Multimodal MRI-based classification of Alzheimer’s disease

Vos, F. de

Citation

Vos, F. de. (2021, December 9). *Multimodal MRI-based classification of Alzheimer’s disease*. Retrieved from https://hdl.handle.net/1887/3245855

Version: Publisher's Version

License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden

Downloaded from: https://hdl.handle.net/1887/3245855

Note: To cite this publication please use the final published version (if applicable).

Multimodal MRI-based classification of Alzheimer's disease

Frank de Vos

ISBN: 978-94-6332-799-2

Cover design, image & lay-out: Esther Beekman (www.estherontwerpt.nl)
Printed by: GVO drukkers B.V., Ede

The research described in this thesis was funded by VICI Grant 016.130.677 of the Netherlands Organisation for Scientific Research (NWO).

Printing of this thesis was supported by Alzheimer Nederland

©2021 Frank de Vos
All rights reserved. No parts of this publication may be reproduced, stored, or transmitted in any form or by any means without permission of the author. The copyrights of published articles have been transferred to the respective journals.

Multimodal MRI-based classification of Alzheimer's disease

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties

te verdedigen op donderdag 9 december 2021
klokke 16:15 uur

door
Frank de Vos
geboren te Amsterdam
in 1985

Promotoren
Prof. dr. S.A.R.B. Rombouts
Prof. dr. M.J. de Rooij

Copromotor
Dr. J. van der Grond

Promotiecommissie

Prof. dr. ir. B.P.F. Lelieveldt	
Prof. dr. K.J. Batenburg	
Prof. dr. J.C. van Swieten	Erasmus MC Rotterdam
Dr. B.M. Tijms	Amsterdam UMC, locatie VUmc

Table of contents

Chapter 1 General introduction 9

Chapter 2 Combining multiple anatomical MRI measures improves Alzheimer's disease classification 17

Chapter 3 A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease 35

Chapter 4 Pre-trained MRI-based Alzheimer's disease classification models to classify memory clinic patients 65

Chapter 5 Predicting future cognitive decline of memory clinic patients using multimodal MRI 91

Chapter 6 General discussion 109

Chapter 7 Bibliography 121

APPENDIX

Nederlandse samenvatting 140
Dankwoord 146
Curriculum vitae 148
List of publications 150

1

General introduction

1.1 Alzheimer's disease

Epidemiology

The number of people living with dementia worldwide was estimated at 50 million in 2018, and it is likely to rise to about 152 million in 2050 (World Alzheimer report 2018). Alzheimer's disease (AD) is the most common form of dementia, accounting for about two third of all dementia cases (World Alzheimer report 2018). AD is a neurodegenerative disease that is clinically characterised by memory loss and declined cognitive functioning. As the disease progresses, these behavioural characteristics get more severe, up until the point that patients are in need for 24h care. Typical AD neurodegeneration starts within the hippocampal area and later spreads out to subcortical brain regions and the medial temporal lobe (Jack et al., 1997; Pini et al., 2016). As the disease progresses further, the atrophy also extents to the frontal areas of the brain (Thompson et al., 2003).

Diagnosis

In the global action plan against dementia, the World Health Organization has specified improved diagnostics as a key area (WHO, 2017). The clinical characteristics of AD manifest relatively late in the disease, and it can only be reliably diagnosed at a stage where irreversible damage has already taken place (Jack et al., 2013). At an earlier phase, the clinical symptoms are often difficult to discriminate AD from normal ageing, or other types of dementia (Sperling et al., 2013). For this reason, these clinical symptoms cannot be used for a timely diagnosis of AD. There is need for reliable AD diagnosis at the beginning of the disease, or ideally before the onset of the disease (Frisoni et al., 2017). This would provide patients with a prognosis so they can prepare for their future trajectory. In addition, early phase AD diagnosis is important for drug research, because early phase AD patients might still be susceptible for drugs (Scheltens et al., 2016).

1.2 Biomarkers

To enable timely diagnosis of AD, there is a need for biomarkers. A number of AD biomarkers have already been proposed. Well-known biomarkers are the proteins B-amyloid and tau (Hardy & Selkoe, 2002), measured in cerebrospinal fluid (CSF) or in the brain using PET scans. The global use of these methods is still limited because of their invasive nature and the high costs associated with PET scanning. Recent innovations have enabled measurement of tau levels via plasma, thereby reducing invasiveness and costs, but this method has not yet been validated for diagnosis in diverse clinical populations (Palmqvist et al., 2020). In addition, many magnetic resonance imaging (MRI) related biomarkers have been proposed. MRI scans visualise neurodegeneration, which is a key feature of AD. Importantly, MRI scans are non-invasive, which is advantageous for day-to-day clinical use.

Anatomical MRI

An important MRI AD biomarker is grey matter atrophy (Frisoni et al., 2010). Grey matter consists of neuronal cell bodies, and neuronal loss in AD is responsible for memory loss and cognitive decline. The location and degree of grey matter atrophy can be accurately visualised with anatomical MRI scans. Anatomical MRI scans improve AD diagnosis over cognitive testing only (Liu et al., 2011), and they are included in diagnostic criteria for the most prevalent non-Alzheimer dementias as well (McKeith et al., 2005; Rascovsky et al., 2011; Roman et al., 1993), reflecting its value for differential diagnosis. Furthermore, grey matter atrophy enables early AD diagnosis, because atrophy in the temporal lobe starts before clinical symptoms occur (Jack et al., 2013). Consequently, medial temporal lobe atrophy is used as a marker for mild cognitive impairment (MCI), the clinical phase that precedes AD, and it is used to predict which MCI patients will convert to AD (Cuingnet et al., 2011; Misra et al., 2009).

Diffusion MRI

Brain damage caused by AD is not limited to grey matter atrophy. It also comprises decreased structural integrity in the white matter (Bozzali et al., 2002; Douaud et al., 2011), which can be visualised with diffusion MRI scans. White matter consists of myelinated axons that transport neuronal signal between grey matter areas. These white matter pathways make up the brain’s structural networks, and they are disrupted in AD (Mito et al., 2018). Furthermore, white matter lesions increase the risk of developing AD (Prins et al., 2004). For these reasons, diffusion MRI scans might complement anatomical MRI scans in the diagnosis of AD.

Resting state functional MRI

In addition to the structural changes, observed in both grey and white matter, AD is also characterised by changes in brain function, measured with resting state functional MRI (fMRI). These changes include decreased activation in the default mode network (Rombouts et al., 2005), altered FC between brain regions (Agosta et al., 2012; Allen et al., 2007; Binnewijzend et al., 2012), and decreased amplitude of low frequency fluctuations (Han et al., 2011), reflecting decreased intensity of spontaneous brain activity. FC may already be altered in early stages of AD, even before the presence of brain atrophy and cognitive decline (Buckner et al., 2005; Sheline and Raichle, 2013). For these reasons, resting state fMRI scans may add complementary information to structural and diffusion MRI scans, which may further improve AD diagnosis.

1.3 Individual classification of Alzheimer's disease

Average group differences vs individual classification

The MRI AD biomarker research discussed in previous paragraphs is mostly based on average group differences, as found in case-control studies. Average group differences are not necessarily useful at the individual level. Many AD biomarkers are present in healthy ageing as well (Salat et al., 1999), and show too much overlap between AD patients and healthy elderly to accurately discriminate AD at the individual level. The focus of MRI AD biomarker research has therefore shifted from the detection of average group differences towards individual classification (Rathore et al., 2017). Individual classification studies usually combine multiple biomarkers that together yield high classification accuracy.

Statistical learning

Statistical learning is a powerful framework for individual classification studies (Klöppel et al., 2008). First, it comprises statistical methods that enable the incorporation of many features into one classification model. This is essential for MRI-based classification studies, because MRI scans yield many features. Second, it incorporates cross-validation. That is, classification models are fitted on training data, and validated on a held-out data set. Cross-validation protects for overfitted models, yielding models that better generalise to out of sample data. In this way, statistical learning enables development of MRI-based classification models that can discriminate AD at the individual level, which is necessary for clinical practice.

Multimodal MRI

Thus far, most MRI-based AD classification studies have used models with only one type of feature. For example, anatomical MRI scans were used to measure only grey matter density features (Beheshti et al., 2016) or only hippocampal shape features (Gerardin et al., 2009). However, these different types of anatomical MRI features possess complementary information, and combining them increases AD classification accuracy (Bron et al., 2015; Wolz et al., 2011; Westman et al., 2013). Moreover, different types of MRI scans contain complementary information as well, and combining these into a multimodal MRI model improves AD classification accuracy even further (Schouten et al., 2016). In this thesis we will derive multiple types of feature from a single MRI scan, and we will combine multiple MRI modalities.

1.4 Aims and outline

The overall aim of this thesis is to develop and evaluate MRI-based models for individual AD classification. These models take MRI features as input, and yield AD probability scores as output. We will develop these models for a wide variety of MRI features, and we will compare those models on AD classification accuracy. This provides information on which MRI features are most informative for AD classification. Furthermore, we will study combinations of features to try to improve accuracy. To this end, we will use statistical learning techniques, because they enable incorporation of many features, and they focus on optimising classification accuracy. In **chapter two,** anatomical MRI scans are used to calculate multiple structural features that are thought to be informative for AD. These features will be compared on AD classification accuracy, and they will be combined into a single model. It is hypothesised that the combination outperforms the separate features. In **chapter three,** a similar approach will be used for resting state fMRI scans. We will use several approaches to calculate FC, and we will derive indirect FC measures, like FC dynamics and graph measures. Again, these features are compared on AD classification accuracy, and they will be combined into a single model. Also here, it is hypothesised that the combination outperforms the separate features.

In order to be clinically useful, AD classification models should generalise well to clinical populations. This is complicated for two reasons. First, research samples that are used to develop the models are often homogeneous, only containing clinically diagnosed AD patients and healthy elderly controls. In contrast, clinical populations are more diverse. They include patients with varying stages of AD progression, as well as MCI patients and preclinical patients that experience memory complaints. In addition, clinical populations include patients with non-AD dementia types as well. Second, MRI scans suffer from technical between-scanner variation (Ewers et al., 2006; Takao et al., 2014; Zhu et al., 2011). An AD classification model that is trained with MRI data from one scanner, is not necessarily useful for MRI data from another scanner. To be clinically useful, AD classification models should be applicable to diverse patient populations, and they should be robust to between-scanner variation. In **chapter four** we will evaluate whether MRI-based models for individual AD classification also discriminate AD in a diverse clinical population. To this end we will train AD classification models on a single-centre data set consisting of AD patients and controls, using features derived from both anatomical MRI, diffusion MRI and resting state fMRI scans. Next, we will use these models to assign AD scores to patients from a multi-centre memory clinic data set including AD patients, MCI patients and patients with subjective memory complaints. We expect that AD patients will receive higher AD scores than MCI patients and patients with subjective memory complaints.

In **chapter five** we will evaluate MRI's predictive accuracy for future cognitive decline. To this end we use the baseline multimodal MRI scans of the memory clinic data set outlined above, to predict two-year follow-up cognitive decline.

2

Combining multiple anatomical MRI measures improves Alzheimer's disease classification

Published in Human Brain Mapping, 2016; 37(5), 1920-1929

Frank de Vos, Tijn M. Schouten, Anne Hafkemeijer, Elise G. P. Dopper,
John C. van Swieten, Mark de Rooij, Jeroen van der Grond
& Serge A. R. B. Rombouts

Abstract

Several anatomical MRI markers for Alzheimer's disease (AD) have been identified. Hippocampal volume, cortical thickness and grey matter density have been used successfully to discriminate AD patients from controls. These anatomical MRI measures have so far mainly been used separately. The full potential of anatomical MRI scans for AD diagnosis might thus not yet have been used optimally. In the current study, we therefore combined multiple anatomical MRI measures in order to improve diagnostic classification of AD. For 21 clinically diagnosed AD patients and 21 cognitively normal controls we calculated i) cortical thickness, ii) cortical area, iii) cortical curvature, iv) grey matter density, v) subcortical volumes and vi) hippocampal shape. These six measures were used separately and combined as predictors in an elastic net logistic regression. We made receiver operating characteristic curves and calculated the area under the curve (AUC) to determine classification performance. AUC values for the single measures ranged from 0.67 (cortical thickness) to 0.94 (grey matter density). The combination of all six measures resulted in an AUC of 0.98. Our results demonstrate that the different anatomical MRI measures contain complementary information. A combination of these measures may therefore improve accuracy of AD diagnosis in clinical practice.

Keywords: Alzheimer's disease, Anatomical MRI, Cortical thickness, Cortical area, Cortical curvature, Grey matter density, Subcortical volumes, Hippocampal shape, Classification

2.1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder that is characterised by both focal and global grey matter atrophy. Hippocampal atrophy is considered to be the hallmark of AD and it is therefore used as a clinical marker (Morra et al., 2009). Grey matter atrophy in AD patients frequently extends to other brain regions, including subcortical structures and the medial temporal lobe (Seeley et al., 2009; Rombouts et al., 2000; Jack et al., 2004). The location and degree of grey matter atrophy can be accurately visualised with anatomical magnetic resonance imaging (MRI) scans, which is used for the clinical diagnosis of AD (Frisoni et al., 2010). Studies using anatomical MRI scans have showed that AD patients have decreased volumes and altered shapes of the hippocampi compared to cognitively normal elderly (Thompson et al., 2004; Scher et al., 2007) and also the volumes of the putamen and thalamus are reduced in AD patients (de Jong et al., 2008). Moreover, reduced cortical thickness (Lerch et al., 2005) as well as widespread grey matter atrophy have been demonstrated in patients with AD compared with controls (Karas et al., 2003).

However, the group differences found in case control studies are not necessarily useful in a clinical setting. Many AD markers have also been observed in healthy ageing (Salat et al., 1999). AD markers are only helpful in a clinical setting if they can accurately discriminate AD patients from non-affected subjects at the individual level. The focus of research on anatomical MRI biomarkers for AD has therefore shifted from the detection of group differences towards disease classification. Cortical thickness, MRI-based morphometry (VBM) and the volume and shape of the hippocampus have been used to discriminate AD patients from controls with moderate to high classification accuracy. (Cuingnet et al., 2010; Davatzikos et al., 2011; Querbes et al., 2009).

Thus far, these anatomical MRI measures have mainly been used separately to classify AD patients. However, the different measures may possess complementary information, and the combination of these measures could therefore increase AD classification accuracy compared to the separate measures. For example, voxel-based cortical thickness (VBCT) and VBM show different aspects of age-associated decline in grey matter (Hutton et al., 2009). Also, VBM, cortical folding and cortical thickness complement each other in showing neurodegenerative changes related to Parkinson's disease (Pereira et al., 2012). Moreover, the combination of different anatomical MRI measures improves AD classification (Wolz et al., 2011; Westman et al., 2013; Bron et al., 2015).

Furthermore, advances in statistical learning have facilitated the integration of different sources of information into a single predictive model (Zou & Hastie, 2005). This enables the incorporation of many predictors into one predictive model by selecting only the relevant information out of these many predictors. These techniques have already been applied in order to separate AD patients from controls or to separate MCI converters from MCI non-converters. (Cui et al., 2011; Dyrba et al., 2015a; Dyrba et al.; 2015b; Schouten et al., 2016; Trzepacz et al., 2014; Teipel et al., 2015; Wee et al., 2013; Zhang et al., 2013). It would therefore make sense to combine all anatomical MRI measures that have been shown to be discriminative for AD into a single model to improve sensitivity and specificity.

In this study we will use anatomical MRI scans from a group of AD patients and a group of cognitively normal controls and calculate several commonly used measures that are informative for AD. These measures are: i) cortical thickness, ii) cortical surface area, iii) cortical curvature, iv) grey matter density, v) the volume of the subcortical structures and vi) the shape of the hippocampus. We will combine all these measures into a single predictive model and calculate its classification performance. We hypothesise that the combination will outperform the separate measures.

2.2 Methods

Participants

Anatomical MRI scans were obtained from 21 probable AD patients (10 females) between ages 50 and 87 (M = 71.7, SD = 9.3) and 21 cognitively healthy controls (10 females) between ages 57 and 80 (M = 68.0, SD = 7.5). The AD patients had an average score on the mini-mental state examination (MMSE) of 23 (SD = 2.4) and the cognitively healthy controls had an average score of 28 (SD = 1.5). All AD patients underwent a standardised dementia screening that included their medical history, informant-based history, physical and neurological examination, and an extensive neuropsychological assessment including the MMSE. Diagnoses were made in a multidisciplinary consensus meeting according to the core clinical criteria of the National Institute on Aging and the Alzheimer's Association workgroup for probable AD (Mckhann et al., 1984; McKhann, 2011). As control subjects, we included cognitively healthy elderly volunteers. This study was approved by the medical ethical committee of the Leiden University Medical Center, and all participants provided written informed consent.

MR image acquisition

All participants were scanned on a Philips 3 Tesla Achieva MRI scanner in the Leiden University Medical Center. Three-dimensional T1-weighted structural scans were acquired with the following parameters: TR = 9.8 ms, TE = 4.6 ms, flip angle = 8, 140 slices, voxel size = 0.88 x 0.88 x 1.20 mm.

Cortical thickness, area, and curvature

To calculate cortical thickness, cortical area, and cortical curvature we processed the T1-weighted images using the Freesurfer software package version 5.3.0 (Dale et al., 1999; Fisch et al., 1999). First, intensity normalisation was applied to the brain-extracted image to create an image with relatively high contrast to noise ratio. This image was used to locate the boundary between grey and white matter. A triangular mesh was then constructed around the white matter surface. This triangular mesh consists of over 160,000 vertices for each hemisphere. To create the grey matter surface, the mesh was deformed outwards so that it closely followed the boundary between grey matter and cerebrospinal fluid (CSF). Cortical thickness was calculated as the distance between the white matter surface and the grey matter surface for each vertex. The image was then registered to the Freesurfer common template, using the image's cortical folding pattern. The neocortex was parcellated into the 68 neocortical regions (34 regions for each hemisphere) of the Desikan-Killiany atlas (Desikan et al., 2006). The thickness of each parcellation unit was calculated as the mean thickness of all the vertices within that parcellation. This yielded 68 cortical thickness features per subject. To calculate cortical surface area we summed the areas of the grey matter mesh triangles for each parcellation, which yielded 68 cortical area features per subject. Cortical curvature was calculated as the mean of the curvature values in the two principal directions of the surface. The curvature of a vertex in these directions was calculated as the inverse of the length of the radius of the osculating circles in these directions (Ronan et al., 2011). For each of the parcellations we averaged the curvature values of the vertices, which yielded 68 cortical curvature features per subject.

Grey matter density of the cortical structures

We calculated grey matter density using VBM in the FMRIB Software Library (FSL version 5.0.7; Ashburner et al., 2000; Smith et al., 2004). First, we segmented the brain-extracted images into grey matter, white matter, and cerebrospinal fluid. Next, we created a study specific grey matter template in two steps. In a first run we affine-registered the grey matter images to the ICBM-152 grey matter template and we averaged the resulting images to create a first-pass template. In a second run we non-linearly registered the grey matter images to the first-pass template and we averaged these images to obtain the final template at 2x2x2 mm^3 resolution in standard space. Finally, we non-linearly registered the grey matter images to the final template and smoothed these images with a Gaussian kernel with sigma = 3 mm. The voxel values in these images range between 0 and 1, representing the percentage of a voxel being grey matter tissue. We averaged the voxel wise values within the 48 regions of the probabilistic Harvard-Oxford cortical atlas. We calculated the weighted averages of the regions, with voxels contributing to the average of a region based on their probability of being part of that region. This yielded 48 grey matter density values per subject.

Subcortical volumes

We calculated the volumes of the subcortical structures using the FMRIB's Integrated Registration and Segmentation Tool (FIRST) in FSL (Patenaude et al., 2011). First, the whole-head images were affine registered to the non-linear MNI-152 template. In a second stage, initialised by the result of the first stage, we used a subcortical mask to achieve a more accurate and robust affine registration. The shapes of the subcortical structures were then modelled by deformable meshes and the boundary voxels were classified as being part of the subcortical structure using structural segmentation (Zhang et al., 2001). Finally, we corrected the subcortical volumes for intracranial volume as obtained by FSL. This yielded 14 subcortical volume features per subject (thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens for both hemispheres).

Hippocampal shape

To calculate hippocampal shape, we used the vertex analysis in FSL (Patenaude et al., 2011). The shape values represent the distance of a vertex on the hippocampal mesh of a specific subject to the mean location of that vertex within the whole sample. A negative value represents a decrease of the size of the hippocampus on that specific location for a subject relative to the mean sample. Vice versa, a positive value represents a relative increase of the size of the hippocampus on that location. In the absence of a sufficiently detailed brain atlas of the human hippocampus we used a data-driven method to reduce the number of features. We ran a principal component analysis on the vertex shape values of both hippocampi and extracted only the first ten components, because these alone explained 86 percent of the variance in hippocampal shape values. This yielded ten hippocampal shape features per subject.

Reference measures: whole brain atrophy and hippocampal volume

To provide a reference for the classification performance of the anatomical MRI measures, we also calculated two simple measures that are commonly used for clinical diagnosis of AD, whole brain atrophy and hippocampal volume (Frisoni et al., 2010). We used Freesurfer to calculate the ratio of total brain volume to intracranial volume as a measure of whole brain atrophy. We used FSL FIRST to calculate the volumes of the left and the right hippocampus.

Statistical analyses

The features of the six anatomical MRI measures were used in an elastic net logistic regression to classify the subjects as either AD or control. Elastic net regression uses penalties to hinder the features from entering the regression model (Zou & Hastie, 2005; Friedman et al., 2012). Thus, only the most relevant predictors will enter the regression model, which is helpful if the number of features outnumbers the number of subjects. Elastic net regression uses a combination of an L1 (LASSO) (Tibshirani, 1996) and L2 (Ridge) (Hoerl and Kennard, 1970) penalty. Therefore, two hyperparameters should be set: the α parameter determines the relative weight of the two different penalties and λ determines the size of those penalties. Elastic net logistic regression has been used for AD classification by Schouten et al. (2016), Trzepacz et al. (2014) and Teipel et al. (2015).

We used cross-validation to ensure that we are not overfitting the prediction models. In our case there are two potential sources of overfitting. We could either include too many predictors in our logistic regression model or we could overestimate the classification accuracy by looping over all the values of the hyperparameters and only pick the best result. To ascertain that we are not subject to any of these two sources of overfitting we used a nested cross-validation approach (Krstajic et al., 2014). We used the inner loop of the nested cross-validation to fit the logistic regression model and the outer loop to tune the hyperparameters. For both the inner and outer loop we used 10-fold cross-validation, thus using 90 percent of the subjects in the training set and 10 percent in the test set, and repeating this 10 times such that all subjects were part of the test set once.

We plotted receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) as a measure of classification accuracy. We repeated the cross-validation procedure 50 times to get a more reliable cross-validation error (Krstajic et al., 2014). We extracted the median AUC value instead of the mean, because we expect that the distribution of AUC values is skewed to the left due to a ceiling effect.

First, we calculated the classification accuracy for each measure separately. Then we calculated the classification accuracy for the combination of all measures and for all pairs of two measures. We used two different methods to combine the features of different measures. In the first method we concatenated the features and we used the concatenated feature set for classification similarly as the single measures. Concatenation is commonly used to combine different sets of features for prediction (De Magalhães Oliveira et al., 2010; Westman et al., 2012). It might however not be the most optimal method for combination, because all sets of features are then weighted equally whereas some sets might be more predictive than others. Therefore, in the second method we used a weighted average of the single measure predictions in order to increase classification performance (Wolpert, 1992; Breiman, 1994). Measures that achieved a higher accuracy were given a larger weight. We used the inverse of the binomial deviance as a measure of accuracy. Binomial deviance is a continuous measure of prediction error and is therefore sensitive for small accuracy differences. We determined the measures' weights for each subject individually. To avoid overfitting we used the binomial deviances within the training set to determine the contributions of the measures for the left out subjects. The weighted average was calculated for all the 50 cross-validation repetitions and the median result was presented.

2.3 Results

Figure 1 shows ROC curves for the classification of AD vs. cognitively normal controls for the six anatomical MRI measures separately, and for the combination of all six measures. The accompanying AUCs, representing a measure of classification accuracy, are shown in Table 1. Table 1 also presents the AUC values for the two reference measures: hippocampal volume and whole brain atrophy. The two reference measures perform reasonably well. Especially, the hippocampal volumes can discriminate well between AD patients and controls. Yet, the grey matter density of the cortical structures and the volumes of subcortical structures discriminate better than the reference measures. Cortical thickness, cortical area, cortical curvature, and hippocampal shape cannot improve over the reference measures. Most importantly however, the combination of all six measures outperforms the separate measures. The weighted combination of all measures discriminates somewhat better than the concatenated combination

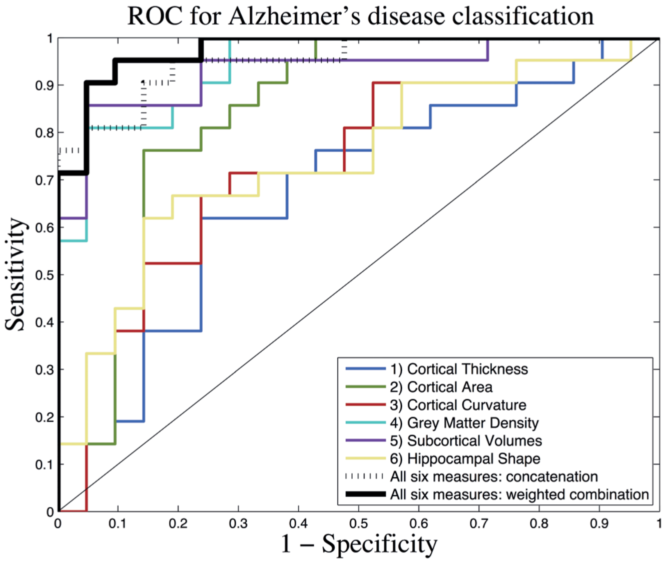


Figure 1. ROC curves for discriminating AD patients from cognitively healthy controls. ROC curves are plotted for the six anatomical measures separately and for the two types of combinations of all six measures.

Table 1. AUC values discriminating AD patients from cognitively healthy controls for the different anatomical MRI measures. Hippocampal volumes and whole brain atrophy are added for comparison, because these are fairly simple measures that are commonly used by clinicians to diagnose AD. The other six measures are used separately and combined to discriminate the two groups. We used two methods for the combination. Both of them outperform the separate measures, and the weighted combination works best.

Anatomical MRI measures	AUC
Reference: hippocampal volumes	0.87
Reference: whole brain atrophy	0.77
1) Cortical thickness	0.67
2) Cortical area	0.85
3) Cortical curvature	0.73
4) Grey matter density	0.94
5) Subcortical volumes	0.93
6) Hippocampal shape	0.74
All six measures: concatenation	0.95
All six measures: weighted combination	0.98

To further investigate the additive value of combining different measures, we calculated the AUCs for all pairs of two measures, using both measure concatenation (Figure 2, left side) and weighed combinations (Figure 2, right side). The tall bars represent the single measure AUCs, like in Table 1. The short bars represent the additive value of a second measure. Note that the additive value can also be negative, when the addition of a second measure worsens the classification accuracy. Both for concatenation and weighted combinations, the AUC of a single measure often improves when a second measure is added. When using concatenation, the highest AUC values are obtained by combining grey matter density with either subcortical volumes or cortical thickness (AUC = 0.98), which is even higher than the concatenation of all six modalities. Using weighted combinations, the highest AUC is obtained by combining grey matter density with subcortical volumes (AUC = 0.98), which is equal to the weighted combination of all six modalities.

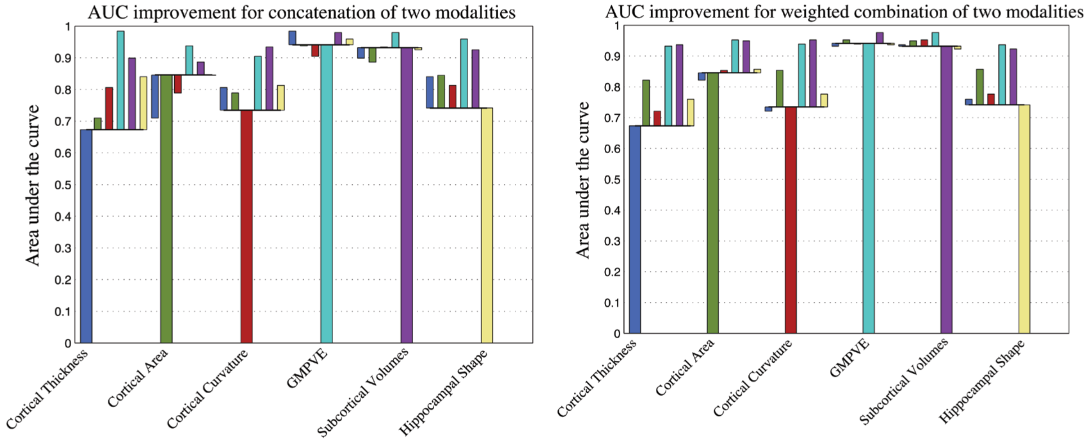


Figure 2. AUC values for all the possible combinations of two measures using feature concatenation (left) and weighted combinations (right). The tall bars represent the single measure AUCs, which are the same as those in Table 1. The short bars represent the additive value of a second measure. The additive values are mostly positive. For example, when cortical area is concatenated with cortical thickness, the AUC increases from 0.67 to 0.71. However, sometimes the additive value a second measure is negative. For example, cortical area is on its self a better predictor (0.85) than concatenated with cortical thickness (0.71).

