



Universiteit
Leiden
The Netherlands

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

Schouten, T.M.

Citation

Schouten, T. M. (2019, September 18). *Classification and early detection of dementia and cognitive decline with magnetic resonance imaging*. Retrieved from <https://hdl.handle.net/1887/78450>

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/1887/78450>

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

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/78450> holds various files of this Leiden University dissertation.

Author: Schouten, T.M.

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

Issue Date: 2019-09-18

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

Tijn Schouten

This PhD thesis was formatted using the L^AT_EX book class with custom adjustments.

The cover was created with adaptive style transfer (Sanakoyeu, Kotovenko, Lang, & Ommer, 2018) by applying the style of Picasso to an ICA component of mean diffusivity. The software is available online: (<https://github.com/CompVis/adaptive-style-transfer>).

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

Printing of this thesis was sponsored by Alzheimer Nederland.

Copyright©2019 Tijn Schouten

All rights reserved. No part of this publication may be reproduced, stored or transmitted in any form or by any means without permission of the author, or, when applicable, of the publisher of scientific papers.

Reference:

Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B. (2018). A Style-Aware Content Loss for Real-time HD Style Transfer, *European Conference on Computer Vision*.

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

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 18 september 2019
klokke 16.15 uur
door

Tijn Milan Schouten

geboren te Maastricht
in 1988

Promotoren:

Prof. dr. S.A.R.B. Rombouts

Prof. dr. M.J. de Rooij

Copromotor:

Dr. J. van der Grond

Leden promotiecommissie:

Prof. dr. ir. B.P.F. Lelieveldt

Prof. dr. ir. W. Kraaij

Dr. ir. A.M. Wink

Dr. K. van Deun

(Amsterdam UMC, locatie VUmc)

(Tilburg University)

Contents

1	General introduction	1
1.1	Introduction	1
1.2	Aims and outline of this thesis	4
I	Alzheimer’s Disease Classification	7
2	Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer’s disease	
	<i>NeuroImage: Clinical, 2016; 11, 46–51</i>	9
2.1	Introduction	11
2.2	Materials and Methods	12
2.3	Results and discussion	17
2.4	Conclusion	24
3	Individual Classification of Alzheimer’s Disease with Diffusion Magnetic Resonance Imaging	
	<i>NeuroImage, 2017: 152, 476–481.</i>	27
3.1	Introduction	29
3.2	Materials and Methods	30
3.3	Classification features	33
3.4	Results and Discussion	36
3.5	Conclusion	40
3.6	Acknowledgements	40

II	Early detection of dementia	43
4	Multiple approaches to diffusion MRI in hereditary cerebral amyloid angiopathy mutation carriers	
	<i>Journal of the American Heart Association, 2019: 8:e011288</i>	45
4.1	Introduction	47
4.2	Methods	48
4.3	Results	52
4.4	Discussion	55
4.5	Sources of Funding	57
4.6	Disclosures	57
5	Longitudinal prediction of cognitive decline using multimodal MRI	59
5.1	Introduction	61
5.2	Materials and methods	62
5.3	Results and Discussion	67
5.4	Conclusion	68
5.5	Acknowledgements	69
6	General Discussion	71
6.1	Alzheimer's Disease Classification	72
6.2	Early detection of cognitive impairment	72
6.3	Cross-validation estimates of classification performance	73
6.4	Strengths and limitations	76
6.5	Future research	78
A	Cross-validation confidence intervals	81
	Bibliography	85
	Nederlandse samenvatting	97
	Dankwoord	101
	Curriculum vitae	103

General introduction

1.1 Introduction

1.1.1 Background

Dementia is a devastating disease that millions of people suffer from. In 2018, 50 million people were suffering from dementia, and this is estimated to increase to 82 million in 2030, and 152 million in 2050 (World Alzheimer Report, 2018). About two-thirds of people with dementia have Alzheimer's Disease (AD), the remainder suffers from vascular dementia, Lewy body dementia, fronto-temporal dementia (FTD) or other less common types of dementia. All variants suffer from irreversible brain cell losses (World Alzheimer report, 2018). Despite many attempts, currently, there is no effective treatment for AD. Two types of drugs are being prescribed, but they only aim to reduce some of the symptoms, and they only work for some of the people. Between 2002 and 2012, only one new treatment for AD was approved for clinical use, which corresponds to a success rate of only 0.4% (Cummings et al., 2014). Possibly, the AD patients that participate in treatment trials are already too far in the disease development for possible treatments to be effective. Reliable early diagnosis of dementia is therefore of paramount importance for finding a cure to prevent or slow down the disease.

It is not yet fully understood what causes AD exactly, but there is substantial evidence that the proteins amyloid β (A_β) and tau are causally related to neurodegeneration in AD patients (Scheltens et al., 2016). A_β is the main component of amyloid plaques, that are found in the brains of AD patients (Hardy and Allsop, 1991; Karran et al., 2011). The other most pronounced hallmark of

AD are neurofibrillary tangles, which are formations of tau inside neurons that are thought to cause neural death (Mudher and Lovestone, 2002). The original amyloid hypothesis postulates a linear causal relation between extracellular amyloid β (A_β) deposits and neural death (Hardy and Allsop, 1991). However, while a strong relationship with the proposed neurodegenerative pathologies exists, much of the variance in cognitive decline remains unexplained, which suggests a multitude of unidentified mechanisms that contribute to dementia (Boyle et al., 2013). Also, many therapeutics that reduce A_β aggregation or production failed as an effective treatment for AD (Karran et al., 2011).

1.1.2 Biomarkers based on group differences

Several biomarkers are being used for diagnosis of AD, mainly focusing on the detection of A_β or by measuring neuronal damage, which is closely associated with tau (Jack et al., 2010). Levels of A_β and tau can be detected in the cerebral spinal fluid (CSF). A_β is a sensitive biomarker for AD, while an AD-like profile of tau and A_β was detected in mild cognitive impairment subjects who later converted to AD (Shaw et al., 2009). Furthermore, neuronal damage can be inferred from measuring metabolism with fluorodeoxyglucose PET (FDG-PET). AD is characterized by a specific pattern of reduced metabolism in the parietotemporal areas, posterior cingulate cortex, and medial temporal lobe (Mosconi et al., 2010). More recently, PET tracers have been developed for A_β , such as the most widely used Pittsburgh Compound-B (PIB), which can be used to determine the location of A_β depositions in the brain. This technique is especially useful for distinguishing AD from other types of dementia (Rowe et al., 2007; Mosconi et al., 2010).

A non-invasive alternative for PET is magnetic resonance imaging (MRI). Arterial spin labeling can provide similar information to FDG-PET, but is less expensive and is easily obtained in the same session as other MRI measures (Wolk and Detre, 2012). Structural MRI (sMRI) can be used to reliably obtain volumetric measurements, which correlate to neuronal numbers (Bobinski et al., 2000). The rate of brain atrophy measured longitudinally with sMRI correlates well to cognitive decline in patients (Fox et al., 1999). It has been hypothesized that resting-state functional MRI (rs-fMRI) might be suitable to detect subtle changes in functional connectivity between brain regions in an earlier, preclinical, phase (Sheline and Raichle, 2013; Jack et al., 2013). Diffusion MRI (dMRI) provides a way to study alterations in the white matter

and has been used to detect alterations in AD and mild cognitive impairment (Douaud et al., 2011). Additionally, dMRI can be used to study structural connectivity between brain regions (Behrens et al., 2007).

1.1.3 Machine Learning Classification

The drawback of studies that focus on group differences is that they are often not suited for individual predictions. If an average group difference for some measure exists, but there is considerable overlap between groups, then the measure will not perform well for individual classification. However, if the sample size is sufficiently large, there may be a highly significant group difference (Arbabshirani et al., 2017). Contrarily, even when groups do not differ on average for some measure, a multivariate combination of measures may still reliably separate groups and make individual predictions. Furthermore, when applied to a new dataset, predictions about the unseen data can be made, which opens up major opportunities for accurate, automated, differential diagnosis.

While MRI research on AD and dementia has traditionally focused on group differences, more recently attention has shifted towards individual classification (Rathore et al., 2017; Arbabshirani et al., 2017). Specifically, machine learning techniques have been applied to MRI data that are aimed to detect multivariate patterns that are specific to a disease.

A large number of studies evaluate classification of Alzheimer’s disease using public databases such as Alzheimer’s Disease Neuroimaging Initiative (ADNI). In an extensive study, Samper-González et al. (2018) evaluated different classification methods based on T1 MRI and PET on a number of open databases including ADNI. They found that PET outperformed MRI, and that out of commonly used classification methods, linear support vector machines and regularized logistic regression performed similarly, and both outperformed random forest. Furthermore, various choices in preprocessing, such as the use of atlases versus voxel-wise, or the size of smoothing kernels, had minimal effect on classification performance.

In addition to using only structural MRI or PET for AD classification, multiple modalities can be combined. By using a combination of structural MRI, PET, levels of A_{β} in the cerebral spinal fluid, and genotype, classification can be improved over using a single modality (Young et al., 2013). Additionally, by identifying subtypes within the heterogeneous group of AD patients, disease progression can be predicted more accurately (Lorenzi et al., 2019).

In order to study how well computer aided classification generalizes to unseen data the CADDementia challenge was organized. The goal of the challenge was for independent teams to provide AD and mild cognitive impairment classification algorithms based on structural MRI, which were subsequently evaluated with data that the organizers held back. The best performing algorithms used voxel-based morphometry, or a combinations of multiple measures derived from structural MRI (Bron et al., 2015).

1.2 Aims and outline of this thesis

From previous research we know that combining data sources, and multiple representations of the data can improve classification performance. In this thesis we aim to extend upon this knowledge by using various types of MRI data and combining MRI modalities and representations of these modalities.

In Part I of this thesis we explore different approaches to classify patients with AD and controls on an individual basis using machine learning with MRI scans. In chapter 2 we combine measures from multiple MRI modalities. In chapter 3 we dive deeper into multiple approaches to analyze diffusion MRI data to explore which diffusion MRI measures are most suitable for AD classification.

Early detection of dementia is an important goal that could help develop treatments. Therefore we explore how our methods perform in cases of early pre-symptomatic dementia in Part II. In chapter 4 we explore a sample of symptomatic and pre-symptomatic hereditary cerebral amyloid angiopathy mutation carriers. These mutation carriers are almost certain to develop a form of dementia similar to cerebral amyloid angiopathy. We compare these mutation carriers to normal controls in a presymptomatic and symptomatic phase. In chapter 5 we perform the prediction of cognitive test scores on a dataset of elderly who are at risk for future cognitive decline. We use baseline multimodal MRI to predict cognitive decline after a follow-up period of four years.

Part I

Alzheimer's Disease Classification

Chapter 2

Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease

Published in NeuroImage: Clinical, 2016; 11, 46–51.

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

Abstract

Magnetic resonance imaging (MRI) is sensitive to structural and functional changes in the brain caused by Alzheimer’s disease (AD), and can therefore be used to help diagnosing the disease. Improving classification of AD patients based on MRI scans might help to identify AD earlier in the disease’s progress, which may be key in developing treatments for AD. In this study we used an elastic net classifier based on several measures derived from the MRI scans of mild to moderate AD patients ($N = 77$) from the prospective registry on dementia study and controls ($N = 173$) from the Austrian stroke prevention family study. We based our classification on measures from anatomical MRI, diffusion weighted MRI and resting state functional MRI. Our unimodal classification performance ranged from an area under the curve (AUC) of 0.760 (full correlations between functional networks) to 0.909 (grey matter density). When combining measures from multiple modalities in a stepwise manner, the classification performance improved to an AUC of 0.952. This optimal combination consisted of grey matter density, white matter density, fractional anisotropy, mean diffusivity, and sparse partial correlations between functional networks. Classification performance for mild AD as well as moderate AD also improved when using this multimodal combination. We conclude that different MRI modalities provide complementary information for classifying AD. Moreover, combining multiple modalities can substantially improve classification performance over unimodal classification.

Key words: Alzheimer’s disease; classification; multimodal; MRI; fMRI; DWI

2.1 Introduction

Early diagnosis is key to the development of treatments for Alzheimer’s disease (AD) (Prince et al., 2011). In this respect it is well recognised that magnetic resonance imaging (MRI) might be highly useful as an early AD biomarker (Jack et al., 2010). Several MRI techniques have been applied successfully to study average group differences between AD patients and controls in voxel based grey matter (Ferreira et al., 2011), white matter (Li et al., 2012), diffusion measures (Douaud et al., 2011), and functional connectivity (Gour et al., 2014; Binnewijzend et al., 2012).

In addition to average group difference in case control studies, similar MRI measures have also been used to predict or classify the disease class (i.e., patient or control) of individuals. This classification based on MRI scans could be helpful in making a reliable diagnosis of AD in the future. Machine learning classification is a suited candidate to make such individual predictions, because it is well equipped to handle high-dimensional data such as those from MRI. Reliable individual classification of AD and controls has already been achieved with MRI measures of grey matter atrophy (Klöppel et al., 2008; Plant et al., 2010a; Cuingnet et al., 2011), white matter integrity (Nir et al., 2014), and brain activity (Lee et al., 2013; Koch et al., 2012).

Some studies suggest that classification of Alzheimer’s disease may further improve when combining several MRI modalities (Mesrob et al., 2012; Sui et al., 2013b), while another recent study found better classification by using a single MRI modality (Dyrba et al., 2015). It is not yet clear which MRI modality or combination of modalities provide the best classification performance of AD patients.

The goal of this study is to perform individual classification of mild to moderate AD from healthy controls, and to combine information from several modalities to improve this individual classification. We compare classification performance for typical measures of grey matter atrophy, white matter integrity, and functional connectivity. Then we investigate whether combining modalities improves classification performance. We test how this multimodal classification model is able to separate patients with mild AD and patients with moderate AD from healthy controls.

2.2 Materials and Methods

2.2.1 Data sample

Participants

Our dataset was collected as a part of the prospective registry on dementia (PRODEM; see also Seiler et al., 2012). Our sample only contained subjects scanned at the Medical University of Graz. The inclusion criteria are: dementia diagnosis according to DSM-IV criteria (American Psychiatric Association, 2000), non-institutionalization and no need for 24-hour care, and availability of a caregiver who agrees to provide information on the patients' and his or her own condition. Patients were excluded from the study if they were unable to sign a written informed consent or if co-morbidities were likely to preclude termination of the study. We conducted our study with the baseline scans from the PRODEM study, and included only patients diagnosed with AD in according the NINCDS-ADRDA Criteria (McKhann et al., 1984), for which anatomical MRI, diffusion MRI, and resting state functional MRI scans were present. Amyloid imaging for additional confirmation of the diagnosis was unavailable in our sample.

The healthy controls were drawn from the Austrian Stroke Prevention Family Study, which is a prospective single-centre community-based follow-up study with the goal of examining the frequency of vascular risk factors and their effects on cerebral morphology and function in the healthy elderly. On the basis of structured clinical interview and a physical and a neurological examination, participants had to be free of overt neurologic or psychiatric findings and had to have no history of a neuropsychiatric disease, including cerebrovascular attacks and dementia. The study protocol was approved by the ethics committee of the Medical University of Graz, Austria, and written informed consent was obtained from all subjects.

This resulted in a dataset of 77 AD patients between ages 47 and 83, of which 39 had mild AD ($\text{MMSE} > 20$), and 38 had moderate AD ($\text{MMSE} \leq 20$) (Pernecky et al., 2006), and 173 healthy controls between ages 47 and 83 (see Table 2.1).

Table 2.1: Demographics for the study population

Demographics	Controls	Mild AD	Moderate AD
Age	66.1 ± 8.71	70.3 ± 7.85	66.9 ± 9.06
Gender, ♂/♀	74 / 99 (57% ♀)	17 / 22 (56% ♀)	14 / 24 (63% ♀)
Education (years)	11.5 ± 2.76	11.6 ± 3.45	10.0 ± 2.79
Disease duration (months)	0.00 ± 0.00	22.6 ± 15.5	30.9 ± 30.7
MMSE	26.7 ± 5.80	24.2 ± 2.07	16.6 ± 2.73
CDR	—	0.72 ± 0.25	0.92 ± 0.39
GDS	2.11 ± 2.15	2.54 ± 2.09	2.74 ± 3.02

Data is represented as mean $\pm$ standard deviation. MMSE = mini mental state exam, CDR = clinical dementia rating, GDS = geriatric depression scale.

MR acquisition

Each participant was scanned on a Siemens Magnetom TrioTim 3T MRI scanner. Anatomical T1-weighted images were acquired with TR = 1900 ms, TE = 2.19 ms, flip angle = 9°, isotropic voxel size of 1 mm. Diffusion images were acquired along 12 non-collinear directions, scanning each direction 4 times with TR = 6700 ms, TE = 95 ms, 50 axial slices, voxel size = $2.0 \times 2.0 \times 2.5$ mm. Resting-state fMRI series of 150 volumes were obtained with TR = 3000 ms, TE = 30 ms, flip angle = 90°, 40 axial slices, with an isotropic voxel size of 3 mm. We instructed participants to lie still with their eyes closed, and to stay awake.

2.2.2 Software

The MRI data were preprocessed using FMRIB Software Library (FSL, version 5.0; Smith et al., 2004; Jenkinson et al., 2012). For all further data analyses we used MATLAB and Statistics Toolbox Release 2015b.

2.2.3 MRI preprocessing

The preprocessing of the anatomical MRI included brain extraction, bias field correction, and non-linear registration to standard MNI152 (Grabner et al., 2006). The preprocessing of the diffusion MRI included brain extraction and correction of eddy currents. For the fMRI data the preprocessing included brain extraction, motion correction (Jenkinson et al., 2002), a temporal high pass filter with a cutoff point of 100 seconds, and 3 mm FWHM spatial smoothing. Additionally, we used the FMRIB’s ICA-based Xnoiseifier (FIX, version 1.06),

with the included standard training data to automatically identify and remove noise components from the fMRI time course (Salimi-Khorshidi et al., 2014).

2.2.4 Anatomical Atlases

In order to compare properties across subjects we used two anatomical atlases (Figure 2.1) included in FSL. For grey matter regions we used the Harvard-Oxford probabilistic anatomical brain atlas (Desikan et al., 2006). Each brain region in this atlas consist of a probability map, where each voxel is assigned a probability of being part of each region. We split the 48 cortical regions of the Harvard-Oxford atlas into left and right hemisphere regions, resulting in 96 cortical regions. The cortical regions were combined with the 14 brain regions from the subcortical atlas, excluding the brain stem because it was not fully scanned for each participant. This resulted in a total of 110 grey matter anatomical regions. For the white matter regions we defined 20 white matter regions using the probabilistic JHU white-matter tractography atlas (Hua et al., 2008). All voxels under 25% probability per region were removed from each of the 110 grey matter, and each of the 20 white matter regions. For the analyses we used the voxel-wise probabilities that survived the thresholding for each region.

