Multimodal MRI-based classification of Alzheimer's disease

Vos, F. de

Predicting future cognitive decline of memory clinic patients using multimodal MRI

Submitted for publication

Frank de Vos, Tijn M. Schouten, Rogier A. Feis, Mark A. van Buchem,
Frans R.J. Verhey, Marcel G. M. Olde Rikkert, Philip Scheltens, Mark de Rooij,
Jeroen van der Grond, Serge A. R. B. Rombouts

Abstract

Memory clinic patients want to get a prognosis of their future cognitive decline, but this cannot be accurately predicted using baseline cognitive tests. Magnetic resonance imaging (MRI) scans show brain abnormalities before the onset of cognitive decline, and can therefore possibly be used for a prognosis. We used a machine learning approach to study prediction accuracy of baseline multimodal MRI scans for two-year follow-up cognitive decline in memory clinic patients. At baseline, we included patients (N = 189) from a Dutch multi-centre memory clinic sample with either subjective memory complaints, mild cognitive impairment, or Alzheimer's disease. At baseline, we acquired structural, diffusion, and resting state functional MRI scans, and the patients underwent neuropsychological testing, including the mini-mental state examination and six tests that measure specific domains of memory and executive functioning. Patients that returned for the two-year follow-up visit (N = 117) underwent neuropsychological testing again, and change scores were used as outcome measures. Anatomical MRI scans were used to calculate grey matter density, cortical thickness, and the volumes of subcortical structures; diffusion MRI scans were used to calculate fractional anisotropy, mean, axial, and radial diffusivity; resting state functional MRI scans were used to calculate a whole brain functional connectivity matrix. The MRI features were combined in group lasso regression models to predict change scores for each neuropsychological test separately. The patients showed significant average decline on the mini-mental state examination (MMSE, M = -1.2, SD = 3.9), the visual association test (M = -0.8, SD = 3.0), and the letter digit substitution test (M = -1.7, SD = 7.2), but not on digit span, Rey auditory verbal learning test, Stroop color and word test, and the trail making test. Two-year follow-up change was predicted above chance level for the MMSE (R^2 = 0.07, *p* - Bonferroni corrected = 0.006), but not for the other neuropsychological tests. The MMSE change prediction model included features from all three MRI modalities. The prediction procedure was complicated by high sample attrition, and limited cognitive decline in the analysed sample. Nevertheless, we showed that baseline MRI scans are predictive for future cognitive decline, but the effect size is small, and hence it is not of direct clinical value.

Keywords: Alzheimer's disease, Mild cognitive impairment, Subjective memory complainers, Anatomical MRI, Diffusion MRI, Resting state fMRI, Cognitive decline, Longitudinal

5.1 Introduction

It is important to provide memory clinic patients with an expected course of their cognitive decline. This will reassure patients with positive prospects, and it will help patients with less fortunate perspectives to prepare for future decline. However, memory clinic patients vary largely, and predicting their cognitive decline based on their current clinical status yields inaccurate results (Eckerström et al., 2017; Wahlund et al., 2003). Magnetic resonance imaging (MRI) scans show brain abnormalities before the occurrence of clinical symptoms (Jack et al., 2013), and they can therefore be useful for prediction of future cognitive decline. For example, anatomical MRI scans predict cognitive decline in cognitively normal elderly (Dumurgier et al., 2017; Pacheco et al., 2015) and in AD patients (Sona et al., 2013).

However, attempts to predict cognitive decline in memory clinic patients using whole brain atrophy rate did not result in above chance accuracy (Sluimer et al., 2008). A single measure of global brain atrophy is probably not refined enough to capture the subtle atrophy patterns that precede cognitive decline. Region specific atrophy measures provide more precise information, and could therefore improve prediction. Indications for this can be found in AD classification studies, where AD classification improves when using region-wise atrophy information over using only global atrophy information (de Vos et al., 2016; Westman et al., 2011). In addition, brain atrophy can be quantified using several approaches (e.g., cortical thickness, grey matter density, and brain volume), and combining these approaches improves AD classification (de Vos et al., 2016; Westman et al., 2013). It is likely that prediction of cognitive decline in memory clinic patients will also benefit from using region specific atrophy measures, and multiple approaches to quantify atrophy.