These statistical models are primarily meant to predict class membership, and not to make claims about which features were most important to distinguish AD patients from controls. If a statistical model is built for prediction, rather than for explanation, one should be careful when interpreting the explanatory part of that model (Shmueli, 2010). We elaborate more on this in the discussion section. Yet, to illustrate the content of the classification models we show the standardised beta values (averaged over the 50 cross-validation repetitions) for the predictors of the grey matter density measure (Figure 3) and the subcortical volumes measure (Figure 4). We used these two measures for illustration, because they discriminate best between AD patients and cognitively normal controls. For regions with a negative beta value, low grey matter density values or low subcortical volumes increase the odds for AD. For regions with a positive beta value, high grey matter density values or high subcortical volumes increase the odds for AD. This might seem contradictionary, because we do not expect to see increased grey matter density values or increased subcortical volumes in AD patients. The interpretation of these effects could be something like: if that region is relatively unaffected (high grey matter density, or large subcortical volume) while some other regions are more affected, this is evidence in favour of AD. We plotted the regions in colour coding for the grey matter density measure (Figure 5) and the subcortical volumes measure (Figure 6).

The grey matter density classification model was mostly driven by decreased grey matter density within the cortical areas in the medial temporal lobes, and to a lesser extent in the occipital and frontal lobes. The regions with large weights include the orbitofrontal cortex, the subcallosal cortex, the insular cortex, and the inferior temporal gyrus. Most regions contribute to the classification model to a certain extent, suggesting that a global pattern of atrophy is predictive for AD. The subcortical volumes classification model was mostly driven by decreased sizes of the hippocampus, putamen and thalamus, and to a lesser extent by decreased sizes of the accumbens and the amygdala.

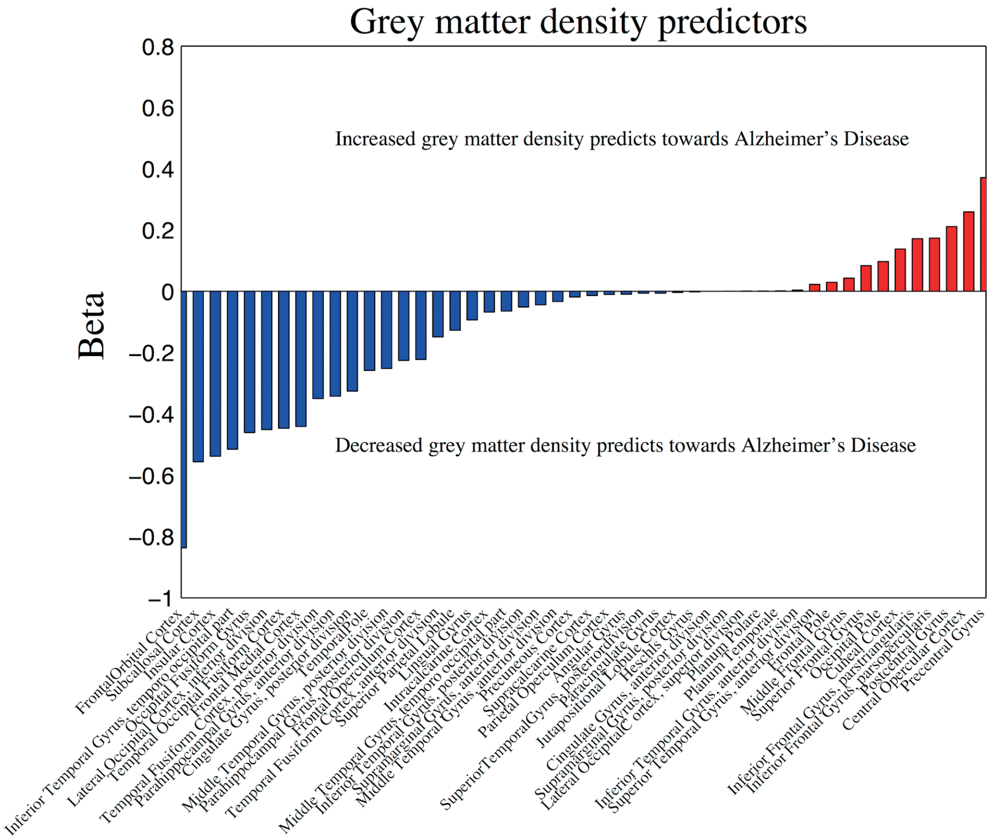


Figure 3. Beta values for the cortical grey matter density estimates. The beta values represent the mean beta values over all the cross-validation folds of all the cross-validation repetitions. The beta values are ordered according to size. Negative beta values are coloured blue, and positive beta values are coloured red. Most beta values are negative, and the meaning of those is that low grey matter density predicts toward AD. Some beta values are positive, which is counterintuitive, but could mean something like: if that region is relatively unaffected (high grey matter density) while some other regions are more affected, this is evidence in favour of AD. Note that grey matter density was only calculated for the cortical regions. The results for the subcortical volumes are presented in Figure 4.

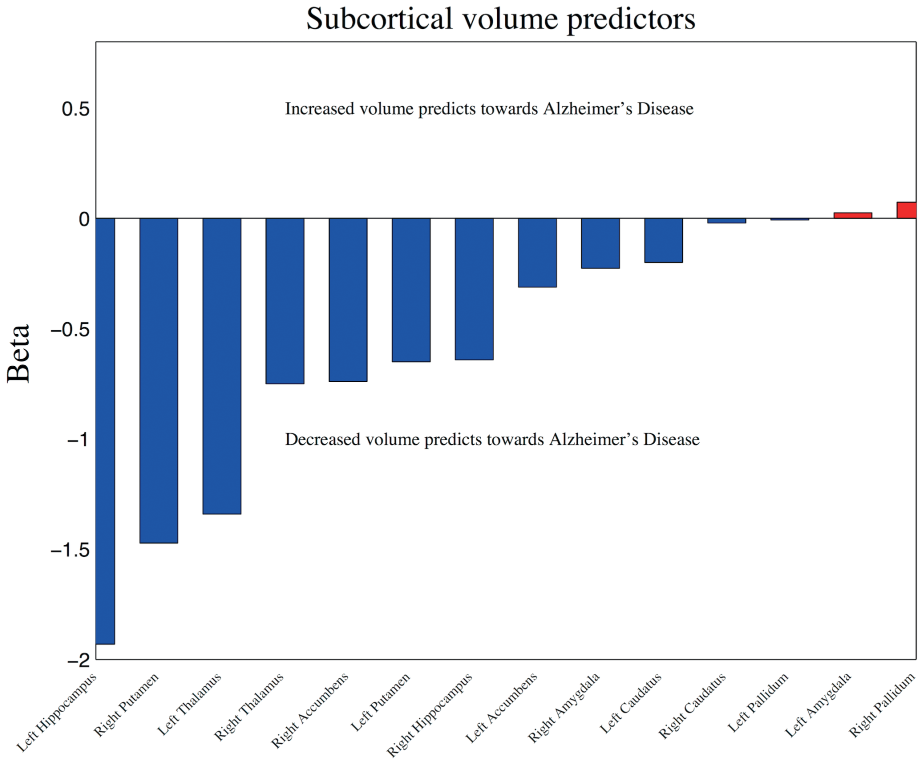


Figure 4. Beta values for the subcortical volumes. The beta values represent the mean beta values over all the cross-validation folds of all the cross-validation repetitions. The beta values are ordered according to size. Negative beta values are coloured blue, and positive beta values are coloured red. Most beta values are negative, and the meaning of those is that a small volume predicts toward AD. Some beta values are positive, which is counterintuitive, but could mean something like: if that region is relatively unaffected (large volume) while some other regions are more affected, this is evidence in favour of AD.

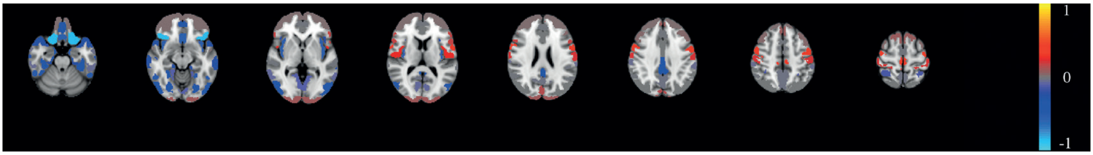


Figure 5. The beta values from Figure 4 are presented here in colour coding. The 'cool' regions correspond with the blue bars and the meaning of those is that low grey matter density predicts toward AD. The 'hot' regions correspond with the red bars and the meaning of those is that high grey matter density predicts toward AD. This is counterintuitive, but could mean something like: if that region is relatively unaffected (high grey matter density) while some other regions are more affected, this is evidence in favour of AD.

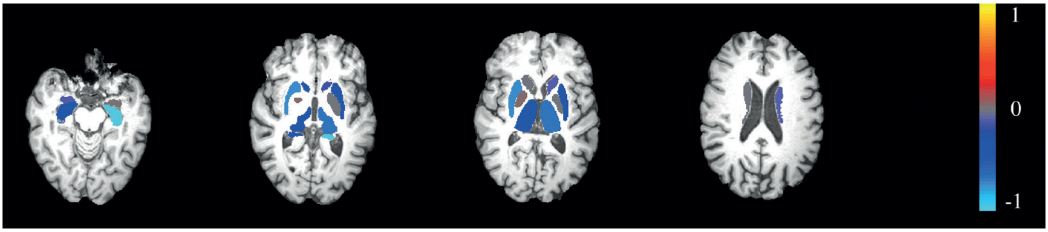


Figure 6. The beta values from Figure 5 are presented here in colour coding. The 'cool' regions correspond with the blue bars and the meaning of those is that a small volume predicts toward AD. The 'hot' regions correspond with the red bars and the meaning of those is that a large volume predicts toward AD. This is counterintuitive, but could mean something like: if that region is relatively unaffected (large volume) while some other regions are more affected, this is evidence in favour of AD.

2.4 Discussion

In this study we used different anatomical MRI measures to separate AD patients from controls. Our main finding is that the combination of the different anatomical MRI measures improves AD classification accuracy over each single measure. The combination of different anatomical MRI measures thus captures more information on grey matter loss than each of the measures separately, and AD classification benefits from this extra information through an automated classification algorithm.

When used separately, all measures were sensitive to the detection of AD. We replicated early findings in which grey matter density values (Cuingnet et al., 2010) and the volumes of the subcortical structures (De Magalhães Oliveira et al., 2010) are highly discriminative for AD. However, the high classification accuracy of cortical surface area compared to cortical thickness is relatively surprising. Previous studies found that cortical grey matter atrophy is primarily reflected in cortical thinning rather than in a decrease of cortical area (Dickerson et al., 2009; Westman et al., 2012).

We have used two different methods to combine the measures. The weighted combination of the measures showed higher accuracy than the concatenation of the measures. Due to a relatively small sample size, it is unclear whether this difference will generalise to other data sets, but it does demonstrate that the method that is used for the combination of the information can influence diagnostic accuracy.

The classification model for the grey matter density estimates (which were only calculated for the cortical regions) was mostly driven by decreased grey matter density within the medial temporal lobes, and to a lesser extent in the occipital and frontal lobes. These findings are in line with other observations of AD atrophy (Karas et al., 2004; Frisoni et al., 2010; Risacher et al., 2010). The subcortical volumes classification model was mostly driven by decreased sizes of the hippocampus, putamen and thalamus, which is also in accordance to previous findings (Leung et al., 2010; de Jong et al., 2008).

It should be noted that we have built predictive models, without the explicit goal to make explanatory models. We should therefore be careful when interpreting the explanatory part of our models (Shmueli, 2010). For example, we have forced sparsity in our models, which hinders potentially relevant features from entering the model. Furthermore, the effect of a feature is conditional on the effect of all the other features in the model, and can therefore not be interpreted independently. Also, multicollinearity is not an issue for prediction models, but it can be problematic for the explanatory part of a model. These constraints should thus be taken into consideration when interpreting the content of the prediction models. For these reasons the effects do not denote one to one relationships between the feature and the probability of being classified as an AD patient, nor do they reflect mean differences between the groups.

We did not use non-linear effects or interaction-effects in our classification algorithm, which improves the reproducibility of our results. More complex classification models might further enhance the classification accuracy. For example, longitudinal AD studies have found that atrophy in some areas in the brain follows a non-linear trend (Chan et al., 2003; Fotenos et al., 2005). On the contrary, more complex classification algorithms can cause overfitting, resulting in a poorer classification performance (Hastie et al., 2009). Neither did we use prior feature selection, because feature selection generally does not improve AD classification results (Chu et al., 2012) and avoiding this extra step eases the reproducibility of our results.

We have used the most up to date versions of Freesurfer and FSL to calculate the structural measures. Freesurfer and FSL have both been validated for the analysis of anatomical MRI scans for both healthy subjects and AD patients. Freesurfer is sensitive to detect cortical thinning in AD patients compared to controls (Redolfi et al., 2015). and it has good scan-rescan reproducibility (Tustison et al., 2014). The results of Freesurfer analyses do however differ between different Freesurfer versions (Groenschild et al., 2012), but the most recent Freesurfer versions correspond better with the results of manual outlining (Clerx et al., 2015). The shape model that is used by FSL First to segment the subcortical structures has been trained on both healthy subjects and pathological brains (including AD patients) to reduce bias towards healthy brains (Patenaude et al., 2011). Furthermore, FSL First calculation of the subcortical volumes is roughly similar to the results of manual outlining for AD patients, MCI patients and controls from the ADNI data set (Mulder et al., 2014). FSL voxel-based morphometry (VBM) is relatively accurate because it makes use of a non-linear registration tool (Callaert et al., 2014). Furthermore, FSL VBM accurately detects hippocampal atrophy, but is biased towards detecting medial temporal lobe atrophy in AD patients (Diaz-de Grenu et al., 2014).

In conclusion, we demonstrated that the combination of i) cortical thickness, ii) cortical area, iii) cortical curvature, iv) grey matter density, v) subcortical volumes and vi) hippocampal shape improves AD diagnosis. The added value of combining different anatomical MRI measures should be considered in AD scanning protocols. It is still common practice to only use the size of the hippocampus or a single measure of whole brain atrophy for AD diagnosis. Our results demonstrate that clinical AD diagnosis could benefit from calculating multiple measures from an anatomical MRI scan and incorporate these all in an automated analysis. Our results further suggest that the grey matter density of the cortical structures and the volumes of the subcortical structures are sufficient for optimal AD classification based on an anatomical MRI scan. These results might also be relevant to studies of early AD diagnosis and other neurodegenerative diseases studies.

2.5 Acknowledgments

This study is supported by grant no. 016130677 of the Netherlands Organisation for Scientific Research (NWO). A. Hafkemeijer is sponsored by NWO grant no. 05613010. E. G. Dopper is sponsored by grants from the Dioraphte Foundation grant no. 09-02-03-00, the association for Frontotemporal Dementia research and NWO grant no. 05613018.

3

A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease

Published in NeuroImage, 2018; 167, 62-72.

Frank de Vos, Marisa Koini, Tijn M. Schouten, Stephan Seiler,
Jeroen van der Grond, Anita Lechner, Reinhold Schmidt, Mark de Rooij,
Serge A. R. B. Rombouts

Abstract

Alzheimer's disease (AD) patients show altered patterns of FC (FC) on resting state functional magnetic resonance imaging (fMRI) scans. It is yet unclear which resting state fMRI measures are most informative for the individual classification of AD patients. We investigated this using resting state fMRI scans from 77 AD patients (MMSE = 20.4 ± 4.5) and 173 controls (MMSE = 27.5 ± 1.8). We calculated i) FC matrices between resting state components as obtained with independent component analysis (ICA), ii) the dynamics of these FC matrices using a sliding window approach, iii) the graph properties (e.g., connection degree, and clustering coefficient) of the FC matrices, and iv) we distinguished five FC states and administered how long each subject resided in each of these five states. Furthermore, for each voxel we calculated v) FC with 10 resting state networks using dual regression, vi) FC with the hippocampus, vii) eigenvector centrality, and viii) the amplitude of low frequency fluctuations (ALFF). These eight measures were used separately as predictors in an elastic net logistic regression, and combined in a group lasso logistic regression model. We calculated the area under the receiver operating characteristic curves (AUC) to determine classification performance. The AUC values ranged between 0.51 and 0.84 and the highest were found for the FC matrices (0.82), FC dynamics (0.84) and ALFF (0.82). The combination of all measures resulted in an AUC of 0.85. We show that it is possible to obtain moderate to good AD classification using resting state fMRI scans. FC matrices, FC dynamics and ALFF are most discriminative and the combination of all the resting state fMRI measures improves classification accuracy slightly.

Keywords: resting state fMRI, Alzheimer's disease, classification, independent component analysis, dual regression, dynamic functional connectivity

3.1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterised by widespread grey matter atrophy (Jack et al., 2004), specifically hippocampal atrophy is considered to be the hallmark of AD (Morra et al., 2009). In order to develop a cure, or to slow down the disease progression, it is essential to diagnose AD in an early stage (Prince et al., 2011).

AD patients differ in their pattern of functional connectivity (FC) as shown by resting state functional magnetic resonance imaging (fMRI) scans. They have decreased FC between the hippocampus and several regions throughout the neocortex (Allen et al., 2007; Wang et al., 2006), reduced FC within the default mode network (Binnewijzend et al., 2012; Greicus et al., 2004), and increased FC within the frontal networks (Agosta et al., 2012). AD patients also have different large-scale FC matrices (Brier et al., 2012) and graph properties derived from these matrices (Sanz-Arigita et al., 2010; Supekar et al., 2008). In addition, AD patients differ in the dynamics of their FC and their dwell time in specific FC states (Jones et al., 2012). Furthermore, AD patients have less signal in the low frequency domain (0 - 0.1 Hz) of their resting state signal (Han et al., 2011).

These FC differences might exist in an early stage of AD, even before the presence of brain atrophy and cognitive decline (Buckner et al., 2005; Sheline and Raichle, 2013). For instance, cognitively normal elderly with increased amyloid binding, an important AD indicator, have decreased FC between the precuneus and several regions within the default mode network, and these effects are similar to those observed in AD patients (Sheline et al., 2010a). Carriers of the APOE ε4 gene, who are at genetic risk for AD, have reduced FC between the precuneus and the hippocampus (Sheline, et al., 2010b), and increased FC within the default mode network (Filippini et al., 2009).

Resting state fMRI might be used for the diagnosis or even early detection of AD and it is important to investigate this potential (Buckner et al., 2005; Sperling, 2011). AD biomarkers can be evaluated using individual classification studies. Resting state fMRI-based AD classification studies have progressed through the use of machine learning techniques. Machine learning techniques enable the incorporation of many predictors into one predictive model and they automatically select the relevant ones. So far, AD has been classified moderately to good using FC matrices (Challis et al., 2015; Chen et al., 2011; Schouten et al., 2016) and their graph properties (Khazaee et al., 2015), FC dynamics (Wee et al., 2016), FC within the default mode network (Koch et al., 2012) and the amplitude of low frequency fluctuations (ALFF; Dai et al., 2012).

It is not known which of these resting state fMRI measures is best for AD classification. Moreover, the combination of different resting state fMRI measures might improve AD classification (Dai et al., 2012; de Vos et al., 2016; Mesrob et al., 2012; Schouten et al., 2016; Sui et al., 2013). In this study, we will use a wide range of resting state fMRI measures in combination with machine learning techniques to classify AD patients and controls. These measures include FC with several resting state networks (RSNs), FC with the hippocampus, FC matrices and their graph properties, FC dynamics, FC states, and the ALFF within the resting state signal. We will determine the most accurately predicting measures and combine them to investigate whether this increases the classification accuracy.

3.2 Materials and methods

Participants

Our dataset consisted of 77 clinically diagnosed probable AD patients and 173 cognitively normal elderly controls (see Table 1). The AD patients were scanned at the Medical University of Graz as a part of the prospective registry on dementia (PRODEM; see also Seiler et al., 2012). The inclusion criteria for PRODEM are: dementia diagnosis according to DSM-IV criteria (American Psychiatric Association, 2000), non-institutionalisation or need for 24-hour care, and the availability of a caregiver who agrees to provide information on the patients' and his or her own condition. Patients were excluded if they were unable to sign an informed consent or if co-morbidities were likely to preclude termination of the study. We used the baseline scans from the PRODEM study, and only included patients that were diagnosed with AD in line with the NINCDS-ADRDA Criteria (McKhann et al., 1984), and for which anatomical MRI and resting state fMRI scans were available. The controls were scanned at the same scanning site, over the same time period, with the same scanning protocol as a part of the Austrian stroke prevention study. The Austrian Stroke Prevention Study is a community-based cohort study on the effects of vascular risk factors on brain structure and function in elderly participants without a history or signs of stroke and dementia on the inhabitants of Graz, Austria (see also Schouten et al., 2016).

MR acquisition

All participants were scanned on a Siemens Magnetom TrioTim 3T MRI scanner. The anatomical T1-weighted images were acquired with the following parameters: TR = 1900 ms, TE = 2.19 ms, flip angle = 9°, and an isotropic voxel size of 1 mm. The resting state fMRI session was conducted, acquiring 150 volumes with TR = 3000 ms, TE = 30 ms, flip angle = 90°, 40 axial slices, with an isotropic voxel size of 3 mm. The participants were instructed to lie still with their eyes closed, and to stay awake.

MRI preprocessing

The MRI data were preprocessed using the FMRIB Software Library (FSL, version 5.0) (Jenkinson et al., 2012; Smith et al., 2004). For the anatomical MRI this included brain extraction, bias field correction, and non-linear registration to standard MNI152 template (Grabner et al., 2006). For the resting state fMRI data this included brain extraction, motion correction, a temporal high pass filter with a cut-off point of 100 seconds, and spatial smoothing. The mean framewise displacement as calculated by MCFLIRT from FSL (Jenkinson et al., 2002) ranges from 0.02 to 0.42 mm (mean = 0.10, SD = 0.06) for the control subjects, and from 0.03 to 0.55 mm (mean = 0.13, SD = 0.11) for the AD patients ($p < 0.05$). To control for head motion, we applied motion correction using MCFLIRT (Jenkinson & Smith, 2002), and regressed the motion parameters out of the fMRI data. Additionally, we used the FMRIB's ICA-based Xnoiseifilter (FIX, version 1.06) to automatically identify and remove noise components from the fMRI data (Salimi-Khorshidi et al., 2014), thereby increasing the signal to noise ratio (Griffanti et al., 2016). For the spatial smoothing, we used a smoothing kernel with a full width half maximum of 3 mm. We performed minimal smoothing, because this is recommended prior to running an independent component analysis (ICA) in order to reduce the probability of finding spurious components (Jenkinson, 2015).

Table 1. Sample demographics

	Controls		**AD[1] patients**		**X^2**
Gender (♂/♀)	**74/99 (57% ♀)**		**31/46 (60% ♀)**		**n.s.[2]**
	min - max	**mean ± SD**	**min - max**	**mean ± SD**	**t-test**
Age	47 - 83	66.1 ± 8.7	47 - 83	68.6 ± 8.6	p<0.05
Education (years)	9 - 18	11.5 ± 2.8	4 - 20	10.8 ± 3.2	n.s.
Disease duration (months)	-	-	2 - 156	26.7 ± 24.5	-
MMSE[3]	22 - 30	27.5 ± 1.8	10 - 28	20.4 ± 4.5	p<0.001
CDR[4]	-	-	0.5 - 2	0.8 ± 0.3	-
GDS[5]	0 - 11	2.1 ± 2.1	0 - 10	2.6 ± 2.6	n.s.

[1]AD = Alzheimer's disease, [2]MMSE = mini-mental state examination, [3]CDR = clinical dementia rating, [4]GDS = geriatric depression scale.

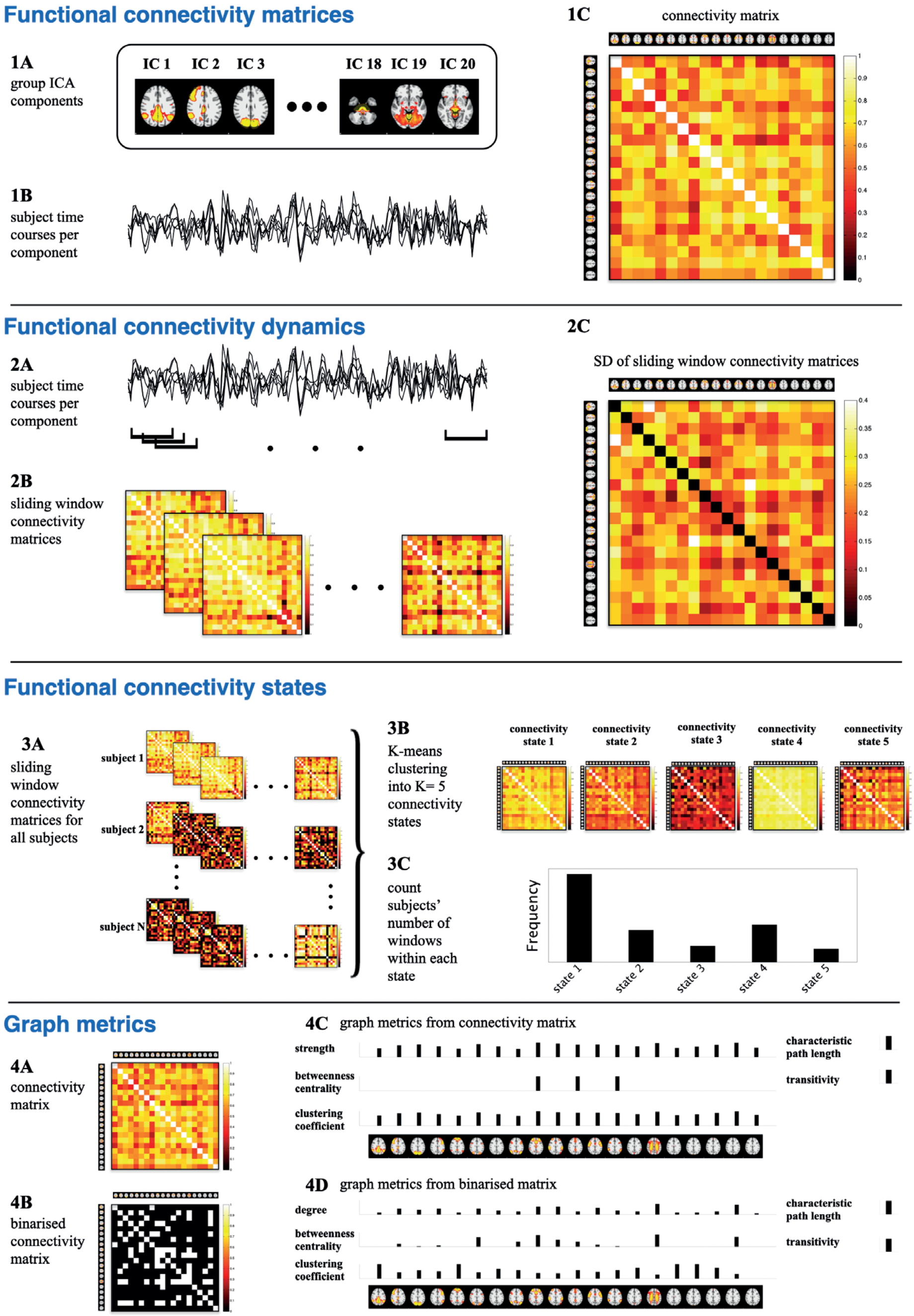
Resting state fMRI measures

We calculated eight types of measures from the resting state fMRI data. For most of those eight types of measures we calculated more than one variety, resulting in a total of 31 measures. These resting state fMRI measures are listed in Table 2, along with the number of values they comprise. These values are used as predictors in the classification analyses. Figure 1 summarises the procedures used to calculate the resting state fMRI measures. A more elaborated description is written below.

Table 2. List of resting state fMRI measures used for Alzheimer's disease classification

Resting state measure	# of predictors
1: FC[1] matrices	
1a. 20 X 20 full correlation	190
1b. 70 X 70 full correlation	2415
1c. 20 X 20 sparse partial correlation	190
1d. 70 X 70 sparse partial correlation	2415
FC dynamics	
2a. SD[2] of 20 X 20 full correlation FC matrix	190
2b. SD of 70 X 70 full correlation FC matrix	2415
2c. SD of 20 X 20 sparse partial correlation FC matrix	190
2d. SD of 70 X 70 sparse partial correlation FC matrix	2415
3: FC states	
3a. FC states of 20 X 20 full correlation FC matrix	5
3b. FC states of 70 X 70 full correlation FC matrix	5
3c. FC states of 20 X 20 partial correlation FC matrix	5
3d. FC states of 70 X 70 partial correlation FC matrix	5
4: Graph metrics	
4a. Graph metrics of 20 X 20 full correlation FC matrix	124
4b. Graph metrics of 70 X 70 full correlation FC matrix	424
4c. Graph metrics of 20 X 20 partial correlation FC matrix	124
4d. Graph metrics of 70 X 70 partial correlation FC matrix	424
5: FC with resting state networks	
5a. FC with visual network 1	190981
5b. FC with visual network 2	190981
5c. FC with visual network 3	190981
5d. FC with default mode network	190981
5e. FC with the cerebellum	190981
5f. FC with sensorimotor network	190981
5g. FC with auditory network	190981
5h. FC with executive control network	190981
5i. FC with frontoparietal network 1	190981
5j. FC with frontoparietal network 2	190981
6: FC with Hippocampus	
6a. FC with left hippocampus	190981
6b. FC with right hippocampus	190981
7: Eigenvector centrality	
Fast eigenvector centrality mapping	190981
8: ALFF[3]	
8a. ALFF	190981
8b. fALFF[4]	190981
All resting state fMRI measures combined	**2,876,251**

[1]FC = functional connectivity, [2]SD = standard deviation, [3]ALFF = amplitude of low frequency fluctuations, [4]fALFF = fractional amplitude of low frequency fluctuations.


