pathology more precisely and objectively offering potential improvements. Transition from visual rating scales to computational imaging biomarkers is not,

however, straightforward in clinical practice as different specialists need to learn interpret such new imaging biomarkers. The purpose of this study is to overcome this challenge making interpretation easier: images are quantified using computational imaging biomarkers, but the results are represented in the scales that specialists are familiar with.

Our first objective is to study whether visual MRI rating scales used in diagnostics of cognitive disorders can be estimated reliably based on a combination of imaging biomarkers. Our second objective is to compare visual ratings with their computed counterparts in separating dementias. Our approach tries to preserve the benefits of quantitative MRI but simultaneously use clinically familiar measures. Computed rating scales may improve underreporting of visual rating scales observed in clinical practice (16) and enable more uniform high-quality reporting even for less experienced readers. These challenges of visual rating are also reflected in (13): 32 % of responders among radiologists reported that they are not fully confident in using visual rating scales in the workup for cognitive disorders. Our hypothesis is that MTA, GCA and Fazekas can be estimated automatically providing useful information for helping in consistent diagnosing of cognitive disorders.

## Materials and methods

The study has been executed in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants.

### Subjects

**Cohorts.** We included subjects from three independent cohorts: 1) Amsterdam dementia cohort (ADC) was used for developing the model. MRI images of 513 subjects were acquired between 2004-2014 (17). 2) PredictND cohort (www.predictnd.eu) was used for external validation**.** MRI images of 672 subjects were included from four memory clinics and acquired between 2015-2016. 3) ADNI cohort was used for external validation. MRI images of 752 subjects were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, www.adni-info.org).

**Clinical workup.** From ADC and PredictND, we included subjects from six diagnostic groups: AD, FTLD, DLB, VaD, mild cognitive impairment (MCI) and subjective cognitive decline which represented cognitively normal (CN) subjects (Table 2). From ADNI, we included AD and CN cases from ADNI-1 and ADNI-2 (Table 2). Probable AD was diagnosed using the NINS-ADRDA criteria; all patients also met the core clinical criteria of the NIA-AA for probable AD (1,18). The Neary and Snowden criteria were used to diagnose FTLD (3). DLB was diagnosed using the McKeith criteria (2) and VaD using the NINDS-AIREN criteria (5). MCI was diagnosed using Petersen's criteria and all patients fulfilled the core clinical criteria of the NIA-AA for MCI (19-20). All clinical diagnoses were made using the standardized multi-disciplinary clinical workup of each clinic.

**Imaging data and visual ratings.** The subjects were scanned using either a 1.5 or 3 T MRI, including a 3-dimensional T1-weighted gradient echo sequence and a fast fluid-attenuated inversion recovery (FLAIR) sequence. ADNI-1 did not contain FLAIR images. Images from >20 different scanner models were used (see more details in Supplement-1). The voxel size varied between 0.4-1.6 × 0.4-1.6 × 0.5-2.2 mm in T1-images and 0.4-1.3 × 0.4-1.3 × 0.6-7.0 mm in FLAIR images. MTA was rated

on coronal T1-weighted images both on the left (MTA-L) and right (MTA-R) sides (14), GCA on axial FLAIR images (15) and WMHs on axial FLAIR images (9-11). As part of standard work-up, all scans of ADC were rated by one of three neuroradiologists, each with >15 years of experience. All readers had gone through a training and were qualified if a weighted kappa of at least 0.80 for MTA, 0.60 for GCA and 0.70 for WMH was obtained (17). In PredictND, one of the clinics (C1) was the same as acquired ADC. In the three other clinics, one expert (C2: 8 years, C3: >15 years, C4: 5 years of experience) rated all images. In ADNI, visual ratings were not available. All raters were blind to clinical diagnosis.

### Estimating visual rating scales using imaging biomarkers

Volumes of brain structures were defined from T1 image segmentations produced by a multi-atlas segmentation algorithm (21-22). WMH segmentation method is described in (22-23). Segmentation methods were fully automatic. The volumes were normalized first for head size (24) and then for age and gender using the method proposed in (25). In addition to volumetry, voxel-based morphometry (VBM) (26) was used to compute gray matter concentrations. A gray matter concentration index was defined reflecting the share of gray matter in a certain region of interest compared with the share in CN subjects. The imaging biomarkers used in this study were 1) volumes of hippocampus and inferior lateral ventricle and concentration index of hippocampal gray matter for estimating MTA 2) volume and concentration index of cortical gray matter for estimating GCA and 3) volumes of white matter hyperintensities and deep white matter hyperintensities for estimating Fazekas.

Computed rating scales were estimated in four steps. 1) Visual rating scale values were first normalized to the same age (70 years) using a linear regression model defined for CN subjects. 2) A linear regression model was used to estimate an age-normalized visual rating scale value (dependent variable) from imaging biomarkers (independent variables). 3) As the relationship between visual rating scales and imaging biomarkers is not necessarily linear, the estimate was finetuned using a partially linear mapping: the median of the estimates (Step 2), defined for all subjects having a certain

visual rating scale value, was mapped to the median of age-normalized visual rating scale values from the same subjects (Step 1). The rating scale values for which only a few measurement values were available (MTA-L=4: 4 subjects, MTA-R=4: 9 subjects, GCA=3: 2 subjects), were excluded to avoid overfitting. 4) The estimate was restricted to the allowed value range of the particular visual scale but keeping the value still as a decimal number.

The model producing the highest Pearson correlation coefficient was selected. Supplement-2 describes the algorithm in detail.

## Statistical Analysis

Area under the curve (AUC) and balanced accuracy (BACC), defined as average of sensitivity and specificity, were used to assess diagnostic accuracy. Original visual scores (not age-normalized) were used in validation if not explicitly stated otherwise. ADC was used to develop the regression model. For internal validation, cross-validation was used: 50 % of ADC subjects were randomly selected for defining the model and the cut-off value maximizing BACC, and the remaining 50 % was used for testing. To obtain more robust performance estimates for correlation and classification accuracy, the selection of the training and test sets was repeated 250 times, and an average was calculated. The independent PredictND and ADNI cohorts were used for external validation to study generalizability.

