Classification and early detection of dementia and cognitive decline with magnetic resonance imaging

Schouten, T.M.

Chapter 5

Longitudinal prediction of cognitive decline using multimodal MRI

Tijn M. Schouten, Frank de Vos, Jessica C. Foster-Dingley, Sanneke van Rooden, Mark J.R.J. Bouts, Rogier A. Feis, Justine E. F. Moonen, Roos C. van der Mast, Mark de Rooij, Jeroen van der Grond, & Serge A.R.B. Rombouts

Abstract

Finding biomarkers for early prediction of cognitive decline is the key to developing treatments for dementia. Magnetic resonance imaging has shown promising results for detection of Alzheimer's disease, conversion from mild cognitive impairment to Alzheimer's disease, and cognitive decline. In this study we use machine learning on multimodal MRI data to predict cognitive decline after four years. We used MRI measures that have been shown to successfully predict Alzheimer's disease and other types of dementia. Specifically, we used grey matter density averaged over pre-defined structural regions, and subcortical volumes from anatomical MRI, fractional anisotropy and mean diffusivity with tract-based spatial statistics from diffusion MRI, functional connectivity between independent components from resting state functional MRI, and small vessel disease markers. We combined these features in a group lasso regression model to predict decline in overall cognition, executive functioning, and memory. The combined MRI model did not contain predictive value for baseline or future cognitive functioning for any of the outcome measures. This suggests that MRI, while successful at disease diagnosis, may not be suitable to predict future cognitive decline.

Key words: cognitive decline; machine learning; MRI; prediction; regression

5.1 Introduction

Finding an early biomarker for dementia or cognition related disorders is the holy grail for developing treatments for Alzheimer's disease and cognitive decline in elderly. Magnetic resonance imaging (MRI) has been studied extensively to find biomarkers associated with dementia. Brain atrophy as measured with structural MRI has been found to be a useful clinical biomarker for Alzheimer's disease (AD) at the mild cognitive impairment (MCI) phase (Frisoni et al., 2010). Resting state functional MRI (rsfMRI) studies also found alterations in functional connectivity associated with dementia (Binnewijzend et al., 2012; Agosta et al., 2012; Greicius et al., 2004). Diffusion MRI revealed alterations in white matter structures for subjective cognitive decline (Li et al., 2016), AD (Medina et al., 2006), MCI (Rose et al., 2006), and fronto-temporal dementia (Zhang et al., 2009).

The most relevant for development of potential early treatments is presymptomatic prediction of diagnoses or future cognitive decline. Some studies show promising potential of MRI for prediction of MCI conversion to Alzheimer's disease (e.g., Cui et al., 2011; Plant et al., 2010b; Adaszewski et al., 2013; Rathore et al., 2017). Some MRI biomarkers, such as cortical thickness and cerebral perfusion measured with arterial spin labelling, have been found to contain some predictive value for future cognitive decline (Dickerson and Wolk, 2012; Chao et al., 2010; Benedictus et al., 2017).

However, prediction of future cognitive decline is challenging. Subjects require long well documented follow-up. Also, the rate of cognitive decline or conversion to a disease diagnosis in a healthy population is low, so that very large samples at baseline are required. Still, selection for potential clinical trials requires that individual predictions can be made with reasonable accuracy. In general, we expect that combining multiple MRI modalities, rather than a single measure, is the most promising approach to make accurate predictions (e.g., Mesrob et al., 2012; Schouten et al., 2016).

In this study we investigate the predictive value of MRI for cognitive decline in elderly individuals with mild cognitive deficits four years after baseline. We will use multimodal machine learning on data from structural, diffusion, and functional MRI in combination with small vessel disease markers. Our sample consisted of elderly individuals that were selected to be at risk of future cognitive decline.

5.2 Materials and methods

5.2.1 Data sample

The data in this study was collected as part of the Discontinuation of Antihypertensive Treatment in Elderly People (DANTE) study; which was an intervention study that investigated the effect of discontinuation of antihypertensive treatment on cognitive functioning (Moonen et al., 2015). Individuals were eligible for inclusion in this study if their mini-mental state examination (MMSE) score was between 21 and 27, and if they were 75 years or older, had a systolic blood pressure of 160 mm Hg or lower, and used antihypertensive treatment. The original study found no effects of discontinuation of antihypertensive treatment on cognitive functioning. Participants had follow-up cognitive assessment after four years. We use the baseline small vascular disease markers and other MRI measurements to predict the cognitive performance differences between baseline and follow-up.