Functional connectivity matrices
1A
group ICA components
IC 1
IC 2
IC 3
IC 18
IC 19
IC 20
1B
subject time courses per component
1C
connectivity matrix
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Functional connectivity dynamics
2A
subject time courses per component
2B
sliding window connectivity matrices
2C
SD of sliding window connectivity matrices
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Functional connectivity states
3A
sliding window connectivity matrices for all subjects
subject 1
subject 2
subject N
3B
K-means clustering into K= 5 connectivity states
connectivity state 1
connectivity state 2
connectivity state 3
connectivity state 4
connectivity state 5
3C
count subjects' number of windows within each state
Frequency
state 1
state 2
state 3
state 4
state 5
Graph metrics
4A
connectivity matrix
4B
binarised connectivity matrix
4C
graph metrics from connectivity matrix
strength
betweenness centrality
clustering coefficient
characteristic path length
transitivity
4D
graph metrics from binarised matrix
degree
betweenness centrality
clustering coefficient
characteristic path length
transitivity

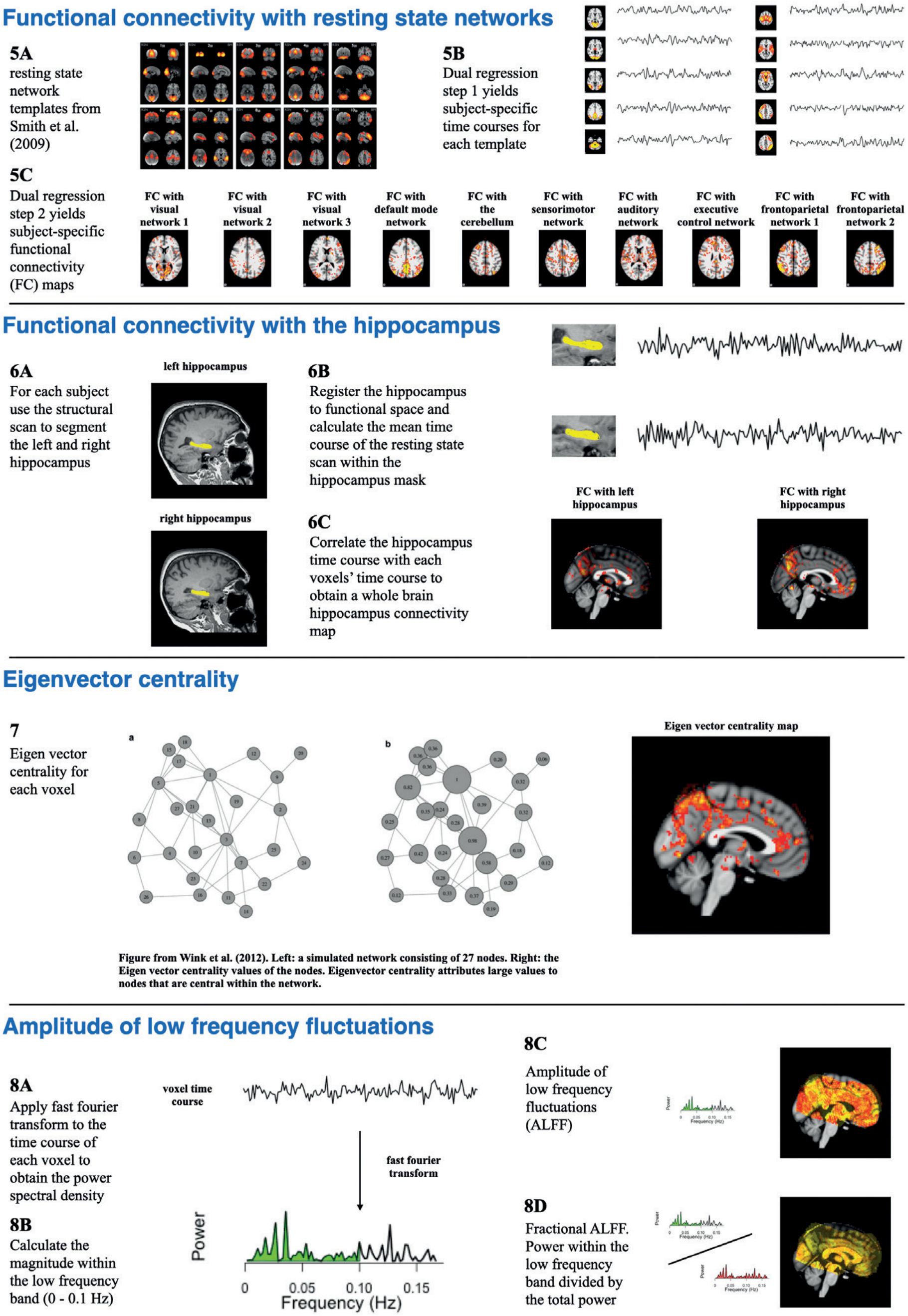

Functional connectivity with resting state networks
5A
resting state network templates from Smith et al. (2009)
5B
Dual regression step 1 yields subject-specific time courses for each template
5C
Dual regression step 2 yields subject-specific functional connectivity (FC) maps
FC with visual network 1
FC with visual network 2
FC with visual network 3
FC with default mode network
FC with the cerebellum
FC with sensorimotor network
FC with auditory network
FC with executive control network
FC with frontoparietal network 1
FC with frontoparietal network 2
Functional connectivity with the hippocampus
6A
For each subject use the structural scan to segment the left and right hippocampus
left hippocampus
right hippocampus
6B
Register the hippocampus to functional space and calculate the mean time course of the resting state scan within the hippocampus mask
6C
Correlate the hippocampus time course with each voxels' time course to obtain a whole brain hippocampus connectivity map
FC with left hippocampus
FC with right hippocampus
Eigenvector centrality
7
Eigen vector centrality for each voxel
a
b
Eigen vector centrality map
Amplitude of low frequency fluctuations
8A
Apply fast fourier transform to the time course of each voxel to obtain the power spectral density
8B
Calculate the magnitude within the low frequency band (0 - 0.1 Hz)
voxel time course
fast fourier transform
Power
0
0.05
0.10
0.15
Frequency (Hz)
8C
Amplitude of low frequency fluctuations (ALFF)
8D
Fractional ALFF. Power within the low frequency band divided by the total power

Figure from Wink et al. (2012). Left: a simulated network consisting of 27 nodes. Right: the Eigen vector centrality values of the nodes. Eigenvector centrality attributes large values to nodes that are central within the network.

Figure 1. The procedures for calculating the eight resting state fMRI modalities.

Functional connectivity matrices

For each participant, we calculated FC between RSNs. We used temporal concatenation ICA in FSL MELODIC (Beckmann & Smith, 2004) to obtain RSNs. First, we registered the functional data of all participants to standard space and concatenated them along the time dimension. We then performed a low and a high dimensional ICA on the concatenated data set, forcing a solution with 20 and 70 components respectively. The components of these two ICA solutions are shown in Figure S2 in the supplementary materials. We registered the resulting ICA component weight maps back to subject space, weighted them by the subject specific grey matter density maps, and multiplied them with the functional data. We then calculated the mean time courses for the components and used these for the FC analysis. We calculated both full and partial correlation matrices. For the partial correlation matrices, we used the graphical lasso algorithm (Friedman et al., 2008) implemented in MATLAB (MATLAB 2013a, The MathWorks Inc., Natick, MA, 2000). We set the λ parameter at 100, because this setting works best in most cases for fMRI functional connectivity (Smith et al., 2011). For each participant, we thus calculated four FC matrices. The two 20 by 20 matrices each contain (20 * 19)/2 = 190 unique elements, and the 70 by 70 matrices each contain (70 * 69)/2 = 2415 unique elements. We used these elements as predictors for classification.

Dynamics of functional connectivity matrices

We also calculated the dynamics of the above-described FC matrices using a sliding window approach (Chang and Glover 2011; Hutchinson et al., 2013; Jones et al., 2012). We used a window size of 33 seconds, similar to Jones et al. (2012) and Rashid et al. (2014), because it was shown that time windows as short as 30 seconds can provide reasonably good connectivity estimates (Shirer et al. 2012). We shifted the windows one volume at a time, resulting in 140 windows (Jones et al., 2012; Rashid et al., 2014). Within each window we calculated the four FC matrices as described in the previous paragraph. Then we calculated the standard deviation of the FC matrices over all the windows. This resulted in four matrices of standard deviations for each subject, with equal size as the FC matrices. We used the elements of these matrices as predictors for classification.

Functional connectivity states

For each of the four types of FC matrices we distinguished five 'FC states' and administered how long each subject resided in each of these five states. Functional connectivity states are patterns of FC that reoccur in time across participants (Allen et al., 2012; Jones et al., 2012; Rashid et al., 2014). In order to determine the FC states, we clustered the sliding window FC matrices using k-means clustering. So, for each of the four types of FC matrices we clustered the 250 (number of subjects) * 140 (number of windows) = 35000 sliding window matrices. We created k = 5 clusters like Jones et al. (2012) and Rashid et al. (2015) and we used the Manhattan distance criterion like Allen et al. (2012). Then, for each participant we counted the number of sliding window matrices that were assigned to each of the five FC states. The five frequency values for each of the four types of FC matrices were used as predictors for classification.

Graph metrics

For each of the four types of FC matrices we calculated commonly used graph metrics. We used both the original and the binarised version of the FC matrices. Binary links denote the presence or absence of connections, while the original values contain information about the connection strengths (Rubinov & Sporns, 2010). Current network methods cannot quantify the role of negative connections in network organisation (Rubinov & Sporns, 2010) and therefore we absolutised the negative links. We binarised the full correlation matrices by maintaining the 20% largest absolute correlations within each matrix (Khazaee et al., 2015). Since the sparse partial correlation matrices are sparse from itself, we did not apply a binarisation threshold, but binarised the matrices by transforming all values greater than zero to 1. We used the Brain Connectivity Toolbox (Rubinov & Sporns, 2010) available for MATLAB (MATLAB 2013a, The MathWorks Inc., Natick, MA, 2000) to calculate the graph metrics. For the original connectivity matrix, we calculated the connection strength, weighted betweenness centrality, and weighted clustering coefficient for every node in the network and the weighted characteristic path length and weighted transitivity for the entire network (Rubinov & Sporns, 2010). For the binarised connectivity matrix we calculated the connection degree, betweenness centrality, and clustering coefficient for every node in the network and the characteristic path length and transitivity for the entire network (Rubinov & Sporns, 2010). So, in total we calculated 10 different graph measures, six measures for every node and four measures for the entire network. This resulted in 6*20 + 4*1 = 124 predictors for the 20*20 FC matrices and 6*70 + 4*1 = 424 predictors for the 70*70 FC matrices.

Whole brain functional connectivity with resting state networks

We calculated whole brain FC with 10 RSNs using the dual regression approach in FSL (Filippini et al., 2009). We used templates that were obtained using an independent data set to increase the reproducibility of our findings (Griffanti, 2016). We used the RSN templates that were obtained using an ICA by Smith et al. (2012). These RSNs are freely available online (http://www.fmrib.ox.ac.uk/analysis/brainmap+rsns/PNAS_Smith09_rsn10.nii.gz) as spatial maps in standard space. These 10 RSNs include three visual networks, the default mode network, the cerebellum, a sensorimotor network, an auditory network, an executive control network and two frontoparietal networks. Additionally, we included the white matter (WM) and cerebrospinal fluid (CSF) maps provided by FSL (Jenkinson et al., 2012; Smith et al., 2004) as confound maps. Those 12 spatial maps (10 RSNs plus two confound maps) were then used in a dual regression analysis. First, for each subject, the 12 spatial maps were regressed (as spatial regressors in a multiple regression) into the subjects' 4D space-time dataset. This results in a set of subject-specific time series, one for each spatial map. Next, those time series were regressed (as temporal regressors, again in a multiple regression) into the same 4D dataset, resulting in 12 subject-specific spatial maps, one for each RSN and one for each of the two confound maps. These subject-specific spatial maps represent whole brain FC with the RSNs. We used the voxel-wise whole brain FC results for the ten RSNs as predictors for classification.

Whole brain functional connectivity with hippocampus

For each participant, we calculated whole brain FC with the left and with the right hippocampus (Allen et al., 2007; Wang et al., 2006). We first calculated the time course of the hippocampus for each participant. To this end we segmented the hippocampus in the anatomical scan using FSL First. We eroded the segmented hippocampus with three voxel layers to ascertain that only hippocampus voxels were included. The eroded hippocampus was then affine registered to the functional data and we calculated the mean time course of the functional data within the hippocampus mask. Then, for each participant we regressed the time course of the hippocampus, along with the mean WM and CSF time courses as confound regressors, into the functional data using multiple regression. This resulted in a whole brain FC map with the hippocampus. We performed this analysis for both the left and the right hippocampus and used the two resulting whole brain FC maps as predictors for classification.

Eigenvector centrality

For each participant, we calculated an eigenvector centrality map. Eigenvector centrality attributes a value to each voxel in the brain such that a voxel receives a large value if it is strongly correlated with many other voxels that are themselves central within the network (Lohmann et al., 2010). We used the fastECM algorithm (Wink et al., 2012; Binnewijzend et al., 2014) to calculate a whole brain eigenvector centrality map in standard space for each participant.

Amplitude of low frequency fluctuations

We calculated ALFF (Biswal et al., 2010; Zang et al., 2007) and fractional ALFF (fALFF) (Zou et al., 2008) for each participant. We used the REST software package (Song et al., 2011) to calculated whole brain ALFF and fALFF maps. ALFF was defined as the power within the 0 - 0.1 Hz frequency band and fALFF was defined as the power within the 0 - 0.1 Hz frequency band divided by the power of the whole frequency spectrum. For standardisation purposes, we divided the voxels' ALFF/fALFF values by the mean ALFF/fALFF within a subjects' whole brain (Zang et al., 2007).

Statistical analyses

For each of the 31 groups of predictors of the eight resting state fMRI modalities we used an elastic net logistic regression model to classify the subjects as either AD or control. Elastic net regression is commonly used for neuroimaging classification studies (Teipel et al., 2017a; Nir et al., 2016; Trzepacz et al., 2016). We used the glmnet package (Friedman et al., 2010; Zou & Hastie, 2005) available for R (R version 3.1.2, R Core Team, 2014). Elastic net regression uses penalties to hinder the predictors from entering the regression model (Friedman et al., 2010; Zou & Hastie, 2005). Thus, only the most relevant predictors will enter the regression model, which is helpful if the number of predictors outnumbers the number of subjects. Elastic net regression uses a combination of an L1 (LASSO) (Tibshirani, 1996) and L2 (Ridge) (Hoerl and Kennard, 1970) penalty. Therefore, two hyper parameters should be set: the α parameter determines the relative weight of the two different penalties and λ determines the size of those penalties. Elastic net logistic regression has already been used for AD classification (de Vos et al., 2016; Schouten et al., 2016; Schouten et al., 2017; Teipel et al., 2015; Trzepacz et al., 2014). For the combined classification model, we concatenated the 31 groups of predictors, resulting in a combined set containing 2,876,251 predictors. These predictors were jointly included in the group lasso model (Simon et al., 2013) and we informed the group lasso with an index vector that indicates the group membership of the predictors. The group lasso is similar to the elastic net, but sparse with respect to groups of predictors. This improves interpretation of the combined model, because a modality is either entirely included or excluded from the prediction model. We used the SGL package (Simon et al., 2013) available for R (R version 3.1.2, R Core Team, 2014).

We used cross-validation to ensure that we are not overfitting the prediction models. In our case there are two potential sources of overfitting. We could either include too many predictors in our logistic regression model or we could overestimate the classification accuracy by looping over all the values of the hyper parameters and only pick the best result. To ascertain that we are not subject to any of these two sources of overfitting we used a nested cross-validation approach (Krstajic et al., 2014). We used the inner loop of the nested cross-validation to tune the hyper parameters and the outer loop to fit and test the logistic regression model. For both the inner and outer loop we used 10-fold cross-validation, thus using 90 percent of the subjects in the training set and 10 percent in the test set, and repeating this 10 times such that all subjects were part of the test set once.

We made receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) as a measure of classification performance. The AUC is invariant to the class distribution (Bradley, 1997; Fawcett, 2004), which is an advantage since the number of control subjects is larger than the number AD patients. We also calculated sensitivity, specificity and balanced accuracy values for those classification cut-offs that resulted in the highest balanced accuracy. We repeated the cross-validation procedure 10 times to get a more reliable cross-validation error (Krstajic et al., 2014) and extracted the mean AUC value.

In order to statistically compare the AUC values, we used bootstrap tests for paired AUCs (Hanley and McNeil, 1983) implemented in the pROC package (Robin et al., 2011) available for R (R version 3.1.2, R Core Team, 2014). For the comparison of the different resting state fMRI measures we used two-sided tests, because we have not formulated any directed hypotheses for these comparisons. To compare the combined model with the single measures we applied one-sided hypothesis tests, because we hypothesised that the combined model would outperform the single measures. We present uncorrected p-values and Bonferroni corrected p-values. The Bonferroni correction was applied separately to the inter measure comparisons and the comparisons of the single measures with the combined model.

3.3 Results

Classification results

Figure 2 shows the AUC values for the 31 different types of resting state fMRI measures and the combined model. Table 3 also presents values for sensitivity, specificity and balanced accuracy. The AUC values range between 0.51 and 0.84. The functional connectivity matrices (AUC values between 0.72 and 0.82) and the FC dynamics

(AUC values between 0.72 and 0.84) distinguish AD patients and controls quite well. Particularly the sparse partial correlations between the 70 ICA components (AUC = 0.82) and the standard deviations of these sparse partial correlations over time (AUC = 0.84) have high AUC values. Also, the ALFF measures are discriminative for AD. ALFF has an AUC value of 0.82, and fALFF has an AUC value of 0.69. The FC states (AUC values between 0.55 and 0.74) and the graph metrics (AUC values between 0.70 and 0.79) have reasonable classification accuracies. Functional connectivity with the 10 RSNs (AUC values between 0.52 and 0.71) mostly performs poorly, except for FC with the default mode network (AUC = 0.70) and the executive control network (AUC = 0.71). FC with the left (AUC = 0.59) and right (AUC = 0.51) hippocampus result in poor classification performances and Eigenvector centrality mapping results in moderate classification performance (AUC = 0.69). As shown in Figure 2 on the right, the combination of all the resting state fMRI measures using the group lasso model results in an AUC value of 0.85, which is higher than any of the measures used alone. Combining resting state fMRI measures thus seems beneficial, although the effect is small.

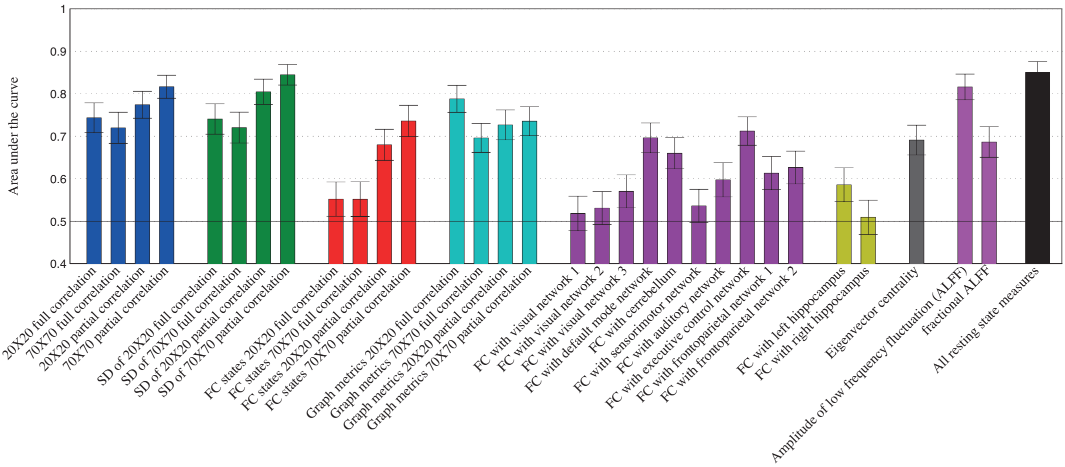


Figure 2. Area under the receiver operating characteristic curve (AUC) values for all the resting state fMRI measures. The wide bar on the right is the AUC value for the combination of all resting state fMRI measures. The error bars represent one standard error above and below the AUC values.

Table 3. Alzheimer's disease classification performance for the resting state fMRI Figure S1. Correlations between the 31 resting state fMRI measures. For each measure, we ran a principal component analysis (PCA) and we cross correlated the component scores of all the 31 first components..

Resting state measure	AUC[1]	Sensitivity	Specificity	Balanced accuracy
1: FC[2] matrices				
1a. 20 X 20 full correlation	0.74	0.73	0.68	0.71
1b. 70 X 70 full correlation	0.72	0.62	0.77	0.69
1c. 20 X 20 sparse partial correlation	0.77	0.68	0.76	0.72
1d. 70 X 70 sparse partial correlation	0.82	0.79	0.71	0.75
FC dynamics				
2a. SD[3] of 20 X 20 full correlation FC matrix	0.74	0.67	0.74	0.7
2b. SD of 70 X 70 full correlation FC matrix	0.72	0.70	0.69	0.69
2c. SD of 20 X 20 sparse partial correlation FC matrix	0.80	0.76	0.76	0.76
2d. SD of 70 X 70 sparse partial correlation FC matrix	0.84	0.83	0.73	0.78
3: FC states				
3a. FC states of 20 X 20 full correlation FC matrix	0.55	0.39	0.75	0.57
3b. FC states of 70 X 70 full correlation FC matrix	0.55	0.54	0.60	0.57
3c. FC states of 20 X 20 partial correlation FC matrix	0.68	0.60	0.71	0.66
3d. FC states of 70 X 70 partial correlation FC matrix	0.74	0.72	0.69	0.70
4: Graph metrics				
4a. Graph metrics of 20 X 20 full correlation FC matrix	0.79	0.79	0.68	0.74
4b. Graph metrics of 70 X 70 full correlation FC matrix	0.70	0.74	0.61	0.68
4c. Graph metrics of 20 X 20 partial correlation FC matrix	0.73	0.75	0.65	0.70
4d. Graph metrics of 70 X 70 partial correlation FC matrix	0.74	0.72	0.69	0.71
5: FC with resting state networks				
5a. FC with visual network 1	0.52	0.46	0.64	0.55
5b. FC with visual network 2	0.53	0.35	0.77	0.56
5c. FC with visual network 3	0.57	0.48	0.68	0.58
5d. FC with default mode network	0.70	0.67	0.66	0.67
5e. FC with the cerebellum	0.66	0.60	0.68	0.64
5f. FC with sensorimotor network	0.54	0.45	0.67	0.56
5g. FC with auditory network	0.60	0.68	0.52	0.60
5h. FC with executive control network	0.71	0.76	0.62	0.69
5i. FC with frontoparietal network 1	0.61	0.50	0.74	0.62
5j. FC with frontoparietal network 2	0.63	0.60	0.65	0.62
6: FC with Hippocampus				
6a. FC with left hippocampus	0.59	0.51	0.66	0.59
6b. FC with right hippocampus	0.51	0.35	0.74	0.55
7: Eigenvector centrality				
Fast eigenvector centrality mapping	0.69	0.66	0.66	0.66
8: ALFF[4]				
8a. ALFF	0.82	0.71	0.82	0.76
8b. fALFF[5]	0.69	0.71	0.61	0.66
All resting state fMRI measures combined	0.85	0.86	0.71	0.79

[1]AUC = area under the receiver operating characteristic curve, [2]FC = functional connectivity, [3]SD = standard deviation, [4]ALFF = amplitude of low frequency fluctuations, [5]fALFF = fractional amplitude of low frequency fluctuations.

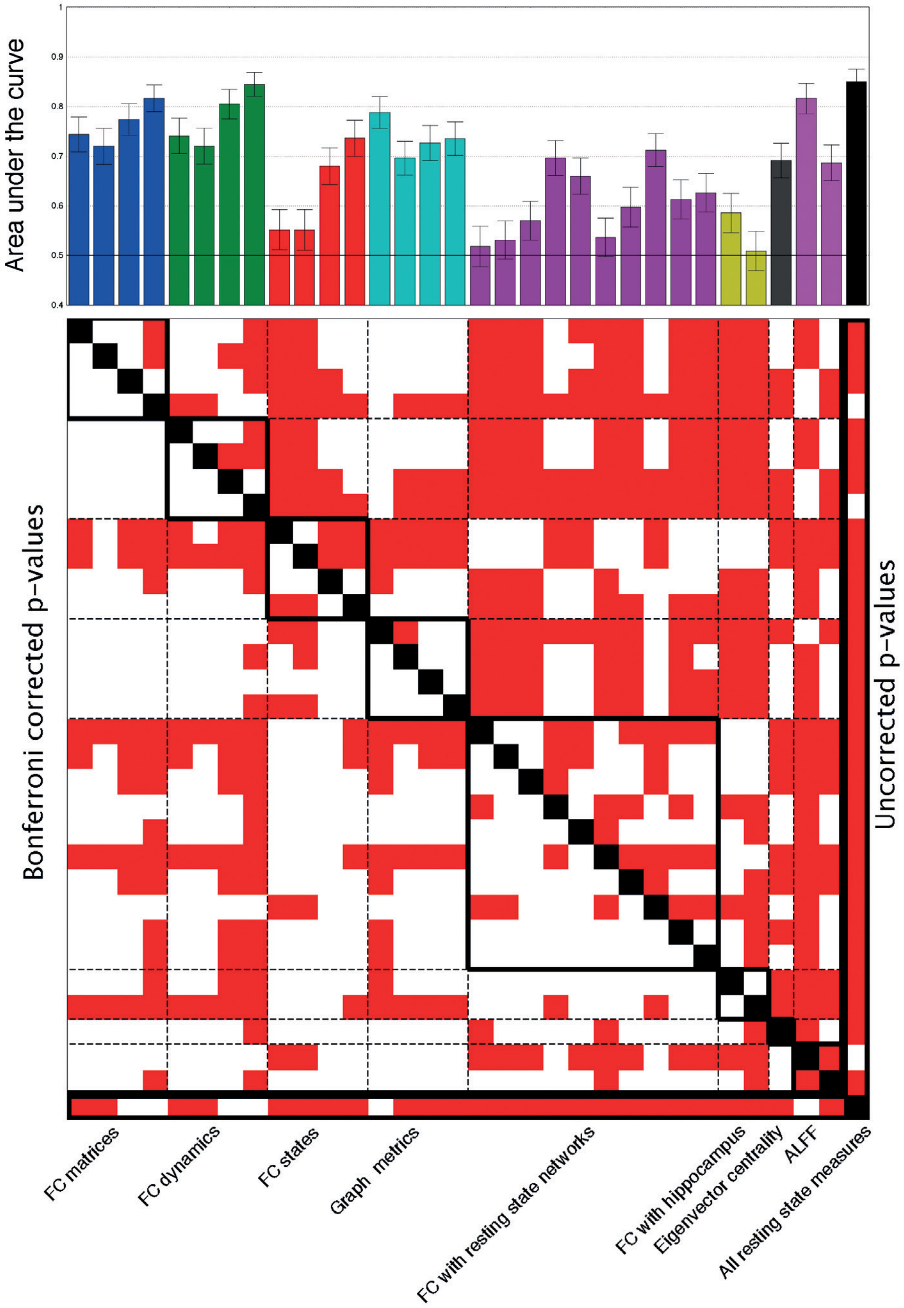


Figure 3. Statistical comparisons between the AUC values. The barplot contains the AUC values for the different resting state fMRI measures and the combined model, together with their standard errors. The matrix contains the results for the statistical comparisons between the AUC's. The top right half of the matrix contains the uncorrected results. The bottom left half of the matrix contains the Bonferroni corrected results. The red coloured elements represent *p* values < 0.05.

Combined classification model

Figure 4 shows the contribution to the combined model for each of the 31 resting state fMRI measures. The y-axis represents the sum of the absolute standardised beta values for all the predictors within a resting state measure. A high value represents an important role for that group of predictors within the combined model. In order to quantify the spread of the contributions we fitted the group lasso model repeatedly on 100 bootstrap samples. The 100 results are represented by the boxplots. In line with the results of the single modalities, the FC matrices and the FC dynamics largely contribute the combined prediction model. There is also some contribution of the FC states and the graph metrics. Remarkably, ALFF hardly contributes, despite its discriminative power when used alone. There is considerable spread in the contribution of the resting state fMRI measures as shown by the 100 bootstrap results. None of the resting state fMRI measures contributes to the group lasso model in each bootstrap sample. However, it remains clear that the FC matrices and FC dynamics are important for the combined prediction model, whereas the other resting state fMRI measures contribute minimally or not at all.

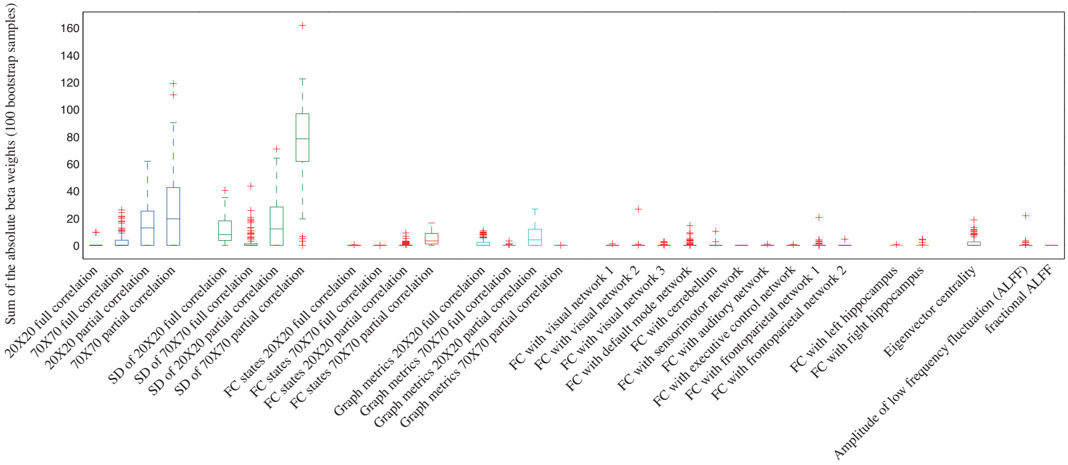


Figure 4. Importance of each resting state measure for the combined model. The combined model is fitted on 100 bootstrap samples to display the spread of the importance's. The importance is quantified by the sum of the absolute beta weights of all the predictors within a resting state measure category.

