◆ Human Brain Mapping 37:1920–1929 (2016) ◆

Combining Multiple Anatomical MRI Measures Improves Alzheimer's Disease Classification

Frank de Vos,[1,2,3]* Tijn M. Schouten,[1,2,3] Anne Hafkemeijer,[1,2,3] Elise G. P. Dopper,[2,4,5] John C. van Swieten,[4,6] Mark de Rooij,[1,3] Jeroen van der Grond,[2] and Serge A. R. B. Rombouts[1,2,3]

[1]*Leiden University, Institute of Psychology, The Netherlands*
[2]*Department of Radiology, Leiden University Medical Center, The Netherlands*
[3]*Leiden Institute for Brain and Cognition, The Netherlands*
[4]*Department of Neurology, Erasmus Medical Center, The Netherlands*
[5]*Department of Neurology, VU Medical Center, The Netherlands*
[6]*Department of Clinical Genetics, VU Medical Center, The Netherlands*

Abstract: Several anatomical MRI markers for Alzheimer's disease (AD) have been identified. Hippocampal volume, cortical thickness, and grey matter density have been used successfully to discriminate AD patients from controls. These anatomical MRI measures have so far mainly been used separately. The full potential of anatomical MRI scans for AD diagnosis might thus not yet have been used optimally. In this study, we therefore combined multiple anatomical MRI measures to improve diagnostic classification of AD. For 21 clinically diagnosed AD patients and 21 cognitively normal controls, we calculated (i) cortical thickness, (ii) cortical area, (iii) cortical curvature, (iv) grey matter density, (v) subcortical volumes, and (vi) hippocampal shape. These six measures were used separately and combined as predictors in an elastic net logistic regression. We made receiver operating curve plots and calculated the area under the curve (AUC) to determine classification performance. AUC values for the single measures ranged from 0.67 (cortical thickness) to 0.94 (grey matter density). The combination of all six measures resulted in an AUC of 0.98. Our results demonstrate that the different anatomical MRI measures contain complementary information. A combination of these measures may therefore improve accuracy of AD diagnosis in clinical practice. *Hum Brain Mapp 37:1920–1929, 2016.*

Key words: Alzheimer's disease; anatomical MRI; cortical thickness; cortical area; cortical curvature; grey matter density; subcortical volumes; hippocampal shape; classification

Contract grant sponsor: Netherlands Organisation for Scientific Research (NWO); Contract grant number: 016130677; 05613010; 05613018; Contract grant sponsor: Dioraphte Foundation; Contract grant number: 09-02-03-00; Contract grant sponsor: Association for Frontotemporal Dementia Research

*Correspondence to: Frank de Vos; Instutute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands. E-mail: f.de.vos@fsw.leidenuniv.nl

Received for publication 21 October 2015; Revised 22 January 2016; Accepted 8 February 2016.

DOI: 10.1002/hbm.23147

Published online 25 February 2016 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

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

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

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

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

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

METHODS

Participants

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

MR Image Acquisition

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

Cortical Thickness, Area, and Curvature

To calculate cortical thickness, cortical area, and cortical curvature, we processed the T1-weighted images using the

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

Grey Matter Density of the Cortical Structures

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

Subcortical Volumes

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

Hippocampal Shape

To calculate hippocampal shape, we used the vertex analysis in FSL [Patenaude et al., 2011]. The shape values represent the distance of a vertex on the hippocampal mesh of a specific subject to the mean location of that vertex within the whole sample. A negative value represents a decrease of the size of the hippocampus on that specific location for a subject relative to the mean sample. Vice versa, a positive value represents a relative increase of the size of the hippocampus on that location. In the absence of a sufficiently detailed brain atlas of the human hippocampus, we used a data-driven method to reduce the number of features. We ran a principal component analysis on the vertex shape values of both hippocampi and extracted only the first 10 components, because these alone explained 86% of the variance in hippocampal shape values. This yielded 10 hippocampal shape features per subject.

