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

Schouten, T.M.

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±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

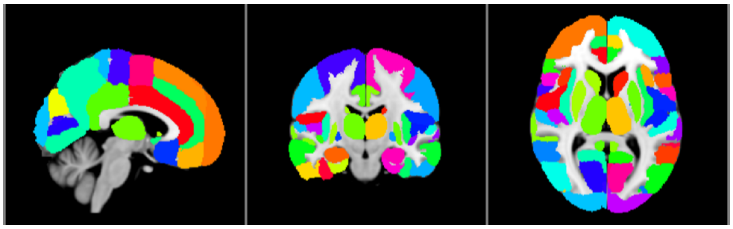

Figure 3.1: Harvard-Oxford cortical and subcortical regions that we used as target and seed nodes for probabilistic tractography. Areas represent the probabilistic regions above the 25% threshold, and then assigned to the highest probability region.

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 ($\frac{110*109}{2} = 5995$ 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.

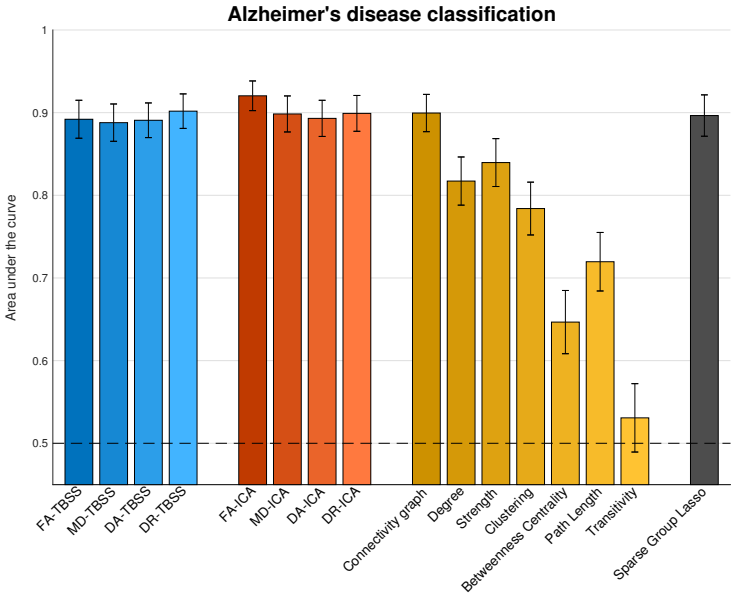

Figure 3.2: Overview of classification results. Bars indicate mean area under the receiver operating characteristics curves over 10 repetitions. The error bars represent standard errors based on 5000 bootstraps.

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

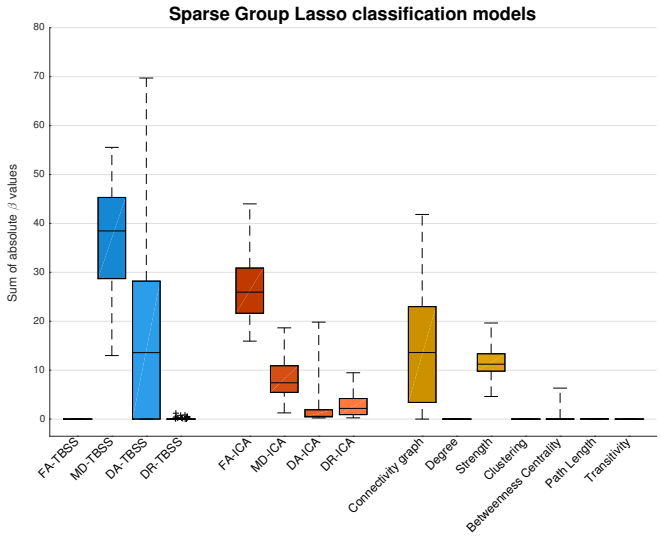

Figure 3.3: Boxplot of the sparse group lasso classification models from 10-fold times 10 repeated cross validation. The bars indicate the spread of the sum of the absolute beta values. The boxplot for DR-TBSS could not be visualised, because the lower 84% of the values was zero. The non-zero values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