Agreement between the visual and computed rating scale values was studied using intra class correlation (ICC) and Kendall W test as described in (7) and (27), respectively.

Statistically significant differences between the groups were analyzed using Mann-Whitney U-test, Chi-squared test, Wilcoxon rank sum test where appropriate and Fisher r-to-z transformation (two-tailed). The difference was considered statistically significant if $p<0.05$. The Matlab toolbox R2016a (The MathWorks Inc) was used to run the data analysis except for ICC for which SPSS version 22 (IBM) was applied.

## Results

### Estimating visual rating scale computationally

Table 3 shows correlation coefficients between visual and computed rating scale values when different imaging biomarkers were used in the model. For MTA, the combination of the hippocampus and inferior lateral ventricle volumes produced the highest correlation. The concentration index of cortical gray matter had the highest correlation coefficient for GCA. The Fazekas score calculated from the volume of deep white matter hyperintensities had the highest correlation coefficient. The correlation coefficients calculated for PredictND remained corresponding to the values obtained for ADC: 0.83/0.78 for MTA-L, 0.83/0.80 for MTA-R, 0.64/0.64 for GCA and 0.76/0.75 for Fazekas in ADC/PredictND. The difference was statistically significant for MTA-L.

Table 3 shows also how rating scales and different imaging biomarkers performed in classifying AD and CN subjects (MTA and GCA) and VaD and non-VaD subjects (Fazekas). For MTA and GCA, BACC was higher for the computed rating scale than for the visual rating scale or any other single imaging biomarker (statistically significant).

Next, agreement between the visual and computational rating scales was studied in detail using data from all diagnostic groups. Figure 1 shows the Box and Whisker plots for the visual and computed ratings in the independent PredictND cohort. The results are presented for each of the four memory clinics (C1 is the same center as acquired ADC). The plots indicate that the computed rating scales generalize relatively well.

The agreement was studied also quantitatively using ICC and Kendal W. ICC was 0.83/0.78 for MTA-L, 0.84/0.80 for MTA-R, 0.64/0.64 for GCA and 0.76/0.75 for Fazekas in ADC/PredictND. If computed scores were rounded to integers, ICC was on average 0.026 smaller. The Kendall W values were 0.89/0.88 for MTA-L, 0.88/0.89 for MTA-R, 0.82/0.82 for GCA and 0.84/0.82 for Fazekas using ADC/PredictND.

More validation results are presented in Supplement-3.

Table 4 shows the computed rating scale models for MTA-L, MTA-R, GCA and Fazekas. The models presented have been defined without cross validation using the whole ADC.

**Visual and computational rating scales in differential diagnostics**

Figure 2 shows BACC for visual and computed MTA (Figure 2a and 2b), GCA (Figure 2c) and Fazekas (Figure 2d) in differential diagnostics of five etiologies (AD, FTLD, DLB, VaD and CN). When BACCs of all 10 disease pairs were compared in both cohorts (10 pairs and 2 cohorts giving 20 accuracy estimates), computed scores provided on average higher accuracies for MTA-L, MTA-R and GCA (statistically significant). For Fazekas, a difference was not found. The highest accuracy was observed for detecting CN subjects from different dementias using computed GCA (0.75-0.86) while the accuracy was around 0.5 for all scales in AD vs. FTLD classification.

For assessing the generalizability in diagnostics, Figure 3 presents ROC curves for the ADC, PredictND and ADNI cohorts. The results indicate that AUC was corresponding to the results obtained in ADC for AD-CN classification. For Fazekas, AUC was smaller in PredictND but a small number of VaD cases may partially explain the difference.

## Discussion

Visual rating scales are used commonly in the diagnostic process of cognitive disorders in Europe. In research, they have been used in numerous studies (28) and supported in different guidelines (6,29). In this work, we studied whether visual rating scales (MTA, GCA and Fazekas) can be estimated computationally. In addition, we compared the performance in differentiating the main types of dementia using visual ratings and their computed counterparts. The use of computed scales based on quantitative imaging biomarkers potentially helps reducing both intra- and inter-rater variability in image interpretation, especially for less experienced raters.

The role of biomarkers is increasing in diagnosing cognitive disorders. For example, the hippocampus volume is a well-established imaging biomarker for Alzheimer's disease. The interpretation of biomarkers is typically based on cut-off values. When using automated image quantification, the challenge is that results are not typically directly comparable between methods making the use of generic cut-offs difficult. Another challenge is how to interpret deviations of the patient value from the cut-off, i.e., assess the clinical meaning of the difference. Representing the values using standardized scales, such as MTA, could help in these both challenges.

When estimating visual MTA, the highest correlations were obtained by combining the volumes of hippocampus and inferior lateral ventricle. For GCA, the concentration index of cortical gray matter was used to compute the rating scale value. The correlation coefficient between the visual and computed GCA was relatively small, only 0.64. The small number of grades (0-3) in the GCA scale explains partly the low correlation. Another potential reason can be the difficulty to evaluate the global cortical atrophy visually. The computed GCA produced good classification results, BACC=0.84, in separating CN subjects from AD subjects, while the value was 0.74 for the visual GCA.

The computed rating scales were validated also in independent cohorts. Correlation coefficients remained at comparable values. In the PredictND cohort, images were rated at four memory clinics

inducing additional heterogeneity to the results and explaining partly the statistically significant difference in the left MTA. The classification performance was stable in all four cohorts except a small decrease of AUC was observed for Fazekas in PredictND.

Agreement was assessed by comparing the ADC and PredictND results with ICC and Kendall W reference values from (7, 27). In (7), the average ICC was computed between four raters (N=80). They reported ICC 0.82 (0.76-0.88) for MTA-L and 0.79 (0.71-0.85) for MTA-R but GCA and Fazekas were not studied. The corresponding values observed in ADC/PredictND were 0.83/0.78 for MTA-L, 0.84/0.80 for MTA-R. In (27), Kendall W was used to measure inter-rater agreement for MTA, GCA and Fazekas. They reported values 0.82 for MTA-L, 0.83 for MTA-R, 0.84 for GCA and 0.92 for Fazekas (N=30). Using ADC/PredictND, the corresponding values were 0.89/0.88 for MTA-L, 0.88/0.89 for MTA-R, 0.82/0.82 for GCA and 0.84/0.82 for Fazekas. A part of the raters in (6, 26) were the same as in ADC and PredictND (C1).