2.2.5 Anatomical features

We identified anatomical features by calculating the grey matter density (GMD), and white matter density (WMD) for each brain voxel (Zhang et al., 2001). For the GMD, we averaged the voxel-wise values for each of the 110 grey matter regions weighted by the voxel-wise region probability. This provided a measure of brain atrophy within grey matter regions. For the WMD, we averaged the voxel-wise values across each of the 20 white matter regions, weighted by voxel-wise region probability. This resulted in a feature vector of 110 average GMDs per subject, and a feature vector of 20 average WMDs per subject.

2.2.6 Diffusion features

We calculated the fractional anisotropy (FA) and mean diffusivity (MD) values for each voxel with dtfit (Basser et al., 1994b). Then we averaged those values for each of the 20 white matter regions, weighted by the region probability,

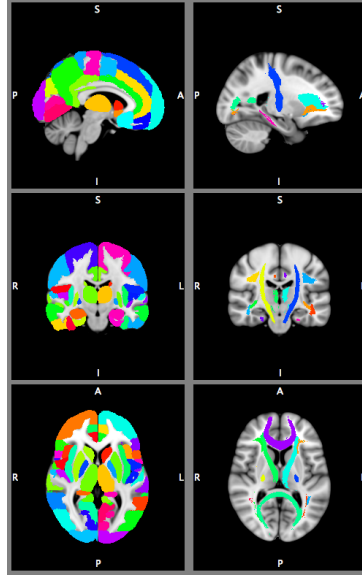


Figure 2.1: Anatomical atlases overlaid on MNI brain template. Left part shows the Harvard-Oxford cortical and subcortical areas. Right part shows the JHU white-matter tractography atlas. The images are thresholded at 25%, and showing the area with the maximum probability for displaying purposes, but the atlases were treated as probabilistic in our analyses.

and partial volume corrected with the WMD, resulting in feature vectors of 20 mean FA and MD values per subject.

2.2.7 Functional Connectivity features

We performed temporal concatenation independent component analysis (ICA) (Beckmann and Smith, 2004) with a relatively high dimensionality fixed at 70 components in order to get a more refined division of functionally coherent areas than with low dimensional ICA (Beckmann, 2012; Smith et al., 2013). We used an ICA threshold of 0.99, meaning that each voxel included in the ICA map was 99 times more likely to be part of the component than to be caused by the Gaussian background noise. Then we calculated the mean time courses for each component for each subject, weighted by the ICA weight map, and partial volume corrected with GMD.

For each component we determined the functional connectivity with every other component. We defined the functional connectivity as the full correla-

tions (FC) or as a sparse L1-regularised partial correlations (PC) between the components' time courses. We calculated the PC using the graphical lasso algorithm (Friedman et al., 2008), with $\lambda = 100$ (Smith et al., 2011). Both functional connectivity measures resulted in a feature vector of $\frac{70 \times 69}{2} = 2415$ (partial) correlations.

2.2.8 Elastic net classification with nested cross-validation

We used the aforementioned six feature vectors from the three modalities with a logistic elastic net regression for classification (Zou and Hastie, 2005; Friedman et al., 2010). We used 10-fold cross-validation to determine the generalisation performance of an elastic net regression models. For each subject this produced a predicted value between 0 and 1, where 0 represents a control subject and 1 represents an AD patient.

The elastic net regression procedure estimates a sparse regression model by imposing a penalty for including features and for the weight of each feature, so that only a subset of the features are included. To determine the parameters for the optimal size of this penalty without overestimating the classification performance we used an additional nested cross-validation loop (Varma and Simon, 2006; Kriegeskorte et al., 2009). In the outer loop we performed 10-fold cross-validation, where 9/10th of the total dataset served as training set, and 1/10th as test set. Then we performed a nested, 10-fold cross-validation on the training set over a grid of parameters to determine the penalty. We used the penalty parameters that resulted in the lowest binomial deviance in the nested loop to train the model on the original training set. This model was used to make predictions for each participant in the test set. This procedure was repeated 10 times so that each participant was part of the test set once. By using this approach we did not use the test set to estimate the model, nor the penalty parameters that we used to train the model. We also included age and sex to the model without any penalty, so that all estimated regression coefficients for the feature weights were conditional on the age and sex of the subject.

To reduce the variability in the classification outcome resulting from the random partitioning in training and test folds we repeated the entire classification procedure 50 times. This allowed us to average out this variability, and report the range of observed outcomes under different train and test set partitioning.

2.2.9 Measuring classification performance

To assess the classification performance we performed receiver operating characteristic (ROC) analyses on the estimated outcomes between 0 and 1 from the elastic net regression. We calculated the ROC curve by shifting the threshold for classifying an individual as AD from 0 to 1, and plotted the true positive rate (sensitivity) versus the false positive rate ($1 - \text{specificity}$) for each intermediate point. The area under this ROC curve (AUC) is a measure of classification performance that is insensitive to the distribution between controls and AD patients (Fawcett, 2006), so that we can take full advantage of the larger number of controls than AD patients in our dataset. We also reported the sensitivity, and specificity values corresponding to the optimal point in the ROC curve, given an equal penalty for a false positive and a false negative prediction, and the class distribution equal to that in our sample. Because we repeated the procedure 50 times, the reported AUCs, sensitivity, and specificity values are the average over the 50 repetitions of the cross-validation procedure.

Additionally, we investigated how well the predicted outcomes were able to separate mild AD from controls, and moderate AD from controls. For this purpose we also assessed the ROC curves for the mild and moderate subgroups versus controls separately.

2.2.10 Combining modalities

After assessing the performance for each individual modality we combined different modalities in order to study possible improvements in classification performance. We took a forward stepwise approach using feature concatenation to combine information from different modalities. We started with the best performing single modality feature. For each step we added each of the remaining modalities to the winning combination from the previous step. We assessed the classification performance for the combined modalities by determining the AUC. We continued the procedure until each of the modalities that we considered had been added.

2.3 Results and discussion

The classification results are summarised in tables 2.2 and 2.3 for the unimodal and stepwise multimodal procedures respectively. The AUC curves for the

Table 2.2: Alzheimer’s patients versus controls classification. The mean, minimum and maximum area under the ROC curve over 50 repetitions are reported, as well as the sensitivity, specificity, and classification accuracy for the optimal point in the ROC. Results are shown for grey matter density (GMD), white matter density (WMD), fractional anisotropy (FA), mean diffusivity (MD), full correlations between ICA components (FC), and regularised partial correlations between ICA components (PC). Multimodal represents the best combination from step 5 of our stepwise multimodal procedure (GMD, WMD, FA, MD, and SPC).

Modality	AUC	min - max	Sensitivity	Specificity	Accuracy
GMD	0.909	(0.901 - 0.915)	0.818	0.899	0.874
WMD	0.850	(0.845 - 0.858)	0.623	0.902	0.816
FA	0.789	(0.784 - 0.796)	0.547	0.885	0.781
MD	0.832	(0.823 - 0.840)	0.537	0.941	0.816
FC	0.760	(0.743 - 0.772)	0.422	0.921	0.767
PC	0.791	(0.778 - 0.803)	0.529	0.859	0.758
Multimodal	0.952	(0.946 - 0.959)	0.826	0.927	0.896

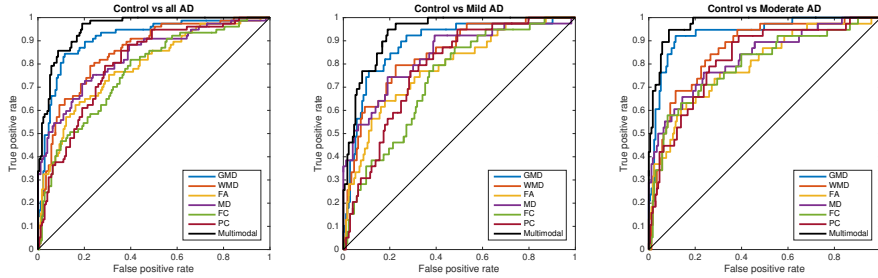


Figure 2.2: Receiver operating characteristic plot for all Alzheimer’s disease patients, mild AD, and moderate AD versus control for elastic net classification with nested cross-validation, for grey matter density (GMD), white matter density (WMD), fractional anisotropy (FA), mean diffusivity (MD), full correlation between independent components (FC), and regularised partial correlation between independent components (PC). Multimodal represents the best combination from step 5 of our stepwise multimodal procedure (GMD, WMD, FA, MD, and PC). The diagonal line represents random classification performance.

unimodal results and the best performing step of the multimodal procedure is depicted in figure 2.2.

2.3.1 Anatomical MRI

The measures derived from the anatomical MRI scan, grey matter density average of Harvard-Oxford regions, and white matter density of JHU tractography regions resulted in an excellent AUC of 0.909 and 0.850 respectively (Table 2.2). The good classification performance for GMD was expected, as AD has traditionally been seen as a grey matter atrophy disease (Frisoni et al., 2010). The classification performance with GMD that we found compares favourably to a recent similar study by Dyrba et al. (2015), who found an AUC of 0.86. While our methods were very similar, we used the Harvard-Oxford atlas to segment our data, and Dyrba et al. (2015) used the AAL atlas. The difference in atlases for segmentation, and our larger sample size might explain the difference in classification performance.

2.3.2 Diffusion weighted MRI

The measures derived from diffusion weighted MRI, fractional anisotropy and mean diffusivity of JHU tractography regions performed very reasonable with an AUC of 0.789 and 0.832 respectively (Table 2.2). This performance was much higher than the AUC between 0.652 and 0.720 that Mesrob et al. (2012) found with combined FA and MD measures, but lower than the 0.86 that Dyrba et al. (2015) found. While Mesrob et al. (2012) examined the DTI measures in grey matter areas, Dyrba et al. (2015) and our study examined the DTI measures in white matter regions, which possibly explains the differences in classification performance.

2.3.3 Functional connectivity

The measures derived from resting state functional MRI resulted in an AUC of 0.760 and 0.791 for full correlations and regularised partial correlations between ICA components respectively (Table 2.2). The higher performance of the regularised partial correlations compared to the full correlations is in line with the simulation study by Smith et al. (2011). Still, this classification performance was relatively poor compared to 0.848 found by Koch et al. (2012), and 0.80 found by Dyrba et al. (2015). Koch et al. (2012) found their result by examining the correlation between ICA components that resulted in the highest discriminative power. Because selecting this best performing correlation was not part of the cross-validation loop, their finding is likely an overestimation of

the out-of-sample generalisability. Dyrba et al. (2015) used predefined components to study the correlations between functional regions, while we used ICA on our own dataset to acquire the components, which might partly explain differences from our findings.

2.3.4 Multimodal

The stepwise procedure that we used to concatenate features from different modalities resulted in an AUC of up to 0.952 (Table 2.3). This result was achieved by starting the procedure with the best performing single modality, GMD. Classification performance improved the most when adding FA (from 0.909 to 0.933 AUC). After that, the best improvement resulted from adding WMD (0.933 to 0.949 AUC). Then, adding PC further improved classification performance marginally (0.949 to 0.951 AUC), which was subsequently improved marginally again by adding MD (from 0.951 to 0.952). Adding the FC to the previous combination decreased the classification performance compared to the previous step (from 0.952 to 0.930 AUC). The resulting best multimodal model containing GMD, FA, WMD, PC, and MD performed well above any of the modalities separately (Figure 2.2).

Our findings are in contrast with the study of Dyrba et al. (2015), who did not find any improved performance by combining similar measures derived from the same MRI modalities. This difference is possibly explained by our larger sample size, allowing many more training examples in each cross-validation fold. Additionally, they used a multi-kernel support vector machine to combine information from different modalities, while we used feature concatenation. Apparently the elastic net classifier that we used in this study is suited to select relevant predictors, even when the feature space increases through concatenation. Still, more advanced methods to combine information from multiple modalities, such as linked ICA (Groves et al., 2011), may benefit even more from the additional information from multiple modalities.

2.3.5 Mild Alzheimer’s disease and moderate Alzheimer’s disease classification

To investigate the results of our classification methods further we assessed the classification performance for mild AD and moderate AD separately. The classification results for mild AD versus controls and moderate AD versus controls

Table 2.3: Multimodal classification performance for the stepwise concatenation procedure. Each step combines the best combination from the previous step with the remaining modalities. The best result occurs with the combination of GMD, FA, WMD, PC, and MD in step 5.

Step\combined with:	GMD	FA	WMD	PC	MD	FC
1: -	0.909	0.789	0.850	0.791	0.832	0.760
2: GMD	-	0.933	0.930	0.926	0.932	0.922
3: GMD+FA	-	-	0.949	0.927	0.934	0.930
4: GMD+FA+WMD	-	-	-	0.951	0.941	0.938
5: GMD+FA+WMD+PC	-	-	-	-	0.952	0.939
6: GMD+FA+WMD+PC+MD	-	-	-	-	-	0.930

can be found in tables 2.4 and 2.5 respectively.

The single modality classification performance for moderate AD (up to 0.933 for GMD) is substantially higher than it is for mild AD (up to 0.886 for GMD). The combination of GMD, FA, WMD, PC, and MD improves the classification performance for both mild AD (from 0.886 for GMD to 0.934 for multimodal) and moderate AD (from 0.933 for GMD to 0.971 for multimodal). This improvement is mainly due to an improved sensitivity, from 0.665 to 0.721 in mild AD, and from 0.777 to 0.813 in moderate AD. At the same time the specificity also marginally improves from 0.920 to 0.935 in mild AD, and from 0.941 to 0.956 in moderate AD.

2.3.6 General discussion

In our method we took much care in the generalisability of our findings by employing a nested cross-validation approach. This approach assured that the class outcomes of the predicted subject was not required to be known when training the model, nor to estimate the model's penalty parameters. Furthermore, none of the feature reduction that we performed relied on observed class difference in our sample, which would result in overestimation of classification performance. Instead we reduced dimensionality by relying on anatomical atlases for the anatomical and diffusion features, and on data-driven unsupervised learning of independent components for the functional features. Further feature reduction was conducted in the model training phase by the elastic net classifier. Again the feature reduction in this phase did not rely on class differences in the test subjects, but only in the training subjects. Additionally,

Table 2.4: Mild AD versus controls classification. Multimodal represents the best combination from step 5 of our stepwise multimodal procedure (GMD, FA, WMD, PC, and MD).

Modality	AUC	min - max	Sensitivity	Specificity	Accuracy
GMD	0.886	(0.878 - 0.897)	0.665	0.920	0.873
WMD	0.841	(0.829 - 0.851)	0.564	0.926	0.859
FA	0.783	(0.779 - 0.790)	0.287	0.974	0.848
MD	0.838	(0.832 - 0.844)	0.369	0.993	0.878
COR	0.728	(0.706 - 0.751)	0.183	0.966	0.822
SPC	0.770	(0.737 - 0.796)	0.176	0.969	0.823
Multimodal	0.934	(0.927 - 0.944)	0.721	0.935	0.896

Table 2.5: Moderate AD versus controls classification. Multimodal represents the best combination from step 5 of our stepwise multimodal procedure (GMD, FA, WMD, PC, and MD).

Modality	AUC	min - max	Sensitivity	Specificity	Accuracy
GMD	0.933	(0.924 - 0.942)	0.777	0.941	0.912
WMD	0.860	(0.853 - 0.866)	0.515	0.936	0.860
FA	0.794	(0.787 - 0.804)	0.361	0.978	0.867
MD	0.826	(0.811 - 0.839)	0.447	0.974	0.879
COR	0.793	(0.769 - 0.823)	0.465	0.944	0.858
SPC	0.812	(0.795 - 0.829)	0.349	0.956	0.847
Multimodal	0.971	(0.964 - 0.975)	0.813	0.956	0.930

because of the relatively large sample size that we used the results were very reliable over different iterations of the cross-validation procedure, increasing our confidence that the results of the procedure generalise well.

Interestingly, the multimodal procedure resulted in the best classification performance when all modalities were combined, except for the full correlation between ICA components. The partial correlations, which were based off of the same components' time-courses, were part of the best multimodal combination. Apparently, the full correlations did not add information to the classification model over what the partial correlations did.

The improvement in classification performance in the multimodal case over the best single modality measure was substantial, especially given the relatively good performance for grey matter density. We found this multimodal improvement in both the mild AD as well as the moderate AD group. Therefore we are optimistic that these findings will apply to even earlier stages of dementia

as well.

2.3.7 Limitations

While we expect that our cross-validation procedure ensured good generalisability of the classification performance, the models that were trained to predict each subject rely heavily on both random and non-random class differences in the training sample. Therefore we cannot reliably differentiate between real and random class differences in the trained models, which is the reason that we have refrained from biological interpretation of model parameters.

Furthermore, even though the general trend in our multimodal procedure suggests that there is added information gained from combining multiple modalities, it is sometimes difficult to draw hard conclusions about which modality improves the classification the most. For example, the improvement from adding FA to GMD resulted in an AUC of 0.933, but adding MD instead resulted in an AUC of 0.932. It would be naive to conclude that the combination of GMD and FA performs better than the combination of GMD and MD. Still, the general finding is that combining modalities with decent individual classification performance improves the classification. More findings from similar research should shed light on what measures result in the most powerful combination to classify AD. Overall the elastic net classification model is very well suitable to build a good model when many features from different modalities are added, which is why the combination of all features, except full correlations, resulted in optimal classification.

In our procedure we have made some choices that could effect the results. We chose the Harvard Oxford atlas to parcellate GMD, and the JHU tracts to parcellate WMD and diffusion measures. Different atlases for parcellation might have produced slightly different results. The 70-dimensionality ICA from which we derived areas for functional connectivity was chosen because they produce a more fine grained representation of functional areas than lower dimensionality ICA. However, the dimensionality of the ICA is a trade-off between detail in the functional areas and the number of correlations, and it is not known what dimensionality is optimal in this trade-off.

The question remains how well our results generalise to cases where the patients' symptoms are less severe, such as in mild cognitive impairment, as well as to early AD diagnosis. The procedures used in this research could serve as a starting point to answer these questions.