Supplementary analyses

Relation between different resting state fMRI measures

To explore relations between the resting state fMRI modalities, we calculated correlations between the 31 different resting state fMRI measures. These are presented in supplementary Figure S1. Straightforward calculation of correlation coefficients between the 31 resting state fMRI measures was not possible, because each resting state fMRI measure contains multiple predictors and the number of predictors is different for every measure. To overcome this problem, we ran a principal component analysis (PCA) for each of the 31 measures and cross correlated the component scores of all the 31 first components. Not surprisingly, resting state fMRI measures within the same modality are generally highly related. In addition, FC matrices, FC dynamics, FC states and ALFF appear to be related to each other.

Functional connections important for classification

To explore which of the ICA components were most important for AD classification, we plotted the mean beta values over all cross-validation folds and cross-validation repetitions for the FC matrices in supplementary Figure S3. Functional connectivity values between higher components have larger beta weights than FC values between lower components. Figure S2 shows that higher components are in fact real functional networks, whereas some lower components are noise components. This suggests that information on real functional networks was contributing to the classifier.

Percentage of non-zero parameters

For each resting state measure, we looked at the percentage of predictors that contributed to the classification model. Figure S4 shows the mean percentage of non-zero parameters over all cross-validation folds and cross-validation repetitions for each resting state measure. The percentages are mostly over 20%, indicating that for most measures many predictors are included in the classification model.

Voxel-wise vs. averaging over regions

The AUC values for resting state fMRI modalities one to four are mostly higher than the AUC values for resting state fMRI modalities five to eight. One notable difference between these two groups is the number of predictors. The number of predictors within resting state fMRI modalities one to four ranges from five to 2415 per category, whereas resting state fMRI modalities five to eight are voxel-wise maps and they contain 190981 predictors per category. To explore the possibility that the number of predictors influences the classification performance, we averaged the voxel-wise maps over the 70 components as obtained by the high dimensional ICA and reran the classification analyses with the reduced number of predictors. Figure S5 shows both the original AUC values and the AUC values after averaging over the 70 components. The differences are small and the ranges of the different cross-validation repetitions are most of the time overlapping. The low classification performance for some of these categories seems not to be caused by the large number of predictors.

Optimal value analysis for the number of ICA components

We investigated the optimal number of ICA components for our connectivity analyses. We ran ICA analyses for 5 to 100 components with steps of five. For each number of components, we calculated connectivity matrices with both full or sparse partial correlations, and the dynamics of these connectivity matrices. The results are plotted in Figure S6. Calculating only five components seems to be too few, but upwards of 10 components the results are too diverse to draw conclusions on the optimal number of components.

3.4 Discussion

In this study, we determined the accuracy of different resting state fMRI measures for the individual classification of AD patients. We used machine learning techniques for efficient use of resting state fMRI measures in prediction models. FC matrices, FC dynamics, and ALFF show best discrimination between AD patients and control subjects. The combination of all the resting state fMRI measures improved the classification accuracy slightly, but not significantly. The FC matrices and the FC dynamics largely contribute to this combined model, whereas the other resting state fMRI measures are mostly redundant. This suggests that only FC matrices and FC dynamics need to be calculated to achieve optimal individual AD classification through a resting state fMRI scan.

FC matrices have been used successfully for AD classification before (Challis et al., 2015; Chen et al., 2011). Our results add to this conclusion and furthermore show that FC as calculated with sparse partial correlation results in higher classification accuracy than FC as calculated with full correlation. Likely, this is due to the fact that sparse partial correlations provide better FC estimates than full correlations (Smith et al., 2011). In addition, FC between 70 components resulted in somewhat higher classification accuracy than FC between 20 components. This is in line with the observation that high dimensional ICA solutions provide a more specific representation of functional regions, and consequently FC between these regions results in better AD classification (Dipasquale et al., 2015). The dynamics of the FC matrices resulted in higher classification accuracy than the FC matrices itself. FC dynamics as opposed to static FC is a relatively unexplored domain in AD, but it has been shown that AD patients differ in their FC dynamics compared to controls (Chen et al., 2016; Jones et al., 2012; Wee et al., 2016). ALFF resulted in good classification accuracy, similar to Dai et al (2012). However, it did not provide additive value over the FC matrices and the FC dynamics for the combined model. For this reason, ALFF does not seem to be necessary for an AD classification model. FC states and graph metrics had reasonable classification accuracies, but they did not provide additive value for the combined model either. Furthermore, these two measures are derived from the FC matrices and FC dynamics, which themselves have higher classification accuracies. Functional connectivity states and graph metrics thus require more work to calculate and they do not seem to be beneficial over the simpler measures.

FC with the ten RSNs resulted mostly in poor classification accuracies. Exceptions are FC with the default mode network and FC with the executive control network. This corresponds to studies reporting abnormal FC in these networks in AD patients (Agosta et al., 2012; Binnewijzend et al., 2012; Greicus et al., 2004). FC with the hippocampus also resulted in poor classification accuracy, despite abnormal hippocampal connectivity patterns observed in AD patients (Allen et al., 2007; Supekar et al., 2008; Wang et al., 2006). These effects are probably not sufficiently consistent for AD classification. Possibly this is due to the fact that the hippocampus is not persistently connected with the cortex, but follows a context dependent connectivity pattern (Huijbers et al., 2012).

Some settings we have not explored. We used an ICA to determine regions as input to the FC analysis (Allen et al., 2012; Hutchison et al., 2013; Jones et al., 2012; Rashid et al, 2014), where others have used the automated anatomical labeling (AAL) atlas (Chen et al., 2011; Wee et al., 2016). We chose an ICA, because it is a data-driven approach that results in spatially independent components well suited for FC analyses. We used a group ICA and imposed the group components onto each subject (Dørum et al, 2017; Miller et al, 2016), because it is important to strive for the same parcellation in each subject in order to compare connectomes across subjects (Smith et al., 2013). We used the ICA components directly as nodes for the FC analysis. Others have used a follow up procedure to split ICA components into multiple nodes and use these nodes as input to the FC analysis (Shirer et al, 2012; Jones et al., 2012; Shaw et al., 2015). We have not explored this option, but we obtained a similar result using the high dimensional ICA solution. When extracting a higher number of components, large networks split into multiple smaller networks. This can be observed in Figure S2. For example, component 1 of the 20 components solution splits into components 1, 6 and 13 of the 70 components solution. For the calculation of FC dynamics, we used blocked sliding windows covering 11 volumes (33 seconds) and we shifted the windows one volume at a time (Jones et al., 2012). Other methods have been reported, using tapered windows (Allen et al., 2012; Wee et al., 2016), different window sizes (Chen et al., 2016; Wee et al., 2016) and larger window shifts (Wee et al., 2016). Further, we quantified FC dynamics by calculating the standard deviations of the sliding window FC estimates. Alternatives are the dwell time in default mode network sub-configurations (Jones et al., 2012), graph measures obtained from the sliding window matrices (Wee et al., 2016), or higher order FC statistics that capture the covariance of the sliding window FC time series (Chen et al., 2016). We have not explored these methodological settings to study their effect on classification accuracy. For the analyses, we sticked to the default settings for that specific analysis as much as possible, and if there was no clear default setting, we based our choices on previous literature. We chose not to optimise the parameter settings within our study, because this would be computationally infeasible. Proper parameter optimisation must be performed using cross-validation, and in our case this would expand the cross-validation analyses considerably, because of the large number of predictors (~2 million) and the high number of parameters that can be optimised. In addition, the resting state fMRI scans used in this study covered 7.5 minutes, which is short for estimation of FC dynamics (Hindriks et al., 2016). It is not known whether classification accuracy improves with longer scan times.

For this study, we used only one sample to both train and test our prediction models. We have carefully used cross-validation techniques to prevent overfitting and obtain realistic accuracy estimates. Nevertheless, when applying these prediction models to other samples scanned at different scanner sites, we might find reduced classification accuracies, because these models are fine-tuned on the current sample. To evaluate the robustness of our classification models they have to be applied to a different sample. In addition, the current sample was not the result of random sampling from a prespecified population. The conclusions from the statistical tests that we have performed therefore only apply to the current sample.

3.5 Conclusion

In conclusion, we demonstrated the use of resting state fMRI scans for individual AD classification. The optimal combination of resting state fMRI measures comprises FC matrices and FC dynamics. These results may direct future studies that use resting state fMRI scans for the classification of patients with preclinical AD or mild cognitive impairment.

3.6 Acknowledgments

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

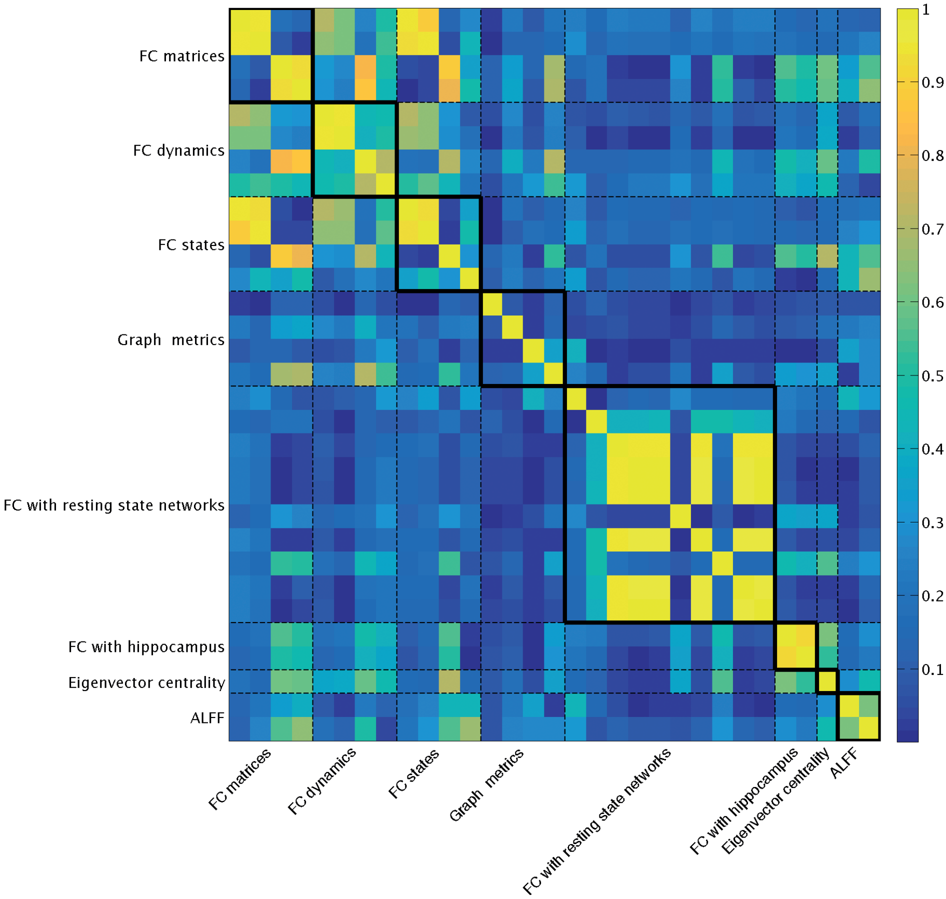


Figure S1. Correlations between the 31 resting state fMRI measures. For each measure, we ran a principal component analysis (PCA) and we cross correlated the component scores of all the 31 first components.

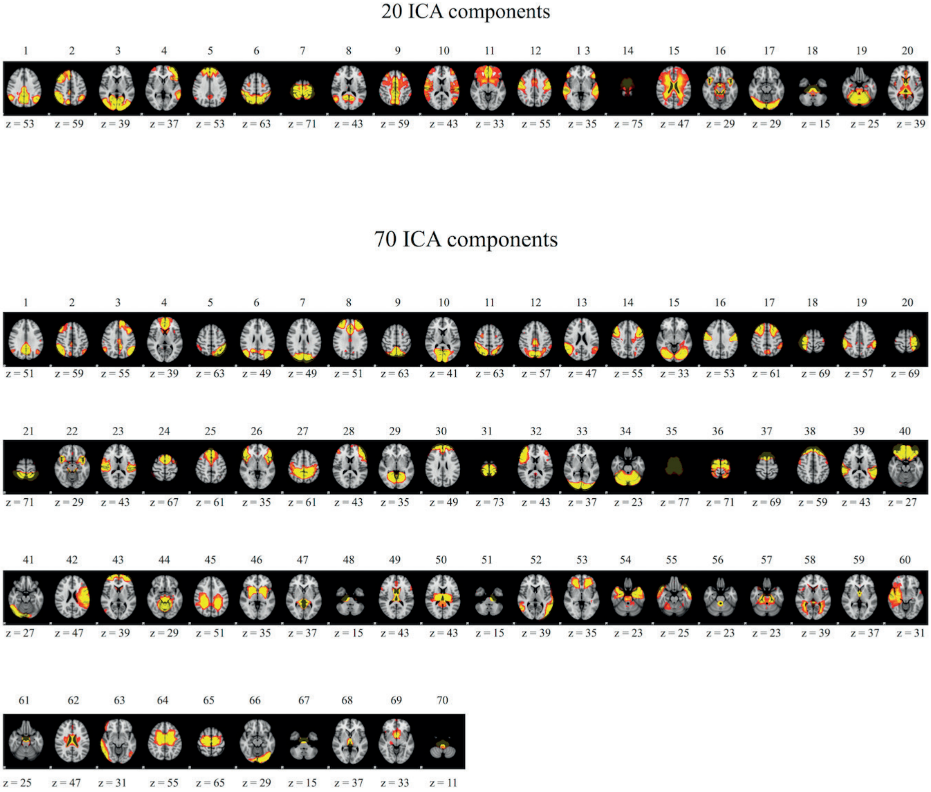

20 ICA components
1 2 3 4 5 6 7 8 9 10 11 12 1 3 14 15 16 17 18 19 20
z = 53 z = 59 z = 39 z = 37 z = 53 z = 63 z = 71 z = 43 z = 59 z = 43 z = 33 z = 55 z = 35 z = 75 z = 47 z = 29 z = 29 z = 15 z = 25 z = 39
70 ICA components
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
z = 51 z = 59 z = 55 z = 39 z = 63 z = 49 z = 49 z = 51 z = 63 z = 41 z = 63 z = 57 z = 47 z = 55 z = 33 z = 53 z = 61 z = 69 z = 57 z = 69
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
z = 71 z = 29 z = 43 z = 67 z = 61 z = 35 z = 61 z = 43 z = 35 z = 49 z = 73 z = 43 z = 37 z = 23 z = 77 z = 71 z = 69 z = 59 z = 43 z = 27
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
z = 27 z = 47 z = 39 z = 29 z = 51 z = 35 z = 37 z = 15 z = 43 z = 43 z = 15 z = 39 z = 35 z = 23 z = 25 z = 23 z = 23 z = 39 z = 37 z = 31
61 62 63 64 65 66 67 68 69 70
z = 25 z = 47 z = 31 z = 55 z = 65 z = 29 z = 15 z = 37 z = 33 z = 11

Figure S2. The 20 and 70 components extracted from the low and high dimensional independent component analysis (ICA).

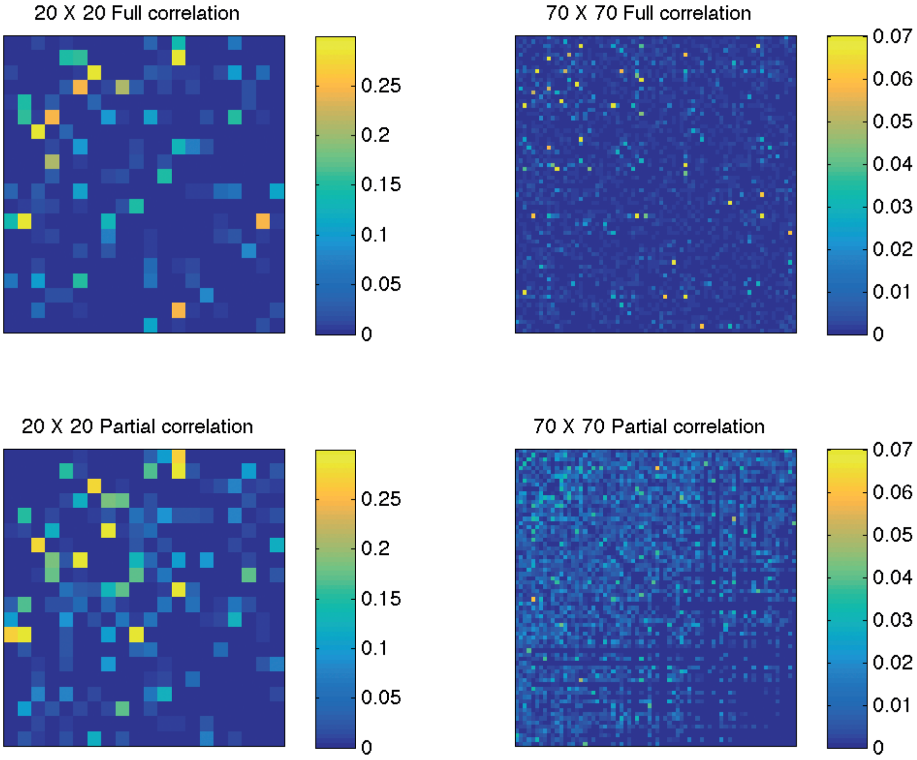


Figure S3. Mean beta values for the functional connectivity matrices. Beta values are averaged over the multiple cross-validation folds and multiple cross-validation repetitions.

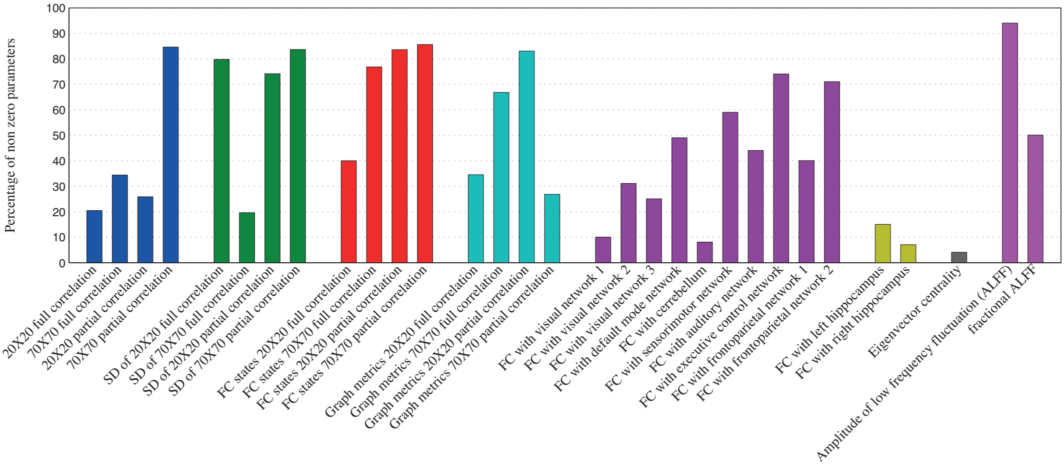


Figure S4. Mean percentage of non-zero parameters over all cross-validation folds and cross-validation repetitions for the 31 resting state fMRI measures.

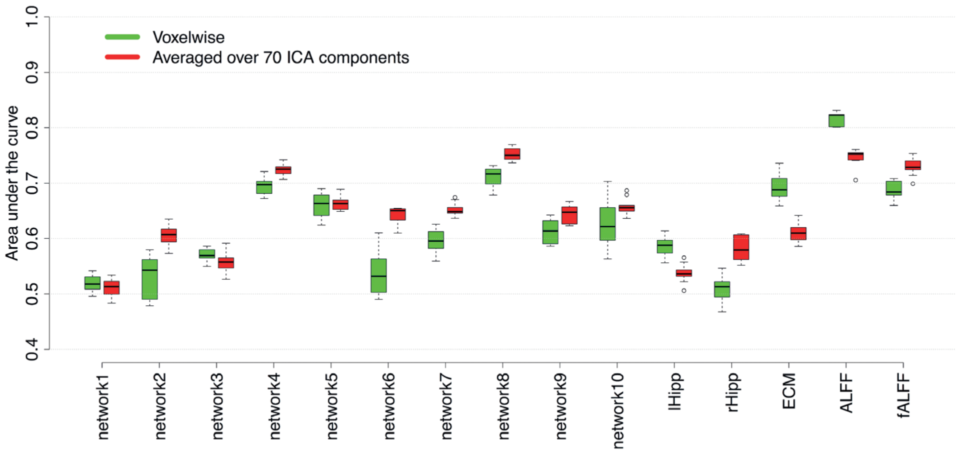
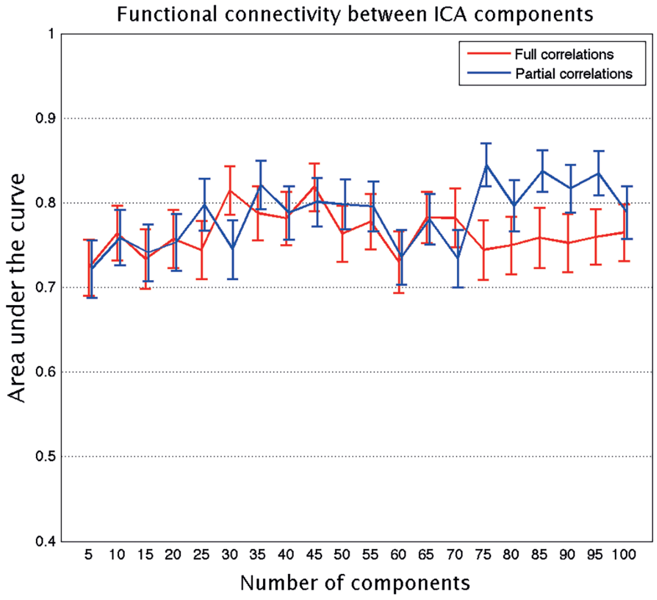


Figure S5. The effect of averaging over regions. The original classification results (green) and the results when the voxel-wise data is averaged over the 70 ICA components (red). The boxplots represent the different cross-validation repetitions.



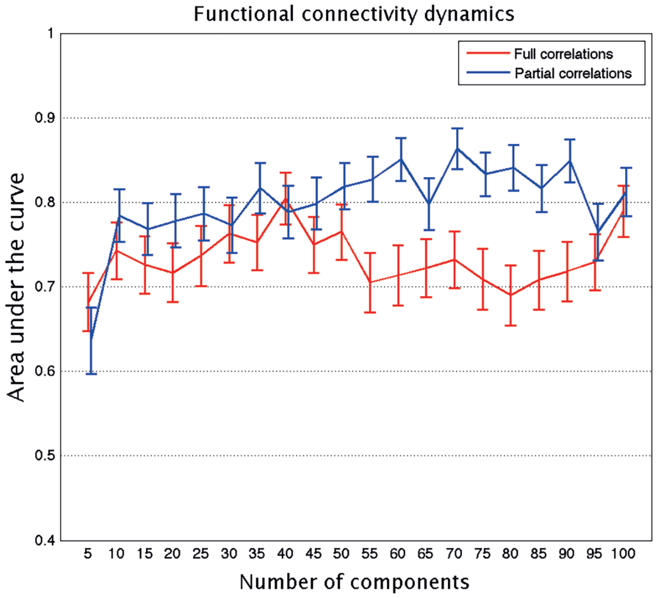


Figure S6. AUC values for the functional connectivity between ICA components (top), and functional connectivity dynamics (bottom) for a range of numbers of ICA components. The error bars represent one standard error above and below the AUC values.

Pre-trained MRI-based Alzheimer's disease classification models to classify memory clinic patients

Published in NeuroImage: Clinical, 2020; 27, 102303.

Frank de Vos, Tijn M. Schouten, Marisa Koini, Mark J.R.J Bouts, Rogier A. Feis,
Anita Lechner, Reinhold Schmidt, 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

Anatomical magnetic resonance imaging (MRI), diffusion MRI and resting state functional MRI (fMRI) have been used for Alzheimer's disease (AD) classification. These scans are typically used to build models for discriminating AD patients from control subjects, but it is not clear if these models can also discriminate AD in diverse clinical populations as found in memory clinics. To study this, we trained MRI-based AD classification models on a single-centre data set consisting of AD patients (N = 76) and controls (N = 173), and used these models to assign AD scores to patients with subjective memory complaints (SMC, N = 67), mild cognitive impairment (MCI) patients (N = 61), and AD patients (N = 61) from a multi-centre memory clinic data set. The anatomical MRI scans were used to calculate grey matter density, subcortical volumes and cortical thickness, the diffusion MRI scans were used to calculate fractional anisotropy, mean, axial and radial diffusivity, and the resting state fMRI scans were used to calculate functional connectivity between resting state networks and amplitude of low frequency fluctuations. Within the multi-centre memory clinic data set we removed scan site differences prior to applying the models. For all models, on average, the AD patients were assigned the highest AD scores, followed by MCI patients, and later followed by SMC patients. The anatomical MRI models performed best, and the best performing anatomical MRI measure was grey matter density, separating SMC patients from MCI patients with an AUC of 0.69, MCI patients from AD patients with an AUC of 0.70, and SMC patients from AD patients with an AUC of 0.86. The diffusion MRI models did not generalise well to the memory clinic data, possibly because of large scan site differences. The functional connectivity model separated SMC patients and MCI patients relatively good (AUC = 0.66). The multimodal MRI model did not improve upon the anatomical MRI model. In conclusion, we showed that the grey matter density model generalises best to memory clinic subjects. When also considering the fact that grey matter density generally performs well in AD classification studies, this feature is probably the best MRI-based feature for AD diagnosis in clinical practice.

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

4.1 Introduction

Early diagnosis of Alzheimer's disease (AD) is important, because it enables patients and caregivers to prepare for disease progression (Prince et al., 2011). It is also beneficial for drug research, because early phase AD patients are more likely to be susceptible to medication (Cummings et al., 2016). Whereas the diagnosis of progressed AD is feasible (Frisoni et al., 2010), early identification of AD is still problematic (Frisoni et al., 2017).

Amyloid and tau pathology are hypothesised to occur early in AD (Jack et al., 2011) and tau-PET and amyloid-PET are hypothesised earliest AD biomarkers (Blennow and Zetterberg, 2018). However, for clinical studies magnetic resonance imaging (MRI) scans are advantageous, because they are often available, they are non-invasive and they are relatively cheap. Further, functional MRI (fMRI) measures have been hypothesised to change in early AD as well (Buckner et al., 2005; Sperling et al., 2011).

MRI has been used to characterise brain changes that occur in AD. Most prominently, AD is characterised by grey matter atrophy, starting in the hippocampus (Morra et al., 2010), and later extending to other brain regions, including subcortical structures and the medial temporal lobe (Jack et al., 2004; Seeley et al., 2009). The location and extent of grey matter atrophy can be determined using anatomical MRI. Brain alterations in AD patients also involves white matter integrity (Douaud et al., 2011), which can be shown by diffusion MRI. In addition, AD patients show altered functional connectivity (FC) between brain regions (Agosta et al., 2012; Binnewijzend et al., 2012), measured using resting state functional MRI (fMRI).

However, these group differences are not necessarily useful in a clinical setting, since many AD markers have also been observed in healthy ageing (Salat et al., 1999). AD markers are only helpful in a clinical setting if they can accurately discriminate AD patients from non-affected subjects at the individual level. The focus of research on MRI biomarkers for AD has therefore shifted from the detection of group differences toward disease classification. MRI-based classification studies have progressed by using machine learning techniques, in which many predictors can be combined into one predictive model. This has led to good AD classification results for anatomical MRI (Cuingnet et al., 2011; Davatzikos et al., 2011; de Vos et al., 2016), diffusion MRI (Dyrba et al., 2013; Schouten et al., 2017) and resting state fMRI (Challis et al., 2015; Chen et al., 2011; de Vos et al., 2018). Moreover, combining these three MRI modalities can further improve the classification accuracy (Schouten et al., 2016).

Although these results are promising, MRI-based classification models still have to surmount at least two problems. First, most MRI-based AD classification studies have used scans of AD patients and healthy elderly controls, and other studies have used scans of mild cognitive impairment (MCI) patients to predict AD conversion (see for an overview Rathore et al., 2017). These models are trained specifically for these classification problems, but it is not clear whether these models can also discriminate AD in diverse clinical populations as found in memory clinics. It is thus important to evaluate the generalisability of MRI-based AD classification models to diverse clinical populations. Second, MRI scans are susceptible to scanner effects (Ewers et al., 2006; Takao et al., 2014; Zhu et al., 2011). This is problematic when a classification model is trained with MRI scans from one scanner, and applied to MRI scans from another scanner. To be clinically useful, AD classification models should be robust to scanner effects.

We will study to which extent MRI-based AD classification models generalise to a diverse patient population. This study is novel on 2 important points. Firstly, we will apply an AD classification model to a group of memory clinic patients, who are prone to AD. This is more clinically relevant than classifying AD from healthy controls, but also much more challenging. Second, we will use both anatomical MRI, diffusion MRI and resting state fMRI scans. This enables a comparison between these imaging modalities, and the use of a multimodal MRI classification model. We will use two different data sets. The first data set consists of AD patients and healthy controls, and will be used for training MRI-based AD classification models. These classification models will then be applied to the second data set, that consists of a diverse patient population collected in four different memory clinics. The memory clinic data set contains AD patients, MCI patients and patients with subjective memory complaints (SMC). We expect that AD patients will have a higher likelihood of being classified as AD patient than both other groups. Furthermore, we expect this to be higher for MCI patients than for SMC patients, because MCI is often an early stage of AD.