Rating scales were tested also in differential diagnostics. High performance was observed in separating cognitively normal subjects from four cognitive disorders, especially for computed GCA. Computed scales produced higher overall accuracy for MTA and GCA than visual scales. This may look unexpected as computed scales estimate visual scales. However, computed scales are in reality volumetry- and VBM-based imaging biomarkers which are just represented in the value range of visual scales. Computed scales preserve the benefits of imaging biomarkers for quantification but provide the benefits of standardized scales for interpretation. Differential diagnostics between AD and FTLD is a clinical challenge, but the performance of the scales was corresponding to guessing both in ADC and PredictND. Although MTA, GCA and Fazekas have been shown to be useful in diagnosing dementia subtypes (30), previous research indicate that MTA is not specific for AD (31), and both AD and FTLD patients have atrophy in the medial temporal lobe (32). Balanced accuracies of 0.77-0.80 have been reported for AD and FTLD classification when using results from the combination of six visual rating scales (7), from the cortical thickness of the left inferior parietal

region (33) and from the ratio of volumes at anterior and posterior brain regions (34). Out of six visual scales used in (7), MTA was found to be the best scale in 4/12 of different diagnostic group comparisons. MTA has been shown to have power also in discriminating also DLB and VaD from AD (BACC=0.93) (35). There are multiple studies showing a concordance or superiority of automated imaging biomarkers compared with visual rating scales (36-40). For improving the diagnostic accuracy further, a richer and more specific set of imaging biomarkers and their combinations could be used (34, 41).

As visual scoring is not very time consuming, it is important that getting computed ratings is automated. Our current image quantification pipeline is fully automatic, and results are available about 30 minutes after image acquisition.

When considering potential clinical use, two issues regarding the representation of computed rating scales need to be considered. First, they were normalized to correspond values at the age of 70 years while clinicians need to normalize age mentally when interpreting visual ratings. Although not consistent with visual ratings today, the use of normalized values might reduce ambiguity in interpreting the values. Second, computed ratings are represented by decimal numbers while few integer values are used in visual ratings. Decimal numbers provide potential benefits, such as ability to assess gradual changes in atrophy. One limitation of the study is that such benefits were not demonstrated. In future studies, a more detailed analysis on the accuracy and consistency of imaging biomarkers, e.g., sensitivity to signal-to-noise ratio, and their impact on rating scales is needed. Another limitation of this study was the small size of the groups with the most severe grades which affects the construction and validation of the model.

In conclusion, differential diagnostics of cognitive disorders is challenging, and the use of quantitative MRI measures can help making image interpretation more objective and uniform. This study suggests that visual ratings scales can be estimated computationally in a reliable way, and these computational scales may improve performance in diagnostics compared with visual scales.

## References


1. McKhann GM, Knopman DS, Chertkow H, et al (2011) The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7(3):263–269
2. McKeith IG, Galasko D, Kosaka K, et al (1996) Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the consortium on DLB international workshop. Neurology 47(5):1113–1124
3. Neary D, Snowden JS, Gustafson L, et al (1998) Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology 51(6):1546–1554
4. Rascovsky K, Hodges JR, Knopman D, et al (2011) Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 134(9):2456–2477
5. Román GC, Tatemichi TK, Erkinjuntti T, et al (1993) Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop. Neurology 43(2):250–260
6. Dubois B, Feldman H, Jacova C (2007) Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurology 6(8):734-746
7. Harper L, Fumagalli GG, Barkhof F, et al (2016) MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases. Brain 139:1211-1225
8. Koedam, EL, Lehmann, M, van der Flier WM, et al (2011) Visual assessment of posterior atrophy development of a MRI rating scale. Eur Radiol 21(12):2618-2625
9. Fazekas F, Chawluk J, Alavi A, Hurtig H, Zimmerman R. (1987) MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. Am J Roentgenol 149(2):351–356
10. O'Brien JT, Erkinjuntti T, Reisberg B, et al (2003) Vascular cognitive impairment. Lancet Neurology 2(2):89-98

11. Pantoni L, Basile AM, Pracucci G, et al (2005) Impact of age-related cerebral white matter changes on the transition to disability. The LADIS Study: rationale, design, and methodology. Neuroepidemiology 24(1-2):51-62
12. Rhodius-Meester H, Benedictus M, Wattjes M, et al (2017) MRI Visual Ratings of Brain Atrophy and White Matter Hyperintensities across the Spectrum of Cognitive Decline Are Differently Affected by Age and Diagnosis. Frontiers in Aging Neuroscience 9:117
13. Vernooij MW, Haller S, Frisoni G, et al (2018) Dementia imaging in Europe: results from the European Society for Neuroradiology (ESNR) Diagnostic Subcommittee Survey. Available via http://ecronline.myesr.org/ecr2018/index.php?p=recorddetail&rid=e426173f-1196-43d1-8024-c90dfc47180e&t=browsesessions#ipp-record-ffb62685-188c-4c9a-9df5-558216a74881 Accessed 17 Aug 2018
14. Scheltens P, Launer LJ, Barkhof F, Weinstein H, van Gool W (1995) Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: interobserver reliability. J Neurol 242(9): 557–560
15. Pasquier F, Leys D, Weerts J, Mounier-Vehier F, Barkhof F, Scheltens P (1996) Inter and intraobserver reproducibility of cerebral atrophy assessment on MRI scans with hemispheric infarcts. Eur Neurol 36 (5):268–272
16. Torisson G, van Westen D, Stavenow L, Minthon L, Londos E (2015) Medial temporal lobe atrophy is underreported and may have important clinical correlates in medical inpatients. BMC Geriatr 15:65
17. van der Flier WM, Pijnenburg YAL, Prins N, et al (2014) Optimizing patient care and research: the Amsterdam Dementia Cohort. J Alzheimers Dis 41:313–327
18. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the

auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34(7):939–944

19. Petersen R. (2004) Mild cognitive impairment as a diagnostic entitity. J Intern Med 256(3):183-194
20. Albert M, Dekosky S, Dickson D, et al (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7:270-279
21. Lötjönen J, Wolz R, Koikkalainen J, Thurfjell L, Waldemar G, Soininen H, Rueckert D (2010) Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49(3):2352-2365
22. Koikkalainen J, Rhodius-Meester H, Tolonen A, et al (2016) Differential diagnosis of neurodegenerative diseases using structural MRI data. NeuroImage Clinical 11:435-449
23. Wang Y, Catindig JA, Hilal S, et al (2012) Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. NeuroImage 60(4):2379-2388
24. Buckner RL, Head D, Parker J, Fotenos AF, Marcus D, Morris JC, Snyder AZ (2004) A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage 23:724-738
25. Cole TJ and Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Statistics in Medicine 11(10):1305-1319
26. Ashburner J, Friston K (2000) Voxel-based morphometry — the methods. NeuroImage 11(6):805–821

27. Wattjes M, Henneman W, van der Flier WM, et al (2009) Diagnostics imaging of patients in a memory clinic: comparison of MR imaging and 64-detector row CT. Radiology 253(1):174-183

28. Kate M, Barkhof F, Boccardi M, et al (2017) Clinical validity of medial temporal atrophy as a biomarker for Alzheimer's disease in the context of a structured 5-phase development framework. Neurobiol Aging 52:167-182

29. Warlaw J, Smith E, Biessels G et al (2013) Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurology 12:822-838

30. Verhagen MV, Guit GL, Hafkamp GJ, Kalisvaart K (2016) The impact of MRI combined with visual rating scales on the clinical diagnosis of dementia: a prospective study. Eur Radiol 26(6):1716-1722

31. Barkhof F, Polvikoski TM, van Straaten EC, et al (2007) The significance of medial temporal lobe atrophy: a postmortem MRI study in the very old. Neurology 69(15):1521-1527

32. van de Pol LA, Hensel A, van der Flier WM, et al (2006) Hippocampal atrophy on MRI in frontotemporal lobar degeneration and Alzheimer's disease. J Neurol Neurosurg 77(4):439-42

33. Canu E, Agosta F, Mandic-Stojmenovic G et al (2017) Multiparametric MRI to distinguish early onset Alzheimer's disease and behavioural variant of frontotemporal dementia. NeuroImage: Clinical 15: 428-438

34. Bruun M, Rhodius-Meester H, Koikkalainen J et al (2018) Evaluating combinations of diagnostic tests to discriminate different dementia types. Alzheimer's & Dementia Diagnosis, Assessment & Disease Monitoring 10: 509-518

35. Burton EJ, Barber R, Mukaetova-Ladinska EB, et al (2009) Medial temporal lobe atrophy on MRI differentiates Alzheimer's disease from dementia with Lewy bodies and vascular cognitive impairment: a prospective study with pathological verification of diagnosis. Brain 132:195-203
36. Bresciani L, Geroldi C, Galluzzi S, et al (2005) Visual assessment of medial temporal atrophy on MR films in Alzheimer's disease: comparison with volumetry. Aging Clin Exp Res 17(1):8-13
37. Cavallin L, Bronge L, Zhang Y, et al (2012) Comparison between visual assessment of MTA and hippocampal volumes in an elderly, non-demented population. Acta Radiol 53(5):573-579
38. Persson K, Barca M, Cavallin L et al (2017) Comparison of automated volumetry of the hippocampus using NeuroQuant® and visual assessment of the medial temporal lobe in Alzheimer's disease. Acta Radiol 59(8): 997-1001
39. Clerx L, van Rossum I, Burns L et al (2013) Measurements of medial temporal lobe atrophy for prediction of Alzheimer's disease in subjects with mild cognitive impairment. Neurobiol Aging 34: 2003-2013
40. van Straaten E, Fazekas F, Rostrup E et al (2006) Impact of white matter hyperintensities scoring method on correlations with clinical data. Stroke 37: 836-840
41. Clerx L, Jacobs H, Burgmans S et al (2013) Sensitivity of different MRI-techniques to assess gray matter atrophy patterns in Alzheimer's disease is region-specific. Current Alzheimer Research 10: 940-951

## Figure legends

**Figure 1**. Box and Whisker plots defined for computed MTA-L (A), MTA-R (B), GCA (C) and Fazekas (D) when defined separately for each of the four memory clinics (C1-C4) in the PredictND cohort. C1 (red) is the same center as acquired the data in the Amsterdam dementia cohort.

**Figure 2.** Balanced accuracy (BACC) computed between all diagnostic classes using different visual (green bars) and computed (yellow bars) rating scales: MTA-L (A), MTA-R (B), GCA (C) and Fazekas (D). The left and right green (yellow) bars contain results from visual (computed) scales using the ADC and PredictND cohorts, respectively. Abbreviations used: MTA = medial temporal lobe atrophy, GCA = global cortical atrophy (prefix 'c' stands for 'computed'), CN = cognitively normal, AD = Alzheimer's disease, FTLD = frontotemporal lobar degeneration, DLB = dementia with Lewy bodies, VaD = vascular dementia, ADC = Amsterdam Dementia Cohort, PND = PredictND cohort.

**Figure 3**. ROC curves for computed MTA-L (A), MTA-R (B), GCA (C) and Fazekas (D) rating scales using the ADC, PredictND, ADNI-1 and ADNI-2 cohorts. Area under the curve (AUC) of the computed scores were 0.88/0.90/0.88/0.91 for MTA-L, 0.88/0.89/0.86/0.90 for MTA-R, 0.92/0.89/0.85/0.89 for GCA and 0.94/0.84/-/- for Fazekas in ADC/PredictND/ADNI-1/ADNI-2 cohorts.

## Table legends

**Table 1**. Details on visual ratings scales used in this study: MTA, Koedam score, CGA and WMH measured by the Fazekas scale. **Footnote:** Abbreviations used: MTA = medial temporal lobe atrophy, GCA = global cortical atrophy, WMH = white matter hyperintensities.