Reference Measures: Whole-Brain Atrophy and Hippocampal Volume

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

Statistical Analyses

The features of the six anatomical MRI measures were used in an elastic net logistic regression to classify the subjects as either AD or control. Elastic net regression uses

penalties to hinder the features from entering the regression model [Zou & Hastie, 2005; Friedman et al., 2010]. Thus, only the most relevant predictors will enter the regression model, which is helpful if the number of features outnumbers the number of subjects. Elastic net regression uses a combination of an L1 (LASSO) [Tibshirani, 1996] and L2 (Ridge) [Hoerl and Kennard, 1970] penalty. Therefore, two hyperparameters should be set: the α parameter determines the relative weight of the two different penalties and λ determines the size of those penalties. Elastic net logistic regression has been used for AD classification by Schouten et al. [2016], Trzepacz et al. [2014], and Teipel et al. [2015].

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

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

First, we calculated the classification accuracy for each measure separately. Then we calculated the classification accuracy for the combination of all measures and for all pairs of two measures. We used two different methods to combine the features of different measures. In the first method, we concatenated the features and we used the concatenated feature set for classification similarly as the single measures. Concatenation is commonly used to combine different sets of features for prediction [Oliviera et al., 2010; Westman et al., 2013]. It might however not be the most optimal method for combination because all sets of features are then weighted equally, whereas some sets might be more predictive than others. Therefore, in the second method, we used a weighted average of the single measure predictions to increase classification performance [Wolpert, 1992; Breiman, 1996]. Measures that achieved a higher accuracy were given a larger weight. We used the inverse of the binomial deviance as a measure of accuracy. Binomial deviance is a continuous measure of prediction error and is therefore sensitive for small accuracy differences. We determined the measures' weights for each subject individually. To avoid overfitting, we used the binomial deviances within the training set to determine the contributions of the measures for the left out subjects. The weighted average was calculated for all the 50 cross-validation repetitions and the median result was presented.

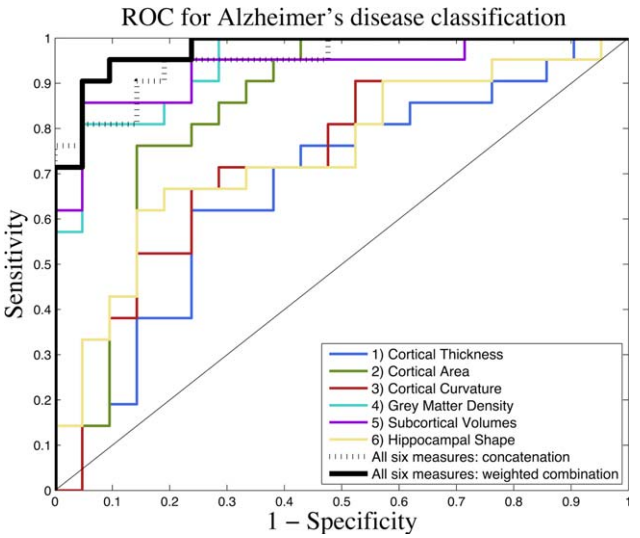

Figure 1.

ROC curves for discriminating AD patients from cognitively healthy controls. ROC curves are plotted for the six anatomical measures separately and for the two types of combinations of all six measures. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

RESULTS

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

TABLE I. AUC values discriminating AD patients from cognitively healthy controls

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

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

To further investigate the additive value of combining different measures, we calculated the AUCs for all pairs of two measures, using both measure concatenation (Fig. 2) and weighed combinations (Fig. 3). The tall bars represent the single measure AUCs, similar to that in Table I. The short bars represent the additive value of a second measure. Note that the additive value can also be negative when the addition of a second measure worsens the classification accuracy. For both concatenation and weighted combinations, the AUC of a single measure often improves when a second measure is added. When using concatenation, the highest AUC values are obtained by combining grey matter density with either subcortical volumes or cortical thickness (AUC = 0.98), which is even higher than the concatenation of all six modalities. Using weighted combinations, the highest AUC is obtained by combining grey matter density with subcortical volumes (AUC = 0.98), which is equal to the weighted combination of all six modalities.