2.4 Conclusion

In our study we found that combining information from anatomical MRI, diffusion weighted MRI, and resting state functional MRI can improve AD classification performance for both mild AD and moderate AD. The best combination in our study consisted of the average grey matter density over anatomical regions, fractional anisotropy, mean diffusivity, and white matter density over white matter tracts, and regularised partial correlations between ICA components. When only a single modality can be considered for classification, grey matter density consistently results in the best classification performance. However, when available there is a clear benefit from incorporating anatomical MRI, diffusion weighted MRI, and resting state functional MRI for diagnostic purposes. Therefore, we recommend that MRI scanning protocols designed for diagnosis of Alzheimer’s disease collect structural, diffusion, and functional MRI. Furthermore, we found that an elastic net classifier is well suited to estimate a predictive model when features from different modalities are combined by simple concatenation.

Acknowledgment

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.

Chapter 3

Individual Classification of Alzheimer's Disease with Diffusion Magnetic Resonance Imaging

Published in NeuroImage, 2017: 152, 476–481.

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

Abstract

Diffusion magnetic resonance imaging (MRI) is a powerful non-invasive method to study white matter integrity, and is sensitive to detect differences in Alzheimer's disease (AD) patients. Diffusion MRI may be able to contribute towards reliable diagnosis of AD. We used diffusion MRI to classify AD patients ($N = 77$), and controls ($N = 173$). We use different methods to extract information from the diffusion MRI data. First, we use the voxel-wise diffusion tensor measures that have been skeletonised using tract based spatial statistics. Second, we clustered the voxel-wise diffusion measures with independent component analysis (ICA), and extracted the mixing weights. Third, we determined structural connectivity between Harvard Oxford atlas regions with probabilistic tractography, as well as graph measures based on these structural connectivity graphs. Classification performance for voxel-wise measures ranged between an AUC of 0.888, and 0.902. The ICA-clustered measures ranged between an AUC of 0.893, and 0.920. The AUC for the structural connectivity graph was 0.900, while graph measures based upon this graph ranged between an AUC of 0.531, and 0.840. All measures combined with a sparse group lasso resulted in an AUC of 0.896. Overall, fractional anisotropy clustered into ICA components was the best performing measure. These findings may be useful for future incorporation of diffusion MRI into protocols for AD classification, or as a starting point for early detection of AD using diffusion MRI.

Key words: Alzheimer's disease; classification; MRI; diffusion; DTI

3.1 Introduction

Reliable and early diagnosis of Alzheimer’s disease (AD) is key to developing a cure for this disease (Prince et al., 2011). Magnetic resonance imaging (MRI) is highly useful as a biomarker for AD, and may be suitable for early detection of AD as well (Jack et al., 2010). Machine learning classification provides a powerful method to make predictions about the disease state of an individual based on MRI scans. So far individual classification studies in AD have mainly focused on anatomical MRI scans (Klöppel et al., 2008; Plant et al., 2010a; Cuingnet et al., 2011; de Vos et al., 2016). Other MRI modalities are increasingly being used for AD classification as well, such as white matter integrity measures (Nir et al., 2014), and functional MRI (Lee et al., 2013; Koch et al., 2012). White matter integrity measures are promising for predicting AD using machine learning classification (Dyrba et al., 2013; O’Dwyer et al., 2012). White matter networks have also been used for classification of mild cognitive impairment, which is often a prodromal state of AD (Wee et al., 2011, 2012). However, multiple measures can be derived from diffusion MRI scans. Traditionally, the diffusion tensor imaging model (Basser et al., 1994a) is applied to the diffusion data to derive voxel-wise measures, such as voxel-wise fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (DA), and radial diffusivity (DR). Additionally, these voxel-wise measures can be clustered into independent components, so that the individuals’ weights for each component can be used to predict AD (Ouyang et al., 2016). Furthermore, structural connectivity networks can be estimated with tractography (Behrens et al., 2007). Graph measures can then be determined based on these structural connectivity networks, such as node strength, degree, clustering, and centrality, as well as average shortest path length, or transitivity of a network (Rubinov and Sporns, 2010). It is not yet known which diffusion MRI measure is most accurate and useful for predicting AD. Moreover, combining multiple MRI-based measures may improve prediction accuracy (Schouten et al., 2016; de Vos et al., 2016; Sui et al., 2013a; Dai et al., 2012).

Here we study AD classification using diffusion MRI measures separately and combined in a comprehensive way. First we explore the predictive performance of voxel-wise diffusion tensor imaging measures using tract based spatial statistics (TBSS) of FA, MD, DA, and DR (Smith et al., 2006). Then we cluster these voxel-wise TBSS measures using independent components analysis (Beckmann, 2012), and use the mixing weights on the components for classifi-

cation (Ouyang et al., 2015). Finally, we study the predictive performance of structural connectivity of probabilistic tractography networks (Behrens et al., 2007), and of graph measures that are based on these structural connectivity networks. Additionally, we explore the combination of all measures using a sparse group lasso.

3.2 Materials and Methods

3.2.1 Data sample

Participants

Our dataset was collected as a part of the prospective registry on dementia (PRODEM; Seiler et al., 2012). Our sample only contained subjects scanned at the Medical University of Graz. The inclusion criteria are: dementia diagnosis according to DSM-IV criteria (American Psychiatric Association, 2000), non-institutionalization and no need for 24-hour care, and availability of a caregiver who agrees to provide information on the patients' and his or her own condition. Patients were excluded from the study if they were unable to sign a written informed consent or if co-morbidities were likely to preclude termination of the study. We conducted our study with the baseline scans from the PRODEM study, and included only patients diagnosed with AD in accordance to the NINCDS-ADRDA Criteria (McKhann et al., 1984), for whom diffusion MRI scans were present.

The controls were drawn from the Austrian Stroke Prevention Family Study, which is a prospective single-center community-based follow-up study with the goal of examining the frequency of vascular risk factors and their effects on cerebral morphology and function in the controls. On the basis of structured clinical interview and a physical and a neurological examination, participants had to be free of overt neurologic or psychiatric findings and had to have no history of a neuropsychiatric disease, including cerebrovascular attacks and dementia. The study protocols were approved by the ethics committee of the Medical University of Graz, Austria, and written informed consent was obtained from all subjects.

This resulted in a dataset of 77 AD patients between ages 47 and 83, and 173 controls between ages 47 and 83 (see Table 3.1).

Table 3.1: Demographics for the study population

Demographics	Control ($N = 173$)	AD ($N = 77$)
Age	66.1 ± 8.71	68.6 ± 8.58
Gender, ♂/♀	74/99 (57.2%♀)	31/46 (59.7%♀)
Education (years)	11.5 ± 2.76	10.8 ± 3.22
Disease duration (months)	0.00 ± 0.00	26.7 ± 24.5
MMSE	27.5 ± 1.83	20.4 ± 4.51
CDR	—	0.82 ± 0.34
GDS	2.11 ± 2.15	2.64 ± 2.57

Data is represented as mean $\pm$ standard deviation. MMSE = mini mental state exam, CDR = clinical dementia rating, GDS = geriatric depression scale.

MRI acquisition

Each participant was scanned on the same Siemens Magnetom TrioTim 3T MRI scanner. Anatomical T1-weighted images were acquired with TR = 1900 ms, TE = 2.19 ms, flip angle = 9°, isotropic voxel size of 1 mm. Diffusion images were acquired along 12 non-collinear directions with a b -value of $1000 \frac{\text{s}}{\text{mm}^2}$. Each direction and a $b = 0$ image was scanned 4 times with TR = 6700 ms, TE = 95 ms, 50 axial slices, voxel size = $2.0 \times 2.0 \times 2.5$ mm.

3.2.2 MRI preprocessing

The MRI data were processed using FMRIB Software Library (FSL, version 5.0; Smith et al., 2004; Jenkinson et al., 2012) unless otherwise specified. For the anatomical MRI this included brain extraction, bias field correction, and non-linear registration to standard MNI152 (Grabner et al., 2006). For the diffusion MRI this included brain extraction and eddy current correction.

3.2.3 Elastic net classification with nested cross-validation

We used the feature vectors derived from the different aforementioned techniques in a logistic elastic net regression model for classification (Zou and Hastie, 2005; Friedman et al., 2010). We used 10-fold cross-validation to determine the generalisation performance of the elastic net regression models. For each subject this produced a probability between 0 and 1 of being classified as an AD patient.

The elastic net imposes a penalty on the regression parameters to ensure that the regression model remains stable even when the number of predictors is larger than the number of observations. Specifically, it uses a combination of a least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), and Ridge penalty (Hoerl and Kennard, 1970). The LASSO penalty enforces sparse solutions, by shrinking many regression parameters to 0. The Ridge penalty smoothly shrinks the size of the regression parameters. The ratio between the two penalties is determined by a hyperparameter α , and the strength of the penalty is determined by a hyperparameter λ . When the values of these hyperparameter are estimated based on the cross-validated classification performance, the out-of-sample classification performance may be overestimated, because a combination of hyperparameters may work particularly well for one specific sample, and may not fully generalise to a different sample (Kriegeskorte et al., 2009). Therefore, we take a nested-cross-validation approach to estimate the hyperparameters (Varma and Simon, 2006), i.e., we perform an additional cross-validation within the training set to estimate the hyperparameters, and then use those settings to train a model on the entire training set in order to predict the test set. The focus of our method is on optimisation of predictive performance and not on model stability. The trade-off of this choice is that the models from the cross-validation folds may differ in sparseness and regularisation, and are therefore not suitable for interpretation (Varoquaux et al., 2016).

To reduce the variability in the classification outcome resulting from the random partitioning in training and test folds we repeated the entire classification procedure 10 times. The reported results are the average over these 10 repetitions.

3.2.4 Combining measures using the Sparse Group Lasso

To explore whether the combination of multiple sets of features improves classification we used the Sparse Group Lasso (SGL; Simon et al., 2013). Sets of features can be grouped together, and the SGL imposes a LASSO penalty between groups, and an elastic net penalty within groups. The resulting models then show sparseness between groups (i.e., the weights of some groups of features are set to zero), while also imposing some sparseness within selected groups (i.e., the weights of some features within a group is set to zero). Like the elastic net, the SGL uses the hyperparameters α to determine the mix be-

tween LASSO and Ridge within the groups, and λ to determine the strength of the penalty. We used the same nested cross-validation approach as in the elastic net procedure to choose λ , but fixed α at 0.05, resulting in a sparse between group and fairly dense within group model. We did not choose α within the nested cross-validation procedure because this was computationally impractical (10-fold, 10-repeats took about 3 weeks to calculate in parallel on a high performance computing cluster using 100 cores for a single α value), and because this procedure does a poor job at model selection (Simon et al., 2013).

3.2.5 Measuring classification performance

To assess the classification performance, we performed receiver operating characteristic (ROC) analyses on the predicted outcomes between 0 and 1 from the elastic net and sparse group lasso regression. We calculated the ROC curve by shifting the threshold for classifying an individual as AD from 0 to 1, and plotted the true positive rate (sensitivity) versus the false positive rate (1 - specificity) for each intermediate point. The area under this ROC curve (AUC) is a measure of classification performance that is insensitive to the distribution between controls and AD patients (Fawcett, 2006), so that we can take full advantage of the larger number of controls than AD patients in our dataset. We performed bootstrapping with 5000 samples to determine the standard error of the AUC. The ROC analyses were performed with the *perfcurve* function in MATLAB R2016b.

3.3 Classification features

3.3.1 Tract-based diffusion tensor features

In order to extract voxel-wise measures from the diffusion images we used tract based spatial statistics (TBSS; Smith et al., 2006). TBSS projects the subjects' diffusion measures onto a mean white matter tract, which can then be used for voxel-wise cross-subjects analyses. Because the values are comparable across subjects we can use these features for individual classification as well. Using TBSS we projected the subjects' fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (DA), and radial diffusivity (DR) onto a mean white matter skeleton that represents the center of the white matter tracts. This resulted in a feature vector with a length of 113282 values per measure for each

individual.

3.3.2 Independent Components Analyses clustered diffusion tensor features

The second method that we employ for classification is independent components analysis (ICA) based classification. We use the same voxel-wise, skeletonized measures from TBSS, but we decompose these voxel maps into a number of independent components using MELODIC (Beckmann, 2012). This resulted in a mixing matrix of one value per component per subject, and their corresponding component weight maps. We use the values from the mixing matrix in the same classification procedure as described previously. The ICA procedure is an unsupervised learning method, that does not require information about the class labels of the individuals. Therefore it was admissible to use ICA as a preprocessing step prior to the cross-validation procedure. We perform this ICA analysis separately for the FA, MD, DA, and DR maps. We call these measures FA-ICA, MD-ICA, DA-ICA, and DR-ICA to distinguish them from the voxel-wise measures.

Independent components analysis does not provide a standardised method to determine the optimal number of components for classification. The preferable method to choose a suitable number of components is to consider number of components as an additional model hyperparameter. This number can then be tuned in the nested cross-validation loop. Unfortunately this was computationally infeasible in our case. Instead we set the number of components to 28, following Ouyang et al. (2015).

3.3.3 Probabilistic tractography based structural connectivity and graph features

In order to perform tractography between comparable regions within each subject we used the Harvard-Oxford anatomical brain atlas (Desikan et al., 2006; Zhan et al., 2015). We split the 48 cortical regions of the Harvard-Oxford atlas into left and right hemisphere regions, resulting in 96 cortical regions. The cortical regions were combined with the 14 brain regions from the subcortical atlas, excluding the brain stem because it was not fully scanned for each participant. This resulted in a total of 110 grey matter anatomical regions. We removed all voxels under 25% probability of being part of any region, and then



Figure 3.1: Harvard-Oxford cortical and subcortical regions that we used as target and seed nodes for probabilistic tractography. Areas represent the probabilistic regions above the 25% threshold, and then assigned to the highest probability region.

assigned each voxel to the region for which its probability was the highest (see Fig. 3.1).

We constructed a structural connectivity network for each individual in order to perform graph analysis. We performed probabilistic tractography between 110 Harvard Oxford Atlas regions using probtrackx from FSL (Behrens et al., 2007; Zhan et al., 2015). The settings that we used were the FSL default settings (curvature threshold = 0.2, maximum number of steps = 2000, step length = 0.5mm). From each voxel within any of the atlas seed regions 100 streamlines were drawn, resulting in a 110 by 110 structural connectivity graph. The graph was made undirected by summing the upper and lower triangles of the connectivity graph, such that the connectivity between regions A and B is the sum of the connections from A to B, and from B to A. Then, in order to normalise the number of streamline counts between two regions, we divided each connection between two regions by the sum of the total number of successfully drawn streamlines from both regions. For each region, this number ranged between 3450 and 241977 streamlines depending on the size of the region and the success rate of drawing a streamline from that region. We used all the elements of the upper triangle of this connectivity graph as features for classification ($\frac{110 \times 109}{2} = 5995$ features).

After constructing the structural connectivity graphs we used the MATLAB implementation of the Brain Connectivity Toolbox (<http://www.brain-connectivity-toolbox.net>; Rubinov and Sporns, 2010) to calculate the strength, degree, clustering, and betweenness centrality for each node in each graph, and the transitivity, and characteristic path length of each graph. This resulted in 110 features per measure for strength, degree, clustering, and betweenness centrality, and a single feature for transitivity, and for path length per individual.

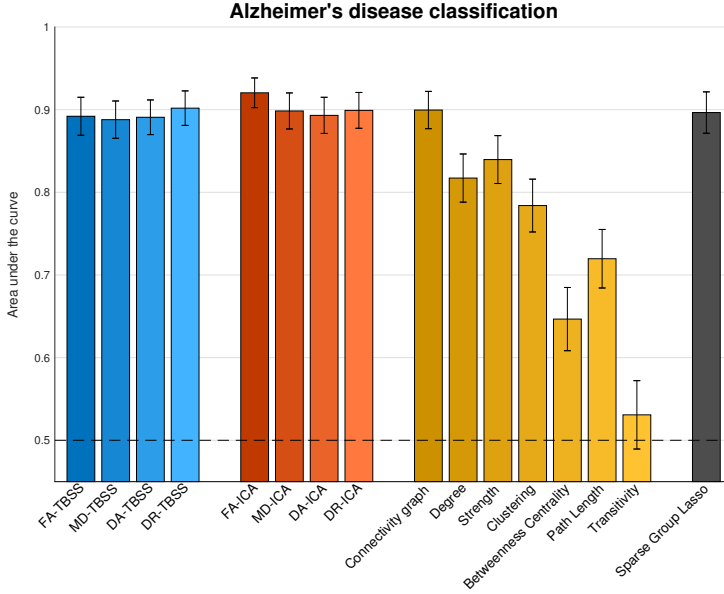


Figure 3.2: Overview of classification results. Bars indicate mean area under the receiver operating characteristics curves over 10 repetitions. The error bars represent standard errors based on 5000 bootstraps.

3.4 Results and Discussion

Detailed results for the classification procedure are summarised in Table 3.2, while an overview of the mean AUCs for each measure is depicted in Figure 3.2.

3.4.1 Classification results of tract-based diffusion tensor features

When using the voxel-wise TBSS measures for classification we found an AUC between 0.888 and 0.902 (Table 3.2). The best single measure performance was achieved with radial diffusivity (DR), closely followed by the other DTI measures.

This method is already commonly used in case control studies with AD or other patient groups, and we show that it is also suitable for individual classification. While DR slightly outperforms the other TBSS measures, the differences are small. It is likely that the differences in performance between the TBSS measures do not generalise to other datasets. Still, TBSS in general

Table 3.2: Alzheimer’s patients versus controls classification using tract-based spatial statistics, ICA-clustered TBSS measures 20 components, graph measures, and all features combined with a sparse group lasso. The mean and the bootstrapped standard error of the areas under the ROC curve over 10 repetitions are reported, as well as the sensitivity, specificity, and classification accuracy for the optimal point in the ROC.