**Table 2.** Characteristics for the Amsterdam Dementia Cohort (ADC), PredictND cohort (PredictND) and ADNI (ADNI-1 and ADNI-2) cohorts. **Footnote**: Abbreviations used: CN = cognitively normal, AD = Alzheimer's disease, FTLD = frontotemporal lobar degeneration, DLB = dementia with Lewy bodies, VaD = vascular dementia, MCI = mild cognitive impairment, OTH = other dementias, MMSE = mini-mental state examination. [a]Statistically significant difference as compared to CN. [b]Statistically significant difference as compared to AD. [c]Statistically significant difference as compared to FTLD. [d]Statistically significant difference as compared to DLB. [e]Statistically significant difference as compared to VaD. [f]Statistically significant difference as compared to MCI. [g]Statistically significant difference as compared to OTH. Bonferroni correction was used in statistical analysis.

**Table 3**. Visual and computed rating scales using different imaging biomarkers in the Amsterdam dementia cohort (ADC). Correlations have been computed using data from all diagnostic groups. MTA and GCA classification results are computed between AD and CN groups and results for Fazekas between VaD and non-VaD. **Footnote:** Abbreviations used: L=left, R=right, MTA = medial temporal lobe atrophy, GCA = global cortical atrophy, VHC = volume of hippocampus, VILV = volume of inferior lateral ventricle, VCO = volume of cortical gray matter, CHC = concentration index of hippocampal gray matter, CCO = concentration index of cortical gray matter, VWMH = volume of white matter hyperintensities, VDWMH = volume of deep white matter hyperintensities, correlation = Pearson correlation coefficient, AUC = area under the curve,

BACC = balanced accuracy. *Difference statistically significant for correlation, AUC and BACC as compared to all other methods.

**Table 4**. Equations for defining computed rating scales. **Footnote:** Abbreviations: L=left, R=right, MTA = medial temporal lobe atrophy, GCA = global cortical atrophy (prefix 'c' stands for 'computed'), VHC = volume of hippocampus, VILV = volume of inferior lateral ventricle, CCO = concentration index of cortical gray matter, VDWMH = volume of deep white matter hyperintensities

## Tables

### Table 1.

Details on visual ratings scales of MTA, Koedam score, CGA and WMH used in this study.

| | |
|---|---|
| **MTA** (14)<br><br>**Scale rated on coronal T1 images:**<br>0= normal<br>1= widened choroid fissure<br>2= increase of widened fissure, widening temporal horn, opening of other sulci<br>3= pronounced volume loss of hippocampus<br>4= end stage atrophy | **Koedam score** (8)<br><br>**Scale rated in sagittal and coronal T1 and axial flair images:**<br>0= no atrophy<br>1= mild atrophy, opening of sulci<br>2= moderate atrophy, volume loss gyri<br>3= severe atrophy; knife blade |
| **GCA** (15)<br><br>**Scale rated on axial flair images:**<br>0= no atrophy<br>1= mild atrophy, opening of sulci<br>2= moderate atrophy, volume loss gyri<br>3= severe atrophy; knife blade | **WMH** (9-11)<br><br>**Scale rated on axial flair images:**<br>0= none or single (max 3) punctate lesions<br>1= multiple (≥3) punctate lesions<br>2= beginning confluent of lesions<br>3= large confluent lesions |

Abbreviations used: MTA = medial temporal lobe atrophy, GCA = global cortical atrophy, WMH = white matter hyperintensities.

**Table 2.**

Characteristics for the Amsterdam Dementia Cohort (ADC), PredictND cohort (PredictND) and ADNI (ADNI-1 and ADNI-2) cohorts.

| **ADC** | All (n=513) | CN (n=75) | AD (n=223) | FTLD (n=62) | DLB (n=40) | VaD (n=19) |
|---|---|---|---|---|---|---|
| Age | 65 ± 7 | 62 ± 7 [b,e] | 66 ± 7 [a,c] | 62 ± 6 [b,e] | 67 ± 9 | 69 ± 6 [a,c] |
| Females | 226 (44%) | 25 (33%) | 120 (54%) [d] | 27 (44%) [d] | 4 (10%) [b,c,f] | 7 (37%) |
| MMSE | 23 ± 5 | 28 ± 1 [b,c,d,e,f] | 21 ± 5 [a,c,f] | 24 ± 5 [a,b,f] | 23 ± 4 [a,f] | 23 ± 5 [a] |
| 1.5T/3T | 114/399 | 14/61 | 53/170 | 16/46 | 10/30 | 4/15 |

| | MCI (n=94) |
|---|---|
| Age | 65 ± 7 |
| Females | 43 (46%) [d] |
| MMSE | 26 ± 2 [a,b,c,d] |
| 1.5T/3T | 17/77 |

| **PredictND** | All (n=672) | CN (n=227) | AD (n=133) | FTLD (n=25) | DLB (n=21) | VaD (n=19) |
|---|---|---|---|---|---|---|
| Age | 69 ± 10 | 64 ± 9 [b,d,e,f,g] | 71 ± 9 [a] | 65 ± 8 [g] | 72 ± 7 [a] | 74 ± 10 [a] |
| Females | 357 (53%) | 144 (63%) [d,f] | 82 (62%) [d,f] | 12 (48%) | 5 (24%) [a,b] | 7 (39%) [a,b] |
| MMSE | 27 ± 3 | 29 ± 1 [b,c,d,e,f,g] | 24 ± 3 [a,f] | 24 ± 4 [a,f] | 25 ± 3 [a] | 24 ± 3 [a,f] |
| 1.5T/3T | 227/445 | 100/127 | 35/98 | 6/19 | 3/18 | 2/17 |

| | MCI (n=131) | OTH (n=116) |
|---|---|---|
| Age | 69 ± 8 [a,g] | 73 ± 9 [a,c,f] |
| Females | 46 (35%) [a,b] | 61 (53%) |
| MMSE | 27 ± 3 [a,b,c,e,g] | 25 ± 4 [a,f] |
| 1.5T/3T | 39/92 | 42/74 |

| **ADNI-1** | All (n=357) | CN (n=169) | AD (n=188) |
|---|---|---|---|
| Age | 76 ± 7 | 76 ± 5 | 75 ± 7 |
| Females | 177 (50 %) | 86 (51 %) | 91 (48 %) |
| MMSE | 26 ± 3 | 29 ± 1 [b] | 23 ± 2 [a] |
| 1.5T/3T | 357/0 | 169/0 | 188/0 |

| **ADNI-2** | All (n=400) | CN (n=257) | AD (n=143) |
|---|---|---|---|
| Age | 73 ± 7 | 73 ± 6 [b] | 75 ± 8 [a] |
| Females | 201 (50 %) | 143 (56 %) [b] | 58 (41 %) [a] |
| MMSE | 27 ± 3 | 29 ± 1 [b] | 23 ± 2 [a] |
| 1.5T/3T | 0/400 | 0/257 | 0/143 |