These statistical models are primarily meant to predict class membership, and not to make claims about which features were most important to distinguish AD patients from controls. If a statistical model is built for prediction, rather than for explanation, one should be careful when interpreting the explanatory part of that model [Shmueli, 2010]. We elaborate more on this in the Discussion section. Yet, to illustrate the content of the classification models, we show the standardized beta values (averaged over the

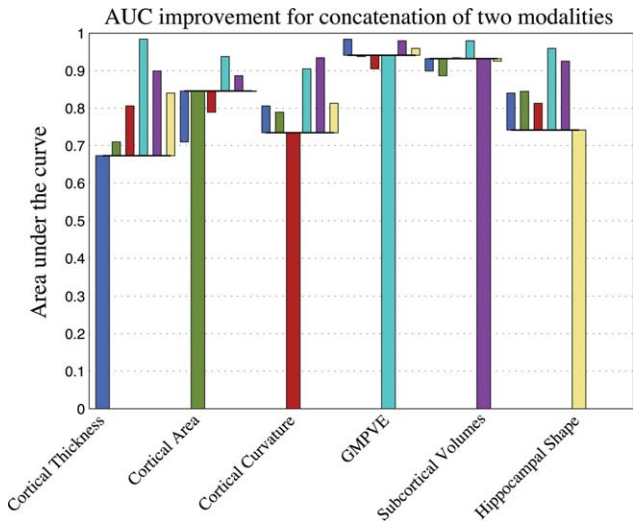

Figure 2.

AUC values for all the possible combinations of two measures using feature concatenation. The tall bars represent the single measure AUCs, similar to those in Table I. The short bars represent the additive value of adding a second measure. The additive values are mostly positive. For example, when cortical area is added to cortical thickness, the AUC increases from 0.67 to 0.71. However, sometimes the additive value of a second measure is negative. For example, cortical area is on itself a better predictor (0.85) than combined with cortical thickness (0.71). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

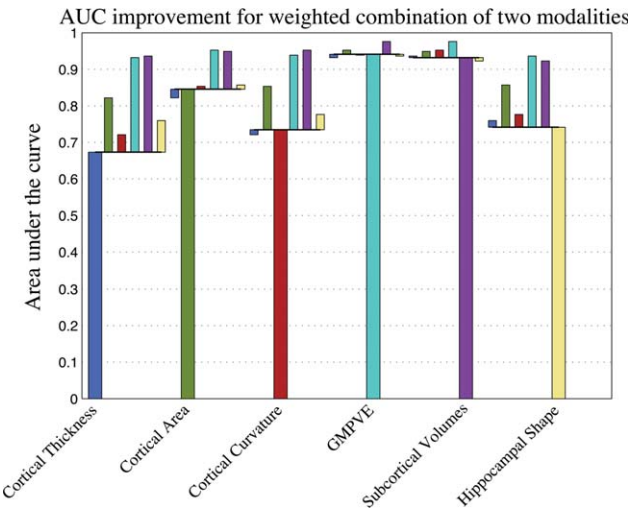

Figure 3.