Brain changes that precede cognitive decline are not limited to atrophy, they also include decreased white matter integrity (Gold et al., 2012), and altered functional connectivity (FC) patterns (Sheline et al., 2010a; Sheline et al., 2010b). Adding this information, by respectively using a diffusion MRI scan and a resting state functional MRI (fMRI) scan, could improve prediction of cognitive decline. In fact, multimodal MRI models improve AD classification accuracy over unimodal MRI models (Schouten et al., 2016).

Multimodal MRI scans contain a wealth of information on brain structure and function, represented by a high number of features. Machine learning techniques are well

equipped to integrate high-dimensional data into a single model to make individual predictions. For example, MRI-based machine learning models separated AD patients from controls with high accuracies (Dyrba et al.; 2015b; Schouten et al., 2016; Wee et al., 2013), and even separated MCI converters from non-converters (Cui et al., 2011; Dyrba et al., 2015a; Trzepacz et al., 2014; Teipel et al., 2015).
The current aim is to study prediction accuracy of baseline multimodal MRI scans for two-year follow-up cognitive decline in memory clinic patients. We will analyse a sample collected at four different memory clinics, consisting of patients with subjective memory complaints (SMC), mild cognitive impairment (MCI) patients, and AD patients.

5.2 Methods

The MRI data from the same sample of participants was used in our earlier work (de Vos et al., 2020). Therefore, we adopted its participants paragraph and the paragraphs that describe MR acquisition, MRI preprocessing, the calculation of MRI features, and scan site correction. The earlier work did however not use the follow-up data that includes the neuropsychological assessment data. Therefore, the paragraphs concerning those data, the missing value analysis, and the statistical analysis were newly written.

Participants

The participants were included as part of the Leiden-Alzheimer research Nederland (LeARN) project (Handels et al., 2012; Jansen et al., 2017). LeARN is a longitudinal multi-centre collaboration of four memory clinics in the Netherlands; Leiden, Maastricht, Nijmegen and Amsterdam, in which participants were followed for two years. At the baseline visit, the participants were scanned with an anatomical, diffusion and resting state fMRI scan, and they underwent a neuropsychological assessment. At the two-year follow-up visit they underwent the same neuropsychological assessment again. The inclusion criteria for LeARN are: subjective and/or objective memory complaints, suspicion of having a primary neurodegenerative disease, a mini-mental state examination ≥ 20, clinical dementia rating between 0 and 1, and the availability of a reliable informer or proxy who visits or contacts the participant at least once a week. At baseline, we included participants for which anatomical MRI, diffusion MRI and resting state fMRI were available, and we excluded participants that were diagnosed with MCI not due to AD or dementia not due to AD (e.g., vascular dementia or frontotemporal dementia). The AD diagnosis was made according to the NINCDS-ADRDA Criteria (McKhann et al., 2011), and the MCI diagnosis was made according to

the core clinical criteria for MCI due to AD (Albert et al., 2011). Participants that did not meet the criteria for either AD or MCI were labeled as patients with subjective memory complaints (SMC). We included 189 participants at baseline, of whom 117 participants returned for the follow-up visit after two years. The data from these 117 participants were used for the statistical analysis (see Table 1 for the sample demographics).

Neuropsychological assessment

To measure cognitive decline, we used a standardised battery of cognitive tests that measure general cognitive functioning (mini-mental state examination), working memory (digit span, Rey auditory verbal learning test, and visual association test), and executive functioning (Stroop colour and word test, trail making test, and letter digit substitution test). These tests were administered by a (neuro)psychologist at baseline and at two-year follow-up (M = 24.9 months, SD = 2.4 months). To quantify cognitive decline, we subtracted the baseline test scores from the follow-up test scores, such that negative scores denote decline and positive scores denote improvement.

Mini-mental state examination

The mini-mental state examination (MMSE; Folstein et al., 1975) is a short but thorough examination of general cognitive functioning. It requires 5-10 minutes to administer and the score range is 0-30.