Out of the 356 participants from the original study, 205 had MRI data. Out of these, 102 had follow-up cognitive performance assessment. We excluded 16 participants because of bad quality MRI scans, due to excessive movement, or artefacts. The final sample consisted of 86 participants for whom good quality anatomical MRI, diffusion MRI, and resting state fMRI data were available. Out of these, four participants had incomplete executive function data, and were excluded from the executive function and compound overall analysis. See table 5.1 for an overview of the sample.

Table 5.1: Baseline demographics for the study population

Characteristic	Baseline sample (N=86)
Age	79.65 (3.30)
Gender, ♂/♀	33 / 53 (62% ♀)
Years of education	9.30 (3.22)
Systolic blood pressure	144.2 (19.17)
Diastolic blood pressure	81.83 (9.807)
Randomisation (yes/no)[a]	42/44 (49%)

Data is represented as mean (standard deviation), or as numbers (percentage)
[a] Randomisation factor in the original DANTE study.

MRI acquisition

The MRI data was acquired at Leiden University Medical Center with a 3T Philips Achieva MR system with a 32-channel head coil. Anatomical T1-weighted images were acquired with TR = 9.7 ms, TE = 4.6 ms, flip angle = 8°, voxel size = $1.17 \times 1.17 \times 1.40$ mm^3. Diffusion images were acquired along 32 measurement directions with TR = 9592, TE = 56 ms, flip angle = 90°, FOV = $220 \times 220 \times 128$ mm, matrix size = 112×112, 64 slices, voxel size = $1.96 \times 1.96 \times 2$ mm^3, b = 1000 mm/s^2. Additionally, a single b = 0 image was acquired. Resting state fMRI series of 200 volumes were obtained with TR = 2.2 ms, TE = 30 ms, flip angle = 80°, 38 axial slices, with a voxel size of $2.75 \times 2.75 \times 2.2$ mm^3. We instructed participants to lie still with their eyes closed, and to stay awake. Fluid attenuated inversion recovery (FLAIR) images were acquired with TR = 11,000 ms, TE = 125 ms, flip angle = 90°, FOV = $220 \times 176 \times 137$ mm^3, matrix size = 320×240, 25 transverse slices, 5 mm slice thickness.

5.2.2 Software

The MRI data were preprocessed using FMRIB Software Library (FSL, version 5.0.8) (Smith et al., 2004; Jenkinson et al., 2012). Python 3.4 was used for whitening the fMRI time courses and calculating sparse partial correlations. R (version 3.2.3) with the gglasso package (Yang and Zou, 2015) was used for fitting the group lasso model.

5.2.3 Preprocessing

The anatomical MRI data preprocessing consisted of brain extraction, bias field correction, and non-linear registration to standard MNI152 (Grabner et al., 2006). The diffusion MRI was brain extracted, motion corrected, and eddy-current corrected using FSL, and denoised using the MRtrix program dwidenoise (Veraart et al., 2016). The resting state fMRI data was motion corrected, intensity normalized, denoised by using Automatic Removal of Motion Artifacts (AROMA), which is an independent component analysis based strategy for automatic removal of movement components from the fMRI time courses (Pruim et al., 2015), temporal high-pass filtered with a cut off point of 100s, and smoothed with a 3 mm FWHM gaussian kernel.

5.2.4 Anatomical MRI features

The anatomical MRI features consisted of subcortical volumes, and cortical grey matter density. The subcortical volumes were calculated with the FMRIB's integrated registration and segmentation tool (FIRST) in FSL (Patenaude et al., 2011). The subcortical volumes were corrected for intracranial brain volume, which we determined by the sum of grey matter and white matter segmentation by FMRIB's automated segmentation tool (FAST; Zhang et al., 2001). This resulted in 14 subcortical volumes per subject (thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens for left and right hemispheres). The cortical grey matter density was determined over the 48 Harvard-Oxford cortical atlas regions that were split into left and right, resulting into 96 regions. The cortical grey matter density was determined by running FSL's voxel based morphometry (VBM) on the preprocessed T1-weighted images, and then calculating the mean grey matter density over the 96 regions, resulting into 96 grey matter density values per person.

5.2.5 Diffusion MRI features

To calculate the diffusion MRI measures we used *dtifit* to determine the fractional anisotropy (FA) and mean diffusivity (MD) per voxel, which we then registered to a white matter skeleton using tract-based spatial statistics. These skeletonised FA and MD were subsequently averaged over 48 regions from the ICBM-DTI-81 white-matter labels atlas (Mori et al., 2005) provided with FSL, resulting in 48 FA and 48 MD values per person.