Measure	AUC $\pm$ SE	Sensitivity	Specificity	Accuracy
FA-TBSS	0.892 $\pm$ 0.023	0.838	0.821	0.826
MD-TBSS	0.888 $\pm$ 0.023	0.844	0.792	0.808
DA-TBSS	0.891 $\pm$ 0.021	0.849	0.804	0.818
DR-TBSS	0.902 $\pm$ 0.021	0.791	0.873	0.848
FA-ICA	0.920 $\pm$ 0.018	0.868	0.844	0.851
MD-ICA	0.898 $\pm$ 0.022	0.842	0.843	0.843
DA-ICA	0.893 $\pm$ 0.022	0.897	0.806	0.834
DR-ICA	0.899 $\pm$ 0.022	0.832	0.844	0.840
Connectivity graph	0.900 $\pm$ 0.023	0.803	0.871	0.850
Degree	0.817 $\pm$ 0.029	0.799	0.740	0.758
Strength	0.840 $\pm$ 0.029	0.766	0.809	0.796
Clustering	0.784 $\pm$ 0.032	0.669	0.795	0.756
Betweenness Centrality	0.647 $\pm$ 0.038	0.595	0.668	0.646
Path Length	0.720 $\pm$ 0.035	0.625	0.727	0.696
Transitivity	0.531 $\pm$ 0.041	0.373	0.772	0.649
Sparse Group Lasso	0.896 $\pm$ 0.025	0.885	0.774	0.808

appears to be a suitable method for individual classification of Alzheimer’s disease.

3.4.2 Classification results of ICA clustered diffusion tensor features

The classification performance of ICA-clustered TBSS measures ranged between 0.893 for DA-ICA, and 0.920 for FA-ICA. The classification performance of MD-ICA (0.896), and DR-ICA (0.899) are very similar to DA-ICA.

The approach of using ICA to cluster diffusion tensor images is not commonly used, but at least one study already showed that the mixing weights of several diffusion components were useful in separating AD from normal controls (Ouyang et al., 2015).

The mixing weights of 28 components resulted in very good classification performance, up to 0.920 for FA-ICA. However, compared to voxel-wise diffusion tensor measures only FA seemed to benefit from ICA clustering. For MD, DA, and DR the classification performance remained virtually unchanged.

Even then, the ICA clustering allows an enormous reduction in the number of features required to describe an individual, from 113282 voxel-wise features to only 28 mixing weights.

One caveat with this method is that it is more difficult to extract these 28 features from an unseen individual, because the entire dataset was used to derive the mixing weights and corresponding component weight maps. One possible method is to spatially regress the feature maps (e.g., FA) of a new individual on the 28 components' weight maps, to find the individuals' mixing weights.

3.4.3 Classification results of Probabilistic tractography based structural connectivity and graph features

For the structural connectivity measures the classification performance ranged between an AUC of 0.531 for transitivity, and 0.840 for strength. Interestingly, the connectivity graph, upon which the graph measures are based, reached an AUC of 0.900, outperforming each graph measure (Table 3.2). Graph measures have been very successful in finding group differences, by summarising graphs into much fewer features than the connectivity matrix. However, in the classification context, where we can use information from the entire graph, the graph measures that we explored do not seem to be beneficial.

3.4.4 Classification results of multiple features combined with the sparse group lasso

The sparse group lasso resulted in good classification performance with an AUC of 0.896. However, this did not outperform the best measure, which was FA-ICA. Nevertheless, the properties of the sparse group lasso allow us to gain valuable insight into which measures are selected for classification, and which measures are left out of the model completely. We explored the sum of the absolute β values for each group of predictors, over the 100 different classification models resulting from 10-fold cross-validation with 10 repetitions (see Fig. 3.3). Here we see that some groups of predictors are always included in the SGL models: MD-TBSS, FA-ICA, MD-ICA, DA-ICA, DR-ICA, and Strength. Other groups of predictors are never included in the SGL models: FA-TBSS, Degree, Clustering, and Transitivity. The rest of the groups are

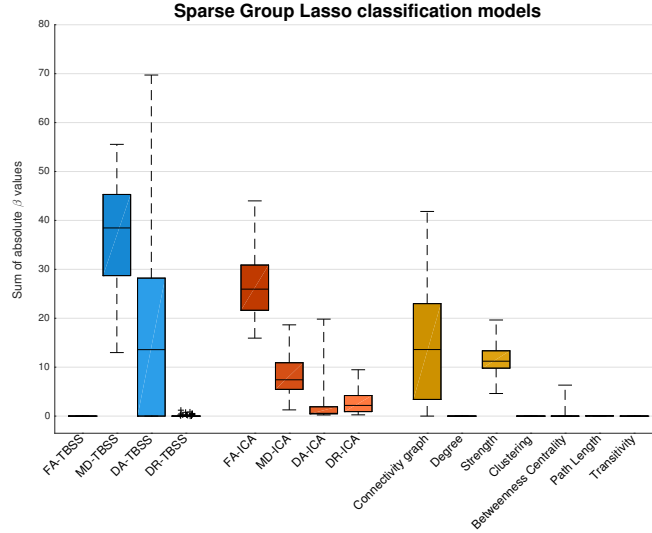


Figure 3.3: Boxplot of the sparse group lasso classification models from 10-fold times 10 repeated cross validation. The bars indicate the spread of the sum of the absolute beta values. The boxplot for DR-TBSS could not be visualised, because the lower 84% of the values was zero. The non-zero values are plotted as + signs.

sometimes included in the models and sometimes set to zero: DA-TBSS, DR-TBSS, Connectivity graph, Betweenness Centrality, and Path Length.

We observe some correspondence with the single measure classification scores (see Fig 3.2). The strongest contribution to the SGL models come from the TBSS and ICA measures, while the Connectivity graph and the Strength are also consistently selected by the SGL. This suggests that there is complementary information in the DTI measures, and the graph measures. At the same time we observe that the very good performing FA-ICA is always selected, but the almost equally well performing FA-TBSS is never selected. The same pattern, albeit it less pronounced, can be seen with MD, DA, and DR. This behaviour of the SGL is expected, as the ICA measures are based upon the TBSS measures, and do not contain complementary information. Unfortunately these mixed results for FA, MD, DA, and DR do not provide a clear winner between the TBSS and ICA approaches in terms of classification performance, but the ICA approach does have the advantage of strong feature reduction.

3.5 Conclusion

Overall, diffusion MRI is a suitable technique for classification of Alzheimer's disease (AD). Fractional anisotropy (FA) is a useful measure to detect AD, and clustering fractional anisotropy into independent components is an especially promising method that had not been fully explored previously. Using probabilistic tractography to determine structural connectivity networks can also result into decent classification performance, especially when the connectivity graph itself is considered instead of the derived graph measures. In this study we explored the possibility of using a sparse group lasso to combine multiple diffusion measures. Although this did not increase classification performance in our sample, it did suggest that FA, MD, DA, and DR could be complemented by Connectivity graphs, and Degree. The sparse group lasso could not unambiguously answer the question of the effectiveness of using ICA with TBSS measures for classification. Specifically, ICA seemed very effective for FA, while the results for MD, DA and DR were mixed. The single best performing measure was FA clustered into independent components. These findings can serve as a starting point to include diffusion MRI in procedures for early AD detection.

3.6 Acknowledgements

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

Part II

Early detection of dementia

Chapter 4

Multiple approaches to diffusion MRI in hereditary cerebral amyloid angiopathy mutation carriers

Published in Journal of the American Heart Association, 2019: 8:e011288.

Tijn M. Schouten, Frank de Vos, Sanneke van Rooden, Mark J.R.J. Bouts, Anna M. van Opstal, Rogier A. Feis, Gisela M. Terwindt, Marieke J.H. Wermer, Mark A. van Buchem, Steven M. Greenberg, Mark de Rooij, Serge A.R.B. Rombouts, & Jeroen van der Grond

Abstract

Background

Cerebral amyloid angiopathy (CAA) is a major cause of lobar intracerebral hemorrhage in the elderly. However, presymptomatic diagnosis of CAA is difficult. Hereditary cerebral hemorrhage with amyloidosis-Dutch type (HCHWA-D) is a rare autosomal dominant disease that leads to pathology similar to sporadic CAA. Presymptomatic HCHWA-D mutation carriers provide a unique opportunity to study CAA related changes before any symptoms have occurred. In this study we investigate early CAA related alterations in the white matter.

Methods and Results

We investigated diffusion MRI (dMRI) data for 15 symptomatic, and 11 presymptomatic HCHWA-D mutation carriers, and 30 noncarrier control subjects using four different approaches. We looked at 1) the relation between age and global dMRI measures for mutation carriers versus controls, 2) voxel-wise diffusion MRI, 3) independent component clustered dMRI measures, and 4) structural connectomics between (pre-)symptomatic carriers and controls. Fractional anisotropy decreased, and mean diffusivity and peak width of the skeletonised mean diffusivity increased significantly stronger over age for mutation carriers compared to controls. Additionally, voxel-wise and independent component-wise fractional anisotropy, and mean diffusivity, and structural connectomics was significantly different between HCHWA-D patients and control subjects, mainly located in the periventricular frontal and occipital regions, and in the occipital lobe. We found no significant differences between presymptomatic carriers and control subjects.

Conclusions

Diffusion MRI is a sensitive technique to detect alterations in symptomatic HCHWA-D carriers, but did not show alterations in presymptomatic carriers. This indicates that diffusion MRI may be less suitable to identify early white matter changes in CAA.

Key words: cerebral amyloid angiopathy; hereditary cerebral amyloid angiopathy; hemorrhage; diffusion mri; magnetic resonance imaging.

4.1 Introduction

Sporadic cerebral amyloid angiopathy (CAA) is a highly prevalent disease in elderly adults that is characterized by deposition of amyloid- β peptides in the media and adventitia of small leptomeningeal and cortical vessels causing hemorrhagic lesions (Vinters, 1987). Although CAA is a major cause of lobar intracerebral hemorrhage (ICH Greenberg, 1998; Charidimou et al., 2012), reliable in vivo diagnosis of sporadic CAA is difficult, especially in the early stage of the disease. Still, from a therapeutic point of view, especially the early-presymptomatic-phase of the disease is of interest. In contrast with the symptomatic phase of the disease which, is characterized by the occurrence of hemorrhagic lesions, no reliable biomarkers for disease severity or progression have been established for the presymptomatic phase. Because accumulation of amyloid- β in the vessel walls and loss of vascular smooth muscle cells are present before symptoms, it has been suggested that in sporadic CAA, ischemia- (Alonzo et al., 1998; Greenberg, 2002; Biffi and Greenberg, 2011; Reijmer et al., 2016b) or hemorrhage (Reijmer et al., 2016b; Wardlaw et al., 2017; Hartz et al., 2012) related alterations in the white matter may already be present in the early, presymptomatic, stage of the disease. Diffusion tensor imaging (DTI) is able to detect altered white matter in probable CAA (Charidimou et al., 2013, 2014; Martinez-Ramirez et al., 2013; Reijmer et al., 2016a, 2017; Salat et al., 2006).

Hereditary cerebral hemorrhage with amyloidosis-Dutch type (HCHWA-D) is an autosomal dominant disease and predominantly occurs in the Netherlands (Levy et al., 1990). The mutation leads to extensive amyloid- β deposition in the meningocortical arterioles. The underlying pathology of these amyloid deposits is similar to that in sporadic CAA with minimal or no neurofibrillary pathology (Bornebroek et al., 1996; Maat-Schieman et al., 1996; Zhang-nunes et al., 2006). Therefore, HCHWA-D may serve as a hereditary CAA model for studying early, presymptomatic, cerebral changes. In the present study we aimed to investigate the potential of diffusion tensor MRI to determine alterations in the white matter of presymptomatic HCHWA-D mutation carriers and symptomatic HCHWA-D patients compared with control participants.

4.2 Methods

The data that support the findings of this study are available from the corresponding author on reasonable request.

To analyze the diffusion MRI (dMRI) data, we applied 4 approaches. First, we explored the differences in the aging effect of DTI parameters between mutation carriers and controls (Grieve et al., 2007). Second, to get a more fine-grained image of differences between presymptomatic carriers and controls and symptomatic carriers and controls, we investigated voxel-wise DTI measures. Third, to increase power, we clustered the voxel-wise diffusion tensor measures into independent components (ICs). Finally, we investigated the structural connectivity between cortical brain areas using structural connectomics.

4.2.1 Participants

Participants were selected via the HCHWA-D patient association in Katwijk (the Netherlands) and the outpatient clinic of the Department of Neurology of the Leiden University Medical Center based on DNA analysis for confirmation of a point mutation in the APP gene (amyloid precursor protein; p.Glu693Gln mutation). Twenty-six DNA-proven HCHWA-D mutation carriers were included in the present study, of which 15 were symptomatic and 11 were presymptomatic. Participants were considered symptomatic if they had reported symptoms associated with HCHWA-D to a general practitioner. Thirty control participants were recruited from individuals at risk for HCHWA-D, for whom one parent had HCHWA-D, and from participant spouses, family members, or friends. All controls were stroke-free and tested genetically negative for HCHWA-D. Investigators remained completely blinded concerning the genetic status of participants during recruitment, MRI, and neurological and psychological exams. At the time of the study, a third of the potential mutation carriers were not aware of their genetic status. Table 4.1 shows the demographic overview of the study sample. The ethics committee of our institution approved the study, and written informed consent was obtained from all participants.

MRI Acquisition and Image Processing

Each participant was scanned at the Leiden University Medical Center on a Philips Achieva 3T MRI scanner (Philips Medical Systems) using an 8-channel

Table 4.1: Baseline characteristics of presymptomatic and symptomatic mutation carriers versus controls.

	Controls (n=30)	Pre-symptomatic carriers (n=11)	Symptomatic carriers (n=15)
Age, y	44.7 (13.7; 34–56)	33.2 (11.9; 22–46)*	55.1 (5.2; 51–60)*
Sex (male/female), n	11/9	3/8	7/8
Systolic blood pressure	129.9 (27.2; 106–142)	124.6 (14.9; 110–133)	144.2 (19.8; 129–168)
Diastolic blood pressure	81.9 (12.5; 72–87)	79.8 (9.6; 73–87)	89.1 (10.4; 78–96)
Mean arterial pressure	97.9 (16.5; 84–108)	94.7 (9.6; 88–102)	107.5 (12.4; 95–117)
Pulse pressure	48.0 (18.9; 34–58)	44.7 (13.9; 33–50)	55.1 (14.7; 43–69)
Hypertension, (%)	20	0	40
Hyperlipidemia, (%)	7	0	33*
Diabetes Mellitus, (%)	0	9	7
Cardiovascular disease, (%)	0	0	7

Data are shown as mean (SD; interquartile range), except as noted.

* indicates significant difference with controls.

head coil. Anatomical 3-dimensional T1-weighted images were acquired with repetition time of 9 ms, echo time of 4.6 ms, flip angle of 8° , and field of view of $224 \times 177 \times 168 \text{ mm}^3$. Diffusion images were acquired along 32 noncollinear directions with a b-value of 1000 s/mm^2 , along with a b=0 image with repetition time of 9033 ms, echo time of 56 ms, flip angle of 90° , 64 axial slices, voxel size of $1.96 \times 2.00 \times 2.00 \text{ mm}^3$, matrix size of $112 \times 110 \times 64$, and field of view of $220 \times 220 \times 128 \text{ mm}^3$.

The dMRI data were preprocessed using the FMRIB software library (FSL; version 5.0.8; Smith et al., 2004; Jenkinson et al., 2012) and MATLAB (R2016b). This consisted of brain extraction, motion correction, and eddy current correction. The FMRIB software library program *dtifit* was used to calculate individual fractional anisotropy (FA) and mean diffusivity (MD). FA images were subsequently nonlinearly registered to the FMRIB58 FA template for individual registration to Montreal Neurological Institute (MNI) space.

Group differences in voxel-wise DTI measures were investigated with tract-based spatial statistics (TBSS) using default settings (Smith et al., 2006). This procedure projects the center of each subject's white matter tracts onto a mean white matter tract. This allows for voxel-wise statistical analyses of FA and MD. The global FA or MD was the average FA or MD value of all voxels that were projected onto the TBSS skeleton. As an additional global measure we investigated the peak width of the skeletonized MD (PSMD) (Baykara et al., 2016), which has been specifically validated as a marker for small vessel disease.

To reduce the dimensionality of the data, we clustered the voxel-wise measures with IC analysis. This procedure identifies clusters of voxels that show similar patterns across participants and characterizes each person with a weight

for each component. Specifically, we used Multivariate Exploratory Linear Optimized Decomposition into Independent Components (MELODIC; Beckmann, 2012) to decompose each of the skeletonized FA and MD maps into 10 ICs (ie, the IC number). This resulted in 10 spatial IC weight maps and, for each participant, a mixing weight per component. The differences in mixing weights could then be compared across groups. This procedure increases power by reducing the dimensionality from $>100\,000$ voxel values to just 10 mixing weights.

The structural connectomics were constructed with probabilistic tractography between Harvard–Oxford cortical and subcortical atlas regions (Desikan et al., 2006; Zhan et al., 2015). For this purpose, the 48 cortical regions of the Harvard–Oxford cortical atlas were split into left and right hemisphere regions, resulting in 96 cortical regions. These were combined with 14 of the 15 regions from the Harvard–Oxford subcortical atlas, excluding the brain stem. These 110 probabilistic gray matter (GM) regions were given thresholds of 25% probability and voxel-wise assigned to the region with the highest probability.

Probabilistic tractography was conducted using anatomically constrained tractography (Smith et al., 2012) with spherical-deconvolution informed filtering of tractograms, or SIFT (Smith et al., 2013; Yeh et al., 2016), as implemented in *mrtrix* (v3.0_RC3-1; Tournier et al., 2012). For anatomically constrained tractography, the T1-weighted image was aligned to the $b=0$ image using rigid-body registration. Partial volume estimates of white matter, cerebrospinal fluid (CSF), and cortical GM were calculated using the FMRIB automated segmentation tool (FAST; Zhang et al., 2001), and subcortical cortical GM estimates—excluding the brain stem—were calculated with the FMRIB integrated registration and segmentation tool (Patenaude et al., 2011). Fiber orientation distributions were estimated using constrained spherical deconvolution with recursive calibration for fiber response function estimation (Tax et al., 2014) and maximum spherical harmonic order of 6. The second-order integration over fiber orientation distributions algorithm (Tournier et al., 2010) was used to generate 5 million streamlines (seeding: GM–white matter interface; step size=1.0 mm; maximum curvature=45° per step; length: 2–200 mm; and fiber orientation distribution threshold=0.0625), which were reduced to 50 000 streamlines using SIFT. Structural connectivity graphs were constructed by summing the regional streamlines that were assigned to the closest Harvard–Oxford GM region within a 2-mm radius of each streamline end point (Smith et al., 2015). This resulted in a 110×110 matrix of SIFT-filtered stream-

line counts between the regions. Subsequently, we used the MATLAB implementation of the brain connectivity toolbox (<http://www.brain-connectivity-toolbox.net>) to calculate the strength, degree, clustering, and betweenness centrality (Rubinov and Sporns, 2010) of each node in the connectivity graph, resulting in 1 value per person per region and 110 values in total per person. In addition, we characterized each participant’s graph with transitivity and global efficiency (Reijmer et al., 2016a), resulting in a single value per participant.