4.2 Methods

Participants

Training data

The training data were collected at the medical university of Graz in Austria, and consisted of 76 clinically diagnosed probable AD patients and 173 cognitively normal elderly controls (see Table 1). The AD patients were part of the prospective registry on dementia (PRODEM; see also Seiler et al., 2012). The inclusion criteria for PRODEM are: dementia diagnosis according to DSM-IV criteria (American Psychiatric Association, 2000), AD diagnosis according to the NINCDS-ADRDA Criteria (McKhann et al., 2011), non-institutionalisation or need for 24-h care, and the availability of a caregiver who agrees to provide information on the patients' and his or her own condition. Patients were excluded if co-morbidities were likely to preclude successful completion of the study. Informed consent was obtained from all patients and their caregivers. We only included patients for which anatomical MRI, diffusion MRI and resting state fMRI were available. The controls were scanned at the same scanning site, over the same period, with the same scanning protocol as the AD patients as a part of the Austrian stroke prevention study. The Austrian Stroke Prevention Study is a community-based cohort study on the effects of vascular risk factors on brain structure and function in elderly participants without a history or signs of stroke and dementia on the inhabitants of Graz, Austria (Schmidt et al., 1994; Freudenberger et al., 2016). Informed consent was obtained from all participants.

Memory clinic data

The memory clinic data (see Table 1) are part of the Leiden-Alzheimer research Nederland (LeARN) project (Handels et al., 2012; Jansen et al., 2017), and consisted of 61 possible or probable AD patients, 61 MCI patients and 67 SMC patients. The AD diagnosis was according to the NINCDS-ADRDA Criteria (McKhann et al., 2011), and the MCI diagnosis was according to the core clinical criteria for MCI due to AD (Albert et al., 2011). Subjects that did not meet the criteria for either AD or MCI were included in the SMC patient group. LeARN is a multi-centre collaboration of four memory clinics in the Netherlands; Leiden, Maastricht, Nijmegen and Amsterdam (see suppl. Table 1 for the demographics stratified over centre). 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 patient at least once a week. We only included patients for which anatomical MRI, diffusion MRI and resting state fMRI were available and excluded patients diagnosed with MCI not due to AD or dementia not due to AD (e.g., vascular dementia or frontotemporal dementia). Informed consent was obtained from both the patient and the informal caregiver.

MR acquisition

The subjects in the training data were scanned on a Siemens TrioTim 3T scanner at the Graz medical centre. The memory clinic subjects were scanned on a Philips Achieva 3T scanner at the Leiden University Medical Center, a Philips Achieva 3T scanner at the Maastricht University Medical Center, a Siemens TrioTim 3T scanner at the Nijmegen University Medical Center and a GE Signa HDxt 3T scanner at the VU university medical center in Amsterdam. The MRI sequence parameter settings are listed in Table 2.

Table 1. Sample demographics.

	Training data		Memory clinic data		
	Controls	AD patients	SMC	MCI	AD patients
N	173	76	67	61	61
Sex (♂/♀)	74/99	30/46	48/19	35/26	34/27
Age	66.1 ± 8.7	68.6 ± 8.6	63.2 ± 10.3	69.7 ± 8.3	72.5 ± 9.2
Years of education	11.5 ± 2.8	10.8 ± 3.2	11.2 ± 3.4	11.2 ± 3.4	10.6 ± 3.5
MMSE	27.5 ± 1.8	20.4 ± 4.5	28.2 ± 1.6	26.9 ± 2.3	24.0 ± 2.7
CDR	-	0.82 ± 0.34	0.34 ± 0.25	0.53 ± 0.15	0.78 ± 0.25
GDS	2.0 ± 2.4	2.7 ± 2.6	3.7 ± 2.8	3.0 ± 2.4	3.2 ± 2.8

Descriptives are presented as frequencies for the categorical variables and as mean ± standard deviation for the other variables. AD = Alzheimer's disease, SMC = Subjective memory complaints, MCI = Mild cognitive impairment, MMSE = mini-mental state examination, CDR = clinical dementia rating, GDS = geriatric depression scale.

Table 2. MRI sequence parameter settings per scan site.

	Slices	TR (ms)	TE (ms)	Flip angle (°)	Matrix size (voxels)	Voxel size (mm)		
anatomical MRI								
Graz	176	1900	2.2	9	256 x 256	1.00 x 1.00 x 1.00		
Leiden	180	9.8	4.6	8	288 x 288	0.78 x 0.78 x 1.00		
Maastricht	180	8.2	3.7	8	240 x 240	1.00 x 1.00 x 1.00		
Nijmegen	192	2300	4.7	12	256 x 256	1.00 x 1.00 x 1.00		
Amsterdam	176	7.8	3.0	12	256 x 256	0.94 x 0.94 x 1.00		
diffusion MRI							Directions[a]	b0 scans
Graz	50	6700	95	90	125 x 125	2.00 x 2.00 x 2.50	12[b]	4
Leiden	70	8250	80	90	128 x 128	2.00 x 2.00 x 2.00	61	1
Maastricht	70	8250	80	90	128 x 128	2.00 x 2.00 x 2.00	61	1
Nijmegen	81	13000	102	90	128 x 128	2.00 x 2.00 x 2.00	30	1
Amsterdam	45	13000	94	90	128 x 128	2.00 x 2.00 x 2.40	30	1
resting state fMRI							Volumes	
Graz	40	3000	30	90	64 x 64	3.00 x 3.00 x 3.00	150	
Leiden	38	2200	30	80	80 x 80	2.75 x 2.75 x 3.00	200	
Maastricht	38	2200	30	80	112 x 112	2.00 x 2.00 x 2.50	200	
Nijmegen	49	2380	30	90	64 x 64	3.50 x 3.50 x 3.50	110	
Amsterdam	34	1800	35	80	64 x 64	3.30 x 3.20 x 3.00	202	

[a] All diffusion directions were acquired with a b value of 1000

[b] The diffusion directions were acquired four times

MRI preprocessing

The MRI data of all subjects were 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

We used both the FSL and Freesurfer software packages to analyse the anatomical MRI scans, because they have different approaches to calculate measures of grey matter atrophy. These approaches are complementary to each other, and combining them improves the accuracy of AD classification (de Vos et al., 2016).

Grey matter density

We used MRI-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 subject.

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 subject (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 subject.

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 subject. Then we projected subjects' 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

Functional connectivity (FC) was calculated between resting state networks (RSNs) as obtained by an independent component analysis (ICA). First, we used only the training sample to obtain 70 RSNs using temporal concatenation ICA in FSL MELODIC (Beckmann & Smith, 2004). Then, for all subjects we registered the ICA component weight maps to subject space, weighted them by the subject 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 $\lambda = 100$ (Smith et al., 2011). For each participant we thus calculated a 70 by 70 sparse partial correlation matrix yielding (70 * 69)/2 = 2415 features.

Amplitude of low frequency fluctuations

To calculate the amplitude of low frequency fluctuations (ALFF; Biswal et al., 2010; Zang et al., 2007), we used the REST software package (Song et al., 2011). ALFF was defined as the power within the 0 - 0.1 Hz frequency band. For standardisation purposes we divided the voxels' ALFF values by the mean ALFF within a subjects' whole brain (Zang et al., 2007). The whole brain voxel-wise ALFF maps consist of 139,712 values.

Correction for age

We regressed out the age effects from the features. To this end we first used the healthy controls from the training sample to estimate 'normal' age effects for all features. Then we used these estimated age effects to regress out the age effects for all subjects.

Correction for scan site within the memory clinic data

We corrected for scan site effects within the memory clinic data 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, making the assumption that sites have both an additive and multiplicative effect on the data. It uses empirical Bayes to improve the estimation of the model parameters. The model furthermore makes the assumption that the expected value of a feature can be modelled by both the site

Table 3. 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
Amplitude of low frequency fluctuations	139,712

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, and MMSE score as factors in the ComBat model.

We did not correct for scan site differences between the training data and the memory clinic data, because the training data consists of different clinical labels (healthy controls and probable AD) than the memory clinic data (SMC, MCI and possible/probable AD). It is therefore not possible to decide whether differences between these data sets should be attributed to scan site differences, or to differences in clinical groups.

Statistical analyses

The nine different MRI feature groups, along with the number of features per group are listed in Table 3. These feature groups were used separately in nine different AD classification models, and combined into an anatomical MRI, diffusion MRI, resting state fMRI and multimodal AD classification model. All features were normalised prior to the statistical analyses.

Penalised logistic regression within the training data

The training data was used to fit AD classification models. We used logistic regression to predict the true class of the subjects. In logistic regression, the outcome variable is dichotomous (0 for healthy controls and 1 for AD patients), and the predicted scores are continuous between 0 and 1. The subjects' predicted scores are adopted as AD scores. To prevent overfitting, we used penalised logistic regression techniques that put penalties on the regression weights, such that only the most relevant features enter the regression model. For the separate feature groups, we used elastic net logistic regression (Friedman et al., 2010; Zou & Hastie, 2005), that uses a combination of an L1 (LASSO; Tibshirani, 1996) and L2 (Ridge; Hoerl and Kennard, 1970) penalty. The L1 penalty tends towards sparse models, including only few features. The L2 penalty tends to include all features, but limits the size of their contributions. Two hyperparameters need to be tuned: the α parameter determines the relative weight of the two different penalties, and λ determines the size of those penalties. For the combined models we used group lasso logistic regression (Simon et al., 2013), which uses an L1 penalty on feature groups and an L2 penalty within the feature groups. The group lasso thereby improves interpretation of the AD classification model, because the L1 penalty on feature groups either entirely includes or excludes feature groups. For the group lasso we only need to tune λ: the size of the penalties.

Cross-validation within the training data

To determine the performance of the AD classification models within the training data, we used nested cross-validation (Krstajic et al., 2014). Nested cross-validation takes into account two potential sources of overfitting. One could either include too many predictors, or overestimate accuracy by looping over all the values of the hyperparameters and only pick the best result. 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 AD classification 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 subjects into folds.

Application to memory clinic data

To determine the performance of the AD classification models on the memory clinic data, we fitted AD classification models on the entire training data using optimal hyperparameter settings. These optimal hyperparameters were determined using a single tenfold cross-validation. The resulting regression models were directly applied to the MRI features of the memory clinic subjects. This yielded AD scores for the memory clinic subjects.

Model evaluation

To evaluate the results, we made receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) as a measure of classification performance. The AUC is invariant to the class distribution (Bradley, 1997), which is an advantage, because within the training data the number of control subjects is larger than the number of AD patients. Within the training data we compared the healthy controls with the AD patients, and within the memory clinic data we pairwise compared the SMC patients, MCI patients and AD patients. The four different patient comparisons, for the nine feature groups plus four combined models, yielded 52 comparisons in total. To test the AUC values against chance, we used a permutation procedure with 10.000 permutations. We combined all 52 comparisons within the same permutation procedure to correct for multiple comparisons. For each permutation we permuted the subjects' labels, and calculated the AUC value for all 52 comparisons. We only registered the maximum of those 52 AUC values, resulting in a permutation distribution of maximum AUC values. The 52 observed AUC values were compared with this distribution, yielding family-wise error corrected *p*-values.

In addition, we calculated sensitivity, specificity, positive predictive values and negative predictive values. We used a cut-off score of 0.5, such that Subjects with Alzheimer's scores below 0.5 were classified as the less severe disease category, and subjects with Alzheimer's scores above 0.5 were classified as the more severe disease category. For example, in the comparison of SMC patients and MCI patients, the former is regarded as the less severe disease category and the latter is regarded as the more severe disease category. To evaluate the classification models in the memory clinic data, 0.5 is not necessarily the optimal cut-off score. For example, the SMC patients and MCI patients are not expected to receive Alzheimer's scores close to either 0 or 1. Consequently, a cut-off score of 0.5 sometimes yields high sensitivity values and low specificity values, or the other way around. In these cases, other cut-off scores might result in a better balance between sensitivity and specificity. We have nevertheless used a fixed cut-off score of 0.5, because it eases the interpretation.

4.3 Results

Correction for scan site

We applied scan site correction to the four memory clinic centres (Fig. S1). 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. These site effects have been removed using the ComBat procedure, leaving no visible site effects between the four memory clinic centres afterwards. We did not correct for scan site differences between the training data and the memory clinic data, because the training data consists of different clinical labels (healthy controls and probable AD) than the memory clinic data (SMC, MCI and possible/probable AD). It is therefore not possible to decide whether differences between these data sets are due to scan site differences, or to differences in clinical groups. The differences between the training data and the corrected test data are largest for the diffusion MRI measures.

Classification results

The single feature classification models and the multiple feature classification models yielded individual AD scores for all participants (Fig. 2 and Fig. 3 respectively). To evaluate these classification models, we calculated AUC values (Table 4), sensitivity and specificity values (Table 5) and positive predictive values and negative predictive values (Table S2).

Training data classification using single features

The median AD score for the AD patients is higher than those of the healthy controls for all single feature classification models (Fig. 2, top row). The AUC values for discriminating between AD patients and controls range between 0.79 for FC and 0.92 for cortical thickness. These AUC values are all above chance level, showing that the classification models work well within the training data itself (Table 4, left side).

Memory clinic data classification using single features

All models, except for the ALFF model, assigned the highest median AD score to the AD patients, followed by the MCI patients and later followed by the SMC patients (Fig. 2, bottom row). The AUC values for the pairwise discrimination between these three groups are depicted in the right side of Table 4. The discrimination between SMC patients and MCI patients is above chance level for grey matter density and FC. The discrimination between MCI patients and AD patients is above chance level for grey matter density, subcortical volumes, and cortical thickness. The discrimination between SMC patients and AD patients is above chance level for grey matter density, subcortical volumes, cortical thickness, FA, MD, DA, and FC (Table 4, right side).

Training data classification using multiple features

In order to increase classification accuracy, the feature groups were combined into an anatomical MRI, diffusion MRI, resting state fMRI, and multimodal MRI model. For all combined classification models, the median AD score for the AD patients is higher than those of the healthy controls (Fig. 3, top row). The AUC values for discriminating between AD patients and controls are higher for the combined models than those for the single feature models. The multimodal model does however not improve upon the combined anatomical MRI model (Table 4, left side).

Memory clinic data classification using multiple features

The combined classification models were also applied to the memory clinic data. All models assigned the highest median AD score to the AD patients, followed by the MCI patients and later followed by the SMC patients (Fig. 3, bottom row). In contrast to the training data, the AUC values of the combined models are most often not higher than the AUC value of the best discriminating single feature group. The AUC only increases when combining the diffusion MRI features in order to classify SMC patients and MCI patients. For all other combined models, the AUC is either the same or lower (Table 4, right side).

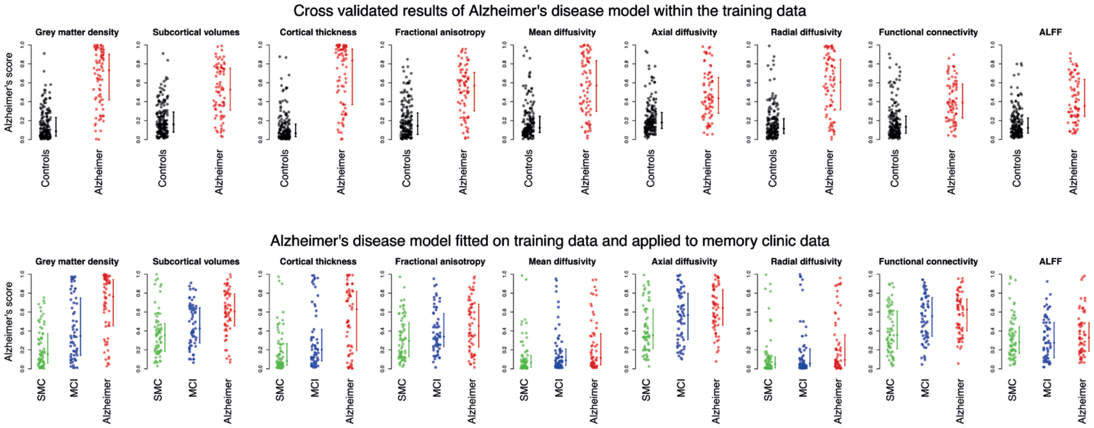


Figure 1. Alzheimer's disease scores for the feature groups. The top row shows the results on the training data, and the bottom row shows the results on the memory clinic data. The error bars represent the median AD score and the interquartile range. SMC = subjective memory complaints, MCI = mild cognitive impairment, ALFF = amplitude of low frequency fluctuations.

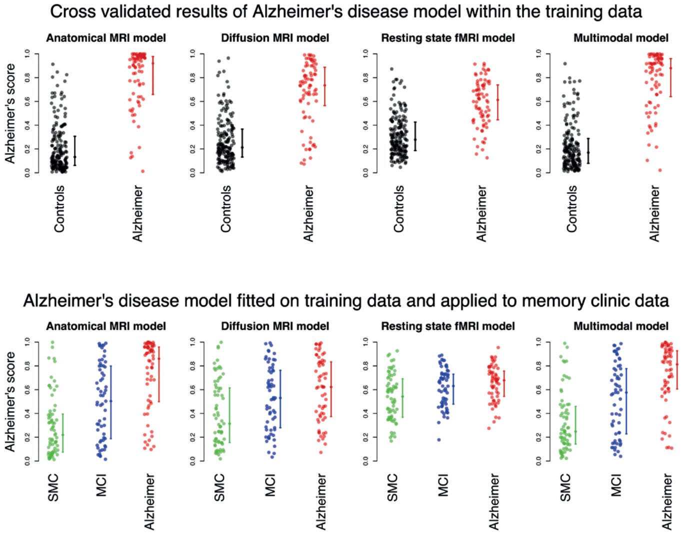


Figure 2. Alzheimer's disease scores for the combined models. The top row shows the results on the training data, and the bottom row shows the results on the memory clinic data. The error bars represent the median AD score and the interquartile range. SMC = subjective memory complaints, MCI = mild cognitive impairment, ALFF = amplitude of low frequency fluctuations.

Table 4. AUC values for the different MRI-based AD classification models.

	Training data	Memory clinic data		
MRI measure	HC vs AD	SMC vs MCI	MCI vs AD	SMC vs AD
Grey matter density	0.91***	0.69**	0.70**	0.86***
Subcortical volumes	0.82***	0.62	0.66*	0.76***
Cortical thickness	0.92***	0.64	0.66*	0.76***
Combined anatomical MRI	**0.94*****	**0.69****	**0.70****	**0.85*****
Fractional anisotropy	0.83***	0.60	0.57	0.65*
Mean diffusivity	0.84***	0.62	0.55	0.66*
Axial diffusivity	0.81***	0.63	0.58	0.72***
Radial diffusivity	0.85***	0.58	0.57	0.64
Combined diffusion MRI	**0.87*****	**0.65***	**0.57**	**0.71*****
Functional connectivity	0.79***	0.66*	0.54	0.71**
ALFF	0.81***	0.49	0.56	0.55
Combined resting state fMRI	**0.85*****	**0.62**	**0.56**	**0.68****
Multimodal MRI	**0.94*****	**0.68****	**0.69****	**0.84*****

HC = healthy controls, AD = Alzheimer's disease, SMC = Subjective memory complaints, MCI = Mild cognitive impairment, ALFF = amplitude of low frequency fluctuations.
$^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

Table 5. Sensitivity / specificity values for the different MRI-based AD classification models.

	Training data	Memory clinic data		
MRI measure	HC vs AD	SMC vs MCI	MCI vs AD	SMC vs AD
Grey matter density	0.66 / 0.96	0.39 / 0.84	0.69 / 0.61	0.69 / 0.84
Subcortical volumes	0.51 / 0.91	0.46 / 0.78	0.70 / 0.54	0.70 / 0.78
Cortical thickness	0.70 / 0.95	0.25 / 0.91	0.54 / 0.75	0.54 / 0.91
Combined anatomical MRI	**0.88 / 0.89**	**0.51 / 0.81**	**0.74 / 0.49**	**0.74 / 0.81**
Fractional anisotropy	0.61 / 0.90	0.28 / 0.75	0.46 / 0.72	0.46 / 0.75
Mean diffusivity	0.57 / 0.92	0.11 / 0.96	0.20 / 0.89	0.20 / 0.96
Axial diffusivity	0.41 / 0.92	0.59 / 0.69	0.69 / 0.41	0.69 / 0.69
Radial diffusivity	0.62 / 0.94	0.11 / 0.90	0.21 / 0.89	0.21 / 0.90
Combined diffusion MRI	**0.75 / 0.87**	**0.56 / 0.63**	**0.61 / 0.44**	**0.61 / 0.63**
Functional connectivity	0.34 / 0.94	0.52 / 0.64	0.59 / 0.48	0.59 / 0.64
ALFF	0.39 / 0.96	0.25 / 0.79	0.25 / 0.75	0.25 / 0.79
Combined resting state fMRI	**0.67 / 0.86**	**0.69 / 0.45**	**0.82 / 0.31**	**0.82 / 0.45**
Multimodal MRI	**0.84 / 0.88**	**0.56 / 0.81**	**0.84 / 0.44**	**0.84 / 0.81**

HC = healthy controls, AD = Alzheimer's disease, SMC = Subjective memory complaints,
MCI = Mild cognitive impairment, ALFF = amplitude of low frequency fluctuations.

Feature group importance

In order to inspect the contribution of the feature groups to the combined models, we plotted their beta values (Fig. 4). The anatomical MRI model takes all three anatomical feature groups into account, and the largest weight is assigned to cortical thickness. The diffusion MRI model takes FA, DA and DR into account, and disregards MD. The largest weight is assigned to DR. The resting state fMRI model takes both FC and ALFF into account, but weighs FC more heavily. The multimodal MRI model relies mostly on the anatomical MRI features, but also includes the DR features.

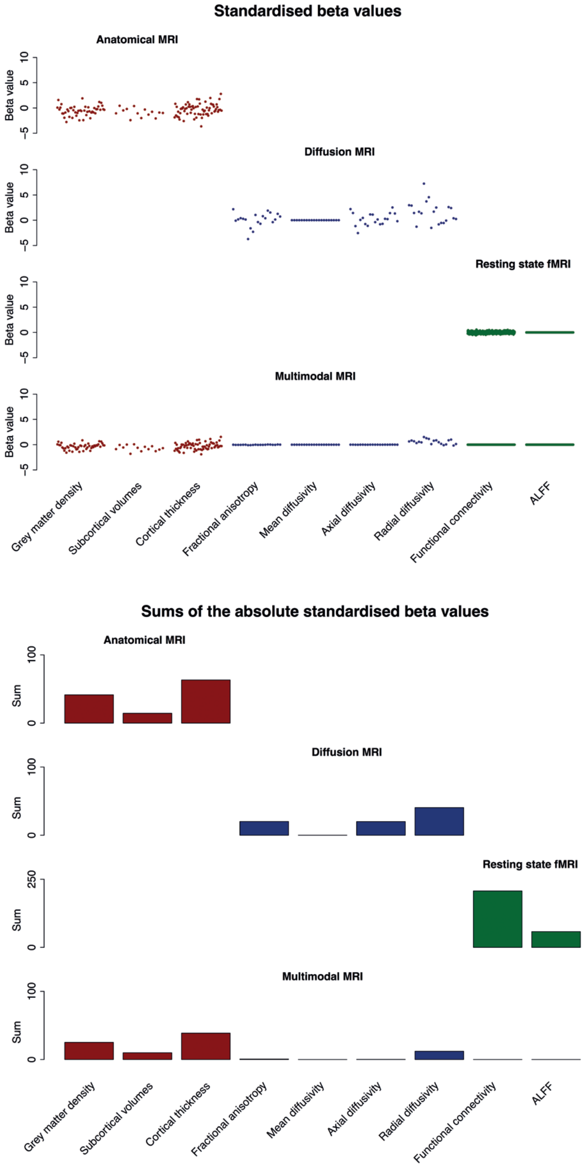


Figure 3. Content of the combined classification models that were fitted on the training data and applied to the memory clinic data. The top panel shows the standardised beta values of the features, and the bottom panel shows the sums of the absolute standardised beta values per feature group. These plots illustrate the importance of the feature groups for the combined models. The anatomical MRI model takes all three anatomical feature groups into account, the diffusion MRI model takes FA, DA and DR into account, the resting state fMRI model takes both functional connectivity and ALFF into account, and the multimodal MRI model relies mostly on the anatomical MRI features. ALFF = amplitude of low frequency fluctuations.

4.4 Discussion

In this study, we evaluated the generalisability of MRI-based AD classification models. To this end, we used a single-centre training data set consisting of AD patients and healthy controls, and a multi-centre application data set consisting of AD patients, MCI patients and SMC patients. First, we showed that within the training data there is good classification performance for both the anatomical MRI, diffusion MRI and resting state fMRI models. When a model was trained on one part of the training data, it generalised well to the other part of the training data. Second, we fitted models on the entire training data, and applied those models to the memory clinic data, resulting in AD scores for the memory clinic subjects. As expected, for all three MRI modalities, the AD patients were on average assigned higher AD scores than MCI patients, and the MCI patients were on average assigned higher AD scores than SMC patients.

There is however large variation in the performance of the different MRI models. The anatomical MRI models generalised best to the memory clinic data. Especially the grey matter density model could differentiate well between all three clinical groups. The cortical thickness model and the subcortical volumes model could differentiate between the AD patients and the other two groups, but not between the SMC patients and MCI patients.

The diffusion MRI models did not perform as well as the anatomical MRI models. Although classification performance was excellent within the training data for all diffusion MRI measures, there was limited generalisation to the memory clinic data. Possibly, this is due to the fact that white matter alterations in AD mostly occur in the late phase of the disease (Clerx et al, 2012). So, white matter changes might be already present in the probable AD patients from the training data, but these changes might not yet be as large in the MCI patients or possible AD patients from the memory clinic data. Another explanation might lie in the scan site differences for the diffusion MRI measures. It is known that technical variabilities across scan sites can have large effects on diffusion MRI scans (Zhu et al., 2011), and also in the current study the four memory clinic centres largely differed on the diffusion MRI measures. These site differences were removed as much as possible using the ComBat procedure (Fortin et al., 2017; Johnson et al., 2007), but they cannot be removed entirely. Furthermore, we did not remove scan site differences between the training data and the memory clinic data, because the subjects within the training data are not comparable with the memory clinic subjects with regard to their clinical labels. It is therefore not possible to decide whether differences between these data sets should be attributed to scan site differences, or to differences in clinical labels. Yet, it is likely that scan site differences exist between the training data and the memory clinic data, and that possibly they have affected the AD scores of the memory clinic subjects. Diffusion MRI have nevertheless been used successfully in a multi-centre AD classification study (Dyrba et al., 2013). However, this study only used probable AD patients and healthy elderly controls, for which differences in white matter are expected to be larger. Furthermore, they used subjects from nine different scan sites, and they achieved the highest accuracy when training and testing was partly done on subjects from the same site. When they trained the model on subjects from eight scan sites, and applied this model on subjects from the ninth scan site, this resulted in lower accuracy.

Regarding the resting state fMRI models, there is a large difference between the FC model and the ALFF model. The FC model is somewhat inferior compared to the structural and diffusion MRI measures within the training data, but it generalises reasonably well to the memory clinic data. This model can differentiate between SMC patients and MCI patients, and between SMC patients and AD patients. The reasonably good generalisation performance of the FC model might partly be explained by the absence of large scan site differences. In addition, alterations in FC likely start in an early phase of AD (Buckner et al., 2005; Sperling et al., 2011), and this might explain why this model could distinguish reasonably well between SMC patients and MCI patients. FC has previously been shown to be successful for the classification of AD patients, MCI patients and controls in a multi-centre setting. However, this was only achieved after employing strict quality measures, including visual inspection of all the data (Teipel et al., 2017b). In the current study this was not much of an issue, possibly because we automatically removed noise components with ICA-AROMA (Pruim et al., 2015), and it has been shown that removing ICA based noise components from resting state fMRI data reduces scan site differences substantially (Feis et al, 2015). In contrast to the FC model, the ALFF model showed very poor generalisation performance. Although the classification performance was good within the training data, this model could not differentiate between any of the three groups within the memory clinic data. This result corresponds to the results of another multi-centre study, in which ALFF showed poor classification performance to classify SMC patients, amnestic MCI patients and AD patients (Teipel et al., 2018).

Combining the MRI features improved the accuracy within the training data, which is a replication of other studies that improved AD classification by combining different

MRI measures from the same imaging modality (de Vos et al., 2016, 2018; Westman et al., 2013), or combining multiple imaging modalities (Dai et al., 2012; Schouten et al., 2016). More importantly, however, this improvement did not translate to the memory clinic data. Some features contributed largely to the combined models, because they had a beneficial effect on AD classification within the training data, but they worsened the results of the combined model on the memory clinic data, because those features did not generalise to the memory clinic data. For example, the combined resting state fMRI model included both FC and ALFF. Within the training data, this combination increased accuracy compared to both of these features alone. However, within the memory clinic data, this combination decreased accuracy compared to using only FC. Probably, this is caused by the poor generalisation performance of ALFF.

The classification accuracies within the memory clinic data were substantially lower than those within the training data for all MRI models. These differences can be caused by multiple factors, and we cannot explicitly attribute these differences to any of these different factors. A factor that has likely been important is the difference in clinical populations. It is easier to distinguish AD patients from healthy elderly controls, as in the training data, than to distinguish AD patients from MCI patients and SMC patients, as in the memory clinic data. In addition, the AD patients in the training data had lower average MMSE scores than the AD patients in the test memory clinic data. The AD patients in the training data were thus clinically more progressed than the AD patients in the memory clinic data. Other factors that might have caused a drop in accuracy from training to test set are scan site differences, differences caused by confounding variables (e.g., age, sex or education) and overfitting on the training data.

We have focused on MRI scans for the AD classification models, although MRI-visible structural and volumetric brain abnormalities occur relatively late in AD (Jack et al., 2010). Amyloid and tau pathology are observable in AD patients well before any pathological change is detectable on a anatomical MRI scan (Jack et al., 2010). For clinical studies however, anatomical MRI scans are advantageous, because they are non-invasive and often available. In addition, there is evidence that functional changes as can be seen on a resting state fMRI scan might already occur in an earlier phase of the disease (Buckner et al., 2005; Sperling et al., 2011). Therefore, resting state fMRI might be sensitive for early detection of AD.