5.2.6 Resting state functional MRI features

The functional connectivity was determined by first performing independent component analysis with 70 components on the preprocessed data (Dipasquale et al., 2015; de Vos et al., 2018). Subsequently, the 70 resulting components were used as spatial regressors into the resting state fMRI volumes for each subject resulting in a time course per component for each subject. These time courses were pre-whitened using a lag 5 autoregressive model. Then we calculated the sparse partial correlations between each pair of the 70 components' time course for each subject using the GraphLassoCV function from the Python toolkit scikit-learn (http://scikit-learn.org/). This resulted into a total of $\frac{70\times(70-1)}{2} = 2415$ partial correlation values per person.

5.2.7 Small vessel disease markers

White matter hyperintensity volume was derived from FLAIR. Cerebral micro bleeds and lacunar infarcts were assessed by visually inspecting FLAIR, T2-weighted, and T1-weighted images (see also Moonen et al., 2015). The features that we used for the subsequent analyses were the log-transformed white matter hyperintensity volume, and two dichotomous variables indicating absence or presence of cerebral micro bleeds and lacunar infarcts.

5.2.8 Cognitive performance measures

We measured cognitive performance at baseline and follow-up using a battery of tests (see table 5.2). Executive functioning was measured with the interference score of the abbreviated Stroop Colour Word Test (Houx et al., 1993), and the difference between the time to complete the Trial Making Test part A and B (TMT delta; Arbuthnott and Frank, 2000). Memory was measured with the immediate and delayed recall of the 15-Word Verbal Learning Test, and the Visual Association Test (Lindeboom et al., 2002).

Table 5.2: Overview of test scores at baseline and after four year follow up

Characteristic	Baseline	Follow up (4-year)
Executive Functioning		
Stroop[a]	28 (18.25 – 42)	37.5 (22.75 – 63.75)
TMT delta[b]	100 (70 – 183)	133 (71 – 182)
Memory		
WVLT Immediate recall[c]	17 (14 – 21)	15 (12 – 18)
WVLT Delayed recall[c]	4 (3 – 6.75)	3.5 (2 – 5)
Visual association test	12 (11 – 12)	11 (9 – 12)
MMSE[d]	27 (26 – 27)	28 (25.25 – 29)

Data is represented as median (Q1 – Q3).
[a] Interference time to complete (seconds).
[b] Time difference between Trial Making Test part A and B
[c] 15-Word Verbal Learning Test
[d] Mini mental state examination.

Because the scale for each test was different, the scores were first z-scored and averaged into a compound executive, and compound memory scores, and subsequently combined into a compound overall score. Follow-up scores were z-scored using the mean and standard deviation of the sub-scores at baseline, and then averaged into compound executive and memory scores, which were subsequently combined into compound overall scores. Delta cognition scores

were defined as the difference between baseline compound scores and follow up scores. Additionally, global cognitive functioning was assessed using the MMSE, and the difference between baseline and follow-up was used as an additional measure of cognitive decline. This resulted in a total of four cognitive measures, 1) compound overall, 2) compound executive functioning, 3) compound memory, 4) MMSE, for which we study the predictive value of baseline MRI data on cognitive measures at baseline and on the decline after follow-up.

5.2.9 Group lasso prediction

To make predictions about baseline scores and future cognitive decline we used a group lasso model. The group lasso fits a regression model with an L1 penalty between groups of predictors and an L2 penalty within groups. This results in a regression model that is sparse between groups of predictors, i.e., some groups are left out of the model, and dense within groups, i.e., if a group is included all predictors have a non-zero β-value. The out of sample predictive performance was determined with 5-fold cross-validation, that we repeated 10 times in order to reduce variance. The amount of shrinkage, determined by the penalty parameter λ, was chosen with 5-fold nested cross-validation within each training set. Age, gender, and randomisation factor from the original DANTE study were regressed out of the predictors for each training set, and applied to each test set.

5.2.10 Measuring predictive performance

As a measure of model fit we used *cross validated* R^2 (R^2_{cv}):

$$R^2_{\text{cv}} = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where PRESS is the predicted error sum of squares, and TSS is the total sum of squares. A R^2_{cv} of 1 represents perfect out-of-sample predictions, whereas a R^2_{cv} of 0 represents random predictions. R^2_{cv} is negative when PRESS is larger than TSS, and thus predictions have larger error than the intercept-only model.

The reported R^2_{cv} was the mean over the 10 cross-validation repetitions. For visualization purposes we selected the repetitions for which the R^2_{cv} were closest to the mean over the 10 repetitions.