Table 1. Sample demographics

	Baseline	Two-year follow-up
N	117	117
Sex (♂/♀)	80/37	80/37
Age	66.7 ± 9.1	68.8 ± 9.1
SMC/MCI/AD/other	47/39/31/0	49/26/34/8
Mini-mental state examination	26.7 ± 2.8	25.5 ± 5.0
Clinical dementia rating	0.55 ± 0.29	0.62 ± 0.51
Geriatric depression scale	3.4 ± 2.8	3.0 ± 2.6

Descriptives are presented as frequencies for the categorical variables and as mean ± standard deviation for the other variables. SMC = Subjective memory complaints, MCI = Mild cognitive impairment, AD = Alzheimer's disease.

Digit span

Digit span is a subtest of the Wechsler adult intelligence scale-III (WAIS-III; Wechsler, 1997). The examinees are required to repeat strings of digits of increasing length in forward and reverse order. They get two tries for each string length, and the test continues until both tries of the same length are missed. The test is scored by summing the correct number of tries. We added the scores for the forward and reverse condition.

Rey auditory verbal learning test

We used the first part of the Rey auditory verbal learning test (RAVLT; Rey, 1958), in which the same list of 15 words is orally presented in five trials. After each trial, the examinee is asked to recall these 15 words. We used the sum score of the five trials, resulting in a possible score range of 0-75.

Visual association test

The visual association test (VAT; Lindeboom et al., 2002) first presents the examinee 12 cards with two objects (e.g., an ape with an umbrella), and then presents 12 cards with only one of those two objects (e.g., only an ape). The examinee is asked to name the missing object for each of these 12 cards. The test is scored by counting the number of correct recalled missing objects, resulting in a possible score range of 0-12.

Stroop colour and word test

The Stroop colour and word test (Stroop, 1935) presents the examinee coloured cards or cards with printed colour names. In three conditions the examinee is asked to name these colours as fast as possible. In the first condition the name of a colour is written in black ink, and in the second condition the card is coloured but contains no text. These two conditions serve as practice rounds for the third condition, which is incongruent. In this third condition, the name of a colour is printed in an inconsistent colour (e.g., the word 'red' printed in blue letters), and the examinee is required to name the colour of the ink instead of reading the word. For the analyses, we divided the total reaction time in the incongruent condition by the total reaction time of the second congruent condition.

Trail making test

The trail making test (TMT; Reitan, 1956) consists of two conditions in which 26 dots need to be connected by a line. In the first condition, the dots are numbered 1 to 26 and they have to be connected in that same order. In the second condition they are numbered 1 to 13 and labeled *a* to *m*, and they need be connected as 1-a-2-b etc. For

the analyses, we divided the time it took to finish the second condition by the time it took to finish the first condition.

Letter digit substitution test

For the letter digit substitution test (LDST; Natu & Agarwal, 1995) the examinee is asked to substitute letters with numbers according to a key that links nine different letters with the numbers 1 to 9. The test is scored by counting the number of correct substitutions within 90 seconds.

MR acquisition

The participants were scanned at four different scan sites in the Netherlands. At the Leiden University Medical Centre and the Maastricht University Medical Centre, the participants were scanned on a Philips Achieva 3T scanner; at the Nijmegen University Medical Centre, they were scanned on a Siemens TrioTim 3T scanner; and at the VU university medical centre in Amsterdam, they were scanned on a GE Signa HDxt 3T scanner. The MRI sequence parameter settings are listed in supplementary Table 1.

MRI preprocessing

The MRI data was preprocessed using the FMRIB Software Library (FSL version 5.0; Jenkinson et al., 2012; Smith et al., 2004). For the anatomical MRI scans, we applied brain extraction and bias field correction. For the diffusion MRI scans, we applied brain extraction and eddy current correction. For the resting state fMRI data, this included brain extraction, motion correction, a temporal high pass filter with a cutoff point of 100 seconds, 3 mm FWHM spatial smoothing, and non-linear registration to standard MNI152 space. Additionally, we used ICA-AROMA to automatically identify and remove noise components from the fMRI time course (Pruim et al., 2015). ICA-AROMA adequately removes motion related noise from fMRI data, without the need for removing volumes with excessive motion (Parkes et al., 2017).