AUC values for all the possible combinations of two measures using weighted combinations. The tall bars represent the single measure AUCs, similar to those in Table I. The short bars represent the additive value of adding a second measure. The additive values are mostly positive. For example, when cortical area is added to cortical thickness, the AUC increases from 0.67 to 0.82. However, sometimes the additive value of a second measure is negative. For example, cortical area is on itself a better predictor (0.85) than combined with cortical thickness (0.82). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

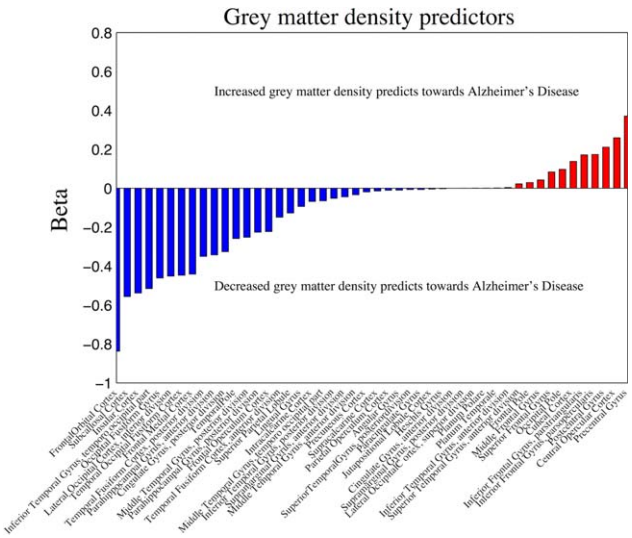

Figure 4.

Beta values for the cortical grey matter density estimates. The beta values represent the mean beta values over all the cross-validation folds of all the cross-validation repetitions. The beta values are ordered according to size. Negative beta values are colored blue, and positive beta values are colored red. Most beta values are negative, and the meaning of those is that low grey matter density predicts toward AD. Some beta values are positive, which is counterintuitive, but could mean something like if that region is relatively unaffected (high grey matter density) while some other regions are more affected, this is evidence in favor of AD. Note that grey matter density was only calculated for the cortical regions. The results for the subcortical volumes are presented in Figure 5. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

50 cross-validation repetitions) for the predictors of the grey matter density measure (Fig. 4) and the subcortical volumes measure (Fig. 5). We used these two measures for illustration because they discriminate best between AD patients and cognitively normal controls. For regions with a negative beta value, low grey matter density values or low subcortical volumes increase the odds for AD. For regions with a positive beta value, high grey matter density values or high subcortical volumes increase the odds for AD. This might seem contradictory because we do not expect to see increased grey matter density values or increased subcortical volumes in AD patients. The interpretation of these effects could be something like if that region is relatively unaffected (high grey matter density or large subcortical volume) while some other regions are more affected, this is the evidence in favor of AD. We plotted the regions in color coding for the grey matter density measure (Fig. 6) and the subcortical volumes measure (Fig. 7).

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

DISCUSSION

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

When used separately, all measures were sensitive to the detection of AD. We replicated previous findings in

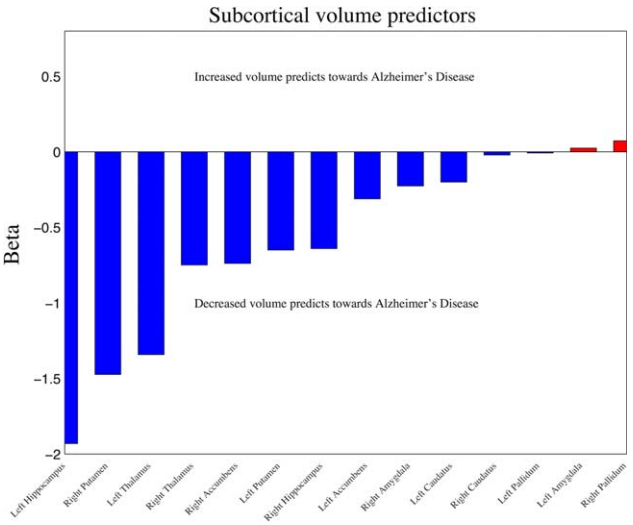

Figure 5.