To correct the data for CSF partial-volume effects, we calculated the proportion of CSF of the total intracranial volume using the segmentations from FAST.

4.2.2 Statistical Analysis

For demographics, the Mann–Whitney U test was used to assess differences in age between groups; univariate general linear modeling analysis was used to assess differences in blood pressure measurements between groups, adjusted for age and sex; and χ^2 tests were used to assess differences in sex and percentage cardiovascular risk factors between groups.

To analyze the DTI data, we first explored the effect of gene presence on the decline in global, whole-brain FA, MD, and PSMD over age, namely, the interaction between age and gene presence. In addition, we explored the presence of a quadratic effect of age on the diffusion measures. To do this, we fitted 3 regression models for each of the 3 DTI measures. As covariates, we included sex and proportional CSF. The reduced model included age and gene presence. The linear age $\times$ gene interaction model also included age $\times$ gene interaction. The quadratic age $\times$ gene interaction model also included age² and age² $\times$ gene. We used partial F tests for testing the increase in explained variance of the linear age $\times$ gene interaction model compared with the reduced model and for the quadratic age $\times$ gene interaction model compared with the linear age $\times$ gene interaction model. To determine significance and to correct for multiple comparisons across all 6 tests (3 measures and 2 comparisons), we used permutation testing with 5000 permutations.

Second, we used a general linear model with age, proportional CSF, and sex as covariates to test the differences between controls and presymptomatic carriers and between controls and symptomatic patients. For the voxel-wise analyses, threshold free cluster enhancement was performed to use spatial neighborhood information (Smith and Nichols, 2009). The same general linear model

was used for the IC analysis clustered mixing matrices and for the graph measures.

To correct for multiple comparisons, we used the permutation analysis of linear models tool (Winkler et al., 2014) with 500 permutations and tail approximation for accelerated permutation inference (Winkler et al., 2016) for the voxel-wise TBSS and 5000 permutations without tail approximation for the other approaches. For each of the latter 3 approaches, the reported P values were familywise error corrected across all tests, contrasts, and modalities within each of 3 approaches: (1) voxel-wise TBSS, (2) IC analysis $\times$ clustered TBSS measures, and (3) probabilistic tractography-based graph measures.

4.3 Results

4.3.1 Global Diffusion Measures Over Age

The age $\times$ gene interaction model explained significantly more variance than the reduced model without gene interaction for FA, MD, and PSMD at $F(1, 50) = 23.51, 14.73, \text{ and } 9.68$, respectively, with corresponding permutation familywise error-corrected P values: $P < 0.001, P = 0.001, P = 0.011$ (see Figure 4.1). Adding quadratic terms did not improve the models for any of the measures, $F(2, 48) = 0.13, 0.19, 0.27$ respectively; all familywise error-corrected P values were not significant.

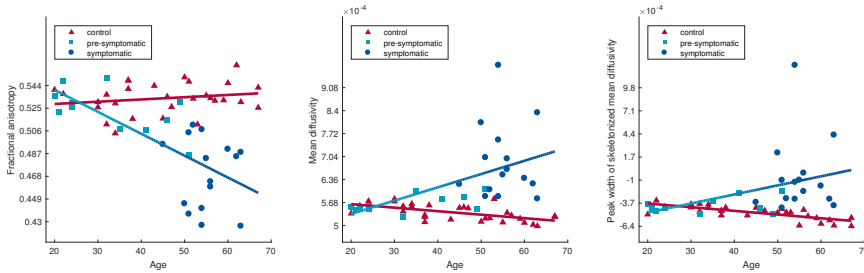


Figure 4.1: Scatterplot of the relation between age and fractional anisotropy (left), mean diffusivity (middle), and peak width of skeletonized mean diffusivity (right) for presymptomatic HCHWA-D mutation carriers, symptomatic HCHWA-D patients and control subjects. The measures are adjusted for gender and proportion CSF. The blue line indicates the trend for mutation carriers (symptomatic and presymptomatic combined), and the red line indicates the trend for controls.

4.3.2 TBSS Analysis

The tract-based spatial statistics results for FA and MD comparing HCHWA-D patients with control participants are shown in Figure 4.2. We found lower FA values in symptomatic HCHWA-D patients compared with control participants, especially in the parietal and occipital lobes. In symptomatic HCHWA-D patients, MD was increased in the white matter, especially in the periventricular frontal and occipital regions and in the centrum semiovale. No differences between presymptomatic HCHWA-D mutation carriers and control participants were found.

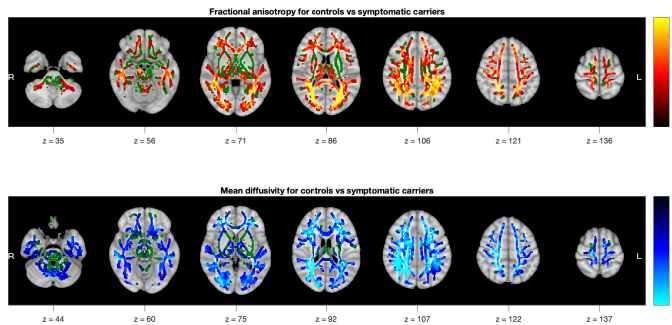


Figure 4.2: Control versus symptomatic carriers' t statistics, $P < 0.05$ family-wise error corrected over modalities, contrasts and voxels. Tract-based spatial statistics skeleton background is plotted in green. The tract-based spatial statistics skeleton was dilated with 1 voxel for visualization purposes.

4.3.3 IC Analysis

Symptomatic carriers and controls differed significantly on 4 IC mixing weights for FA (see Figure 4.3, left). The bars represent the average mixing weight of each component for each individual group, corrected for age, sex, and proportional CSF. This analysis shows that, for both FA values, the components that differed significantly had high component importance mainly in the occipital and parietal lobes. For MD, 3 ICs were identified that differed significantly between symptomatic carriers and controls (Figure 4.3, right). Overall, high component importance was located mainly in periventricular frontal and occipital regions and the occipital lobe. No significant differences between presymptomatic mutation carriers and control participants were found.

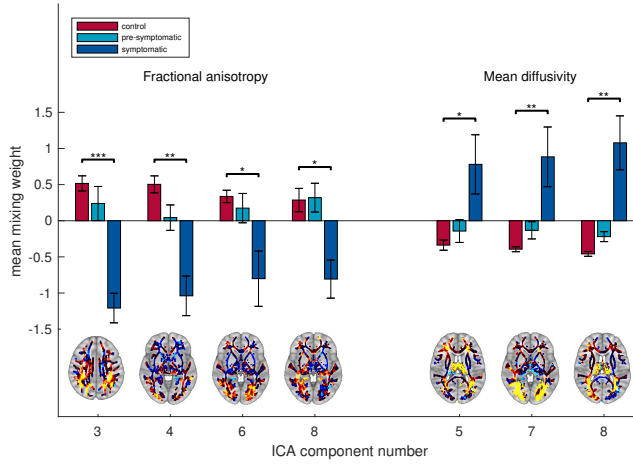


Figure 4.3: Independent component mixing weights for fractional anisotropy and mean diffusivity z-scored for displaying purposes. Means adjusted for age, gender and proportion CSF. The TBSS skeleton was dilated with 1 voxel for visualization purposes. Red-yellow indicates positive values of the weight maps (yellow is higher), while blue-cyan indicates negative values of the weight maps (cyan is more negative). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (Familywise error corrected for components, measures, and contrasts). ICA indicates independent component analysis.

4.3.4 Structural Connectomics Analyses

Figure 4.4 shows the results of the structural connectomics analyses between brain areas. This analysis reveals structural connectivity of cortical GM structures rather than intrinsic white matter tract information. Our analyses showed that the degree of structural connectivity is lower in 7 cortical regions for symptomatic carriers compared with controls (see Figure 4.4, left). In addition, in 1 of these regions, the left cuneal cortex, the clustering was higher for symptomatic carriers compared with controls (see Figure 4.4, middle). Furthermore, the betweenness centrality was higher for symptomatic carriers compared with controls in 2 regions (see Figure 4.4, right). We found no differences in strength, transitivity, or global efficiency. None of the measures showed significant differences between presymptomatic carriers and controls.

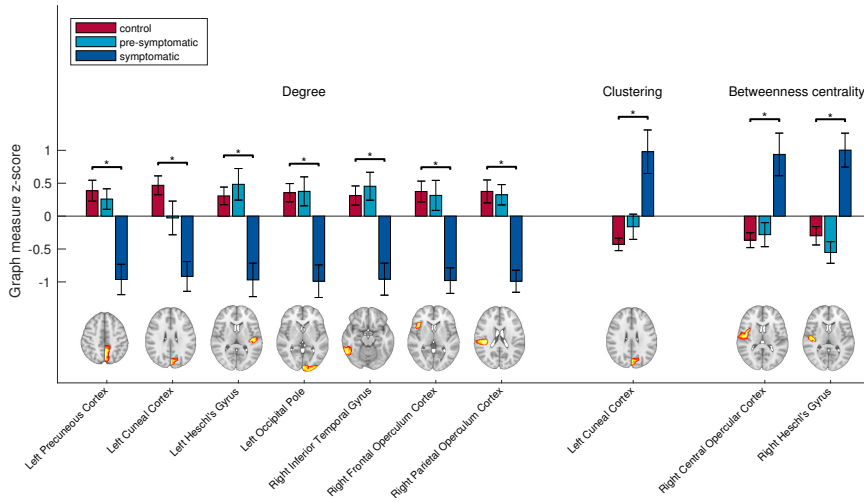


Figure 4.4: Significant structural connectomics measures averaged per group and z scores for visualization. Means adjusted for age, sex, and proportional cerebral spinal fluid. * $P < 0.05$ (familywise error corrected for regions, graph measures, and contrasts).

4.4 Discussion

Our data show that differences in DTI and DTI-related connectivity parameters between HCHWA-D patients and control participants are mainly located in the periventricular frontal and occipital regions and in the occipital lobe. We found no significant differences in any of these parameters between presymptomatic HCHWA-D mutation carriers and control participants.

In our first analysis, we found a greater decrease in global FA and a greater increase in global MD and PSMD over age for mutation carriers than for controls. White matter integrity appears largely unaffected in the presymptomatic phase and then deteriorates in a later stage, when mutation carriers become symptomatic. PSMD has been validated as a marker for small vessel disease (Baykara et al., 2016), but it was remarkably similar to global MD in our findings.

All our dMRI analyses were sensitive to differences between controls and symptomatic carriers, whereas no significant differences were found between presymptomatic mutation carriers and control participants. The symptomatic carriers had reduced FA and increased MD in widespread areas of the brain, covering almost the entire white matter skeleton. Reduced FA and increased

MD is associated with altered white matter. The effects were most pronounced in the periventricular frontal and occipital regions and in the occipital lobe of the brain. The same trend arose from IC analysis—clustered diffusion tensor measures. ICs that contributed strongly toward the posterior regions of white matter showed a significant difference between symptomatic carriers and controls. This finding is also indicative of altered white matter, specifically in the posterior regions of the brain. The finding that the occipital region is most severely affected is in line with previous findings that the occipital lobe is a predilection site of CAA (Duan et al., 2006; Zhu et al., 2012). Structural connectomics revealed a lower connection degree for symptomatic carriers compared with controls in several cortical, mainly posterior, regions, indicating fewer connections between these regions and the rest of the brain. Furthermore, clustering was higher in one of these regions. This finding is to be expected because this region has fewer neighbors, so it is more likely that these neighbors would be connected to each other. Furthermore, a decrease in degree and an increase in clustering is typical for a hierarchical network, and the structural connectivity graph has been found to show these characteristics (Bassett et al., 2008). Furthermore, betweenness centrality was higher in 2 regions for symptomatic carriers compared with controls. It is likely that because of the decrease in connections in the symptomatic patients, many of the shortest paths between nodes must pass through relatively unaffected regions, resulting in higher betweenness centrality in these regions. These findings are in line with previous research that found that structural networks are altered in patients with CAA (Reijmer et al., 2015).

Importantly, these changes in DTI measures were not yet visible in the presymptomatic phase of the diseases. Interestingly, a previous study has shown that in presymptomatic mutation carriers, the volume of white matter hyperintensities on fluid-attenuated inversion recovery imaging is slightly higher than in control participants (Van Rooden et al., 2016). This discrepancy likely is not caused by lower sensitivity of dMRI compared with white matter hyperintensity volume but rather by differences in approach. Because we used multiple approaches to analyzing the dMRI data, the statistical power was reduced by multiple comparison corrections. Still, the absence of significant differences between controls and presymptomatic carriers may be explained by the finding that white matter is still largely unaffected until a later stage of the disease. In patients with CAA, brain network connectivity in patients wors-

ened measurably over 1.3-year follow-up (Reijmer et al., 2016a). This result is in line with our finding that the decline in global diffusion measures manifests mostly at a later age. This may suggest that the high sensitivity of dMRI in symptomatic carriers may not extrapolate to earlier, presymptomatic carriers or early cases of CAA.

An important limitation of the present study is sample size. It is possible that the small sample size did not provide enough power to detect significant differences between presymptomatic carriers and controls. The rareness of the point mutation in the APP gene causing HCHWA-D makes it challenging to collect larger samples. Still, follow-up studies in a larger cohort of presymptomatic mutation carriers are needed to confirm that differences in dMRI-related connectivity are late markers for CAA. Another limitation is the limited number of diffusion directions and the relatively low b-value of 1000 s/mm² in the DWI protocol. A higher number of diffusion directions and a b-value of ≈ 3000 s/mm² would result in more accurate estimation of the fiber orientation distribution and more accurate construction of the structural connectomics, which may provide more power to detect group differences.

4.5 Sources of Funding

This study is supported by VICI Grant 016.130.677 of the Netherlands Organisation for Scientific Research. This study was supported by US National institutes of Health grant R01NS070834.

4.6 Disclosures

None

Chapter 5

Longitudinal prediction of cognitive decline using multimodal MRI

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

Abstract

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

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

5.1 Introduction

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

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

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

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

5.2 Materials and methods

5.2.1 Data sample

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

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

Table 5.1: Baseline demographics for the study population

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

Data is represented as mean (standard deviation), or as numbers (percentage)

^a Randomisation factor in the original DANTE study.

MRI acquisition

The MRI data was acquired at Leiden University Medical Center with a 3T Philips Achieva MR system with a 32-channel head coil. Anatomical T1-weighted images were acquired with TR = 9.7 ms, TE = 4.6 ms, flip angle = 8°, voxel size = $1.17 \times 1.17 \times 1.40$ mm³. Diffusion images were acquired along 32 measurement directions with TR = 9592, TE = 56 ms, flip angle = 90°, FOV = $220 \times 220 \times 128$ mm, matrix size = 112×112 , 64 slices, voxel size = $1.96 \times 1.96 \times 2$ mm³, b = 1000 mm/s². Additionally, a single b = 0 image was acquired. Resting state fMRI series of 200 volumes were obtained with TR = 2.2 ms, TE = 30 ms, flip angle = 80°, 38 axial slices, with a voxel size of $2.75 \times 2.75 \times 2.2$ mm³. We instructed participants to lie still with their eyes closed, and to stay awake. Fluid attenuated inversion recovery (FLAIR) images were acquired with TR = 11,000 ms, TE = 125 ms, flip angle = 90°, FOV = $220 \times 176 \times 137$ mm³, matrix size = 320×240 , 25 transverse slices, 5 mm slice thickness.

5.2.2 Software

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

5.2.3 Preprocessing

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

5.2.4 Anatomical MRI features

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

5.2.5 Diffusion MRI features

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

5.2.6 Resting state functional MRI features

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

5.2.7 Small vessel disease markers

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

5.2.8 Cognitive performance measures

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

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

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

Data is represented as median (Q1 – Q3).

^a Interference time to complete (seconds).

^b Time difference between Trial Making Test part A and B

^c 15-Word Verbal Learning Test

^d Mini mental state examination.

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

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

5.2.9 Group lasso prediction

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

5.2.10 Measuring predictive performance

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

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

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

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

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

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

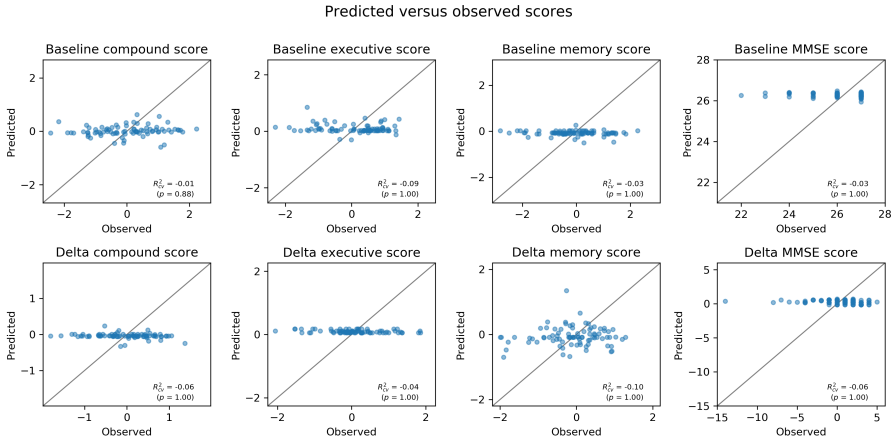


Figure 5.1: Scatterplots for predicted scores versus observed scores. Observed scores are z-scored compound scores at baseline (top), and at four year follow up (bottom). Predicted values are multimodal MRI predictions with a group lasso after cross-validation. MMSE = mini mental state examination. p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

5.3 Results and Discussion

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

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

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

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

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

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

5.4 Conclusion

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

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

5.5 Acknowledgements

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

Chapter 6

General Discussion

6.1 Alzheimer's Disease Classification

In part I of this thesis we have explored various approaches to MRI based classification of AD. In chapter 2 we used features from structural MRI, diffusion MRI, and resting-state functional MRI to classify AD. We found that using gray matter density from structural MRI alone resulted in good classification performance. When we combined measures from other modalities, the classification performance increased further, with measures from each modality included in the best performing combination.

In chapter 3 we used the same dataset to explore diffusion MRI in more detail. Instead of region-wise, we used voxel-wise diffusion MRI measures, that were registered with tract-based spatial statistics. This resulted in classification performance that was similar to the performance of structural MRI. Furthermore, using independent component analysis to cluster the diffusion measures to only 28 components resulted in similar or even slightly improved performance. In this case however, combining multiple measures from diffusion MRI did not improve performance further.