Anatomical MRI features

Grey matter density

We used voxel-based morphometry (VBM; Ashburner et al., 2000) in FSL (Jenkinson et al., 2012; Smith et al., 2004) to calculate grey matter density. This includes segmentation of the brain-extracted images into grey matter, white matter, and cerebrospinal fluid (CSF), and non-linear registration of the grey matter images to the ICBM-152 grey

matter template. We then calculated weighted averages of the voxel-wise grey matter density values within the 48 regions of the probabilistic Harvard-Oxford cortical atlas, yielding 48 grey matter density values per participant.

Subcortical volumes

We used the FMRIB's Integrated Registration and Segmentation Tool (FIRST; Patenaude et al., 2011) to calculate the volumes of the subcortical structures and we corrected the volumes for intracranial volume. This yielded 14 subcortical volume features per participant (thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens for both hemispheres).

Cortical thickness

We used the Freesurfer software package (Dale et al., 1999; Fisch et al., 1999) to calculate cortical thickness. This includes intensity normalisation of the brain-extracted image to obtain an image with high contrast to noise ratio. This image is used to locate the boundaries between grey matter, white matter and CSF. Subsequently, a triangular mesh is constructed around the white matter surface, and this mesh is deformed outwards to create a grey matter surface that closely follows the boundary between grey matter and CSF. Cortical thickness is defined as the distance between the white matter surface and the grey matter surface. The image is registered to the Freesurfer common template using the image's cortical folding pattern, and the neocortex is parcellated into the 68 neocortical regions (34 regions for each hemisphere) of the Desikan-Killiany atlas (Desikan et al., 2006). This yielded 68 cortical thickness features per participant.

Diffusion MRI features

We used the diffusion MRI scans to calculate fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (DA), and radial diffusivity (DR). First, we used DTIFIT in FSL (Jenkinson et al., 2012; Smith et al., 2004) to fit a diffusion tensor model at each voxel to calculate voxel-wise FA, MD, DA and DR images for each participant. Then we projected participants' FA, MD, DA and DR images onto the FMRIB58_FA mean FA image using tract-based spatial statistics (TBSS; Smith et al., 2006). Finally, we calculated weighted averages of the FA, MD, DA and DR values within the 20 regions of the probabilistic JHU white-matter tractography atlas, yielding 20 features for FA as well as MD, DA and DR.

Resting state fMRI features

Functional connectivity was calculated between resting state networks (RSNs) as obtained by an independent component analysis (ICA). The ICA was previously run in our earlier work with 76 AD patients and 173 elderly controls (de Vos et al., submitted). We had used temporal concatenation ICA in FSL MELODIC (Beckmann & Smith, 2004) to obtain 70 ICA components. For the current study, we transformed the weight maps of these ICA components to subject space, weighted them by the participant specific grey matter density maps, and multiplied them with the functional data. Subsequently, we calculated the mean time courses for the 70 components and used these for the FC analysis. We calculated sparse partial correlations using the Graphical Lasso algorithm (Friedman et al., 2008), with regularisation parameter λ = 100 (Smith et al., 2011). For each participant we thus calculated a 70 by 70 sparse partial correlation matrix yielding (70 * 69)/2 = 2415 features.

Correction for scan site effects

We corrected for scan site effects using ComBat (Johnson et al., 2007). ComBat is validated for anatomical MRI data (Fortin et al., 2018), diffusion MRI data (Fortin et al., 2017), and resting state fMRI data (Yu et al., 2018). ComBat fits a linear model of location and scale for each feature, assuming that sites have both an additive and multiplicative

Table 2. MRI features

	# of features
Anatomical MRI features	
Grey matter density	48
Subcortical volumes	14
Cortical thickness	68
Diffusion MRI features	
Fractional anisotropy	20
Mean diffusivity	20
Axial diffusivity	20
Radial diffusivity	20
Resting state fMRI features	
Functional connectivity	2,415

effect on the data. It uses empirical Bayes to improve the estimation of the model parameters. The model furthermore assumes that the expected value of a feature can be modelled by both the site effect, and biological and demographical factors. ComBat thus removes the unwanted site effects, while it preserves the variation that is associated with the biological and demographical factors. We included age, sex, years of education, clinical label at baseline, and MMSE score at baseline as factors in the ComBat model. To achieve most accurate removal of the scan site effects, we included all 189 participants that were available at the baseline visit in the ComBat procedure.