Beta values for the subcortical volumes. The beta values represent the mean beta values over all the cross-validation folds of all the cross-validation repetitions. The beta values are ordered according to size. Negative beta values are colored blue, and positive beta values are colored red. Most beta values are negative, and the meaning of those is that a small volume predicts toward AD. Some beta values are positive, which is counterintuitive, but could mean something like if that region is relatively unaffected (large volume) while some other regions are more affected, this is evidence in favor of AD. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

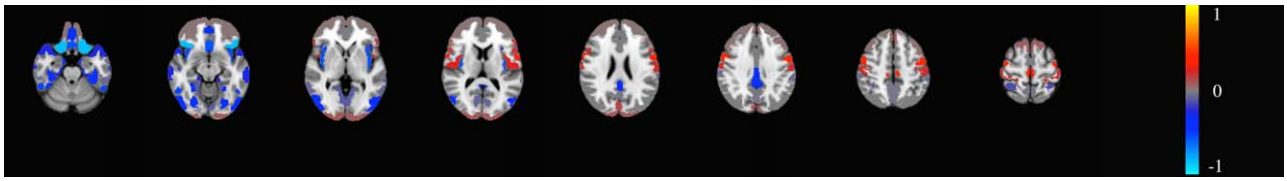

Figure 6.

The beta values from Figure 4 are presented here in color coding. The "cool" regions correspond with the blue bars and the meaning of those is that low grey matter density predicts toward AD. The "hot" regions correspond with the red bars and the meaning of those is that high grey matter density predicts toward AD. This is counterintuitive, but could mean something like if that region is relatively unaffected (high grey matter density) while some other regions are more affected, this is evidence in favor of AD. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

which grey matter density values [Cuingnet et al., 2011] and the volumes of the subcortical structures [Oliviera et al., 2010] are highly discriminative for AD. However, the high classification accuracy of cortical surface area compared to cortical thickness is relatively surprising. Previous studies found that cortical grey matter atrophy is primarily reflected in cortical thinning rather than in a decrease of cortical area [Dickerson et al., 2009; Westman et al., 2013].

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

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

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

We did not use nonlinear effects or interaction effects in our classification algorithm, which improves the reproducibility of our results. More complex classification models

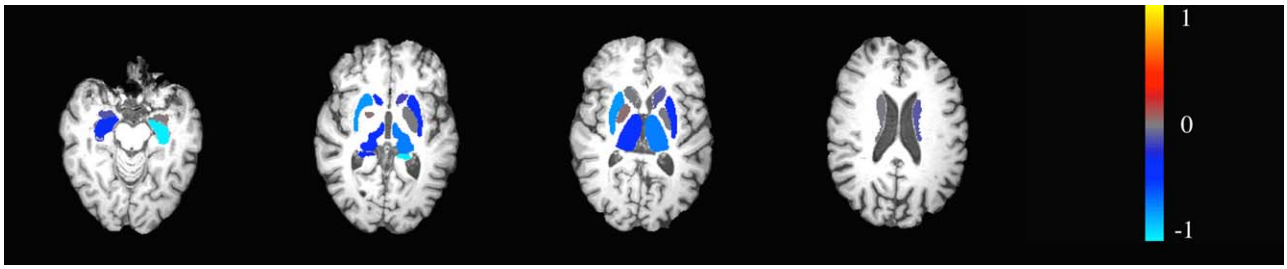

Figure 7.

The beta values from Figure 5 are presented here in color coding. The "cool" regions correspond with the blue bars and the meaning of those is that a small volume predicts toward AD. The "hot" regions correspond with the red bars and the meaning of those is that a large volume predicts toward AD. This is counterintuitive, but could mean something like if that region is relatively unaffected (large volume) while some other regions are more affected, this is evidence in favor of AD. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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

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

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

REFERENCES

Ashburner J, Friston KJ (2000): Voxel-based morphometry–the methods. Neuroimage 11:805–821.

Breiman L (1996): Stacked regressions. Mach Learn 24:49–64.