Abbreviations used: CN = cognitively normal, AD = Alzheimer's disease, FTLD = frontotemporal lobar degeneration, DLB = dementia with Lewy bodies, VaD = vascular dementia, MCI = mild cognitive impairment, OTH = other dementias, MMSE = mini-mental state examination.

[a]Statistically significant difference as compared to CN.

[b]Statistically significant difference as compared to AD.

[c]Statistically significant difference as compared to FTLD.

[d]Statistically significant difference as compared to DLB.

[e]Statistically significant difference as compared to VaD.

[f]Statistically significant difference as compared to MCI.

[g]Statistically significant difference as compared to OTH.

Bonferroni correction was used in statistical analysis.

**Table 3**.

Visual and computed rating scales using different imaging biomarkers in the Amsterdam dementia cohort (ADC).

| | **MTA-L** | **VHC-L** | **VILV-L** | **CHC-L** | **VHC&VILV-L** |
|---|---|---|---|---|---|
| correlation | - | 0.62±0.03 | 0.80±0.02 | 0.76±0.02 | 0.83±0.01* |
| AUC | 0.82±0.03 | 0.85±0.02 | 0.82±0.03 | 0.83±0.02 | 0.88±0.02* |
| BACC | 0.77±0.03 | 0.74±0.03 | 0.74±0.03 | 0.76±0.03 | 0.79±0.03* |

| | **MTA-R** | **VHC-R** | **VILV-R** | **CHC-R** | **VHC&VILV-R** |
|---|---|---|---|---|---|
| correlation | - | 0.55±0.03 | 0.83±0.02 | 0.76±0.02 | 0.83±0.02* |
| AUC | 0.79±0.03 | 0.84±0.02 | 0.82±0.03 | 0.83±0.02 | 0.88±0.02* |
| BACC | 0.72±0.03 | 0.78±0.03 | 0.73±0.03 | 0.77±0.03 | 0.81±0.03* |

| | **GCA** | **VCO** | **CCO** |
|---|---|---|---|
| correlation | - | 0.46±0.03 | 0.64±0.03* |
| AUC | 0.76±0.03 | 0.89±0.02 | 0.92±0.02* |
| BACC | 0.74±0.03 | 0.80±0.03 | 0.84±0.03* |

| | **Fazekas** | **VWMH** | **VDWMH** |
|---|---|---|---|
| correlation | - | 0.75±0.02 | 0.76±0.01* |
| AUC | 0.88±0.04 | 0.96±0.01* | 0.94±0.02 |
| BACC | 0.79±0.05 | 0.89±0.06* | 0.85±0.06 |

Note: MTA and GCA classification results are computed between AD and CN groups and results for Fazekas between VaD and non-VaD. Abbreviations used: L=left, R=right, MTA = medial temporal lobe atrophy, GCA = global cortical atrophy, VHC = volume of hippocampus, VILV = volume of inferior lateral ventricle, VCO = volume of cortical gray matter, CHC = concentration index of hippocampal gray matter, CCO = concentration index of cortical gray matter, VWMH = volume of white matter hyperintensities, VDWMH = volume of deep white matter hyperintensities, correlation = Pearson correlation coefficient, AUC = area under the curve, BACC = balanced accuracy.

*Difference statistically significant for correlation, AUC and BACC as compared to all other methods.

**Table 4**.

Equations for defining computed rating scales.

| Visual score | Computed rating scale value* |
|---|---|
| MTA-L | y = 2.1 - 0.7·VHC + 0.9·VILV<br>cMTA-L = 1.6·y - 0.7, if y < 1.1<br>cMTA-L = 1.7·y - 0.8, if y > 1.1 and y < 1.6<br>cMTA-L = 1.4·y - 0.4, if y > 1.6 |
| MTA-R | y = 1.4 - 0.4·VHC + 0.8·VILV<br>cMTA-R = 2.2·y - 1.2, if y < 1.0<br>cMTA-R = 1.7·y - 0.6, if y > 1.0 and y < 1.6<br>cMTA-R = 1.1·y + 0.3, if y > 1.6 |
| GCA | y = 0.5 + 0.03·CCO<br>cGCA = 2.3·y - 1.3, if y < 1.0<br>cGCA = 2.0·y - 1.0, if y > 1.0 |
| Fazekas | y = 0.8 + 0.4·log(VDWMH)<br>cFazekas = 2.2·y - 1.3, if y < 1.1<br>cFazekas = 1.5·y - 0.5, if y > 1.1 and y < 1.7<br>cFazekas = 1.8·y - 1.1, if y > 1.7 |

* If needed, the final values of computed scores are cut to make them correspond the range of the visual rating scale value.