Missing value analysis

The data contained missing values. This was either due to drop out of the patients, or because patients were unable to execute a specific test. The missing values can either be missing completely at random (MCAR), representing the situation where missingness of data is unrelated to any variable, missing at random (MAR) where missingness of data is only related to the observed variables, or missing not at random (MNAR) where missingness of data is related to the values of the missing data itself. We used Little's test (Little, 1988) to test the null hypothesis that the missing values are MCAR. If Little's test rejects the MCAR hypothesis, the missingness is either MAR or MNAR, but there is no way of deciding between these two alternatives. Within Little's test we included age, sex, years of education, and the baseline and follow-up measures of the seven cognitive tests.

Statistical analyses

The eight different MRI feature groups, along with the number of features per group are listed in Table 2. These feature groups were used jointly to predict cognitive decline. We studied prediction of decline for each neuropsychological test separately.

Group lasso regression

We used group lasso regression (Yuan and Lin, 2006) to predict cognitive decline. Group lasso is a form of penalised regression that takes the group structure of the predictors into account. It uses an L1 penalty (LASSO; Tibshirani, 1996) on the feature groups. This enforces sparseness by either entirely including or excluding a group of predictors, which facilitates the interpretation of the prediction model. It uses an L2 penalty (Ridge; Hoerl and Kennard, 1970) within the feature groups that tends to include all features of the included groups, but limits the size of their contributions. For the group lasso we need to tune the hyperparameter λ: which determines the size of the penalties.

Nested cross-validation

We used nested cross-validation to determine the prediction error (Krstajic et al., 2014). Nested cross-validation takes into account two potential sources of overfitting: too liberally including predictors, and overfitting on the hyperparameter. To ascertain that one is not subject to any of these two sources of overfitting, nested cross-validation uses an inner loop to tune the hyperparameters and an outer loop to train and test the prediction model. For both the inner and outer loop we used 10-fold cross-validation. We repeated this procedure 10 times to reduce the variance resulting from the random partitioning of the participants into folds.

Model evaluation

To evaluate prediction performance, we calculated R squared (R^2) values using the following formula

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

where denotes the sums of squares of the residual error, SS_{tot} denotes the total sums of squares, y denotes the observed values, y denotes the predicted values, and y denotes the sample mean. An R^2 of 1 indicates perfect model fit, whereas an R^2 of 0 indicates that the model predicts no better than the sample mean. Negative R^2 values occur when the model predicts worse than the sample mean. This result is counter intuitive, but can occur when using cross-validation and there is little to no relation between the predictors and the outcome variable (van Loon et al., 2020).

To test the R^2 values against chance, we used a permutation procedure with 10.000 permutations. We permuted the order of the participants' true outcome, while maintaining the order of the participants' predictions. For each permuted version of the data we calculated the R^2 value, resulting in an empirical distribution of R^2 values. To determine the p-value, we compared the observed R^2 value to this distribution. The permutation procedure was executed separately for each outcome measure, and we applied the Bonferroni correction to correct for multiple comparisons.