Bron EE, Smits M, van der Flier WM, Vrenken H, Barkhof F, Scheltens P, Papma JM, Steketee RME, Orellana CM, Meijboom R, Pinto M, Meireles JR, Garrett C, Bastos-Leite AJ, Abdulkadir A, Ronneberger O, Amoroso N, Bellotti R, Cárdenas-Peña D, Álvarez-Meza AM, Dolph CV, Iftekharuddin KM, Eskildsen SF, Coupé P, Fonov VS, Franke K, Gaser C, Ledig C, Guerrero R, Tong T, Gray KR, Moradi E, Tohka J, Routier A, Durrleman S, Sarica A, Di Fatta G, Sensi F, Chincarini A, Smith GM, Stoyanov ZV, Sørensen L, Nielsen M, Tangaro S, Inglese P, Wachinger C, Reuter M, van Swieten JC, Niessen WJ, Klein S (2015): Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. Neuroimage 111:562–579.

Callaert DV, Ribbens A, Maes F, Swinnen SP, Wenderoth N (2014): Assessing age-related gray matter decline with voxel-based morphometry depends significantly on segmentation and normalization procedures. Front Aging Neurosci 6:124

Chan D, Janssen JC, Whitwell JL, Watt HC, Jenkins R, Frost C, Rossor MN, Fox NC (2003): Change in rates of cerebral atrophy over time in early-onset Alzheimer's disease: Longitudinal MRI study. Lancet 362:1121–1122.

Chu C, Hsu AL, Chou KH, Bandettini P, Lin C (2012): Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 60:59–70.

Clerx L, Gronenschild EHBM, Echavarri C, Verhey F, Aalten P, Jacobs HIL (2015): Can FreeSurfer compete with manual volumetric measurements in Alzheimer's disease? Curr Alzheimer Res 12:358–367.

Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, Zhu W, Park M, Jiang T, Jin JS (2011): Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PLoS One 6:e21896

Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O (2011): Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage 56:766–781.

Dale aM, Fischl B, Sereno MI (1999): Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9: 179–194.

Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ (2011): Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging 32: 2322.e19–2322.e27.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006): An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980.

Diaz-de-Grenu LZ, Acosta-Cabronero J, Chong YFV, Pereira JMS, Sajjadi Sa, Williams GB, Nestor PJ (2014): A brief history of voxel-based grey matter analysis in Alzheimer's disease. J Alzheimers Dis 38:647–659.

Dickerson BC, Feczko E, Augustinack JC, Pacheco J, Morris JC, Fischl B, Buckner RL (2009): Differential effects of aging and Alzheimer's disease on medial temporal lobe cortical thickness and surface area. Neurobiol Aging 30:432–440.

Dyrba M, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, Kirste T, Teipel SJ (2015a): Predicting prodromal Alzheimer's disease in subjects with mild cognitive

impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. J Neuroimaging 25:738–747.

Dyrba M, Grothe M, Kirste T, Teipel SJ (2015b): Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. Hum Brain Mapp 36:2118–2131.

Fischl B, Sereno MI, Dale aM (1999): Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9:195–207.

Fotenos AF, Snyder AZ, Girton LE, Morris JC, Buckner RL (2005): Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 64:1032–1039.

Friedman J, Hastie T, Tibshirani R (2010): Regularization paths for generalized linear models via coordinate descent. J Stat Softw 30:1–3.

Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM (2010): The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6:67–77.

Gronenschild EHBM, Habets P, Jacobs HIL, Mengelers R, Rozendaal N, van Os J, Marcelis M (2012): The effects of FreeSurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One 7:e38234

Hastie T, Tibshirani R, Friedman J (2008): *The Elements of Statistical learning*. Retrieved from http://link.springer.com/book/10.1007%2F978-0-387-84858-7.

Hoerl AE, Kennard RW (1970): Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67.

Hutton C, Draganski B, Ashburner J, Weiskopf N (2009): A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. Neuroimage 48:371–380.

Jack CR, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, Knopman DS, Boeve BF, Ivnik RJ, Smith GE, Cha RH, Tangalos EG, Petersen RC (2004): Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 62:591–600.