Abbreviations: L=left, R=right, MTA = medial temporal lobe atrophy, GCA = global cortical atrophy (prefix ‘c’ stands for ‘computed’), VHC = volume of hippocampus, VILV = volume of inferior lateral ventricle, CCO = concentration index of cortical gray matter, VDWMH = volume of deep white matter hyperintensities

## Figures

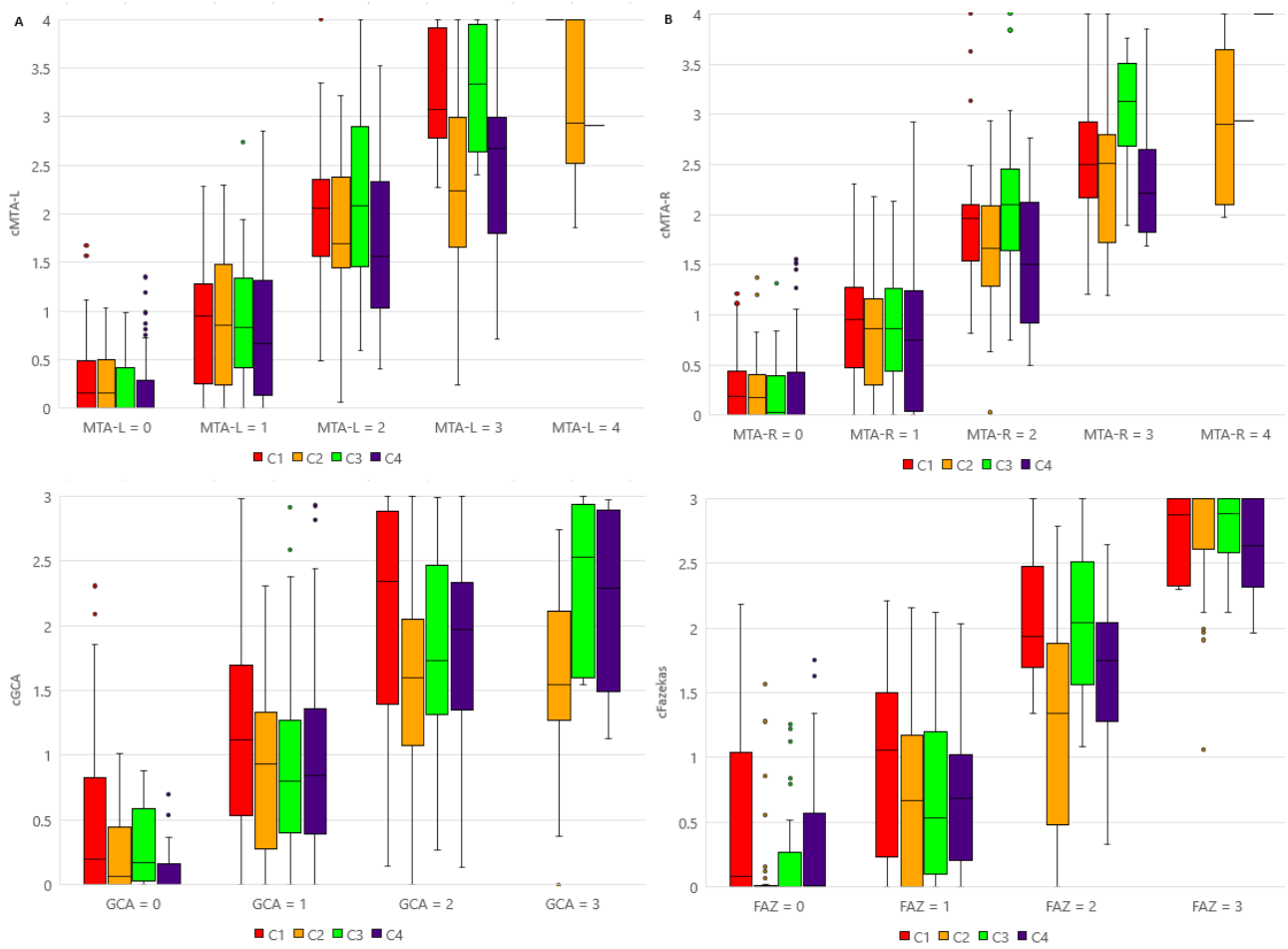


**Figure 1**. Box and Whisker plots computed for cMTA-L (A), cMTA-R (B), cGCA (C) and cFazekas (D) when defined separately for each of the four memory clinics (C1-C4) in the PredictND cohort. C1 (red) is the same center as acquired the data in the Amsterdam dementia cohort.