We have only studied AD classification, while memory clinics are confronted with non-AD types of dementia as well. In future efforts, to create clinically valuable classification models for more dementia types, it is important to also include non-AD types of dementia.

In conclusion, we studied the generalisation performance of single-centre MRI-based AD classification models to a multi-centre memory clinic data set. The anatomical MRI models generalised best to the memory clinic data, and grey matter density was the best performing anatomical MRI measure. The diffusion MRI models did not generalise well, possibly due to large scan site effects on the diffusion MRI measures, or because white matter alterations mostly occur in progressed AD (Clerx et al, 2012). The FC model showed reasonable performance for identifying prodromal AD stages, but it was still inferior to the grey matter density model. Moreover, the multimodal MRI model did not improve upon the anatomical MRI model.

4.5 Funding

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

Supplementary table 1. Sample demographics for the memory clinic centers.

	Memory clinic data			
	Leiden	Maastricht	Nijmegen	Amsterdam
N	40	68	43	38
SMC / MCI / AD	12 / 13 / 15	31 / 24 / 13	16 / 13 / 14	8 / 11 / 19
Sex (♂/♀)	20 / 20	42 / 26	25 / 18	30 / 8
Age	70.9 ± 9.0	66.6 ± 11.6	71.6 ± 9.0	65.0 ± 7.5
MMSE	26.4 ± 2.5	27.5 ± 2.6	25.7 ± 2.8	25.4 ± 3.1
CDR	0.59 ± 0.30	0.52 ± 0.17	0.49 ± 0.37	0.63 ± 0.33
GDS	3.9 ± 3.3	3.3 ± 2.6	2.8 ± 1.8	3.5 ± 2.9

Descriptives are presented as frequencies for the categorical variables and as mean _ standard deviation for the other variables. SMC = Subjective memory complainers, MCI = Mild cognitive impairment, AD = Alzheimer's disease, MMSE = mini mental state examination, CDR = clinical dementia rating, GDS = geriatric depression scale.

Supplementary table 2. Positive predictive values / negative predictive values for the different MRI-based AD classification models.

	Training data	Memory clinic data		
	HC vs AD	SMC vs MCI	MCI vs AD	SMC vs AD
Grey matter density	0.81 / 0.88	0.61 / 0.69	0.65 / 0.64	0.76 / 0.79
Subcortical volumes	0.71 / 0.72	0.62 / 0.65	0.62 / 0.61	0.74 / 0.74
Cortical thickness	0.83 / 0.85	0.58 / 0.71	0.65 / 0.69	0.73 / 0.85
Combined anatomical MRI	0.88 / 0.78	0.66 / 0.70	0.61 / 0.59	0.77 / 0.78
Fractional anisotropy	0.75 / 0.73	0.51 / 0.50	0.59 / 0.62	0.60 / 0.62
Mean diffusivity	0.74 / 0.76	0.53 / 0.70	0.54 / 0.63	0.58 / 0.80
Axial diffusivity	0.67 / 0.70	0.64 / 0.63	0.55 / 0.54	0.69 / 0.67
Radial diffusivity	0.77 / 0.81	0.51 / 0.50	0.55 / 0.65	0.55 / 0.65
Combined diffusion MRI	0.81 / 0.71	0.59 / 0.58	0.52 / 0.52	0.62 / 0.60
Functional connectivity	0.63 / 0.69	0.58 / 0.57	0.53 / 0.53	0.62 / 0.60
ALFF	0.68 / 0.82	0.52 / 0.52	0.50 / 0.50	0.52 / 0.52
Combined rs-fMRI	0.76 / 0.67	0.57 / 0.53	0.57 / 0.54	0.63 / 0.57
Combined multimodal MRI	0.86 / 0.76	0.68 / 0.72	0.64 / 0.60	0.82 / 0.80

HC = healthy controls, AD = Alzheimer's disease, SMC = Subjective memory complainers, MCI = Mild cognitive impairment, ALFF = amplitude of low frequency fluctuations.

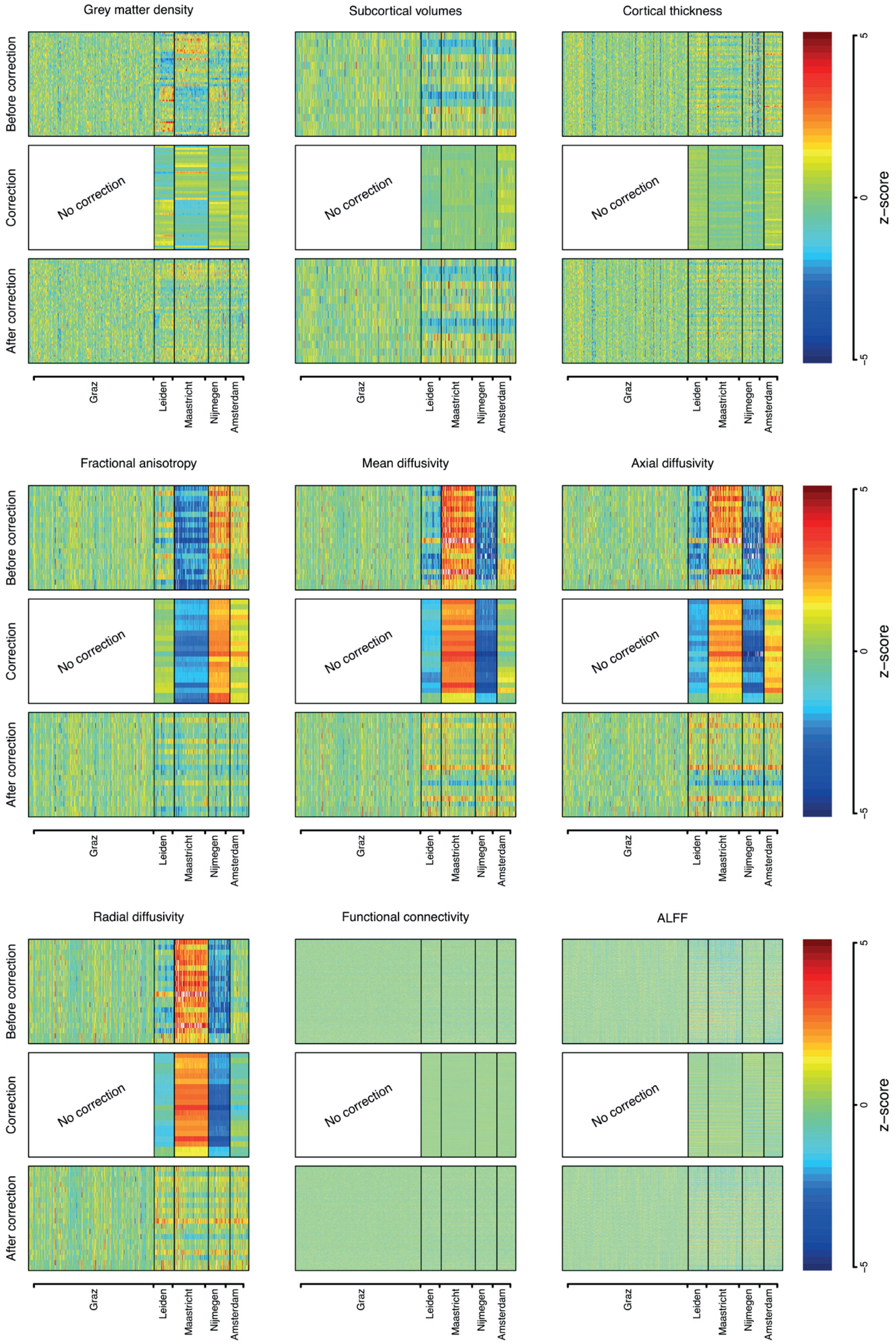

Grey matter density
Subcortical volumes
Cortical thickness
Fractional anisotropy
Mean diffusivity
Axial diffusivity
Radial diffusivity
Functional connectivity
ALFF
Before correction
Correction
After correction
No correction
Graz
Leiden
Maastricht
Nijmegen
Amsterdam
5
0
-5
z-score

Figure S1. The effect of scan site correction within the four Dutch memory clinic centres. The top row shows the feature values before scan site correction, the middle row shows the amount of scan site correction, and the bottom row shows the feature values after scan site correction. The matrix rows represent the features, and the matrix columns represent the subjects. Before correction there are large site effects, but after correction these site effects have largely disappeared. We did not apply scan site correction to the training data from Graz, because that data set consists of different clinical labels (healthy controls and probable AD) than the memory clinic data (SMC, MCI and possible/probable AD). It is therefore not possible to decide whether differences between these data sets should be attributed to scan site differences, or to differences in clinical groups.

5

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 $\lambda = 100$ (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

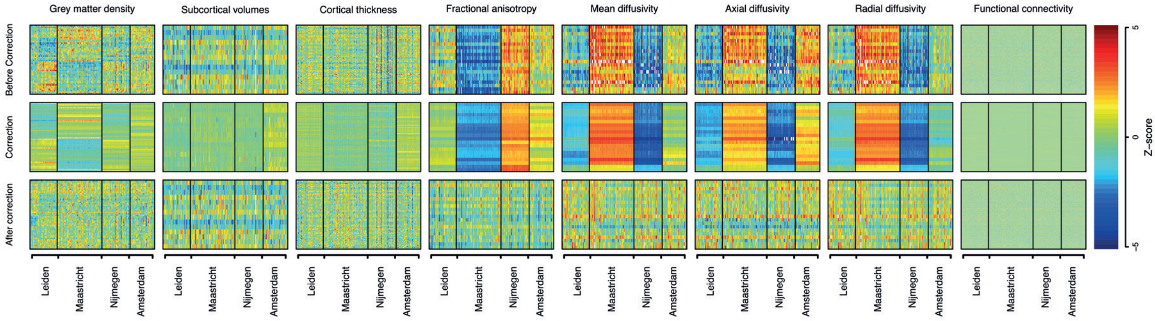


Figure 1. Scan site correction for the MRI data. The top row shows the feature values before scan site correction, the middle row shows the amount of scan site correction that was applied, and the bottom row shows the feature values after scan site correction. The matrix rows represent the features, 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

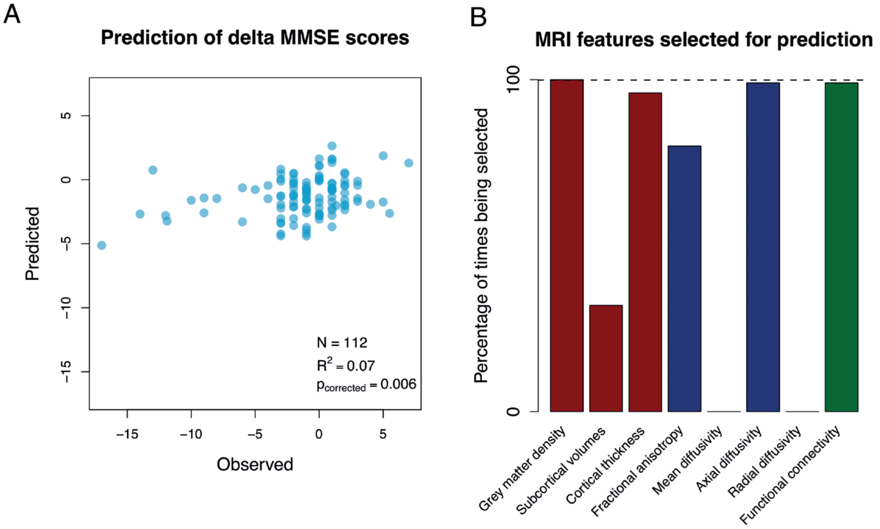


Figure 2. Observed versus predicted change scores on the mini-mental state examination (MMSE) after a two-year follow-up period. Negative scores denote decline and positive scores denote improvement (**panel A**). MRI features used for the prediction of MMSE decline. The group lasso regression method 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).

6

General discussion

6.1 Summary of results

In this thesis we studied magnetic resonance imaging based (MRI-based) Alzheimer's disease (AD) classification. We derived a wide range of features from anatomical MRI, diffusion MRI and resting state functional MRI (fMRI) scans. This is important, because MRI scans contain a wealth of information that might help clinicians to improve AD diagnosis. However, anatomical MRI scans are usually only used to inspect specific AD characteristics like the size and shape of the hippocampus, and diffusion MRI and resting state fMRI scans are not at all used in clinical practice. We also studied the use of MRI scans for early identification of AD, and for future prediction of cognitive decline. This is important, because there is need for reliable AD diagnosis at the early symptomatic phase, or ideally even before the onset of the disease (Frisoni et al., 2017). This would provide patients with a prognosis so they can prepare for their future trajectory. In addition, early phase AD diagnosis is important for drug research, because early phase AD patients might still be susceptible for drugs (Scheltens et al., 2016).

We have tried to answer several hypotheses. In **chapter two** we used anatomical MRI scans to calculate multiple anatomical features, and used those to classify AD patients and healthy elderly controls. The grey matter density of the cortical structures and the volumes of subcortical structures yielded the highest classification scores. Moreover, combining multiple types of features increased accuracy. To put these results in perspective, we compared them with hippocampal volume and whole brain atrophy, because these are well established AD biomarkers. The combination of different anatomical MRI features yielded higher accuracies than these two simpler measures. This suggests that anatomical MRI scans contain more information than can be captured by simple measures only, and extracting this extra information benefits AD classification. These improvements are substantial, and might be considered for clinical use.

In **chapter three** we took a similar approach to resting state fMRI scans. We used those scans to calculate multiple feature types, and to classify AD patients and healthy elderly controls. We calculated functional connectivity (FC) using multiple approaches, like large scale connectivity matrices and whole brain voxel-wise seed correlation analyses. In addition, we derived indirect FC measures, including FC dynamics and graph measures. We also quantified the amplitude of low frequency fluctuations that reflect neuronal signal. The FC matrices, FC dynamics and the amplitude of low frequency fluctuations yielded the highest classification scores. Also here, The combination of different types of features increased accuracy.

In **chapter four** we evaluated our MRI-based models for individual AD classification in diverse clinical populations from various memory clinics. This is an important step towards clinical applicability, because it reflects the day-to-day challenge of clinicians. This effort is difficult for at least two reasons. First, our AD classification models are developed using only AD patients and healthy elderly controls, and therefore they do not capture the full heterogeneity found in diverse clinical populations. Second, MRI scans collected at different sites can be largely different due to different scanner vendors and scanner settings (Ewers et al., 2006; Takao et al., 2014; Zhu et al., 2011). Consequently, MRI-based AD classification models do not always translate well to MRI scans acquired from another scanner.

To evaluate our MRI-based AD classification models for diverse clinical populations, we trained them on a single-centre data set consisting of AD patients and controls, and used them to assign AD probability scores to AD patients, mild cognitive impairment (MCI) patients and SMC patients from a multi-centre memory clinic data set. We used models based on anatomical MRI, diffusion MRI and resting state fMRI scans. For all three MRI modalities, the AD patients were on average assigned higher AD probability scores than MCI patients, and the MCI patients were on average assigned higher AD probability scores than SMC patients. However, for some models the three clinical groups show large overlap in AD probability scores. The anatomical MRI models generalised best to the memory clinic data. Especially the grey matter density model could differentiate well between all three groups. However, the diffusion MRI model did not generalise well to the multi-centre memory clinic data set, probably due to large scanner site differences in the diffusion MRI data. The resting state fMRI model generalised reasonably well to the memory clinic data, but it was inferior to the anatomical MRI model.

In **chapter five** we evaluated MRI's predictive accuracy for future cognitive decline. To this end, baseline multimodal MRI scans were used to predict two-year follow-up cognitive decline. Change in general cognitive functioning as measured by the mini-mental state examination (MMSE) was predicted above chance level, but the effect size was small. Change on six other tests, that measure specific aspects of memory and executive functioning, was not predicted above chance level. Our effort was complicated by a lack of observed cognitive decline. At average, there was significant decline on only three out of seven cognitive tests, and the amount of decline for these tests was little. The lack of observed cognitive decline may have been caused by the large drop-out rate. Patients that decline more are probably more likely to drop out.

6.2 Strengths and limitations

Comprehensive comparison of MRI measures

A major strength of this thesis is the use of both anatomical MRI, diffusion MRI and resting state fMRI scans for AD classification. Moreover, for each scan type we used an extensive list of features. Especially for resting state fMRI scans we derived a large number of different features, including FC networks and its graph measures, the dynamics of these networks over time, FC within resting state networks, and the amplitude of low frequency fluctuations. For the purpose of AD classification, this has not been done as extensively before. This enabled a comparison between different MRI modalities, and between the different features per modality. Furthermore, we could combine all these different features, and we showed that they can complement each other for AD classification. Our results can aid researchers in choosing relevant MRI measures for AD research.

Extensive evaluation of generalisability

We took a careful approach to evaluate our AD classification models. Importantly, we used cross-validation to ensure that our models are not the result of overfitting. More specifically, we employed nested cross-validation. In nested cross-validation, there is an outer cross-validation loop that prevents ordinary overfitting. Ordinary overfitting is the practice of fitting too complex models that do not generalise to external data. Additionally, there is an inner loop that prevents over-hyping. Over-hyping is overfitting on the hyperparameter by trying out many different hyperparameter values and reporting the best result. Whereas overfitting is a problem that most researchers are aware of nowadays, over-hyping is still happening a lot, making results overoptimistic (Hosseini et al., 2020). We put effort in preventing both these types of overfitting to ensure that our results represent honest estimates of AD classification accuracy.

Although cross-validation uses ‘out of sample’ data for model evaluation, these data are still similar to the training data, because they originate from the same data set. Cross-validation does not ensure external validity beyond data similar to the analysed sample. In the case of MRI-based AD classification, there are two important issues that can limit external validity. First, the patient population used for training a model might be too homogeneous. In chapter two and three we used data sets containing only diagnosed AD patients and healthy elderly controls. However, in a clinical setting, the patient group is more diverse, consisting of patients ranging from mild cognitive complaints to progressed dementia. Classification models trained on a too homogeneous sample might not generalise well to a clinically diverse sample. Second, MRI scans acquired at different MRI scanners can have different characteristics, due to different scanner vendors or due to different acquisition settings (Ewers et al., 2006; Takao et al., 2014; Zhu et al., 2011). The MRI scans used in chapter two and three were acquired at a single scan site, and the AD classification models might not generalise well to MRI scans from different scanners.

For these reasons, AD classification models should also be evaluated on external data. Therefore, in **chapter four** we evaluated external validity of our AD classification models on a multi-centre memory clinic data set. This data set was collected at four different scan sites, and consisted of clinically diverse patients. This evaluation is important for judging the clinical applicability of MRI-based AD classification models.

Classification methods

To create classification models, we used penalised logistic regression methods. These methods are well suited for MRI-based AD classification (Schouten et al., 2016; Teipel et al., 2015). They use penalties to hinder predictors entering the regression model. Predictors will only be included in the model if their added value for prediction outweighs the penalty. Consequently, only the most relevant predictors will enter the regression model (Friedman et al., 2010; Zou and Hastie, 2005). This is crucial for MRI-based classification studies, because in these studies the number of predictors largely outnumbers the number of subjects. This comes with the risk of including too many predictors and overfitting the model.

Penalised logistic regression methods only include main effects of the predictors, and they only estimate (log)linear effects between the predictors and the outcome variable. Some other methods also include interaction effects or non-linear effects. These methods can potentially include more information, which could increase accuracy. For example, random forests (Breiman, 2001a) include both non-linear effects and interaction effects, and they generally show high classification accuracy (Fernández-Delgado et al., 2014). An even more flexible method is deep learning. Deep learning is gaining increasing popularity, because it outperforms other classification methods in a variety of tasks (LeCun et al., 2015), and more recently in medical image analysis (Esteva et al., 2017; Vieira et al., 2017). It is unclear whether deep learning is also beneficial for MRI-based AD classification. Due to its flexibility, it requires many subjects (e.g., thousands) to train the classification models, especially when there are many predictors. For MRI-based AD classification studies this is problematic, because they typically use many predictors while not many subjects are available. Nevertheless, deep learning was successfully applied to the ADNI data set for AD classification based on anatomical MRI scans. An accuracy of 99% was reported for classifying AD patients vs healthy elderly controls (Basaia et al., 2019). On the other hand, deep learning did not outperform penalised logistic regression in various MRI-based classification tasks, even though 9,300 MRI scans were used (Schulz et al., 2019). However, it was argued that this was due to using pre-engineered features, depriving deep learning of its main advantage - representation learning (Abrol et al. 2021).

In conclusion, it is yet unclear whether MRI-based AD classification could benefit from deep learning. However, it is certainly worth investigating more. Due to data sharing initiatives (e.g., ADNI; Jack et al., 2008) and publicly available large data sets (e.g., UK biobank; Alfaro-Almagro et al., 2018), it has become easier to apply deep learning to MRI data.

Model interpretation

In this thesis we have mainly focused on the accuracy of our classification models, but less so on the content of these models. To some extent, we interpreted the models, by summarising which features were most important. However, the interpretation of these models has some limitations. For example, we have forced sparsity in our models, which hinders potentially relevant features from entering the model. Also, when highly correlated features are all informative, only one of those features is necessary for the model. Consequently, training the model on slightly different training sets can result in largely different models. Furthermore, the effects of the features are conditional on the effect of all the other features in the model, and should therefore be interpreted in that relative context. When many features are involved, this further complicates interpretation.

A focus on classification accuracy usually comes at the cost of interpretability. This is referred to as the distinction between ‘predicting’ and ‘explaining’ (Breiman, 2001b; Shmueli, 2010). In statistical modelling one normally chooses to focus on either one of these two goals. We choose to focus on ‘predicting’, because classification accuracy is most important for individual patient classification. However, preferably the predictive mode provides some explanation as well (Ribeiro et al., 2016). This is important because clinicians are more willing to accept a model-based diagnosis if they know what the model bases its diagnosis on. Therefore, one should make accurate models, and subsequently try to make these models interpretable (Breiman, 2001b). This could increase the likeliness of AD classification models to be used by clinicians.

Sample sizes

In this thesis we used medium sized samples (N = 42 for chapter 2, N = 250 for chapter 3, and N = 189 for chapter 4 and 5), which are considered small sample sizes for machine learning studies. Therefore, the uncertainty of the accuracy estimates is relatively high. For example, for sample sizes of 300 the discrepancy between accuracy measured by cross-validation and expected accuracy on new data is estimated between 4% and 6% (Varoquaux, 2018). Consequently, it is difficult to evaluate small effect sizes. In this thesis, this is not a problem for comparing the accuracies against chance, because many reported AUC values were over .80, which is much higher than chance level. However, comparisons of accuracies against each other mostly concern small differences. For example, in chapter 2 the AUC values of the anatomical MRI measures range from 0.67 to 0.94, and the AUC of the combined model is 0.98. In chapter 3 the AUC values of the resting state fMRI measures range from 0.51 To 0.85, and the AUC of the combined model is 0.85. The increases of the combined models over the single measures are small, and we can therefore not rule out that these increases are due to chance.

Still, we think it is useful that we have chosen these data sets for our analyses, because they have never been analysed in this context before. Most other AD classification studies have used the ADNI data set (Jack et al., 2008), and it is essential to use other data sets as well to not blind-stare on the results on the ADNI data set alone. Nevertheless, we acknowledge that our data sets are not large enough for firm conclusions, and we advise to replicate our results on a larger data set. This can be accomplished using either the ADNI data set, or large imaging population studies like the UK biobank (Alfaro-Almagro et al., 2018).

6.3 Future research

Early phase AD identification

In this thesis we included clinically diagnosed AD patients. This diagnosis requires at least cognitive and behavioural symptoms, which are known to start in a late phase of the disease (Jack et al., 2013). Consequently, the included AD patients are relatively progressed. Brain changes, as observed on MRI scans, are known to start already before cognitive and behavioural symptoms occur (Jack et al., 2013). Therefore, MRI scans might be used to diagnose AD in an earlier phase. In fact, anatomical MRI scans have been used to predict future conversion to AD in MCI patients (Davatzikos et al., 2011).

In line with this thesis, MCI to AD conversion prediction could be improved by calculating multiple anatomical MRI measures, or by using multimodal MRI models. Also, this would enable a comparison between the three MRI modalities on predictive accuracy of AD conversion. In this thesis, anatomical MRI was unquestionably superior to diffusion MRI and resting state fMRI for AD classification. However, for the prediction of AD conversion, this may be different since resting state fMRI has been put forward as a potential early AD marker (Buckner et al., 2005; Sheline and Raichle, 2013). However, for MCI to AD conversion prediction this has not yet been investigated thoroughly using a wide range of resting state fMRI measures in combination with machine learning. In chapter 3 we presented the methodology for such a study.

Taking into account the heterogeneity of clinical populations

In this thesis we set the focus on classifying AD vs. healthy elderly controls. In chapter 2 and 3 we developed AD classification models using only AD patients and healthy elderly controls. In chapter 4 we extended our focus, and applied these models to a memory clinic data set including SMC patients, MCI patients and AD patients. But still, this data set does not reflect the heterogeneity of clinical populations as seen in memory clinics. These populations consist of patients suffering from other dementia types as well, like frontotemporal dementia, Lewy body dementia or vascular dementia. In addition, some of these patients may experience their symptoms due to non-dementia causes, like depression, medication or alcohol abuse.

In order to be clinically useful, the full heterogeneity of clinical populations should be taken into account. Therefore, MRI-based AD classification should proceed from AD vs. control classification to classification of AD in a more diverse sample. This task is more complex, and requires larger data sets, including a wider variety of patient types.

Combining MRI measures with other biomarkers

The goal of our studies was to extensively evaluate MRI scans for AD classification. We did not study the inclusion of other types of biomarkers in our models. In future efforts, our models can be extended with other biomarkers that are known to be predictive for AD. In a first step (cerebrospinal fluid) CSF markers, PET scans and cognitive measures could be added, because these are proven to be discriminative for AD (Jack et al., 2016). In a second step one could add markers that are less discriminative, but could still be of additive value, like genetic markers (Genin et al., 2011) or metabolic markers (de Leeuw et al., 2017).

Scanner site differences

A challenge for MRI-based classification is the presence of scanner variability. It is known that technical variabilities across scan sites can have large effects on MRI scans. This is especially so for diffusion MRI scans (Zhu et al., 2011), but it is also the case for anatomical MRI (Takao et al., 2014) and fMRI (Feis et al., 2015). Consequently, MRI models based on scans from one scanner are not easily applied to MRI scans from another scanner. Partly, this problem can be solved by applying data harmonisation methods (Fortin et al., 2017; Fortin et al., 2018; Yu et al., 2018), as done in chapter 4 and 5 in this thesis. However, this only works for scanner sites for which MRI scans are available when fitting the MRI-based classification model. The created model will not take into account scan site effects of other scanner sites. Consequently, the use of the model is limited to MRI scans of scanners that were used when creating the model.

To facilitate clinical usefulness of MRI-based classification, care should be taken to standardise MRI scan protocols as much as possible, such that later data harmonisation is not or less needed. This is difficult though, because standardisation is only possible to a certain extent, and limited by the type of scanner hardware. Moreover, as shown by the UK Biobank data, even when using identical hardware for MRI acquisition at each site, site is still an important confounder (Alfaro-Almagro et al., 2021). This shows the persistent character of scan site effects, and indicates the necessity to incorporate this effect in MRI-based classification models.

6.4 Conclusion

This thesis includes two important results. First, AD classification accuracy increases when combining multiple types of measures from a single MRI scan. This is the case for both anatomical MRI scans and resting state fMRI scans. For anatomical MRI scans, this implies that a combination of measures may increase accuracy of AD diagnoses in clinical practice. Usually, anatomical MRI scans are only used to inspect specific AD characteristics like the size and shape of the hippocampus, or to rule out other causes for symptoms like a brain tumour or other dementia types. This thesis shows that more information can be extracted from a single anatomical MRI scan, and this may add to clinical AD diagnosis. The same applies to resting state fMRI scans. Combining multiple types of measures from these scans increases AD classification accuracy. This result is not of direct clinical importance, but it is important for further research. Resting state fMRI scans have potential for early AD diagnosis (Viviano & Damoiseaux, 2020), and to further investigate this potential we advise to combine multiple types of resting state fMRI measures.

Second, MRI-based AD classification models are not only useful for the data set that was used for training the model. They also generalise to some extent to heterogeneous patient populations, and to MRI scans acquired at different scan sites. This result is important, because MRI-based classification models are usually trained on a single-centre data set, including only AD patients and healthy elderly controls. The external validity of these models was yet unclear. This result adds to the translation of MRI-based AD classification to clinical practice.

7

Bibliography

Abou Elseoud, A., Littow, H., Remes, J., Starck, T., Nikkinen, J., Nissilä, J., … Kiviniemi, V. (2011). Group-ICA Model Order Highlights Patterns of Functional Brain Connectivity. *Frontiers in Systems Neuroscience*, *5*(June), 37. https://doi.org/10.3389/fnsys.2011.00037

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., & Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications*, *12*(1), 1–17.

Agosta, F., Pievani, M., Geroldi, C., Copetti, M., Frisoni, G. B., & Filippi, M. (2012). Resting state fMRI in Alzheimer's disease: beyond the default mode network. *Neurobiology of Aging*, *33*(8), 1564–1578. https://doi.org/10.1016/j.neurobiolaging.2011.06.007

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., … Phelps, C. H. (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 : The Journal of the Alzheimer's Association*, *7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., … Vallee, E. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, *166*, 400–424.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L. R., Bastiani, M., Miller, K. L., … Smith, S. M. (2021). Confound modelling in UK Biobank brain imaging. *NeuroImage*, *224*, 117002.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, *24*(3), 663–676. https://doi.org/10.1093/cercor/bhs352

Allen, G., Barnard, H., McColl, R., Hester, A. L., Fields, J. A., Weiner, M. F., … Cullum, C. M. (2007). Reduced hippocampal functional connectivity in Alzheimer disease. *Archives of Neurology*, *64*(10), 1482–1487. https://doi.org/10.1001/archneur.64.10.1482

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry--the methods. *NeuroImage*, *11*(6), 805–821. https://doi.org/10.1006/nimg.2000.0582

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., … Initiative, A. D. N. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, *21*, 101645.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *Medical Imaging, IEEE Transactions On*, *23*(2), 137–152. https://doi.org/10.1109/TMI.2003.822821

Beheshti, I., Demirel, H., & Initiative, A. D. N. (2016). Feature-ranking-based Alzheimer's disease classification from structural MRI. *Magnetic Resonance Imaging*, *34*(3), 252–263.