5.3 Results

Correction for scan site

Before correction, there are large site effects for the diffusion MRI features, moderate site effects for the anatomical MRI features, and no visible site effects for the resting state fMRI features. The site effects have been removed using the ComBat procedure, leaving no visible site effects afterwards (Fig. 1)

Missing value analysis

We included 189 participants at baseline, of whom 117 participants returned for the follow-up visit after two years. The data from these 117 participants were used for the statistical analyses. However, not all these 117 participants finished all neuropsychological tests at both time points. There were 112 available participants for the MMSE and the digit span test, 109 for the RAVLT, 111 for the VAT, 96 for the Stroop task, 88 for the TMT, and 89 for the LDST (Table 3). Little's test indicated that the missing values are not missing completely at random, X^2 (538) = 782.20, $p < 0.001$. In addition, we compared the baseline characteristics of the 117 participants that returned for the follow-up visit, and the 72 participants that did not. The results are presented in suppl. Table 2. The group that returned for the follow-up visit had of a higher percentage of male participants ($p<0.05$) and higher average scores on the digit span test ($p<0.001$), the Rey auditory verbal learning test ($p<0.01$), and the visual association test ($p<0.01$).

Cognitive decline

The participants showed a significant average decline on the MMSE, the VAT and the LDST, but not on the four other tests (Table 3).

MRI-based prediction of cognitive decline

Two-year follow-up decline in test scores was predicted above chance level for the MMSE ($R^2 = 0.07$, p-corrected = 0.006; Fig 2, panel A). For the other neuropsychological tests, the R^2 values are negative (see Table 4), indicating that the model predicts worse than the sample mean. This result is counter intuitive, but can occur when using cross-validation and there is little to no relation between the predictors and the outcome variable (van Loon et al., 2020).

Contribution of MRI features to the prediction of MMSE decline

To determine the contribution of the different MRI feature groups to the prediction of MMSE decline, we counted the number of times that an MRI feature group was

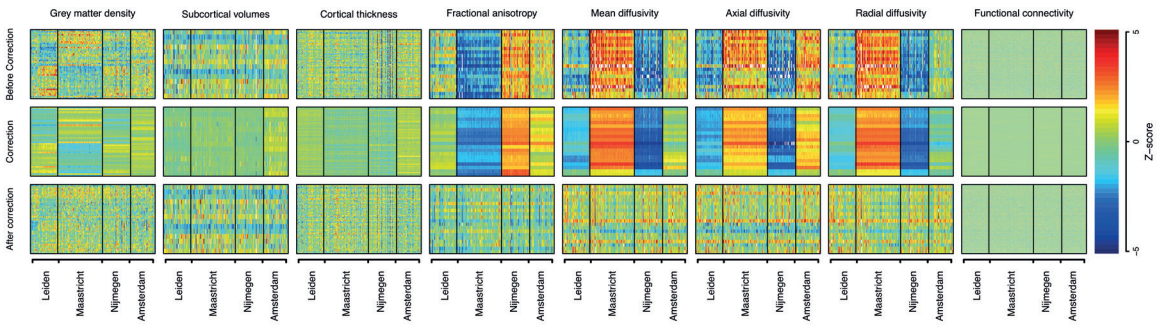

Figure 1. Scan site correction for the MRI data. The top row shows the feature values before scan site correction, the middle row shows the amount of scan site correction that was applied, and the bottom row shows the feature values after scan site correction. The matrix rows represent the features, and the matrix columns represent the participants from the four different scan sites. Before correction there are large site effects, but after correction these site effects have largely disappeared.

selected within the different prediction models (Fig 2, panel B). In total, we fitted 100 prediction models (10 outer folds 10 cross-validation repetitions). All those prediction models selected the grey matter density features (100 times), and most prediction models selected the axial diffusivity (99 times) and FC features (99 times). Most prediction models thus selected features of all three scan types, which likely reflects the added value of multimodal MRI scans.