6.2 Early detection of cognitive impairment

While automatic classification of AD can be useful in a diagnosis setting, early prediction is arguably more important for development of treatments. In part II of this thesis we explored if MRI can also predict symptoms before they have occurred.

In chapter 4 we had a sample of HCHWA-D pre-symptomatic and symptomatic mutation carriers. This study aimed to explore if we can detect differences between these groups and controls using a variety of diffusion MRI measures. For the symptomatic patients, we found that each diffusion measure was able to detect extensive differences compared to controls. For the pre-symptomatic group, none of the measures showed a significant difference. The global diffusion measures showed a significantly different slope over age for mutation carriers compared to controls, indicating a different course of the diffusion measures as subjects grow older. Notably, the pre-symptomatic carriers resemble the controls on global diffusion measures, while the measures deteriorate after subjects become symptomatic.

In chapter 5 we used successful techniques from part I of this thesis in an

attempt to predict cognitive decline after a four year follow up in subjects at risk for future cognitive decline. Successfully predicting which subjects are most likely to decline in the future could help to determine whom to target for possible treatments. In addition to structural, diffusion, and resting-state functional MRI, we also used small vessel disease markers. However, our model was not able to make predictions about future cognitive decline. While MRI has proven to be very powerful in delineating groups that differ substantially in cognitive ability, such as AD from controls, it may not be sensitive enough to detect subtle differences that precede cognitive decline.

MRI is a powerful tool to reliably classify subjects with AD, but with the same techniques, it is not possible to detect future cognitive decline in our samples.

6.3 Cross-validation estimates of classification performance

In chapters 2,3, and 5 we have used cross-validation. In this section I will explain why cross-validation is used, and what some of its caveats are.

There is an important distinction between (1) model assessment, and (2) model selection (Stone, 1974), and for both of these cross-validation can be used.

Traditionally, model fitting uses the same observations to estimate the model and to assess the model fit. When the number of predictors (p) is larger than the number of observations, or the model allows for complex, non-linear relations, there could be an infinite number of possible models that fit the data perfectly. However, when the model becomes more complex than the true (unknown) function underlying the data then the model will increasingly fit noise. This is called overfitting. A model that is overfitted does not generalize well to new observations, because it is fitted in a large part to noise. On the other hand, when the model is too simple it won't be able to fit the data well, this is called underfitting. For (1) model assessment, we use generalization performance measured with cross-validation; part of the data is left out, the test set, and the remaining data is used for model fitting, the training set. Then, we apply the fitted model to the test set to get an estimate of how well the model works when applied to different data than it was trained on. To maximally use the data we rotate the test set, such that all data has been part of the test set

once.

However, when the number of predictors is large, or the model is flexible, the training data can easily overfit on the training set, which results in poor performance on the test set. Therefore we also want to do (2) model selection of a model with an appropriate complexity within our training set. For this, we can also use cross-validation. Within the training set, we perform an additional, nested, cross-validation for many different model complexity levels, and select a model complexity that performs well within the training set. Using that model complexity level we can use the entire training set to fit a model, which can then be applied to the test set. Because many model complexity levels may be tested within the training set it is likely that the best-performing complexity level in the training set depends in part on chance and does not generalize to the test set. To mitigate this, Breiman et al. (1984) suggests choosing the least complex model that performs within 1 standard error of the best performing model (1-SE rule). This parsimonious approach favors simpler models over complex models to avoid overfitting. However, this approach is arbitrary because there is no guarantee that the model complexity does not become too low. Also, the standard error within the training set cannot be estimated reliably. Therefore, we have instead opted to use the best performing model complexity in the training set to apply to the test set in chapters 2,3 and 5.

Cross-validation yields an estimate of the generalization performance that is almost unbiased. The only bias that occurs is due to the training set being smaller than the entire dataset. However, the estimated generalization performance does have variance resulting from several sources (Dietterich, 1998):

- Random sample from a population
- Random errors in class labels (such as misdiagnosis)
- Random selection of the training data
- Randomness in the learning algorithm (such as starting seed)

By repeating the cross-validation procedure multiple times, using a different train/test set partitioning, and a random seed every time we can average out the latter two sources of variance. This is repeated cross-validation. The first two sources of variance remain. Unfortunately, there are currently no methods that allow accurate estimation of this variance, which makes it difficult to perform statistical tests between two cross-validated generalization performance

estimates. For example, in chapter 2 we found improved performance for the multimodal model compared to the gray matter density, but we are not able to estimate the likelihood that this difference results from variance caused by random sampling, and random errors in class labels, or if it is a true difference. We did perform repeated cross-validation to average out the variance from a random selection of the training data, and randomness in the learning algorithm, and included the minimum and maximum classification performance across these repetitions in chapter 2 (see Table 2.2). In chapter 3 we used bootstrapping to estimate confidence intervals for classification performance (see Table 3.2). However, the predicted outcomes that were bootstrapped all depend on models fitted to largely overlapping training data. Therefore these confidence intervals do not take variance from random samples from a population and random errors in class labels into account and are therefore also not suitable for making a statistical comparison between classification performance of different measures. In chapter 5 we tested if the relation between predicted cognitive decline and observed cognitive decline is stronger than chance levels. For this purpose, it was not required to estimate the standard error of the generalization performance, but instead, we could perform a permutation test to compare the observed relation to the relation under the null hypothesis. In chapters 2 and 3 a permutation test could have been performed, but this would test the classification performance against chance levels instead of making a comparison between competing modalities, while the purpose of these studies was to explore the latter.

A recent study estimates the size of the error bars of cross-validated accuracy (Varoquaux, 2018). In a simulation experiment Varoquaux shows that with a true generalization accuracy of 75%, the 90% confidence intervals (CI) are as large as 75% $\pm$ 10% with a sample size of 100. With a sample size of 200 to 300, these CIs are about 75% $\pm$ 5-6%. To explore how these results generalize to the findings in this thesis I replicated Varoquaux's simulations adjusted to resemble the conditions of this thesis. Specifically, instead of accuracy, I used AUC. I tuned the distances between the simulated class distributions in such a way that the classifier reached an AUC of approximately 90% for a sample size of 250. Instead of balanced classes, I simulated the same 77/173 ratio as we have in chapters 2 and 3, and I used 10 repeated 10 fold cross-validation. The resulting 90% CI is 5.0% below and 4.3% above 90%, similar to the findings of Varoquaux (2018) (see Appendix A for details).

The implications of this for our research is that we should interpret the differences between measures with similar classification performance with care. In chapter 2 we have found several multimodal combinations with substantial performance improvement over gray matter density alone. We can be reasonably confident that there is a performance improvement from using multiple modalities. However, because several multimodal combinations perform very similarly, the differences between them are likely due to chance. In chapter 3 we also see very similar performance between all TBSS measures, ICA clustered measures, connectivity graph, and all measures combined. The differences between these are likely within the margin of error, which makes it difficult to differentiate between these methods based on classification performance. Notably, class imbalance in our simulated dataset results in very similar confidence intervals compared to using balanced data, suggesting that performance is robust to imbalanced data with a ratio of over 2 to 1. This finding is confirmed by Samper-González et al. (2018), who find that moderately imbalanced classes (2 to 1) does not impair performance, but very strong class imbalance (6 to 1) impairs performance.

6.4 Strengths and limitations

A major strength of our research is that we took much care to ensure the generalizability of our results by employing a repeated nested cross-validation approach. An important pitfall in neuroimaging research is using the same data to both construct a hypothesis, and to test the hypothesis, which is referred to as circular analysis or double dipping (Kriegeskorte et al., 2009). In machine learning research this is often expressed as a spill-over from the test data to the training procedure, for example, by performing feature selection that is based on class labels before cross-validation. A slightly more innocent example is to use performance on the test set to estimate the model complexity. We have however used a nested cross-validation approach, where we estimate the model complexity within the training set.

Despite our precautions, one limitation of our research involves this pitfall of spill-over from information between training and test sets. In chapter 2 we used a step-wise procedure to combine features, based on the classification performance in the previous step. However, in order to perform each next step it was first required to assess the performance of the previous step,

which required the class labels. The classification performance of the ultimate combination is therefore almost certainly an overestimation. We do however see that the combination of the best single measure, gray matter density, with each of the remaining measures, improves classification performance by roughly the same amount. Therefore we still support the overall conclusion that multimodal MRI likely has performance benefits over single modality MRI, but we recommend against using a step-wise procedure. Since this research we have instead opted for using a (sparse) group lasso in chapters 3 and 5 to combine features. The (sparse) group lasso selects groups of features automatically within a single model fitting procedure and therefore eliminates the need to explore all combinations of features based on their performance.

Another strength of our research is the data samples that we had access to. A recent review on neuroimaging based classification of AD (Rathore et al., 2017), where our research from chapter 2 was featured, illustrated that the majority of AD classification research uses the Alzheimer's disease neuroimaging initiative (ADNI; Weiner et al., 2015) dataset. ADNI is a valuable, rich, multi-center dataset, but it is essential that methods are validated on different datasets as well. In this review, our dataset from Graz University was the largest multimodal MRI AD/control dataset that did not come from ADNI.

In chapter 4 we studied a dataset of rare HCHWA-D mutation carriers. Because these mutation carriers are almost certain to develop symptoms similar to cerebral amyloid angiopathy (CAA), this allowed us to look into the future of the pre-symptomatic carriers. Furthermore, the symptomatic carriers are younger than typical CAA patients, which allowed us to study CAA like symptoms without the confounding effect of age. The rareness of the mutation imposed the greatest limitation for this research. While our sample is the largest HCHWA-D sample on which diffusion MRI has been extensively studied to date, the power was still limited.

We studied another interesting dataset in chapter 5. Here we aimed to predict cognitive decline after a four-year follow-up. This was a challenging research question because this requires a model that is very sensitive to subtle differences that indicate a future cognitive decline. Using proven MRI techniques, with relatively limited sample size, this was not feasible. A limitation of this sample was the restricted range of cognitive ability at baseline. This choice was made because these people were expected to be at risk for future cognitive decline. However, this may have limited the variance at baseline re-

quired to differentiate between people with a greater or smaller risk of future cognitive decline.

6.5 Future research

In our research, we show that accurate AD classification with MRI is possible in a sample of AD patients and controls. To be useful in a diagnostic setting it is important that the models are not only sensitive but also specific to AD. Future research could explore how the models respond to the MRI scans of people with different types of dementia, or who are in an earlier stage of AD and extend the models to perform multi-class differential diagnosis.

There is limited information to be gained by continuing the path of AD classification with slightly different MRI features or classification models, especially when applied to thoroughly explored datasets such as ADNI. The relatively large, but difficult to estimate, errors for samples smaller than thousands make it nearly impossible to differentiate between the majority of similarly performing measures. Instead, the focus should be on the more difficult task of early prediction. In our research, we have experienced that this is going to require more than was needed for AD classification. To have a fighting chance at accurate early prediction models with clinical applications, we require large diverse multi-center datasets. An extremely promising initiative that is going to enable this is the UK Biobank (Miller et al., 2016), who aim to acquire high-quality multimodal imaging data from 100.000 participants and track their health records.

Another promising development in the research field is deep learning. Deep learning has been extremely successful in many fields of machine learning, including computer vision and image processing (Lecun et al., 2015). In addition to potential performance improvements over traditional machine learning techniques (Vieira et al., 2017), deep learning models can be trained to require little or no preprocessing of MRI data, which makes them very suitable for fast application to new data. Deep learning models are generally complex and have many tunable parameters. It is therefore especially important that datasets large enough to train these models become available.

Another important future research field is clustering. In supervised learning, we are always limited by the accuracy of labels. While great effort is put into the construction of reliable criteria for diagnosis, e.g. by the National Institute

of Aging and Alzheimer's Association (NIA-AA; McKhann et al., 2011), there always remains some uncertainty in the labels. Also, there could be subtypes within existing diagnoses. While the interpretation of clustering is particularly challenging, it has great potential to unravel these unidentified classes.

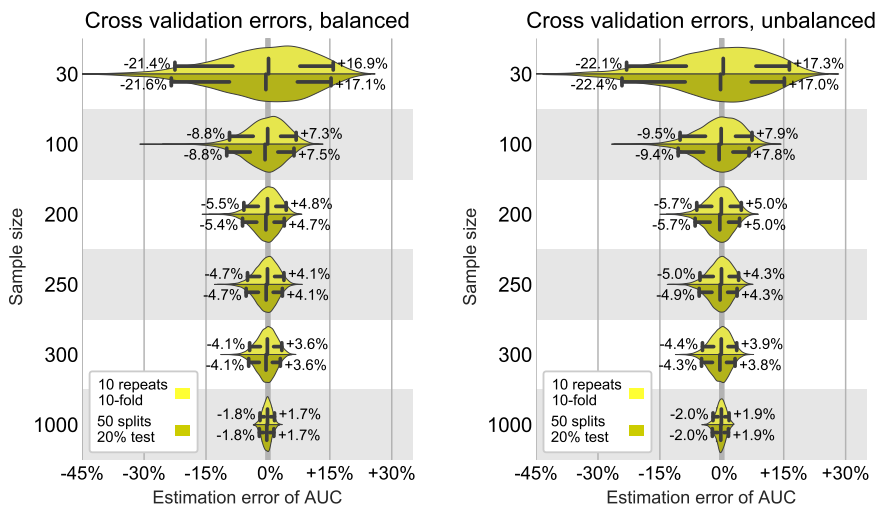
Appendix A

Cross-validation confidence intervals

In section 6.3 I discussed the unknown confidence intervals of cross-validation estimates of generalization performance. Here I have replicated the simulation by Varoquaux (2018). The goal of this study was to estimate the confidence intervals of a typical neuroimaging study that uses machine learning predictions. In order to see how these findings apply to our research I have adjusted the simulation to resemble the situation in this thesis more closely. Specifically, instead of using accuracy, I use AUC. To allow this, I used the distances to the decision boundary for the support vector classifier instead of predicted class. The separation between classes have been tuned to approximate 90% AUC, whereas the original study used 75% accuracy. Furthermore, as we used repeated cross-validation in our studies, I used repeated 10 fold cross-validation with 10 repeats instead of leave-one-out. This allows an additional comparison between repeated cross-validation that we used, and repeated hold-out testing that was used in the original simulation.

The resulting 90% confidence intervals are similar in size, but slightly skewed to the left due to a ceiling effect of the AUC compared to the results in Varoquaux (2018). The differences between repeated 10-fold cross-validation and repeated hold-out testing are small (see Figure A.1a).

To explore how these findings generalize to unbalanced data, I have repeated the simulation with a 77/173 ratio between classes, the same as the ratio in chapters 2 and 3. The resulting confidence intervals are slightly wider. In the case of 250 samples they increase from $[-4.7\%, 4.1\%]$ to $[-5.0\%, 4.3\%]$ (see Figure A.1b).



(a) Balanced 50/50 class labels

(b) Unbalanced 77/173 class labels

Figure A.1: Differences between validation performance on a sample of 10000 and estimated generalization performance by 10 repeats 10 fold cross-validation or 50 repeats 80/20 split hold-out validation. The decision function of a support vector classifier was used without tuning hyperparameters. Whiskers represent 5th and 95th percentile values, and the black vertical stripe indicates the median. The simulation was repeated 1000 times for sample size of 1000, and 10000 times for smaller sample sizes.

Bibliography

- Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., and Draganski, B. (2013). How early can we predict Alzheimer’s disease using computational anatomy? *Neurobiology of Aging*, 34(12):2815–2826.
- Agosta, F., Pievani, M., Geroldi, C., Copetti, M., Frisoni, G. B., and Filippi, M. (2012). Resting state fMRI in Alzheimer’s disease: Beyond the default mode network. *Neurobiology of Aging*, 33(8):1564–1578.
- Alonzo, N. C., Hyman, B. T., Rebeck, G. W., and Greenberg, S. M. (1998). Progression of cerebral amyloid angiopathy: accumulation of amyloid-beta40 in affected vessels. *Journal of neuropathology and experimental neurology*, 57(4):353–359.
- American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders, Fourth Edition, Text Revision*. American Psychiatric Association, Washington, DC.
- Apostolova, L. G., Hwang, K. S., Kohannim, O., Avila, D., Elashoff, D., Jack, C. R., Shaw, L., Trojanowski, J. Q., Weiner, M. W., and Thompson, P. M. (2014). ApoE4 effects on automated diagnostic classifiers for mild cognitive impairment and Alzheimer’s disease. *NeuroImage: Clinical*, 4:461–472.
- Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137–165.
- Arbuthnott, K. and Frank, J. (2000). Trail Making Test, Part B as a Measure of Executive Control: Validation Using a Set-Switching Paradigm. *Journal of Clinical and Experimental Neuropsychology*, 22(4):518–528.
- Basser, P. J., Mattiello, J., and LeBihan, D. (1994a). Estimation of the effective self-diffusion tensor from the NMR spin echo.
- Basser, P. J., Mattiello, J., and LeBihan, D. (1994b). MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267.
- Bassett, D. S., Bullmore, E., Verchinski, B. a., Mattay, V. S., Weinberger, D. R., and Meyer-lindenberg, A. (2008). Hierarchical Organization of Human Cortical Networks in Health and Schizophrenia. *J Neurosci*, 28(37):9239–9248.
- Baykara, E., Gesierich, B., Adam, R., Tuladhar, A. M., Biesbroek, J. M., Koek, H. L., Ropele, S., Jouvent, E., Chabriat, H., Ertl-Wagner, B., Ewers, M., Schmidt, R., de Leeuw, F. E., Biessels, G. J., Dichgans, M., and Duering, M. (2016). A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. *Annals of Neurology*, 80(4):581–592.
- Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage*, 62(2):891–901.