To determine the significance of the results we used permutation testing. We permuted the predicted outcomes to calculate the null distribution of R^2_{cv}

for each of the four cognitive decline measures, with 5000 random permutations, including the original data order (Smyth and Phipson, 2010). For each random permutation we determined the maximum over the four cognitive measures, for baseline and follow up predictions, and compared our observed R^2_{cv} values to this distribution. This resulted in a p-value that is corrected for the eight multiple comparisons.

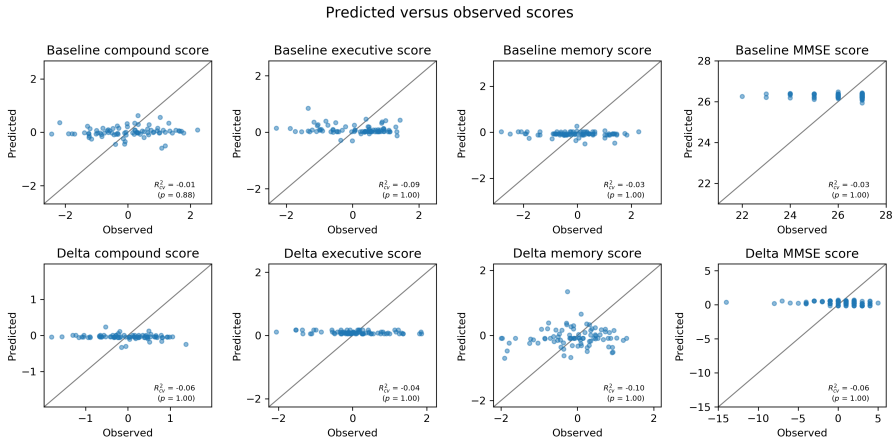

Figure 5.1: Scatterplots for predicted scores versus observed scores. Observed scores are z-scored compound scores at baseline (top), and at four year follow up (bottom). Predicted values are multimodal MRI predictions with a group lasso after cross-validation. MMSE = mini mental state examination. p-values are family-wise error corrected over the eight tests. $R^2_{\text{cv}} = 0$ represents intercept-only level performance.

5.3 Results and Discussion

See figure 5.1 for the results. We did not find any predicted cognitive decline score for baseline or follow up that was significantly above chance level. Predictions were relatively closely centered around the mean, and were not predictive for the actual scores.

In all cases, the predicted scores were slightly negatively correlated with the actual scores, resulting in a negative R^2_{cv} value. This counter-intuitive finding is the result of using cross-validation. When a test set is removed from the entire dataset the remaining training set is slightly biased in the opposite direction of the test set. This effect is negligible in the case where good predictions can

be made, but when poor predictions are made the effect becomes noticeable.

The inability to make above chance predictions is surprising, because similar methods applied to Alzheimer's disease (Schouten et al., 2016, 2017; de Vos et al., 2018), fronto-temporal dementia (Bouts et al., 2018), and mild cognitive impairment (Apostolova et al., 2014) resulted in accurate classification.

There are some limitations that could partly explain these findings. The data sample had a restricted range of cognitive scores at baseline; only participants with an MMSE between 21 and 27 were admitted to the study. These criteria were chosen so that participants were at risk for cognitive decline in the near future. However, this selection criteria may have limited the variance at baseline, which made it more difficult to construct a model that is sensitive to the range of cognitive scores.

Another limitation is the relatively long follow up period of four years. A previous study could make predictions about cognitive decline after 18 months (Woodard et al., 2010), but a four year follow up may be too far in the future to make accurate predictions. Furthermore, the long follow up time made the dropout rate relatively high; only 102 out of 205 participants had four year follow up. Additionally, it is likely that people who had serious cognitive decline during the follow up time were more likely to drop out than people who remained stable, which may have reduced the variance of cognitive decline scores.

5.4 Conclusion

In this study we used a variety of MRI measures, that have proven to successfully classify dementia, to predict future cognitive decline after a four year follow-up. We derived grey matter density and subcortical volumes from anatomical MRI, fractional anisotropy and mean diffusivity from diffusion MRI, functional connectivity from resting state functional MRI, and small vessel disease markers. These measures were combined with a group lasso regression in a machine learning approach to predict decline in overall cognition, executive functioning, and memory. We found no predictive value for the combination of MRI measures for any future cognitive decline measure. Also, cognitive scores at baseline could not be predicted from the MRI data. Overall, these results suggest that MRI measures that can successfully distinguish patients from controls, are not necessarily predictive for cognitive scores in a mildly cognitively

impaired population, or for cognitive decline after a four year follow-up.

5.5 Acknowledgements

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