Table 3. Neuropsychological test results at baseline and two-year follow-up

		Baseline	Two-year follow-up	Change
	N	M ± SD	M ± SD	M ± SD
Mini-mental state examination	112	26.7 ± 2.8	25.5 ± 5.1	-1.2 ± 3.9**
Digit span	112	13.3 ± 3.3	13.5 ± 3.9	0.2 ± 2.8
Rey auditory verbal learning test	109	32.5 ± 11.0	32.0 ± 12.9	-0.5 ± 7.5
Visual association test	111	9.9 ± 3.1	9.1 ± 3.9	-0.8 ± 3.0**
Stroop colour and word test	96	1.9 ± 0.5	1.9 ± 0.6	0.0 ± 0.5
Trail making test	88	2.6 ± 1.4	2.8 ± 1.3	0.2 ± 1.3
Letter digit substitution test	89	39.8 ± 10.5	38.1 ± 13.0	-1.7 ± 7.2*

Two-sided paired sample *t*-tests were performed to test the average change in test scores against 0. The *p*-values are not corrected for multiple comparisons. *$p<0.05$, **$p<0.01$

Table 4. Results for MRI-based prediction of cognitive decline

	N	R^2	p	p-corrected
Mini-mental state examination	112	0.07	0.001	0.006
Digit span	112	-0.06	0.26	1.00
Rey auditory verbal learning test	109	-0.03	0.08	0.56
Visual association test	111	-0.08	0.98	1.00
Stroop colour and word test	96	-0.08	0.90	1.00
Trail making test	88	-0.11	0.99	1.00
Letter digit substitution test	89	-0.09	0.24	1.00

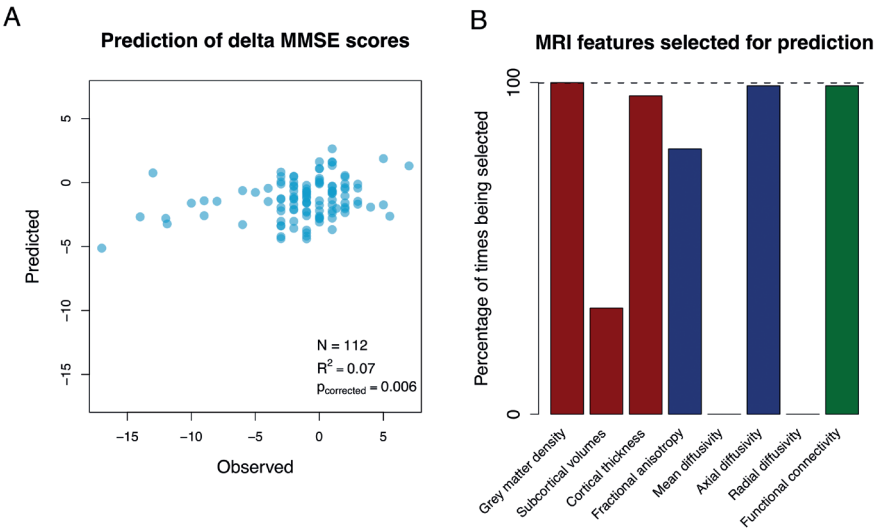

Figure 2. Observed versus predicted change scores on the mini-mental state examination (MMSE) after a two-year follow-up period. Negative scores denote decline and positive scores denote improvement (**panel A**). MRI features used for the prediction of MMSE decline. The group lasso regression method selects groups of features for the prediction model. In total, 100 different prediction models were fitted during the cross-validation procedure (10 outer folds 10 repetitions). The barplot shows the percentage of times that a certain MRI feature group was selected for the prediction model. The anatomical MRI features are presented in red, the diffusion MRI features are presented in blue, and the resting state fMRI features are presented in green. All prediction models contained the grey matter density features, and most prediction models contained the axial diffusivity and functional connectivity features (**panel B**).

5.4 Discussion

The current aim was to study prediction accuracy of baseline MRI scans for future cognitive decline in memory clinic patients. We calculated features from anatomical MRI, diffusion MRI and resting state fMRI scans and used those in a machine learning approach to predict two-year follow up change scores on seven different neuropsychological tests. Change in general cognitive functioning as measured by the MMSE was predicted above chance level, but we could not predict change on the six other neuropsychological tests that measured specific domains of memory and executive functioning. The model that predicted change in MMSE score selected features from both the structural, diffusion and resting state fMRI scans. This might indicate the added value of multimodal MRI for the prediction of cognitive decline, and it would add to research that shows that there is complementary information in multimodal MRI scans for AD classification (Mesrob et al., 2012; Schouten et al., 2016). However, we only used multimodal MRI prediction models, and did not try unimodal

MRI prediction models. We have thus not statistically compared the prediction performances of the multimodal MRI model with unimodal MRI models. Therefore, it is not certain that there is significant added value in multimodal MRI over unimodal MRI for the prediction of cognitive decline.