-
- Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152.
- Behrens, T. E. J., Berg, H. J., Jbabdi, S., Rushworth, M. F. S., and Woolrich, M. W. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34(1):144–155.
- Benedictus, M. R., Leeuwis, A. E., Binnewijzend, M. A., Kuijer, J. P., Scheltens, P., Barkhof, F., van der Flier, W. M., and Prins, N. D. (2017). Lower cerebral blood flow is associated with faster cognitive decline in Alzheimer’s disease. *European Radiology*, 27(3):1169–1175.
- Biffi, A. and Greenberg, S. M. (2011). Cerebral amyloid angiopathy: a systematic review. *Journal of clinical neurology (Seoul, Korea)*, 7(1):1–9.
- Binnewijzend, M. A. A., Schoonheim, M. M., Sanz-Arigita, E., Wink, A. M., van der Flier, W. M., Tolboom, N., Adriaanse, S. M., Damoiseaux, J. S., Scheltens, P., van Berckel, B. N. M., and Barkhof, F. (2012). Resting-state fMRI changes in Alzheimer’s disease and mild cognitive impairment. *Neurobiology of aging*, 33(9):2018–2028.
- Bobinski, M., de Leon, M. J., Wegiel, J., Desanti, S., Convit, A., Saint Louis, L. A., Rusinek, H., and Wisniewski, H. M. (2000). The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer’s disease. *Neuroscience*, 95(3):721–5.
- Bornebroek, M., Haan, J., Maat-Schieman, M. L., Van Duinen, S. G., and Roos, R. a. (1996). Hereditary cerebral hemorrhage with amyloidosis-Dutch type (HCHWA-D): I–A review of clinical, radiologic and genetic aspects. *Brain pathology (Zurich, Switzerland)*, 6(2):111–4.
- Bouts, M. J., Möller, C., Hafkemeijer, A., van Swieten, J. C., Doppler, E. G., van der Flier, W. M., Vrenken, H., Wink, A. M., Pijnenburg, Y. A., Scheltens, P., Barkhof, F., Schouten, T. M., de Vos, F., Feis, R. A., van der Grond, J., de Rooij, M., and Rombouts, S. A. (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.
- Boyle, P. A., Wilson, R. S., Yu, L., Barr, A. M., Honer, W. G., Schneider, J. A., and Bennett, D. A. (2013). Much of late life cognitive decline is not due to common neurodegenerative pathologies. *Annals of Neurology*, 74(3):478–489.
- Breiman, L., Friedman, J., Stone, C., and Olshen, R. A. (1984). *Classification and regression trees*. Taylor & Francis.
- Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez-Meza, A. M., Dolph, C. V., Iftekharuddin, K. M., Eskildsen, S. F., Coupé, P., Fonov, V. S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K. R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G. M., Stoyanov, Z. V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J. C., Niessen, W. J., and Klein, S. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, 111:562–579.
- Chao, L. L., Buckley, S. T., Kornak, J., Schuff, N., Madison, C., Yaffe, K., Miller, B. L., Kramer, J. H., and Weiner, M. W. (2010). ASL perfusion MRI predicts cognitive decline and conversion from MCI to dementia. *Alzheimer Disease and Associated Disorders*, 24(1):19–27.

- Charidimou, A., Gang, Q., and Werring, D. J. (2012). Sporadic cerebral amyloid angiopathy revisited: recent insights into pathophysiology and clinical spectrum. *Journal of neurology, neurosurgery, and psychiatry*, 83(2):124–137.
- Charidimou, A., Jaunmuktane, Z., Baron, J.-c., and Jäger, R. (2014). White matter perivascular spaces. *Neurology*, 82:57–62.
- Charidimou, A., Meegahage, R., Fox, Z., Peeters, A., Vandermeeren, Y., Laloux, P., Baron, J.-c., Jäger, H. R., and Werring, D. J. (2013). Enlarged perivascular spaces as a marker of underlying arteriopathy in intracerebral haemorrhage: a multicentre MRI cohort study. *Journal of neurology, neurosurgery, and psychiatry*, 84(6):624–9.
- Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., and Jin, J. S. (2011). Identification of conversion from mild cognitive impairment to alzheimer’s disease using multivariate predictors. *PLoS ONE*, 6(7):2–11.
- Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.-O., Chupin, M., Benali, H., and 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.
- Cummings, J. L., Morstorf, T., and Zhong, K. (2014). Alzheimer’s disease drug-development pipeline: Few candidates, frequent failures. *Alzheimer’s Research and Therapy*, 6(4):1–7.
- Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., and 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.
- de Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., Schmidt, R., de Rooij, M., and Rombouts, S. A. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer’s disease. *NeuroImage*, 167(December 2016):62–72.
- de Vos, F., Schouten, T. M., Hafkemeijer, A., Dopper, E. G. P., van Swieten, J. C., de Rooij, M., van der Grond, J., and Rombouts, S. A. R. B. (2016). Combining multiple anatomical MRI measures improves Alzheimer’s disease classification. *Human Brain Mapping*.
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and 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.
- Dickerson, B. C. and Wolk, D. a. (2012). MRI cortical thickness biomarker predicts AD-like CSF and cognitive decline in normal adults. *Neurology*, 78(2):84–90.
- Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms.
- Dipasquale, O., Griffanti, L., Clerici, M., Nemni, R., Baselli, G., and 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):1–7.
- Douaud, G., Jbabdi, S., Behrens, T. E. J., Menke, R. a., Gass, A., Monsch, A. U., Rao, A., Whitcher, B., Kindlmann, G., Matthews, P. M., and 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.
- Duan, J.-H., Wang, H.-Q., Xu, J., Lin, X., Chen, S.-Q., Kang, Z., and Yao, Z.-B. (2006). White matter damage of patients with Alzheimer’s disease correlated with the decreased cognitive function. *Surgical and radiologic anatomy : SRA*, 28(2):150–156.

-
- Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., Meindl, T., Pievani, M., Bokde, A. L. W., Fellgiebel, A., Filippi, M., Hampel, H., Kl?ppel, S., Hauenstein, K., Kirste, T., and 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).
- Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. (2015). Automated detection of functional and structural disconnection in AD using multiple kernels SVM. *Hum. Brain Mapp.*, 36(6):2118–2131.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.
- Ferreira, L. K., Diniz, B. S., Forlenza, O. V., Busatto, G. F., and Zanetti, M. V. (2011). Neurostructural predictors of Alzheimer’s disease: a meta-analysis of VBM studies. *Neurobiology of aging*, 32(10):1733–1741.
- Fox, N. C., Scahill, R. I., Crum, W. R., and Rossor, M. N. (1999). Correlation between rates of brain atrophy and cognitive decline in AD. *Neurology*, 52(8):1687–1687.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–441.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22.
- Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature reviews. Neurology*, 6(2):67–77.
- Gour, N., Felician, O., Didic, M., Koric, L., Gueriot, C., Chanoine, V., Confort-Gouny, S., Guye, M., Ceccaldi, M., and Ranjeva, J. P. (2014). Functional connectivity changes differ in early and late-onset alzheimer’s disease. *Human Brain Mapping*, 35(7):2978–2994.
- Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and 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.
- Greenberg, S. M. (1998). Cerebral amyloid angiopathy: Prospects for clinical diagnosis and treatment. *Neurology*, 51(3):690–694.
- Greenberg, S. M. (2002). Cerebral amyloid angiopathy and vessel dysfunction. *Cerebrovascular Diseases*, 13(SUPPL. 2):42–47.
- Greicius, M. D., Srivastava, G., Reiss, A. L., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer ’ s disease from healthy aging : Evidence from functional MRI. *PNAS*, 101(13):4637–4642.
- Grieve, S. M., Williams, L. M., Paul, R. H., Clark, C. R., and Gordon, E. (2007). Cognitive aging, executive function, and fractional anisotropy: a diffusion tensor MR imaging study. *AJNR Am J Neuroradiol*, 28(2):226–235.
- Groves, A. R., Beckmann, C. F., Smith, S. M., and Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *NeuroImage*, 54(3):2198–2217.
- Hardy, J. and Allsop, D. (1991). Amyloid deposition as the central event in the aetiology of Alzheimer’s disease. *Trends in Pharmacological Sciences*, 12:383–388.

- Hartz, A. M., Bauer, B., Soldner, E. L., Wolf, A., Boy, S., Backhaus, R., Mihaljevic, I., Bogdahn, U., Klünemann, H. H., Schuierer, G., and Schlachetzki, F. (2012). Amyloid- β contributes to blood-brain barrier leakage in transgenic human amyloid precursor protein mice and in humans with cerebral amyloid angiopathy. *Stroke*, 43(2):514–523.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.
- Houx, P. J., Jolles, J., and Vreeling, F. W. (1993). Stroop interference: Aging effects assessed with the stroop color-word test.
- Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D. S., Calabresi, P. a., Pekar, J. J., van Zijl, P. C. M., and Mori, S. (2008). Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *NeuroImage*, 39(1):336–347.
- Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., Shaw, L. M., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Pankratz, V. S., Donohue, M. C., and Trojanowski, J. Q. (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., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade. *The Lancet Neurology*, 9:119–128.
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2):825–841.
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *NeuroImage*, 62(2):782–790.
- Karran, E., Mercken, M., and Strooper, B. D. (2011). The amyloid cascade hypothesis for Alzheimer’s disease: An appraisal for the development of therapeutics. *Nature Reviews Drug Discovery*, 10(9):698–712.
- Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., and Frackowiak, R. S. J. (2008). Automatic classification of MR scans in Alzheimer’s disease. *Brain : a journal of neurology*, 131:681–689.
- Koch, W., Teipel, S., Mueller, S., Benninghoff, J., Wagner, M., Bokde, A. L. W., Hampel, H., Coates, U., Reiser, M., and 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.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540.
- Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Lee, M. H., Smyser, C., and Shimony, J. (2013). Resting-state fMRI: a review of methods and clinical applications. *American Journal of Neuroradiology*, 35(11):1866–1872.
- Levy, E., Carman, M. D., Fernandezmadrid, I. J., Power, M. D., Lieberburg, I., Van-duinen, S. G., Bots, G., Luyendijk, W., and Frangione, B. (1990). Mutation of the Alzheimers-Disease Amyloid Gene in Hereditary Cerebral-Hemorrhage, Dutch Type. *Science*, 248(4959):1124–1126.

-
- Li, J., Pan, P., Huang, R., and Shang, H. (2012). A meta-analysis of voxel-based morphometry studies of white matter volume alterations in Alzheimer's disease. *Neuroscience and Biobehavioral Reviews*, 36(2):757–763.
- Li, X.-y., Tang, Z.-c., Sun, Y., Tian, J., Liu, Z.-y., and Han, Y. (2016). White matter degeneration in subjective cognitive decline: a diffusion tensor imaging study. *Oncotarget*, 7(34).
- Lindeboom, J., Schmand, B., Tulner, L., Walstra, G., and Jonker, C. (2002). Visual association test to detect early dementia of the Alzheimer type. *Journal of Neurology Neurosurgery and Psychiatry*, 73(2):126–133.
- Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., and Ourselin, S. (2019). Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, 190(May 2017):56–68.
- Maat-Schieman, M. L., van Duinen, S. G., Bornebroek, M., Haan, J., and Roos, R. a. (1996). Hereditary cerebral hemorrhage with amyloidosis-Dutch type (HCHWA-D): II—A review of histopathological aspects. *Brain pathology (Zurich, Switzerland)*, 6(2):115–20.
- Martinez-Ramirez, S., Pontes-Neto, O. M., Dumas, A. P., Auriel, E., Halpin, A., Quimby, M., Gurol, M. E., Greenberg, S. M., and Viswanathan, A. (2013). Topography of dilated perivascular spaces in subjects from a memory clinic cohort. *Neurology*, 80(17):1551–1556.
- McKhann, G. M., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. 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(7):939–944.
- McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., and 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 & Dementia*, 7(3):263–269.
- Medina, D., DeToledo-Morrell, L., Urresta, F., Gabrieli, J. D., Moseley, M., Fleischman, D., Bennett, D. A., Leurgans, S., Turner, D. A., and Stebbins, G. T. (2006). White matter changes in mild cognitive impairment and AD: A diffusion tensor imaging study. *Neurobiology of Aging*, 27(5):663–672.
- Mesrob, L., Sarazin, M., Hahn-barma, V., Souza, L. C. D., Dubois, B., Gallinari, P., and Kinkingnéhun, S. (2012). DTI and Structural MRI Classification in Alzheimer's Disease. *Advances in Molecular Imaging*, 2:12–20.
- Miller, K. L., Alfaro-Almagro, F., Bangarter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536.
- Moonen, J. E. F., Foster-Dingley, J. C., de Ruijter, W., van der Grond, J., Bertens, A. S., van Buchem, M. A., Gussekloo, J., Middelkoop, H. A., Wermer, M. J. H., Westendorp, R. G. J., de Craen, A. J. M., and van der Mast, R. C. (2015). Effect of Discontinuation of Antihypertensive Treatment in Elderly People on Cognitive Functioning - the DANTE Study Leiden. *JAMA Internal Medicine*, 175(10):1622.
- Mori, S., Wakana, S., Van Zijl, P. C. M., and Nagae-Poetscher, L. M. (2005). *MRI atlas of human white matter*. Elsevier.

- Mosconi, L., Berti, V., Glodzik, L., Pupi, A., De Santi, S., and de Leon, M. J. (2010). Pre-Clinical Detection of Alzheimer's Disease Using FDG-PET, with or without Amyloid Imaging. *Journal of Alzheimer's Disease*, 20(3):843–854.
- Mudher, A. and Lovestone, S. (2002). Alzheimer's disease - Do tauists and baptists finally shake hands? *Trends in Neurosciences*, 25(1):22–26.
- Nir, T. M., Villalon-Reina, J. E., Prasad, G., Jahanshad, N., Joshi, S. H., Toga, A. W., Bernstein, M. a., Jack, C. R., Weiner, M. W., and Thompson, P. M. (2014). Diffusion weighted imaging-based maximum density path analysis and classification of Alzheimer's disease. *Neurobiology of Aging*, 36:1–9.
- O'Dwyer, L., Lamberton, F., Bokde, A. L. W., Ewers, M., Faluyi, Y. O., Tanner, C., Mazoyer, B., O'Neill, D., Bartley, M., Collins, D. R., Coughlan, T., Prvulovic, D., and Hampel, H. (2012). Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. *PLoS ONE*, 7(2).
- Ouyang, X., Chen, K., Yao, L., Wu, X., Zhang, J., Li, K., Jin, Z., and Guo, X. (2015). Independent Component Analysis-Based Identification of Covariance Patterns of Microstructural White Matter Damage in Alzheimer's Disease. *PLoS ONE*, 10(3):e0119714.
- Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56(3):907–922.
- Pernecky, R., Wagenpfeil, S., Komossa, K., Grimmer, T., Diehl, J., and Kurz, A. (2006). Mapping scores onto stages: mini-mental state examination and clinical dementia rating. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry*, 14(2):139–44.
- Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A. W., Hampel, H., and Ewers, M. (2010a). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage*, 50(1):162–174.
- Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A. W., Hampel, H., and Ewers, M. (2010b). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage*, 50(1):162–174.
- Prince, M., Bryce, R., and Ferri, C. (2011). World Alzheimer Report 2011. Technical report, Alzheimer's Disease International, London.
- Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., and Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, 112:267–277.
- Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 155(March):530–548.
- Reijmer, Y. D., Fotiadis, P., Charidimou, A., Veluw, S. J. V., Xiong, L., Riley, G. A., Martinez-ramirez, S., Schwab, K., Viswanathan, A., Gurol, M. E., and Greenberg, S. M. (2017). Relationship Between White Matter Connectivity Loss and Cortical Thinning in Cerebral Amyloid Angiopathy. *Human Brain Mapping*, 38(March):3723–3731.
- Reijmer, Y. D., Fotiadis, P., Martinez-Ramirez, S., Salat, D. H., Schultz, A., Shoamanesh, A., Ayres, A. M., Vashkevich, A., Rosas, D., Schwab, K., Leemans, A., Biessels, G. J., Rosand, J., Johnson, K. A., Viswanathan, A., Gurol, M. E., and Greenberg, S. M. (2015). Structural network alterations and neurological dysfunction in cerebral amyloid angiopathy. *Brain*, 138(1):179–188.

-
- Reijmer, Y. D., Fotiadis, P., Riley, G. A., Xiong, L., Charidimou, A., Boulouis, G., Ayres, A. M., Schwab, K., Rosand, J., Gurol, M. E., Viswanathan, A., and Greenberg, S. M. (2016a). Progression of Brain Network Alterations in Cerebral Amyloid Angiopathy. *Stroke*, 47(10):2470–2475.
- Reijmer, Y. D., Van Veluw, S. J., and Greenberg, S. M. (2016b). Ischemic brain injury in cerebral amyloid angiopathy. *Journal of Cerebral Blood Flow and Metabolism*, 36(1):40–54.
- Rose, S. E., McMahon, K. L., Janke, A. L., O’Dowd, B., De Zubizaray, G., Strudwick, M. W., and Chalk, J. B. (2006). Diffusion indices on magnetic resonance imaging and neuropsychological performance in amnesic mild cognitive impairment. *Journal of Neurology, Neurosurgery and Psychiatry*, 77(10):1122–1128.
- Rowe, C. C., Ng, S., Ackermann, U., Gong, S. J., Pike, K., Savage, G., Cowie, T. F., Dickinson, K. L., Maruff, P., Darby, D., Smith, C., Woodward, M., Merory, J., Tochon-Danguy, H., O’Keefe, G., Klunk, W. E., Mathis, C. A., Price, J. C., Masters, C. L., and Villemagne, V. L. (2007). Imaging β -amyloid burden in aging and dementia. *Neurology*, 68(20):1718–1725.
- Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–69.
- Salat, D. H., Smith, E. E., Tuch, D. S., Benner, T., Pappu, V., Schwab, K. M., Gurol, M. E., Rosas, H. D., Rosand, J., and Greenberg, S. M. (2006). White Matter Alterations in Cerebral Amyloid Angiopathy Measured by Diffusion Tensor Imaging. *Stroke*, 37(7):1759 LP – 1764.
- Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90:449–468.
- Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M. O., Durrleman, S., Evgeniou, T., and Colliot, O. (2018). Reproducible evaluation of classification methods in Alzheimer’s disease: Framework and application to MRI and PET data. *NeuroImage*, 183(August):504–521.
- Scheltens, P., Blennow, K., Breteler, M. M. B., Strooper, B. D., Frisoni, G. B., Salloway, S., and Flier, W. M. V. D. (2016). Alzheimer’s disease. *The Lancet*, 388(10043):505–517.
- 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., and Rombouts, S. A. (2017). Individual Classification of Alzheimer’s Disease with Diffusion Magnetic Resonance Imaging. *NeuroImage*, 152(March):476–481.
- 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., and 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.
- Seiler, S., Schmidt, H., Lechner, A., and Benke, T. (2012). Driving cessation and dementia: results of the prospective registry on dementia in Austria (PRODEM). *PloS one*, 7(12):1–6.
- Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V. M., and Trojanowski, J. Q. (2009). Cerebrospinal fluid biomarker signature in alzheimer’s disease neuroimaging initiative subjects. *Annals of Neurology*, 65(4):403–413.