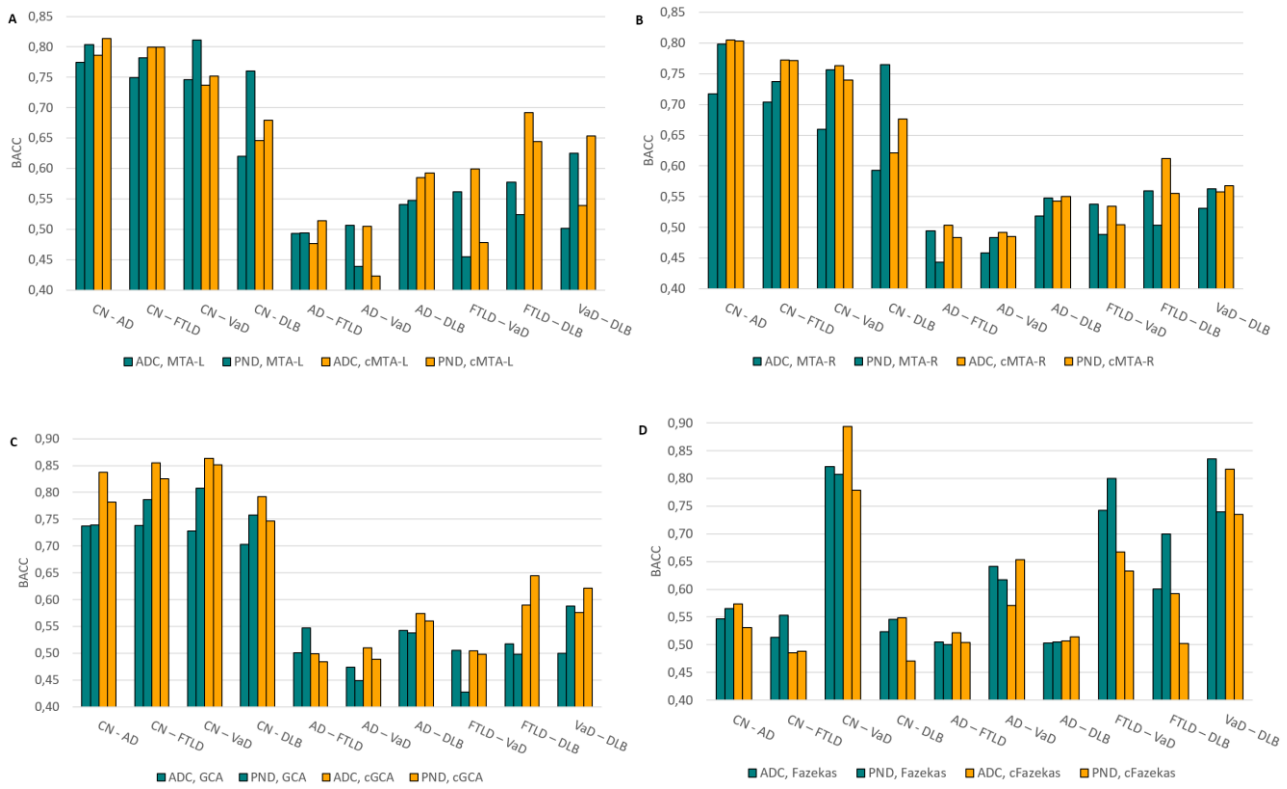


**Figure 2.** Balanced accuracy (BACC) computed between all diagnostic classes using different visual (green bars) and computed (yellow bars) rating scales: MTA-L (A), MTA-R (B), GCA (C) and Fazekas (D). The left and right green (yellow) bars contain results from visual (computed) scales using the ADC and PredictND cohorts, respectively. Abbreviations used: MTA = medial temporal lobe atrophy, GCA = global cortical atrophy (prefix 'c' stands for 'computed'), CN = cognitively normal, AD = Alzheimer's disease, FTLD = frontotemporal lobar degeneration, DLB = dementia with Lewy bodies, VaD = vascular dementia, ADC = Amsterdam Dementia Cohort, PND = PredictND cohort.

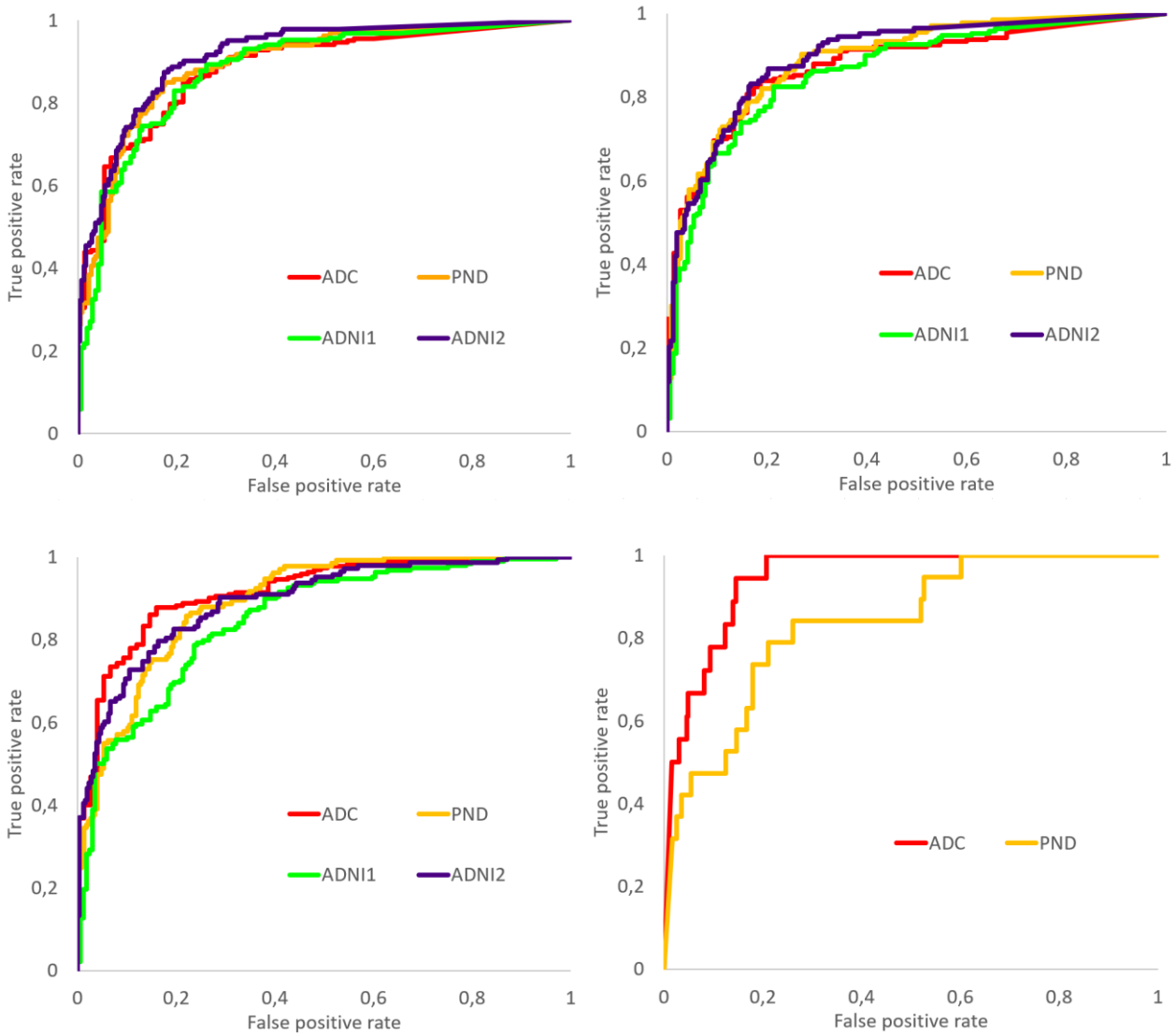


**Figure 3**. ROC curves for computed MTA-L (A), MTA-R (B), GCA (C) and Fazekas (D) rating scales using the ADC, PredictND, ADNI-1 and ADNI-2 cohorts. Area under the curve (AUC) of the computed scores were 0.88/0.90/0.88/0.91 for MTA-L, 0.88/0.89/0.86/0.90 for MTA-R, 0.92/0.89/0.85/0.89 for GCA and 0.94/0.84/-/- for Fazekas in ADC/PredictND/ADNI-1/ADNI-2 cohorts.