Furthermore, we did not inspect the contributions of specific brain regions or specific connections between brain regions. In the current study it is difficult to provide information on that scale, because of the high dimensionality of the MRI data. We have for example forced sparsity in the prediction models, which hinders potentially relevant brain regions from entering the model. Also, the effect of a brain region is conditional on the effect of all the other brain regions in the model, and can therefore not be interpreted independently. Furthermore, due to multicollinearity, otherwise important brain regions can become redundant in light of other brain regions, and consequently be left out the model. For these reasons we chose not to inspect the contribution of specific brain regions or specific connections between brain regions.

The effort of predicting cognitive decline has been complicated by a lack of observed cognitive decline in the analysed sample. At average, there was significant decline on the MMSE, the VAT and the LDST, but the amount of decline was little. For example, the average decline on the MMSE was only 1.2, whereas a change of at least 4 is considered to be a reliable change (Tombaugh, 2005). In the current sample only 14 participants (12.5%) declined more than 4 points on the MMSE. For the digit span test, the RAVLT, the Stroop colour and word test and the TMT there was no significant average decline.

Possibly, there was more actual cognitive decline, but it did not fully appear due to missing data. The number of missing values is large, and we showed that these missing values are not missing completely at random. The analysed part of the sample, as compared to participants that dropped out, had higher average baseline scores on three out of seven cognitive tests. In this case, missing at random would be the most fortunate scenario, in which the missing of the data points is independent of the actual value of that data point, and cognitive decline itself would not have caused dropout. However, this scenario is unlikely. For example, patients that decline more can be less motivated to further participate, or death may cause early dropout. Therefore, the missing values are most probable missing not at random. This scenario would affect the results most, because the observed cognitive decline would then be an underestimation of the cognitive decline within the entire sample. This makes

prediction of cognitive decline more difficult, and the predictive accuracy of baseline MRI that we observed would be an underestimation. Nevertheless, baseline MRI scans were predictive for MMSE change scores within the analysed part of the sample. It shows that baseline MRI scans do contain information on future cognitive decline, but the quality of the prediction is perhaps affected by the large amount of missing data.

The lack of observed cognitive decline could also have been caused by practice effects. Neuropsychological tests are vulnerable to practice effects (Calamia et al., 2012), and these practice effects complicate the detection of cognitive decline (Machulda et al., 2013). Instead, we could have aimed to predict conversion from one diagnosis to another. Clinical diagnoses combine information on cognitive tests, with observed behaviour and biological markers, and therefore more reliably characterise cognitive status. However, the memory clinic patients in this sample vary in their initial diagnoses, and the follow-up period was only two years. Hence, the observed conversions are too few and too diverse to use for statistical analysis. Therefore, neuropsychological testing was more suitable to detect cognitive decline within the current data set.

In conclusion, we used baseline MRI scans of memory clinic patients to predict their two-year follow-up cognitive decline. We were able to predict decline in general cognitive functioning as measured by the MMSE, but we could not predict decline on specific domains of memory and executive functioning. Our procedure was extensive, because we used a wide range of measures from both anatomical MRI, diffusion MRI and resting state fMRI scans. Yet, the extent to which baseline MRI could predict future cognitive decline was limited. The prediction procedure has been complicated by a lack of observed cognitive decline, which was probably caused by the large number of participants that dropped out. However, drop-out is common to longitudinal studies among elderly, and it is not easy to prevent this. In addition, neuropsychological tests are vulnerable to practice effects, and it might be better feasible to predict future conversions to new clinical diagnoses. However, memory clinic patients have diverse clinical trajectories, and studying clinical conversions in this population requires a large number of participants and a sufficiently long follow-up period.

5.5 Funding

This study is supported by VICI grant no. 016.130.677 of the Netherlands Organisation for Scientific Research (NWO).