- Sheline, Y. I. and Raichle, M. E. (2013). Resting state functional connectivity in preclinical Alzheimer's disease. *Biological Psychiatry*, 74(5):340–347.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.
- Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2012). Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *NeuroImage*, 62(3):1924–1938.
- Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2015). The effects of SIFT on the reproducibility and biological accuracy of the structural connectome. *NeuroImage*, 104:253–265.
- Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., and Behrens, T. E. J. (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., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:S208–S219.
- Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891.
- Smith, S. M. and Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98.
- Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Uğurbil, K., and Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in cognitive sciences*, 17(12):666–682.
- Smyth, G. K. and Phipson, B. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):Article 39.
- Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.
- Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. a., Yu, Q., Clark, V. P., Castro, E., White, T., Mueller, B. a., Ho, B. C., Andreasen, N. C., and Calhoun, V. D. (2013a). Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia. *NeuroImage*, 66:119–132.
- Sui, J., He, H., Yu, Q., Chen, J., Rogers, J., Pearlson, G. D., Mayer, A., Bustillo, J., Canive, J., and Calhoun, V. D. (2013b). Combination of Resting State fMRI, DTI, and sMRI Data to Discriminate Schizophrenia by N-way MCCA + jICA. *Frontiers in human neuroscience*, 7(May):1–14.
- Tax, C. M., Jeurissen, B., Vos, S. B., Viergever, M. A., and Leemans, A. (2014). Recursive calibration of the fiber response function for spherical deconvolution of diffusion MRI data. *NeuroImage*, 86:67–80.

-
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- Tournier, J.-D., , F. Calamante, and a. Connelly (2010). Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. *Ismrm*, 88(2003):2010.
- Tournier, J. D., Calamante, F., and Connelly, A. (2012). MRtrix: Diffusion tractography in crossing fiber regions. *International Journal of Imaging Systems and Technology*, 22(1):53–66.
- Van Rooden, S., Van Opstal, A. M., Labadie, G., Terwindt, G. M., Wermer, M. J. H., Webb, A. G., Middelkoop, H. A. M., Greenberg, S. M., Van Der Grond, J., and Van Buchem, M. A. (2016). Early Magnetic Resonance Imaging and Cognitive Markers of Hereditary Cerebral Amyloid Angiopathy. *Stroke*, 47(12):3041–3044.
- Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7:91.
- Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180(June 2017):68–77.
- Varoquaux, G., Raamana, P., Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2016). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *arXiv:1606.05201 [stat.ML]*, 145(October):1–14.
- Veraart, J., Novikov, D. S., Christiaens, D., Ades-aron, B., Sijbers, J., and Fieremans, E. (2016). Denoising of diffusion MRI using random matrix theory. *NeuroImage*, 142:394–406.
- Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience and Biobehavioral Reviews*, 74:58–75.
- Vinters, H. V. (1987). Cerebral amyloid angiopathy. A critical review. *Stroke; a journal of cerebral circulation*, 18(2):311–24.
- Wardlaw, J. M., Makin, S. J., Valdés Hernández, M. C., Armitage, P. A., Heye, A. K., Chappell, F. M., Muñoz-Maniega, S., Sakka, E., Shuler, K., Dennis, M. S., and Thrippleton, M. J. (2017). Blood-brain barrier failure as a core mechanism in cerebral small vessel disease and dementia: evidence from a cohort study. *Alzheimer’s and Dementia*, 13(6):634–643.
- Wee, C. Y., Yap, P. T., Li, W., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., and Shen, D. (2011). Enriched white matter connectivity networks for accurate identification of MCI patients. *NeuroImage*, 54:1812–1822.
- Wee, C. Y., Yap, P. T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., and Shen, D. (2012). Identification of MCI individuals using structural and functional connectivity networks. *NeuroImage*, 59:2045–2056.
- Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Luthman, J., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L., Shen, L., Schwarz, A., Toga, A. W., and Trojanowski, J. Q. (2015). 2014 Update of the Alzheimer’s Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer’s & Dementia*, 11(6):e1–e120.
- Winkler, A. M., Ridgway, G. R., Douaud, G., Nichols, T. E., and Smith, S. M. (2016). Faster permutation inference in brain imaging. *NeuroImage*, 141:502–516.

- Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92:381–397.
- Wolk, D. A. and Detre, J. A. (2012). Arterial spin labeling MRI. *Current Opinion in Neurology*, 25(4):421–428.
- Woodard, J. L., Seidenberg, M., Nielson, K. a., Smith, J. C., Antuono, P., Durgerian, S., Guidotti, L., Zhang, Q., Butts, A., Hantke, N., Lancaster, M., and Rao, S. M. (2010). Prediction of cognitive decline in healthy older adults using fMRI. *Journal of Alzheimer's disease : JAD*, 21(3):871–885.
- Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141.
- Yeh, C. H., Smith, R. E., Liang, X., Calamante, F., and Connelly, A. (2016). Correction for diffusion MRI fibre tracking biases: The consequences for structural connectomic metrics. *NeuroImage*, 142:150–162.
- Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., and Ourselin, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2(1):735–745.
- Zhan, L., Zhou, J., Wang, Y., Jin, Y., Jahanshad, N., Prasad, G., Nir, T. M., Leonardo, C. D., Ye, J., and Thompson, P. M. (2015). Comparison of 9 tractography algorithms for detecting abnormal structural brain networks in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7(MAR):1–19.
- Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57.
- Zhang, Y., Schuff, N., Du, A. T., Rosen, H. J., Kramer, J. H., Gorno-Tempini, M. L., Miller, B. L., and Weiner, M. W. (2009). White matter damage in frontotemporal dementia and Alzheimers disease measured by diffusion MRI. *Brain*, 132(9):2579–2592.
- Zhang-nunes, S. X., Maat-schieman, M. L. C., Duinen, S. G. V., Roos, R. A. C., and Frosch, M. P. (2006). The Cerebral β -Amyloid Angiopathies : Hereditary and Sporadic. *Brain Pathol*, 16(1):30–39.
- Zhu, Y.-C., Chabriat, H., Godin, O., Dufouil, C., Rosand, J., Greenberg, S. M., Smith, E. E., Tzourio, C., and Viswanathan, A. (2012). Distribution of white matter hyperintensity in cerebral hemorrhage and healthy aging. *Journal of neurology*, 259(3):530–6.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Samenvatting

Dementie is een verwoestende ziekte waar wereldwijd miljoenen mensen aan leiden. De meest voorkomende oorzaak van dementie is de ziekte van Alzheimer. Vele pogingen om een behandeling tegen alzheimer te ontwikkelen zijn tot op heden niet succesvol gebleken. De medicatie die goedgekeurd is voor de behandeling van alzheimer bestrijdt slechts symptomen. Het is mogelijk dat het onderzoek naar effectieve behandelmethoden wordt bemoeilijkt doordat deelnemers aan onderzoek reeds in een stadium van de ziekte verkeren waarin de hersenschade onomkeerbaar is. Om een kans te hebben om effectieve behandelingen te ontwikkelen is het belangrijk om dementie in een vroeg stadium te detecteren.

Hoe de ziekte van Alzheimer precies ontstaat is niet bekend, maar er zijn sterke vermoedens dat de eiwitten amyloid β (A_β) en tau een direct verband hebben met de neurodegeneratie in alzheimerpatiënten. De hoeveelheid van deze eiwitten verklaart echter slechts een beperkt deel van de variantie in cognitieve achteruitgang. Bovendien zijn therapieën gericht op het afbreken van A_β niet effectief gebleken als behandelmethode tegen de ziekte van Alzheimer.

Om alzheimer betrouwbaar te diagnosticeren zijn er verschillende biomarkers in gebruik. Deze biomarkers zijn ruwweg in te delen in twee groepen: 1) biomarkers die gericht zijn op het meten van A_β en 2) biomarkers die gericht zijn op het meten van neuronale schade, welke sterk samenhangt met tau. Neuronale schade kan worden afgeleid van het metabolisme van het brein, welke kan worden gemeten met fluorodeoxyglucose positron emission tomography (FDG-PET). Recent kan ook de hoeveelheid A_β gemeten en gelokaliseerd worden door middel van positron emission tomography (PET) met de tracer Pittsburgh Compound-B.

Nadelen van PET zijn de hoge kosten en de noodzaak om een radioactieve tracer toe te dienen. Een goedkoper en minder invasief alternatief is mag-

netic resonance imaging (MRI). Structurele MRI (sMRI) kan worden toegepast om hersenvolume te meten, wat een maat geeft van neuronale schade. Met functionele MRI in rusttoestand (rs-fMRI) kunnen de functionele verbindingen tussen hersengebieden bestudeerd worden. Met diffusie MRI kunnen afwijkingen in de witte stof en de structurele verbindingen tussen hersengebieden in kaart worden gebracht.

Traditioneel alzheimeronderzoek is voornamelijk gericht op het vinden van groepsverschillen tussen patiënten en controles. Deze vorm van onderzoek is waardevol voor het verkrijgen van inzicht in hoe de ziekte zich in de hersenen ontwikkelt. Hierdoor is bekend dat bijvoorbeeld de hippocampus in het bijzonder aangedaan is bij patiënten met de ziekte van Alzheimer. Recent onderzoek is deels verschoven naar individuele classificatie. Hierbij wordt er gebruikgemaakt van machine learning om patronen in de MRI data te zoeken die specifiek zijn voor een ziektebeeld. Het verschil met traditioneel onderzoek is dat er hiermee uitspraken gedaan kunnen worden over een individu, bijvoorbeeld voor automatische diagnose, of voor voorspellingen over toekomstige ontwikkelingen.

In deel I van dit proefschrift hebben we verschillende MRI modaliteiten gebruikt voor individuele classificatie van alzheimerpatiënten. In hoofdstuk 2 gebruikten we sMRI, diffusie MRI, en rs-fMRI om alzheimerpatiënten te classificeren. Door gebruik te maken van grijze stof dichtheid konden we de patiënten betrouwbaar onderscheiden van controles. Wanneer we deze maat combineerden met maten van andere MRI scans namen de classificatieprestaties toe. De combinatie die het beste presteerde maakte gebruik van maten die afkomstig waren van alle drie de gebruikte MRI scans. In hoofdstuk 3 gebruikten we dezelfde dataset om diffusie MRI in meer detail te bestuderen. In plaats van gebieden op te delen in anatomische regio's hebben we de scans op voxelniveau geanalyseerd. De behaalde classificatieaccuratesse was hiermee vergelijkbaar met sMRI. Door gebruik te maken van independent component analysis kon de data gereduceerd worden tot 28 componenten, zonder op classificatieprestaties in te boeten. Het combineren van verschillende diffusie MRI maten zorgde er niet voor dat de prestaties toenamen.

In deel II hebben we ons gericht op vroegere vormen van cognitieve achteruitgang. In hoofdstuk 4 onderzochten we een steekproef van pre-symptomatische en symptomatische dragers van een gen dat de Katwijkse ziekte (hereditary cerebral hemorrhage with amyloidosis, Dutch type [HCHWA-D]) veroorzaakt. Dragere van dit gen zullen vrijwel zeker een vorm van amyloïdangiopathie on-

twikkelen. We vergeleken deze gendragers met controles. We vonden hier sterke globale verschillen in diffusie MRI tussen symptomatisch patiënten en controles. Tussen pre-symptomatische patiënten en controles vonden we geen significante verschillen. Bovendien vonden we een verschil tussen patiënten en controles in het verloop van diffusiematen wanneer de proefpersonen ouder worden. Terwijl jonge gendragers nauwelijks verschilden van jonge controles, bestonden deze verschillen in toenemende mate bij oudere proefpersonen.

In hoofdstuk 5 onderzochten we een steekproef van ouderen met een verhoogd risico op cognitieve achteruitgang. We gebruikten de MRI data van deze proefpersonen om een voorspelling te doen van hun cognitieve achteruitgang vier jaar na afname van de scan. Het voorspellen van toekomstige cognitieve achteruitgang is belangrijk voor de selectie van proefpersonen om mogelijke nieuwe behandelingen te testen. Om de voorspelling te maken gebruikten we de maten die effectief waren in het classificeren van alzheimer in deel I. Naast structurele, diffusie, en resting state functionele MRI gebruikten we tevens markers voor afwijkingen in de kleine bloedvaten. Ondanks de goede prestaties van deze methoden bij het classificeren van alzheimer waren deze niet in staat om toekomstige cognitieve achteruitgang te voorspellen. Mogelijk is de gevoeligheid van MRI onvoldoende om de subtiele verschillen te detecteren die duiden op toekomstige achteruitgang.

De studies tezamen laten zien dat MRI een krachtige en veelzijdige methode is waarmee betrouwbare individuele uitspraken gedaan kan worden over patiënten in relatief vergevorderde stadia van dementie. We laten zien dat met name maten van structurele en diffusie MRI goed in staat zijn om patiënten van controles te onderscheiden, en dat het combineren van verschillende MRI modaliteiten de prestaties kunnen verbeteren. Verder laten we zien dat diffusie MRI een geschikte methode is voor zowel individuele classificatie van alzheimerpatiënten, alsmede voor het detecteren van groepsverschillen tussen symptomatische HCHWA-D patiënten en controles. Dezelfde methodologie is echter niet in staat om ook verschillen te bespeuren bij pre-symptomatische patiënten. Daarnaast is het voorspellen van toekomstige cognitieve achteruitgang een meer uitdagend probleem, waar de huidige methoden niet gevoelig genoeg voor zijn.

Dankwoord

Graag bedank ik mijn promotoren, Serge Rombouts, Jeroen van der Grond, en Mark de Rooij voor de diverse en intensieve begeleiding. Serge, bedankt voor het houden van het overzicht en het zetten van de koers. Jeroen, bedankt voor de klinische en praktische insteek, en de overvloed aan ideeën. Mark, bedankt voor het waarborgen van wetenschappelijke integriteit, en de inbreng van, soms decennia lang vergeten, literatuur. Naast de goed begeleiding heb ik ook veel plezier met jullie gehad.

Ik ben zeer dankbaar voor de proefpersonen van de verschillende studies die veel tijd hebben geofferd en ongemak hebben ondergaan om een onbaatzuchtige bijdrage aan de wetenschap te leveren.

Mijn dank gaat ook zeer uit naar de mensen die de data hebben verzameld. Een belangrijke en tijdsintensieve klus die mij bespaard is gebleven. Ik weet dat er vele mensen bij betrokken zijn geweest, maar dank hiervoor in het bijzonder Marisa Koini, Anne Hafkemeijer, Sanneke van Rooden, en Jessica Foster-Dingley.

Mijn collega's van Methodologie en Statistiek. In het bijzonder Frank de Vos, waarmee ik tijdens onze promotietrajecten zo intensief mee samengewerkt heb dat velen ons louter als duo kennen. Bernadet, bedankt voor de gezelligheid tijdens en na werk, en voor alle adviezen en praktische tips omtrent het promoveren. De andere leden van de VICI-groep, vooral Mark Bouts en Rogier Feis, voor de waardevolle discussies.

Mijn vrienden, voor het aanhoren van onbegrijpelijke vaktaal en toch de moeite doen om er iets van te maken, in het bijzonder Pepijn en Daniël.

Graag bedank ik ook mijn familie, mijn ouders, Sonja en Jos, en mijn zus Myrthe, die altijd onvoorwaardelijke trots heeft gehad voor haar broertje.

Marjan, bedankt voor de steun en liefde. Sebastiaan en Hannah, bedankt voor jullie blijde koppies en enthousiaste onthaal.

Curriculum vitae

Tijn Schouten was born on January 27, 1988 in Maastricht, the Netherlands. He graduated from high school at Sint-Maartenscollege in Maastricht in 2007.

He received his bachelor's degree in Psychology from Groningen University in 2011. He then received his research master's degree in Behavioural and Cognitive Neurosciences from Groningen University in 2013.

Then he started his PhD research that led to this thesis under supervision of prof. dr. Serge Rombouts, prof. dr. Mark de Rooij, and dr. Jeroen van der Grond. This research was performed in collaboration between the Leiden University Medical Center department of Radiology, and the department of Methodology and Statistics, institute of Psychology, faculty of Social and Behavioural Sciences, Leiden University.

Tijn is currently employed as a postdoctoral researcher at the Methodology and Statistics department of Leiden University, where his research is focussed on application of deep learning models on MRI data for classification of Alzheimer's disease.

List of publications

2019

Schouten, T. M., de Vos, F., Van Rooden, S., Bouts, M. 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., and 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(e011288):1–8.

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

Feis, R. A., Bouts, M. J., De Vos, F., **Schouten, T. M.**, Panman, J. L., Jiskoot, L. C., Dopper, E. G., Van Der Grond, J., Van Swieten, J. C., and Rombouts, S. A. (2019). A multimodal MRI-based classification signature emerges just prior to symptom onset in frontotemporal dementia mutation carriers. *Journal of Neurology, Neurosurgery and Psychiatry*, pages 1–8.

2018

de Vos, F., Koini, M., **Schouten, T. M.**, Seiler, S., van der Grond, J., Lechner, A., Schmidt, R., de Rooij, M., and Rombouts, S. A. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer’s disease. *NeuroImage*, 167(December 2016):62–72.

Bouts, M. J., Möller, C., Hafkemeijer, A., van Swieten, J. C., Dopper, E. G., van der Flier, W. M., Vrenken, H., Wink, A. M., Pijnenburg, Y. A., Scheltens,

P., Barkhof, F., **Schouten, T. M.**, de Vos, F., Feis, R. A., van der Grond, J., de Rooij, M., and Rombouts, S. A. (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.

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

2017

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., and Rombouts, S. A. (2017). Individual Classification of Alzheimer's Disease with Diffusion Magnetic Resonance Imaging. *NeuroImage*, 152(March):476–481.

2016

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., and 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.

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

2015

Hafkemeijer, A., Möller, C., Dopper, E. G. P., Jiskoot, L. C., **Schouten, T. M.**, van Swieten, J. C., van der Flier, W. M., Vrenken, H., Pijnenburg, Y. A. L., Barkhof, F., Scheltens, P., van der Grond, J., and Rombouts, S. A. R. B. (2015). Resting state functional connectivity differences between behavioral variant frontotemporal dementia and Alzheimer's disease. *Frontiers in Human Neuroscience*, 9(September):474.

2014

- Schouten, T. M.**, Bolderdijk, J. W., and Steg, L. (2014). Framing car fuel efficiency: linearity heuristic for fuel consumption and fuel-efficiency ratings. *Energy Efficiency*, 7(5):891–901.