Binnewijzend, M. A. A., Schoonheim, M. M., Sanz-Arigita, E., Wink, A. M., van der Flier, W. M., Tolboom, N., … Barkhof, F. (2012). Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment. *Neurobiology of Aging*, *33*(9), 2018–2028. https://doi.org/10.1016/j.neurobiolaging.2011.07.003

Binnewijzend, M. A. A., Adriaanse, S. M., Van der Flier, W. M., Teunissen, C. E., de Munck, J. C., Stam, C. J., … Wink, A. M. (2014). Brain network alterations in Alzheimer's disease measured by Eigenvector centrality in fMRI are related to cognition and CSF biomarkers. *Human Brain Mapping*, *35*(5), 2383–2393. https://doi.org/10.1002/hbm.22335

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739. https://doi.org/10.1073/pnas.0911855107

Blennow, K., & Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: current status and prospects for the future. *Journal of Internal Medicine*, *284*(6), 643–663.

Bozzali, M., Falini, A., Franceschi, M., Cercignani, M., Zuffi, M., Scotti, G., … Filippi, M. (2002). White matter damage in Alzheimer's disease assessed in vivo using diffusion tensor magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, *72*(6), 742–746.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64. https://doi.org/10.1007/BF00117832

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Brier, M. R., Thomas, J. B., Snyder, A. Z., Benzinger, T. L., Zhang, D., Raichle, M. E., … Ances, B. M. (2012). Loss of intranetwork and internetwork resting state functional connections with Alzheimer's disease progression. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(26), 8890–8899. https://doi.org/10.1523/JNEUROSCI.5698-11.2012

Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., … Klein, S. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, *111*, 562–579. https://doi.org/10.1016/j.neuroimage.2015.01.048

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., … Mintun, M. A. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *J. Neurosci.*, *25*(34), 7709–7717. https://doi.org/10.1523/JNEUROSCI.2177-05.2005

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570.

Callaert, D. V, Ribbens, A., Maes, F., Swinnen, S. P., & Wenderoth, N. (2014). Assessing age-related gray matter decline with voxel-based morphometry depends significantly on segmentation and normalization procedures. *Frontiers in Aging Neuroscience*, *6*(June), 124. https://doi.org/10.3389/fnagi.2014.00124

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., & Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage*, *112*, 232–243. https://doi.org/10.1016/j.neuroimage.2015.02.037

Chan, D., Janssen, J. C., Whitwell, J. L., Watt, H. C., Jenkins, R., Frost, C., … Fox, N. C. (2003). Change in rates of cerebral atrophy over time in early-onset Alzheimer's disease: Longitudinal MRI study. *Lancet*, *362*, 1121–1122. https://doi.org/10.1016/S0140-6736(03)14469-8

Chang, C., & Glover, G. H. (2011). Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage*, *50*(1), 81–98. https://doi.org/10.1016/j.neuroimage.2009.12.011.Time-frequency

Chen, G., Ward, B. D., Xie, C., Li, W., Wu, Z., Jones, J. L., … Li, S.-J. (2011). Classifi cation of Alzheimer Disease , Mild Cognitive Impairment , and Normal Cognitive Status with Large-Scale Network Analysis Based on Resting-State. *Neurology*, *259*(1), 213–221. https://doi.org/10.1148/radiol.10100734/-/DC1

Chen, X., Zhang, H., Gao, Y., Wee, C. Y., Li, G., & Shen, D. (2016). High-order resting-state functional connectivity network for MCI classification. *Human Brain Mapping*, *37*(9), 3282–3296. https://doi.org/10.1002/hbm.23240

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., & Lin, C. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, *60*(1), 59–70. https://doi.org/10.1016/j.neuroimage.2011.11.066

Clerx, L., Gronenschild, E. H. B. M., Echavarri, C., Verhey, F., Aalten, P., & Jacobs, H. I. L. (2015). Can FreeSurfer Compete with Manual Volumetric Measurements in Alzheimer's Disease? *Current Alzheimer Research*, *12*(4), 358–367.

Clerx, L., Visser, P. J., Verhey, F., & Aalten, P. (2012). New MRI markers for Alzheimer's disease: a meta-analysis of diffusion tensor imaging and a comparison with medial temporal lobe measurements. *Journal of Alzheimer's Disease*, *29*(2), 405–429.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., … Jin, J. S. (2011). Identification of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Multivariate Predictors. *PLoS ONE*, *6*(7), e21896. https://doi.org/10.1371/journal.pone.0021896

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., … Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, *56*(2), 766–781. https://doi.org/10.1016/j.neuroimage.2010.06.013

Cummings, J., Aisen, P. S., Dubois, B., Frölich, L., Jack, C. R., Jones, R. W., … Scheltens, P. (2016). Drug development in Alzheimer's disease: The path to 2025. *Alzheimer's Research and Therapy*, *8*(1), 1–12. https://doi.org/10.1186/s13195-016-0207-9

Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., & He, Y. (2012). Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). *NeuroImage*, *59*(3), 2187–2195. https://doi.org/10.1016/j.neuroimage.2011.10.003

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

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., & Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, *32*(12), 2322.e19-2322.e27. https://doi.org/10.1016/j.neurobiolaging.2010.05.023

De Jong, L. W., Van Der Hiele, K., Veer, I. M., Houwing, J. J., Westendorp, R. G. J., Bollen, E. L. E. M., … Van Der Grond, J. (2008). Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: An MRI study. *Brain*, *131*(12), 3277–3285. https://doi.org/10.1093/brain/awn278

de Leeuw, F. A., Peeters, C. F. W., Kester, M. I., Harms, A. C., Struys, E. A., Hankemeier, T., … Scheltens, P. (2017). Blood-based metabolic signatures in Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *8*, 196–207.

De Magalhães Oliveira, P. P., Nitrini, R., Busatto, G., Buchpiguel, C., Sato, J. R., & Amaro, E. (2010). Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. *Journal of Alzheimer's Disease*, *19*(4), 1263–1272. https://doi.org/10.3233/JAD-2010-1322

de Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., … Rombouts, S. A. R. B. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *NeuroImage*, *167*(November 2017), 62–72. https://doi.org/10.1016/j.neuroimage.2017.11.025

de Vos, F., Schouten, T. M., Koini, M., Bouts, M. J. R. J., Feis, R. A., Lechner, A., … Rikkert, M. G. M. O. (2020). Pre-trained MRI-based Alzheimer's disease classification models to classify memory clinic patients. *NeuroImage: Clinical*, 102303. https://doi.org/https://doi.org/10.1016/j.nicl.2020.102303

de Vos, F., Schouten, T. M., Hafkemeijer, A., Dopper, E. G. P., van Swieten, J. C., de Rooij, M., … Rombouts, S. A. R. B. (2016). Combining multiple anatomical MRI measures improves Alzheimer's disease classification. *Human Brain Mapping*, *37*(5), 1920–1929. https://doi.org/10.1002/hbm.23147

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Diaz-de-Grenu, L. Z., Acosta-Cabronero, J., Chong, Y. F. V., Pereira, J. M. S., Sajjadi, S. a, Williams, G. B., & Nestor, P. J. (2014). A brief history of voxel-based grey matter analysis in Alzheimer's disease. *Journal of Alzheimer's Disease : JAD*, *38*(3), 647–659. https://doi.org/10.3233/JAD-130362

Dickerson, B. C., Feczko, E., Augustinack, J. C., Pacheco, J., Morris, J. C., Fischl, B., & Buckner, R. L. (2009). Differential effects of aging and Alzheimer's disease on medial temporal lobe cortical thickness and surface area. *Neurobiology of Aging*, *30*(3), 432–440. https://doi.org/10.1016/j.neurobiolaging.2007.07.022

Dipasquale, O., Griffanti, L., Clerici, M., Nemni, R., Baselli, G., & Baglio, F. (2015). High-Dimensional ICA Analysis Detects Within-Network Functional Connectivity Damage of Default-Mode and Sensory-Motor Networks in Alzheimer's Disease. *Frontiers in Human Neuroscience*, *9*(February), 43. https://doi.org/10.3389/fnhum.2015.00043

Dørum, E. S., Kaufmann, T., Alnæs, D., Andreassen, O. A., Richard, G., Kolskår, K. K., … Westlye, L. T. (2017). Increased sensitivity to age-related differences in brain functional connectivity during continuous multiple object tracking compared to resting-state. *NeuroImage*, *148*(October 2016), 364–372. https://doi.org/10.1016/j.neuroimage.2017.01.048

Douaud, G., Jbabdi, S., Behrens, T. E. J., Menke, R. A., Gass, A., Monsch, A. U., … Smith, S. (2011). DTI measures in crossing-fibre areas: Increased diffusion anisotropy reveals early white matter alteration in MCI and mild Alzheimer's disease. *NeuroImage*, *55*(3), 880–890. https://doi.org/10.1016/j.neuroimage.2010.12.008

Dumurgier, J., Hanseeuw, B. J., Hatling, F. B., Judge, K. A., Schultz, A. P., Chhatwal, J. P., … Hyman, B. T. (2017). Alzheimer's disease biomarkers and future decline in cognitive normal older adults. *Journal of Alzheimer's Disease*, *60*(4), 1451–1459.

Dyrba, M., Barkhof, F., Fellgiebel, A., Filippi, M., Hausner, L., Hauenstein, K., … group, E. study. (2015a). Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. *Journal of Neuroimaging*, *25*(5), 738–747.

Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., … Teipel, S. J. (2013). Robust Automated Detection of Microstructural White Matter Degeneration in Alzheimer's Disease Using Machine Learning Classification of Multicenter DTI Data. *PLoS ONE*, *8*(5), e64925. https://doi.org/10.1371/journal.pone.0064925

Dyrba, M., Grothe, M., Kirste, T., & Teipel, S. J. (2015b). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Human Brain Mapping*, *36*(6), 2118–2131. https://doi.org/10.1002/hbm.22759

Eckerström, M., Göthlin, M., Rolstad, S., Hessen, E., Eckerström, C., Nordlund, A., … Sacuiu, S. (2017). Longitudinal evaluation of criteria for subjective cognitive decline and preclinical Alzheimer's disease in a memory clinic sample. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *8*, 96–107.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Ewers, M., Teipel, S. J., Dietrich, O., Schönberg, S. O., Jessen, F., Heun, R., … Hampel, H. (2006). Multicenter assessment of reliability of cranial MRI. *Neurobiology of Aging*, *27*(8), 1051–1059. https://doi.org/10.1016/j.neurobiolaging.2005.05.032

Fawcett, T. (2004). ROC Graphs : Notes and Practical Considerations for Researchers. *ReCALL*, *31*(HPL-2003-4), 1–38. https://doi.org/10.1.1.10.9777

Feis, R. A., Smith, S. M., Filippini, N., Douaud, G., Dopper, E. G. P., Heise, V., … Mackay, C. E. (2015). ICA-based artifact removal diminishes scan site differences in multi-center resting-state fMRI. *Frontiers in Neuroscience*, *9*(OCT). https://doi.org/10.3389/fnins.2015.00395

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Filippini, N., Macintosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., … Mackay, C. E. (2009). Distinct patterns of brain activity in young carriers of the APOE- e4 allele. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(17), 7209–7214. https://doi.org/10.1073/pnas.0811879106

Fischl, B., Sereno, M. I., & Dale, a M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*(2), 195–207. https://doi.org/10.1006/nimg.1998.0396

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., … Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*(June 2017), 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., … Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*(August), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Fotenos, A. F., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology*, *64*(6), 1032–1039. https://doi.org/10.1212/01.WNL.0000154530.72969.11

Freudenberger, P., Petrovic, K., Sen, A., Töglhofer, A. M., Fixa, A., Hofer, E., … Schmidt, R. (2016). Fitness and cognition in the elderly The Austrian Stroke Prevention Study. *Neurology*, *86*(5), 418–424.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *30*(1), 1–3. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/#

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Review Neurology*, *6*(2), 67–77. https://doi.org/10.1038/nrneurol.2009.215

Frisoni, G. B., Boccardi, M., Barkhof, F., Blennow, K., Cappa, S., Chiotis, K., … Gietl, A. (2017). Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *The Lancet Neurology*, *16*(8), 661–676.

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., … Berr, C. (2011). APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Molecular Psychiatry*, *16*(9), 903–907.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., … Colliot, O. (2009). Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, *47*(4), 1476–1486. https://doi.org/10.1016/j.neuroimage.2009.05.036

Gold, B. T., Johnson, N. F., Powell, D. K., & Smith, C. D. (2012). White matter integrity and vulnerability to Alzheimer's disease: preliminary findings and future directions. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1822*(3), 416–422.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., & Collins, D. L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv*, *9*, 58–66. https://doi.org/10.1007/11866763_8

Greicius, M. D., Srivastava, G., Reiss, A. L., & Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(13), 4637–4642. https://doi.org/10.1073/pnas.0308627101

Griffanti, L., Rolinski, M., Szewczyk-Krolikowski, K., Menke, R. A., Filippini, N., Zamboni, G., … Mackay, C. E. (2016). Challenges in the reproducibility of clinical studies with resting state fMRI: An example in early Parkinson's disease. *NeuroImage*, *124*, 704–713. https://doi.org/10.1016/j.neuroimage.2015.09.021

Han, Y., Wang, J., Zhao, Z., Min, B., Lu, J., Li, K., … Jia, J. (2011). Frequency-dependent changes in the amplitude of low-frequency fluctuations in amnestic mild cognitive impairment: A resting-state fMRI study. *NeuroImage*, *55*(1), 287–295. https://doi.org/10.1016/j.neuroimage.2010.11.059

Handels, R. L., Aalten, P., Wolfs, C. A., OldeRikkert, M., Scheltens, P., Visser, P. J., … Verhey, F. R. (2012). Diagnostic and economic evaluation of new biomarkers for Alzheimer's disease: the research protocol of a prospective cohort study. *BMC Neurology*, *12*(1), 72. https://doi.org/10.1186/1471-2377-12-72

Hanley, J. a, & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839–843. https://doi.org/10.1148/radiology.148.3.6878708

Hardy, J., & Selkoe, D. J. (2002). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*, *297*(5580), 353–356.

Hindriks, R., Adhikari, M. H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N. K., & Deco, G. (2016). Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *NeuroImage*, *127*, 242–256. https://doi.org/10.1016/j.neuroimage.2015.11.055

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*.

Huijbers, W., Vannini, P., Sperling, R. a., C.M., P., Cabeza, R., & Daselaar, S. M. (2012). Explaining the encoding/retrieval flip: Memory-related deactivations and activations in the posteromedial cortex. *Neuropsychologia*, *50*(14), 3764–3774. https://doi.org/10.1016/j.neuropsychologia.2012.08.021

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., … Chang, C. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, *80*, 360–378. https://doi.org/10.1016/j.neuroimage.2013.05.079

Hutton, C., Draganski, B., Ashburner, J., & Weiskopf, N. (2009). A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage*, *48*(2), 371–380. https://doi.org/10.1016/j.neuroimage.2009.06.043

Ittner, L. M., & Götz, J. (2011). Amyloid-β and tau--a toxic pas de deux in Alzheimer's disease. *Nature Reviews. Neuroscience*, *12*(2), 65–72. https://doi.org/10.1038/nrn2967

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., … Ward, C. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27*(4), 685–691.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., … Weigand, S. D. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, *12*(2), 207–216.

Jack, C. R., Shiung, M. M., Gunter, J. L., O'Brien, P. C., Weigand, S. D., Knopman, D. S., … Petersen, R. C. (2004). Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology*, *62*(4), 591–600. https://doi.org/10.1212/01.WNL.0000110315.26026.EF

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., … Knopman, D. S. (2016). A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*, *87*(5), 539–547.

Jack, C. R., Petersen, R. C., Xu, Y. C., Waring, S. C., O'Brien, P. C., Tangalos, E. G., … Kokmen, E. (1997). Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, *49*(3), 786–794.

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., … Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, *9*(1), 119–128. https://doi.org/10.1016/S1474-4422(09)70299-6

Jansen, W. J., Handels, R. L. H., Visser, P. J., Aalten, P., Bouwman, F., Claassen, J., … Ramakers, I. H. G. B. (2017). The Diagnostic and Prognostic Value of Neuropsychological Assessment in Memory Clinic Patients. *Journal of Alzheimer's Disease*, *55*(2), 679–689. https://doi.org/10.3233/JAD-160126

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

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Jones, D. T., Vemuri, P., Murphy, M. C., Gunter, J. L., Senjem, M. L., Machulda, M. M., … Jack, C. R. (2012). Non-stationarity in the "resting brain's" modular architecture. *PLoS ONE*, *7(6)*, e39731. https://doi.org/10.1371/journal.pone.0039731

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., … Fischl, B. (2009). MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, *46*(1), 177–192. https://doi.org/10.1016/j.neuroimage.2009.02.010

Karas, G. B., Burton, E. J., Rombouts, S. a R. B., Van Schijndel, R. a., O'Brien, J. T., Scheltens, P., … Barkhof, F. (2003). A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *NeuroImage*, *18*(4), 895–907. https://doi.org/10.1016/S1053-8119(03)00041-7

Karas, G. B., Scheltens, P., Rombouts, S. a R. B., Visser, P. J., Van Schijndel, R. a., Fox, N. C., & Barkhof, F. (2004). Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, *23*(2), 708–716. https://doi.org/10.1016/j.neuroimage.2004.07.006

Khazaee, A., Ebrahimzadeh, A., & Babajani-Feremi, A. (2015). Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clinical Neurophysiology*, *126*(11), 2132–2141. https://doi.org/10.1016/j.clinph.2015.02.060

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., … Frackowiak, R. S. J. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain : A Journal of Neurology*, *131*(Pt 3), 681–689. https://doi.org/10.1093/brain/awm319

Koch, W., Teipel, S., Mueller, S., Benninghoff, J., Wagner, M., Bokde, A. L. W., … Meindl, T. (2012). Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiology of Aging*, *33*(3), 466–478. https://doi.org/10.1016/j.neurobiolaging.2010.04.013

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(1), 1–15. https://doi.org/10.1186/1758-2946-6-10

Lancaster, M. A., Seidenberg, M., Smith, J. C., Nielson, K. A., Woodard, J. L., Durgerian, S., & Rao, S. M. (2016). Diffusion tensor imaging predictors of episodic memory decline in healthy elders at genetic risk for alzheimer's disease. *Journal of the International Neuropsychological Society*, *22*(10), 1005–1015.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J., & Evans, A. C. (2005). Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. *Cerebral Cortex*, *15*(7), 995–1001. https://doi.org/10.1093/cercor/bhh200

Leung, K. K., Barnes, J., Ridgway, G. R., Bartlett, J. W., Clarkson, M. J., Macdonald, K., … Ourselin, S. (2010). Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, *51*(4), 1345–1359. https://doi.org/10.1016/j.neuroimage.2010.03.018

Lindeboom, J., Schmand, B., Tulner, L., Walstra, G., & Jonker, C. (2002). Visual association test to detect early dementia of the Alzheimer type. *Journal of Neurology, Neurosurgery & Psychiatry*, *73*(2), 126–133.

Liu, Y., Paajanen, T., Zhang, Y., Westman, E., Wahlund, L.-O., Simmons, A., … Tsolaki, M. (2011). Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups—the AddNeuroMed study. *Neurobiology of Aging*, *32*(7), 1198–1206.

Lohmann, G., Margulies, D. S., Horstmann, A., Pleger, B., Lepsien, J., Goldhahn, D., … Turner, R. (2010). Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PLoS ONE*, *5(4)*, e10232. https://doi.org/10.1371/journal.pone.0010232

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., … Petersen, R. C. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *The Clinical Neuropsychologist*, *27*(8), 1247–1264.

McKeith, I. G., Dickson, D. W., Lowe, J., Emre, M., O'brien, J. T., Feldman, H., … Perry, E. K. (2005). Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology*, *65*(12), 1863–1872.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., … Phelps, C. H. (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 and Dementia*, *7*(3), 263–269. https://doi.org/10.1016/j.jalz.2011.03.005

Mckhann, G., Drachman, D., & Folstein, M. (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. *Neurology*, *34*, 939–944. https://doi.org/10.1212/WNL.34.7.939

Mesrob, L., Sarazin, M., Hahn-Barma, V., Souza, L. C. De, Dubois, B., Gallinari, P., & Kinkingnéhun, S. (2012). DTI and Structural MRI Classification in Alzheimer's Disease. *Advances in Molecular Imaging*, *02*(02), 12–20. https://doi.org/10.4236/ami.2012.22003

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Misra, C., Fan, Y., & Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *Neuroimage*, *44*(4), 1415–1422.

Mitchell, A. J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research*, *43*(4), 411–431.

Mito, R., Raffelt, D., Dhollander, T., Vaughan, D. N., Tournier, J.-D., Salvado, O., … Connelly, A. (2018). Fibre-specific white matter reductions in Alzheimer's disease and mild cognitive impairment. *Brain*, *141*(3), 888–902.

Morra, J. H., Tu, Z., Apostolova, L. G., Amity, E., Avedissian, C., Madsen, S. K., … Weiner, M. W. (2010). Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. *Brain*, *30*(9), 2766–2788. https://doi.org/10.1002/hbm.20708.Automated

Mulder, E. R., de Jong, R. a., Knol, D. L., van Schijndel, R. a., Cover, K. S., Visser, P. J., … Vrenken, H. (2014). Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. *NeuroImage*, *92*, 169–181. https://doi.org/10.1016/j.neuroimage.2014.01.058

Natu, M. V, & Agarwal, A. K. (1995). Digit letter substitution test (DLST) as an alternative to digit symbol substitution test (DSST). *Human Psychopharmacology: Clinical and Experimental*, *10*(4), 339–343.

Pacheco, J., Goh, J. O., Kraut, M. A., Ferrucci, L., & Resnick, S. M. (2015). Neurobiology of Aging Greater cortical thinning in normal older adults predicts later cognitive impairment. *Neurobiology of Aging*, *36*(2), 903–908. https://doi.org/10.1016/j.neurobiolaging.2014.08.031

Palmqvist, S., Janelidze, S., Quiroz, Y. T., Zetterberg, H., Lopera, F., Stomrud, E., … Leuzy, A. (2020). Discriminative Accuracy of Plasma Phospho-tau217 for Alzheimer Disease vs Other Neurodegenerative Disorders. *JAMA*.

Parkes, L., Fulcher, B., Yücel, M., & Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, *171*(July 2017), 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073

Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, *56*(3), 907–922. https://doi.org/10.1016/j.neuroimage.2011.02.046

Patterson, C. (2018). World Alzheimer report 2018: the state of the art of dementia research: new frontiers. *Alzheimer's Disease International (ADI): London, UK*, 32–36.

Pereira, J. B., Ibarretxe-Bilbao, N., Marti, M.-J., Compta, Y., Junqué, C., Bargallo, N., & Tolosa, E. (2012). Assessment of cortical degeneration in patients with Parkinson's disease by voxel-based morphometry, cortical folding, and cortical thickness. *Human Brain Mapping*, *33*(11), 2521–2534. https://doi.org/10.1002/hbm.21378

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., … Frisoni, G. B. (2016). Brain atrophy in Alzheimer's disease and aging. *Ageing Research Reviews*, *30*, 25–48.

Prince, M., Bryce, R., & Ferri, C. (2011). *World Alzheimer Report 2011: The benefits of early diagnosis and intervention*.

Prins, N. D., van Dijk, E. J., den Heijer, T., Vermeer, S. E., Koudstaal, P. J., Oudkerk, M., … Breteler, M. M. B. (2004). Cerebral white matter lesions and the risk of dementia. *Archives of Neurology*, *61*(10), 1531–1534.

Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage*, *112*, 267–277.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., … Celsis, P. (2009). Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, *132*(8), 2036–2047. https://doi.org/10.1093/brain/awp105

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., … Onyike, C. U. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, *134*(9), 2456–2477.

Rashid, B., Damaraju, E., Pearlson, G. D., & Calhoun, V. D. (2014). Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects. *Frontiers in Human Neuroscience*, *8*(November), 897. https://doi.org/10.3389/fnhum.2014.00897

Rathore, S., Habes, M., Aksam, M., Shacklett, A., & Davatzikos, C. (2017). NeuroImage A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer ' s disease and its prodromal stages. *NeuroImage*, *155*(April), 530–548. https://doi.org/10.1016/j.neuroimage.2017.03.057

Redolfi, A., Manset, D., Barkhof, F., Wahlund, L.-O., Glatard, T., Mangin, J.-F., & Frisoni, G. B. (2015). Head-to-Head Comparison of Two Popular Cortical Thickness Extraction Algorithms: A Cross-Sectional and Longitudinal Study. *Plos One*, *10*(3), e0117692. https://doi.org/10.1371/journal.pone.0117692

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276.

Rey, A. (1958). *L'examen clinique en psychologie*. Oxford: Universitaries De France.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Risacher, S. L., Shen, L., West, J. D., Kim, S., McDonald, B. C., Beckett, L. a., … Saykin, A. J. (2010). Longitudinal MRI atrophy biomarkers: Relationship to conversion in the ADNI cohort. *Neurobiology of Aging*, *31*(8), 1401–1418. https://doi.org/10.1016/j.neurobiolaging.2010.04.029

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Román, G. C., Tatemichi, T. K., Erkinjuntti, T., Cummings, J. L., Masdeu, J. C., Garcia, J. H., … Hofman, A. (1993). Vascular dementia: diagnostic criteria for research studies: report of the NINDS-AIREN International Workshop. *Neurology*, *43*(2), 250.

Rombouts, S. a R. B., Barkhof, F., Goekoop, R., Stam, C. J., & Scheltens, P. (2005). Altered resting state networks in mild cognitive impairment and mild Alzheimer's disease: An fMRI study. *Human Brain Mapping*, *26*(4), 231–239. https://doi.org/10.1002/hbm.20160

Rombouts, S. a, Barkhof, F., Witter, M. P., & Scheltens, P. (2000). Unbiased whole-brain analysis of gray matter loss in Alzheimer's disease. *Neuroscience Letters*, *285*(3), 231–233.

Ronan, L., Pienaar, R., Williams, G., Bullmore, E., Crow, T. J., Roberts, N., … Fletcher, P. C. (2011). Intrinsic Curvature: a Marker of Millimeter-Scale Tangential Cortico-Cortical Connectivity? *International Journal of Neural Systems*, *21*(5), 351–366. https://doi.org/10.1142/S0129065711002948

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, *52*(3), 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

Salat, D. H., Kaye, J. a, & Janowsky, J. S. (1999). Prefrontal gray and white matter volumes in healthy aging and Alzheimer disease. *Archives of Neurology*, *56*(3), 338–344. https://doi.org/10.1001/archneur.56.3.338

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., & Smith, S. M. (2014). Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, *90*, 449–468. https://doi.org/10.1016/j.neuroimage.2013.11.046

Sanz-Arigita, E., Schoonheim, M. M., Damoiseaux, J., Rombouts, S. A., Maris, E., Barkhof, F., … Stam, C. J. (2010). Loss of "small-world" networks in Alzheimer's disease: graph analysis of FMRI resting-state functional connectivity. *PloS One*, *5(11)*, e13788. https://doi.org/10.1371/journal.pone.0013788

Scheltens, P., Blennow, K., Breteler, M. M., de Strooper, B., Frisoni, G. B., Salloway, S., & Van der Flier, W. M. (2016). Alzheimer's disease. *The Lancet*, *388*, 505–517.

Scher, a. I., Xu, Y., Korf, E. S. C., White, L. R., Scheltens, P., Toga, a. W., … Launer, L. J. (2007). Hippocampal shape analysis in Alzheimer's disease: A population-based study. *NeuroImage*, *36*(1), 8–18. https://doi.org/10.1016/j.neuroimage.2006.12.036

Schmidt, R., Lechner, H., Fazekas, F., Niederkorn, K., Reinhart, B., Grieshofer, P., … Dusek, T. (1994). Assessment of Cerebrovascular Risk Profiles in Healthy Persons: Definition of Research Goals and the Austrian Stroke Prevention Study (ASPS). *Neuroepidemiology*, *13*(6), 308–313. https://doi.org/10.1159/000110396

Schouten, T. M., Koini, M., de Vos, F., Seiler, S., van der Grond, J., Lechner, A., … Rombouts, S. A. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage: Clinical*, *11*, 46–51. https://doi.org/10.1016/j.nicl.2016.01.002

Schouten, T. M., Koini, M., Vos, F. De, Seiler, S., Rooij, M. De, Lechner, A., … Rombouts, S. a. R. B. (2017). Individual Classification of Alzheimer's Disease with Diffusion Magnetic Resonance Imaging. *NeuroImage*, *152*(October 2016), 476–481. https://doi.org/10.1016/j.neuroimage.2017.03.025

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., … Bzdok, D. (2019). Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets. *BioRxiv*, 757054.

Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., & Greicius, M. D. (2009). Neurodegenerative Diseases Target Large-Scale Human Brain Networks. *Neuron*, *62*(1), 42–52. https://doi.org/10.1016/j.neuron.2009.03.024

Seiler, S., Schmidt, H., Lechner, A., Benke, T., Sanin, G., Ransmayr, G., … Schmidt, R. (2012). Driving Cessation and Dementia: Results of the Prospective Registry on Dementia in Austria (PRODEM). *PLoS ONE*, *7(12)*, e52710. https://doi.org/10.1371/journal.pone.0052710

Shaw, E. E., Schultz, A. P., Sperling, R. a, & Hedden, T. (2015). Functional Connectivity in Multiple Cortical Networks Is Associated with Performance Across Cognitive Domains in Older Adults. *Brain Connectivity*, *5*(8), 505–516. https://doi.org/10.1089/brain.2014.0327

Sheline, Y. I., Morris, J. C., Snyder, A. Z., Price, J. L., Yan, Z., D'Angelo, G., … Mintun, M. A. (2010a). APOE4 allele disrupts resting state fMRI connectivity in the absence of amyloid plaques or decreased CSF Aβ42. *J Neurosci*, *30*(50), 17035–17040. https://doi.org/10.1523/JNEUROSCI.3987-10.2010

Sheline, Y. I., & Raichle, M. E. (2013). Resting state functional connectivity in preclinical Alzheimer's disease. *Biological Psychiatry*, *74*(5), 340–347. https://doi.org/10.1016/j.biopsych.2012.11.028

Sheline, Y. I., Raichle, M. E., Snyder, A. Z., Morris, J. C., Head, D., Wang, S., & Mintun, M. A. (2010b). Amyloid Plaques Disrupt Resting State Default Mode Network Connectivity in Cognitively Normal Elderly. *Biological Psychiatry*, *67*(6), 584–587. https://doi.org/10.1016/j.biopsych.2009.08.024

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., & Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral Cortex*, *22*(1), 158–165. https://doi.org/10.1093/cercor/bhr099

Shmueli, G. (2011). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group Lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245. https://doi.org/10.1080/10618600.2012.681250

Sluimer, J. D., van der Flier, W. M., Karas, G. B., Fox, N. C., Scheltens, P., Barkhof, F., & Vrenken, H. (2008). Whole-brain atrophy rate and cognitive decline: longitudinal MR study of memory clinic patients. *Radiology*, *248*(2), 590–598.

Smith, S. M., Woolrich, M. W., Ugurbil, K., Moeller, S., Xu, J., Glasser, M. F., … Yacoub, E. S. (2012). Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, *109*(8), 3131–3136. https://doi.org/10.1073/pnas.1121329109

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., … Matthews, P. M. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, *31*(4), 1487–1505.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., … Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23 Suppl 1*, S208-19. https://doi.org/10.1016/j.neuroimage.2004.07.051

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., … Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, *54*(2), 875–891. https://doi.org/10.1016/j.neuroimage.2010.08.063

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., … Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, *17*(12), 666–682. https://doi.org/10.1016/j.tics.2013.09.016

Sona, A., Ellis, K. A., & Ames, D. (2013). Rapid cognitive decline in Alzheimer's disease: A literature review. *International Review of Psychiatry*, *25*(6), 650–658. https://doi.org/10.3109/09540261.2013.859128

Song, X.-W., Dong, Z.-Y., Long, X.-Y., Li, S.-F., Zuo, X.-N., Zhu, C.-Z., … Zang, Y.-F. (2011). REST: A Toolkit for Resting-State Functional Magnetic Resonance Imaging Data Processing. *PLoS ONE*, *6*(9), e25031. https://doi.org/10.1371/journal.pone.0025031

Sperling, R. (2011). The potential of functional MRI as a biomarker in early Alzheimer's disease. *Neurobiology of Aging*, *32*(SUPPL. 1), S37–S43. https://doi.org/10.1016/j.neurobiolaging.2011.09.009

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., … Montine, T. J. (2011). Toward defining the preclinical stages of 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), 280–292.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Sui, J., He, H., Yu, Q., Chen, J., Rogers, J., Pearlson, G. D., … Calhoun, V. D. (2013). Combination of Resting State fMRI, DTI, and sMRI Data to Discriminate Schizophrenia by N-way MCCA + jICA. *Frontiers in Human Neuroscience*, *7*(May), 235. https://doi.org/10.3389/fnhum.2013.00235

Supekar, K., Menon, V., Rubin, D., Musen, M., & Greicius, M. D. (2008). Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Computational Biology*, *4*(6), e1000100. https://doi.org/10.1371/journal.pcbi.1000100

Takao, H., Hayashi, N., & Ohtomo, K. (2014). Effects of study design in multi-scanner voxel-based morphometry studies. *Neuroimage*, *84*, 133–140.

Teipel, S. J., Grothe, M. J., Metzger, C. D., Grimmer, T., Sorg, C., Ewers, M., … Dyrba, M. (2017a). Robust detection of impaired resting state functional connectivity networks in Alzheimer's disease using elastic net regularized regression. *Frontiers in Aging Neuroscience*, *8*(JAN), 1–9. https://doi.org/10.3389/fnagi.2016.00318

Teipel, S. J., Kurth, J., Krause, B., & Grothe, M. J. (2015). The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment — Beyond classical regression. *NeuroImage: Clinical*, *8*, 583–593. https://doi.org/10.1016/j.nicl.2015.05.006

Teipel, S. J., Metzger, C. D., Brosseron, F., Buerger, K., Brueggen, K., Catak, C., … Dyrba, M. (2018). Multicenter Resting State Functional Connectivity in Prodromal and Dementia Stages of Alzheimer's Disease. *Journal of Alzheimer's Disease*, *64*(3), 801–813. https://doi.org/10.3233/jad-180106

Teipel, S. J., Wohlert, A., Metzger, C., Grimmer, T., Sorg, C., Ewers, M., … Dyrba, M. (2017b). Multicenter stability of resting state fMRI in the detection of Alzheimer's disease and amnestic MCI. *NeuroImage: Clinical*, *14*, 183–194. https://doi.org/10.1016/j.nicl.2017.01.018

Thompson, P. M., Hayashi, K. M., de Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., … Toga, A. W. (2003). Dynamics of gray matter loss in Alzheimer's disease. *The Journal of Neuroscience*, *23*(3), 994–1005.

Thompson, P. M., Hayashi, K. M., De Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., … Toga, A. W. (2004). Mapping hippocampal and ventricular change in Alzheimer disease. *NeuroImage*, *22*(4), 1754–1766. https://doi.org/10.1016/j.neuroimage.2004.03.040

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, Vol. 58, pp. 267–288. https://doi.org/10.2307/2346178

Tombaugh, T. N. (2005). Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Archives of Clinical Neuropsychology*, *20*(4), 485–503.

Trzepacz, P. T., Hochstetler, H., Yu, P., Castelluccio, P., Witte, M. M., Dellagnello, G., & Degenhardt, E. K. (2016). Relationship of Hippocampal Volume to Amyloid Burden across Diagnostic Stages of Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders*, *41*(1–2), 68–79. https://doi.org/10.1159/000441351

Trzepacz, P. T., Yu, P., Sun, J., Schuh, K., Case, M., Witte, M. M., … Hake, A. (2014). Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to alzheimer's dementia. *Neurobiology of Aging*, *35*(1), 143–151. https://doi.org/10.1016/j.neurobiolaging.2013.06.018

Tustison, N. J., Cook, P. a., Klein, A., Song, G., Das, S. R., Duda, J. T., … Avants, B. B. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage*, *99*, 166–179. https://doi.org/10.1016/j.neuroimage.2014.05.044

van Loon, W., Fokkema, M., Szabo, B., & de Rooij, M. (2020). Stacked penalized logistic regression for selecting views in multi-view learning. *Information Fusion*, *61*(March), 113–123. https://doi.org/10.1016/j.inffus.2020.03.007

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58–75.

Viviano, R. P., & Damoiseaux, J. S. (2020). Functional neuroimaging in subjective cognitive decline: current status and a research path forward. *Alzheimer's Research & Therapy*, *12*(1), 1–18.

Wahlund, L., Pihlstrand, E., & Jönhagen, M. E. (2003). Mild cognitive impairment: experience from a memory clinic. *Acta Neurologica Scandinavica*, *107*, 21–24.

Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., … Li, K. (2006). Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *NeuroImage*, *31*(2), 496–504. https://doi.org/10.1016/j.neuroimage.2005.12.033

Wechsler, D. (1997). *WMS-III Administration and Scoring Manual*. San Antonio, TX: The Psychological Corporation.

Wee, C. Y., Yang, S., Yap, P. T., & Shen, D. (2016). Sparse temporally dynamic resting-state functional connectivity networks for early MCI identification. *Brain Imaging and Behavior*, *10*(2), 342–356. https://doi.org/10.1007/s11682-015-9408-2

Wee, C.-Y., Yap, P.-T., & Shen, D. (2013). Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Human Brain Mapping*, *34*(12), 3411–3425. https://doi.org/10.1002/hbm.22156

Westman, E., Aguilar, C., Muehlboeck, J. S., & Simmons, A. (2013). Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. *Brain Topography*, *26*(1), 9–23. https://doi.org/10.1007/s10548-012-0246-x

Westman, E., Simmons, A., Zhang, Y., Muehlboeck, J.-S., Tunnard, C., Liu, Y., … Vellas, B. (2011). Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. *Neuroimage*, *54*(2), 1178–1187.

Wink, A. M., de Munck, J. C., van der Werf, Y. D., van den Heuvel, O. A., & Barkhof, F. (2012). Fast Eigenvector Centrality Mapping of Voxel-Wise Connectivity in Functional Magnetic Resonance Imaging: Implementation, Validation, and Interpretation. *Brain Connectivity*, *2*(5), 265–274. https://doi.org/10.1089/brain.2012.0087

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D. P., Rueckert, D., … Lötjönen, J. (2011). Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS ONE*, *6*(10), e25446. https://doi.org/10.1371/journal.pone.0025446

World Health Organization. (2017). *Global action plan on the public health response to dementia 2017–2025*.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., Mcinnis, M., Fava, M., … Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, (March). https://doi.org/10.1002/hbm.24241

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zang, Y.-F., He, Y., Zhu, C.-Z., Cao, Q.-J., Sui, M.-Q., Liang, M., … Wang, Y.-F. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain & Development*, *29*(2), 83–91. https://doi.org/10.1016/j.braindev.2006.07.002

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., & Initiative, A. D. N. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, *55*(3), 856–867.

Zhang, Y., Schuff, N., Camacho, M., Chao, L. L., Fletcher, T. P., Yaffe, K., … Weiner, M. W. (2013). MRI markers for mild cognitive impairment: comparisons between white matter integrity and gray matter volume measurements. *PloS One*, *8*(6), e66367. https://doi.org/10.1371/journal.pone.0066367

Zhu, T., Hu, R., Qiu, X., Taylor, M., Tso, Y., Yiannoutsos, C., … Schifitto, G. (2011). Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study. *Neuroimage*, *56*(3), 1398–1411.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, Q. H., Zhu, C. Z., Yang, Y., Zuo, X. N., Long, X. Y., Cao, Q. J., … Zang, Y. F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *Journal of Neuroscience Methods*, *172*(1), 137–141. https://doi.org/10.1016/j.jneumeth.2008.04.012

Appendix

Nederlandse samenvatting

Dankwoord

Curriculum vitae

List of publications

Nederlandse samenvatting

De ziekte van Alzheimer

In 2018 werd het wereldwijde aantal dementiepatiënten geschat op 50 miljoen, en naar verwachting zal dit aantal stijgen tot 152 miljoen in 2050 (World Alzheimer report 2018). De meest voorkomende oorzaak van dementie is de ziekte van Alzheimer. Deze ziekte wordt gekenmerkt door geheugenverlies en achteruitgang in de cognitieve functies veroorzaakt door aantasting van de hersenen (neurodegeneratie). De neurodegeneratie begint al ruim voordat de eerste cognitieve symptomen van de ziekte van Alzheimer zich manifesteren. Daarom is het in kaart brengen van de neurodegeneratie geschikter voor het vroegtijdig stellen van de diagnose, dan te varen op de – later optredende – cognitieve symptomen. Kenmerkend voor de ziekte van Alzheimer is atrofie van de grijze stof. Deze atrofie bevindt zich in eerste instantie in de regio van de hippocampus. Later breidt het zich uit naar de rest van de hersenen.

MRI-scans

Neurodegeneratie van grijze stof kan goed in kaart worden gebracht met een anatomische magnetic resonance imaging (MRI)-scan. Anatomische MRI-scans worden om deze reden door artsen gebruikt voor het ondersteunen van een Alzheimer diagnose.

Naast de atrofie van de grijze stof, worden bij de ziekte van Alzheimer ook de wittestofbanen aangetast. Wittestofbanen verbinden de grijzestofgebieden voor het overbrengen van informatie. In de grijzestofgebieden wordt informatie verwerkt, de witte stof is cruciaal voor de *snelheid* waarmee iemand die informatie kan verwerken. De wittestofbanen kunnen in kaart worden gebracht met een diffusiegewogen MRI-scan. Dit type MRI-scan wordt niet gebruikt voor de klinische diagnose van de ziekte van Alzheimer, maar biedt wel aanvullende informatie over neurodegeneratie. Dit type scan zou een anatomische MRI-scan daarom kunnen aanvullen bij de diagnose van de ziekte van Alzheimer.

Daarnaast wordt bij de ziekte van Alzheimer de functionele connectiviteit tussen hersengebieden aangetast. De functionele connectiviteit tussen hersengebieden wordt gedefinieerd als synchrone patronen van hersenactiviteit. Dit kan worden gemeten met een functionele MRI-scan. Dit type MRI-scan zou daarom ook bij kunnen dragen aan de diagnose van de ziekte van Alzheimer.

MRI-scans voor Alzheimerdiagnose

Omdat MRI-scans neurodegeneratie en verandering van hersenfunctie bij Alzheimerpatiënten goed in kaart kunnen brengen, kunnen zij potentieel gebruikt worden voor het vroegtijdig en nauwkeurig diagnosticeren van deze ziekte. In de klinische praktijk wordt tot nog toe alleen gebruikt gemaakt van anatomische MRI-scans om de diagnose te ondersteunen. Er wordt dan gekeken naar enkele kenmerken van de ziekte, zoals atrofie van de hippocampus en globale atrofie, maar deze kenmerken zijn op zichzelf niet onderscheidend genoeg voor een betrouwbare Alzheimerdiagnose. De ziekte van Alzheimer gaat echter ook gepaard met andere, subtielere veranderingen. Hoewel deze veranderingen op het oog niet makkelijk waar te nemen zijn, zou ook deze informatie uit een MRI-scan gebruikt kunnen worden om de Alzheimerdiagnostiek te verbeteren. Daarnaast zou een combinatie van verschillende typen MRI-scans tot een nog nauwkeuriger diagnose kunnen leiden.

MRI-scans voor Alzheimerclassificatie

In eerder Alzheimeronderzoek werden met name de gemiddelde verschillen in neurodegeneratie tussen Alzheimerpatiënten en gezonde ouderen in kaart gebracht. Er treedt bij normale veroudering ook neurodegeneratie op en deze overlapt deels met de neurodegeneratie die bij de ziekte van Alzheimer wordt gezien. Als er naar slechts één herseneigenschap wordt gekeken, is het door die overlap moeilijk om een nauwkeurig onderscheid te maken tussen Alzheimerpatiënten en gezonde ouderen. Daarom is de focus van het MRI-onderzoek in de afgelopen jaren verschoven naar classificatiestudies. Dit type onderzoek heeft als doel om met behulp van statistische technieken verschillende groepen zo nauwkeurig mogelijk van elkaar te onderscheiden, op basis van meerdere herseneigenschappen tegelijk. Dat maakt het mogelijk om van een enkele MRI-scan verschillende subtiele vormen van hersenatrofie te berekenen en deze allemaal mee te nemen in één classificatiemodel. Daarnaast kan informatie van verschillende typen MRI-scans gecombineerd worden in een multimodaal MRI-model. In dit proefschrift worden eerst twee onderzoeken beschreven waarin met behulp van bovengenoemde techniek een zo nauwkeurig mogelijk classificatiemodel wordt gemaakt voor de ziekte van Alzheimer. Vervolgens wordt in het derde onderzoek de klinische toepasbaarheid van deze modellen onderzocht. In het vierde onderzoek worden MRI-scans gebruikt om toekomstige cognitieve achteruitgang te voorspellen.

Combineren anatomische MRI-maten

In het eerste onderzoek gebruikten we anatomische MRI-scans om verschillende anatomische eigenschappen te berekenen en daarmee Alzheimerpatiënten van gezonde ouderen te kunnen onderscheiden. De dichtheid van de grijze stof van corticale hersengebieden en de volumes van subcorticale hersengebieden bleken hiervoor het meest geschikt. Het combineren van verschillende eigenschappen bleek de nauwkeurigheid van het classificatiemodel te verhogen. We vergeleken deze resultaten met een classificatie op basis van alleen het volume van de hippocampus en een globale maat voor hersenatrofie. Deze worden namelijk vaak gebruikt bij het diagnosticeren van Alzheimerpatiënten. Het combineren van verschillende anatomische MRI-maten leverde een nauwkeuriger classificatie op dan deze twee traditionele hersenmaten.

Combineren functionele MRI-maten

In het tweede onderzoek gebruikten we eenzelfde benadering voor functionele MRI-scans. We berekenden op verschillende manieren de functionele connectiviteit tussen hersengebieden om daarmee Alzheimerpatiënten van gezonde patiënten te kunnen onderscheiden. Ook in dit geval bleek een combinatie van verschillende maten het meest nauwkeurige classificatiemodel op te leveren.

Klinische toepasbaarheid Alzheimerclassificatiemodellen

In het derde onderzoek evalueerden we in hoeverre op MRI gebaseerde Alzheimerclassificatiemodellen toepasbaar zijn in de klinische praktijk. Dit is een uitdaging om tenminste twee redenen. Deze classificatiemodellen zijn namelijk ontwikkeld om Alzheimerpatiënten van gezonde ouderen te onderscheiden, maar klinische populaties zijn heterogener. Patiënten kunnen milde cognitieve klachten hebben als gevolg van het voorstadium van de ziekte van Alzheimer, of als gevolg van een ander type dementie. Daarnaast gebruiken ziekenhuizen niet allemaal dezelfde MRI-scanners en de MRI-scanners kunnen verschillen in software-instellingen. Een Alzheimerclassificatiemodel dat is gemaakt met MRI-scans van ziekenhuis A is daarom niet per definitie toepasbaar op MRI-scans van ziekenhuis B.

We testten de Alzheimerclassificatiemodellen daarom op diverse patiëntgroepen uit geheugenklinieken van vier verschillende ziekenhuizen. We analyseerden modellen op basis van anatomische MRI-scans, diffusie MRI-scans en functionele MRI-scans. Alle drie de MRI-modaliteiten bleken de patiënten met de ziekte van Alzheimer boven kansniveau te kunnen onderscheiden van proefpersonen met andere geheugenklachten. De modellen op basis van anatomische MRI-scans bleken het beste te presteren en de diffusiegewogen MRI-modellen konden de verschillende proefpersonen het minst goed onderscheiden.

Voorspellen toekomstige cognitieve achteruitgang

In het vierde onderzoek onderzochten we of multimodale MRI-scans cognitieve achteruitgang over een periode van 2 jaar konden voorspellen. De scans bleken een verandering in het cognitief functioneren - gemeten met de MMSE - boven kansniveau te kunnen voorspellen. Verandering van cognitief functioneren gemeten met zes andere testen, die ieder een specifiek aspect van het geheugen en het executief functioneren meten, kon met de scans echter niet boven kansniveau voorspeld worden.

Conclusie

De eerste conclusie die uit dit proefschrift getrokken kan worden is dat het combineren van verschillende typen informatie uit eenzelfde MRI-scan de nauwkeurigheid van Alzheimerclassificatie kan vergroten. Dit geldt zowel voor anatomische MRI-scans als voor functionele MRI-scans. Voor anatomische MRI-scans heeft dit implicaties voor de klinische praktijk. Er wordt nu al gebruik gemaakt van anatomische MRI-scans ter ondersteuning van een Alzheimerdiagnose, maar deze scans worden blijkbaar nog niet optimaal benut. Voor functionele MRI-scans heeft dit niet direct een klinische implicatie, want deze scans zijn (nog) niet geschikt voor een Alzheimerdiagnose. Dit type scan is echter veelbelovend voor vroegdiagnostiek, want de ziekte van Alzheimer gaat gepaard met een verschil in hersenfunctie voordat atrofie zichtbaar is. Daarnaast bleek dat MRI-classificatiemodellen tot op zekere hoogte kunnen worden gebruikt om Alzheimer te diagnosticeren binnen een heterogene patiëntpopulatie. Deze generalisatie naar de klinische praktijk ging het best met anatomische MRI-scans en minder goed met diffusiegewogen- en functionele MRI-scans. Verder blijken MRI-scans in enige mate voorspellend te zijn voor toekomstige cognitieve achteruitgang.

Dankwoord

Voor de totstandkoming van dit proefschrift wil ik graag enkele personen bedanken. Ik was gezegend met drie fijne (co)promotoren die er altijd voor me waren en waarmee ik met veel plezier heb samengewerkt. Serge, bedankt voor de goede begeleiding. Ik vind het knap hoe je overzicht hebt kunnen houden op de haalbaarheid en relevantie van dit proefschrift, en als het moest ook tot in detail mee hebt kunnen denken over complexe methodologie. Daarnaast wil ik Mark bedanken voor het altijd kritisch blijven reflecteren op de wetenschappelijke integriteit van mijn proefschrift. Jeroen, bedankt voor je empathische begeleiding in tijden dat het schrijven aan het proefschrift moeizaam ging. Het was verder goed dat je me af en toe in de war maakte met al je ideeën, zodat ik uit mijn tunnelvisie kwam.

Tijn, jij hebt veel bijgedragen aan dit proefschrift. Ik denk met veel plezier terug aan onze discussies over relevante en minder relevante zaken. Ik heb veel van je geleerd en ben blij dat je als paranimf ook bij het laatste gedeelte van mijn promotietraject betrokken bent.

Daarnaast wil ik iedereen van het 'fMRI team' bedanken. Bernadet, bedankt voor je rol als gids tijdens de 3 oktoberfeesten. Mark, Christiane, Rogier en Jeffrey; het was leuk om met jullie samen te werken en congressen te bezoeken.

Ik wil de hele M&S sectie bedanken voor de gezellige lunches en de gemoedelijke sfeer. Ik kijk ernaar uit om in de toekomst met jullie samen te blijven werken op het gebied van onderzoek en onderwijs.

Mijn dank gaat verder uit naar alle proefpersonen die dit proefschrift mogelijk hebben gemaakt. Ook wil ik de collega's bedanken die de data verzameld hebben, met name Marisa Koini en Anne Hafkemeijer.

Daarnaast wil ik twee vrouwen bedanken die belangrijk zijn geweest in het begin van mijn wetenschappelijke carrière. Marijke Engels-Freeke, bedankt voor het vertrouwen dat je mij hebt gegeven om als bachelorstudent statistiekwerkgroepen te verzorgen. Ik vond dit enorm leuk en het heeft mij gemotiveerd om in het academische onderwijs verder te gaan. Hilde Huizenga, ik wil je bedanken omdat je mij tijdens mijn master hebt gemotiveerd voor statistisch onderzoek met fMRI-data.

Papa, mama, ik wil jullie bedanken omdat jullie altijd vertrouwen in me hebben gehad. Door deze steun heb ik keuzes kunnen maken die goed bij me passen. Ook wil ik vrienden en familie bedanken die niet direct hebben bijgedragen aan dit proefschrift, maar er wel voor me zijn geweest in deze periode. In het bijzonder ben ik mijn broertje Mark dankbaar, omdat je altijd voor me klaar staat. Ik ben blij dat je me straks bijstaat als paranimf.

Lorette, bedankt voor je vele steun tijdens het schrijven van mijn proefschrift. Wat vind ik het fijn om samen met jou twee dochters te hebben. Ik heb veel zin in onze toekomst. Emma & Josephine, bedankt dat jullie mij een heel gelukkige vader maken.

Curriculum Vitae

Frank de Vos was born in 1985 in Amsterdam. He graduated from the Montessori Lyceum Amsterdam in 2003. He started studying physics at the University of Twente. In 2004 he switched to studying Psychology at the university of Amsterdam, because he realised he was more interested in humans than in matter. His Bachelor's degree focused on statistical methods for Psychology. During the Bachelor program, Frank participated in the European Erasmus program and went to Bordeaux to study half a year at the Université Victor Ségalen. During his Master program, he studied another six months abroad; at the University of Texas (Austin). There he wrote his thesis in the lab of Russel Poldrack, on the statistical analysis of fMRI data. After graduating in 2012, Frank started working at the Methods and statistics unit of the institute of psychology at Leiden University. Under the supervision of Mark de Rooij, he evaluated methods for the analysis of longitudinal fMRI data. This was followed by the current PhD research under the supervision of Serge Rombouts. Frank is now appointed as an assistant professor at the Methods and statistics unit. He teaches various statistics courses for Bachelor and Master students, and performs research on the statistical analysis of MRI data.

List of publications

2020

1. de Vos, F., Schouten, T. M., Koini, M., Bouts, M. J. R. J., Feis, R. A., Lechner, A., Schmidt, R., van Buchem, M. A., Verhey, F. R. J., Olde Rikkert, M. G. M., Scheltens, P., de Rooij, M., van der Grond, J., & Rombouts, S. A. R. B. (2020). Pre-trained MRI-based Alzheimer's disease classification models to classify memory clinic patients. *NeuroImage: Clinical*, 102303.

2. Feis, R. A., van der Grond, J., Bouts, M. J. R. J., Panman, J. L., Poos, J. M., Schouten, T. M., de Vos, F., Jiskoot, L. C., Dopper, E. G. P., van Buchem, M. A., van Swieten, J. C., & Rombouts, S. A. R. B. (2020). Classification using fractional anisotropy predicts conversion in genetic frontotemporal dementia, a proof of concept. *Brain Communications*, *2*(2).

2019

3. Bouts, M. J. R. J., Van Der Grond, J., Vernooij, M. W., Koini, M., Schouten, T. M., de Vos, F., Feis, R. A., Cremers, L. G. M., Lechner, A., Schmidt, R., de Rooij, M., Niessen, W. J., Ikram, M. A., & Rombouts, S. A. R. B. (2019). Detection of mild cognitive impairment in a community-dwelling population using quantitative, multiparametric MRI-based classification. *Human Brain Mapping*, *40*(9), 2711–2722.

4. Feis, R. A., Bouts, M. J. R. J., de Vos, F., Schouten, T. M., Panman, J. L., Jiskoot, L. C., Dopper, E. G. P., van der Grond, J., van Swieten, J. C., & Rombouts, S. A. R. B. (2019). A multimodal MRI-based classification signature emerges just prior to symptom onset in frontotemporal dementia mutation carriers. *Journal of Neurology, Neurosurgery & Psychiatry*, *90*(11), 1207–1214.

5. Schouten, T. M., de Vos, F., van Rooden, S., Bouts, M. J. R. J., van Opstal, A. M., Feis, R. A., Terwindt, G. M., Wermer, M. J. H., van Buchem, M. A., Greenberg, S. M., de Rooij, M., Rombouts, S. A. R. B., & van der Grond, J. (2019). Multiple approaches to diffusion magnetic resonance imaging in hereditary cerebral

amyloid angiopathy mutation carriers. *Journal of the American Heart Association*, *8*(3), e011288.

6. van Duijvenvoorde, A. C. K., Westhoff, B., de Vos, F., Wierenga, L. M., & Crone, E. A. (2019). A three-wave longitudinal study of subcortical–cortical resting-state connectivity in adolescence: Testing age-and puberty-related changes. *Human Brain Mapping*, *40*(13), 3769–3783.

2018

7. Bouts, M. J. R. J., Möller, C., Hafkemeijer, A., van Swieten, J. C., Dopper, E., van der Flier, W. M., Vrenken, H., Wink, A. M., Pijnenburg, Y. A. L., Scheltens, P., Barkhof, F., Schouten, T. M., de Vos, F., Feis, R. A., van der Grond, J., de Rooij, M., & Rombouts, S. A. R. B. (2018). Single subject classification of Alzheimer's disease and behavioral variant frontotemporal dementia using anatomical, diffusion tensor, and resting-state functional magnetic resonance imaging. *Journal of Alzheimer's Disease*, *62*(4), 1827–1839.

8. de Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., Schmidt, R., de Rooij, M., & Rombouts, S. A. R. B. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *Neuroimage*, *167*, 62–72.

9. Feis, R. A., Bouts, M. J. R. J., Panman, J. L., Jiskoot, L. C., Dopper, E. G. P., Schouten, T. M., de Vos, F., van der Grond, J., van Swieten, J. C., & Rombouts, S. A. R. B. (2018). Single-subject classification of presymptomatic frontotemporal dementia mutation carriers using multimodal MRI. *NeuroImage: Clinical*, *20*, 188–196.

2017

10. Klaassens, B. L., van Gerven, J., van der Grond, J., de Vos, F., Möller, C., & Rombouts, S. A. R. B. (2017). Diminished posterior precuneus connectivity with the default mode network differentiates normal aging from Alzheimer's disease. *Frontiers in Aging Neuroscience*, *9*, 97.

11. Schouten, T. M., Koini, M., de Vos, F., Seiler, S., de Rooij, M., Lechner, A., Schmidt, R., van den Heuvel, M., van der Grond, J., & Rombouts, S. A. R. B. (2017). Individual classification of Alzheimer's disease with diffusion magnetic resonance imaging. *Neuroimage*, *152*, 476–481.

2016

12. de Vos, F., Schouten, T. M., Hafkemeijer, A., Dopper, E. G. P., van Swieten, J. C., de Rooij, M., van der Grond, J., & Rombouts, S. A. R. B. (2016). Combining multiple anatomical MRI measures improves Alzheimer's disease classification. *Human Brain Mapping*, *37*(5), 1920–1929.

13. Schouten, T. M., Koini, M., de Vos, F., Seiler, S., van der Grond, J., Lechner, A., Hafkemeijer, A., Möller, C., Schmidt, R., & de Rooij, M. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage: Clinical*, *11*, 46–51.

2011

14. Weeda, W. D., de Vos, F., Waldorp, L. J., Grasman, R., & Huizenga, H. M. (2011). arf3DS4: An integrated framework for localization and connectivity analysis of fMRI data. *Journal of Statistical Software*, *44*(14), 1–33.