De Jong LW, Van Der Hiele K, Veer IM, Houwing JJ, Westendorp RGJ, Bollen ELEM, De Bruin PW, Middelkoop HaM, Van Buchem Ma, Van Der Grond J (2008): Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: An MRI study. Brain 131:3277–3285.

Karas GB, Burton EJ, Rombouts SaRB, Van Schijndel Ra, O'Brien JT, Scheltens P, McKeith IG, Williams D, Ballard C, Barkhof F (2003): A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18:895–907.

Karas GB, Scheltens P, Rombouts SaRB, Visser PJ, Van Schijndel Ra, Fox NC, Barkhof F (2004): Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. Neuroimage 23:708–716.

Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014): Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 6:1–15. Journal of Cheminformatics.

Lerch JP, Pruessner JC, Zijdenbos A, Hampel H, Teipel SJ, Evans AC (2005): Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. Cereb Cortex 15:995–1001.

Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, Schuff N, Fox NC, Ourselin S (2010): Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51:1345–1359.

Mckhann G, Drachman D, Folstein M (1984): Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force. Neurology 34:939–944.

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (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 Dement 7:263–269.

Morra JH, Tu Z, Apostolova LG, E A, Avedissian C, Madsen SK, Parikshak N, Hua X, Toga AW, Jr CRJ, Schuff N, Weiner MW (2009): Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Brain 30:2766–2788.

Mulder ER, de Jong Ra, Knol DL, van Schijndel Ra, Cover KS, Visser PJ, Barkhof F, Vrenken H (2014): Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92:169–181.

Oliveira DMPP, Nitrini R, Busatto G, Buchpiguel C, Sato JR, Amaro E (2010): Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. J Alzheimer's Dis 19:1263–1272.

Patenaude B, Smith SM, Kennedy DN, Jenkinson M (2011): A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56:907–922.

Pereira JB, Ibarretxe-Bilbao N, Marti M-J, Compta Y, Junqué C, Bargallo N, Tolosa E (2012): Assessment of cortical degeneration in patients with Parkinson's disease by voxel-based morphometry, cortical folding, and cortical thickness. Hum Brain Mapp 33:2521–2534.

Querbes O, Aubry F, Pariente J, Lotterie J-A, Démonet J-F, Duret V, Puel M, Berry I, Fort J-C, Celsis P (2009): Early diagnosis of Alzheimer's disease using cortical thickness: Impact of cognitive reserve. Brain 132:2036–2047.

Redolfi A, Manset D, Barkhof F, Wahlund L-O, Glatard T, Mangin J-F, Frisoni GB (2015): Head-to-head comparison of two popular cortical thickness extraction algorithms: A cross-sectional and longitudinal study. PLoS One 10:e0117692

Risacher SL, Shen L, West JD, Kim S, McDonald BC, Beckett La, Harvey DJ, Jack CR, Weiner MW, Saykin AJ (2010): Longitudinal MRI atrophy biomarkers: Relationship to conversion in the ADNI cohort. Neurobiol Aging 31:1401–1418.

Rombouts Sa, Barkhof F, Witter MP, Scheltens P (2000): Unbiased whole-brain analysis of gray matter loss in Alzheimer's disease. Neurosci Lett 285:231–233.

Ronan L, Pienaar R, Williams G, Bullmore E, Crow TJ, Roberts N, Jones PB, Suckling J, Fletcher PC (2011): Intrinsic curvature: A marker of millimeter-scale tangential cortico-cortical connectivity? Int J Neural Syst 21:351–366.

Salat DH, Kaye Ja, Janowsky JS (1999): Prefrontal gray and white matter volumes in healthy aging and Alzheimer disease. Arch Neurol 56:338–344.

Scher aI, Xu Y, Korf ESC, White LR, Scheltens P, Toga aW, Thompson PM, Hartley SW, Witter MP, Valentino DJ, Launer LJ (2007): Hippocampal shape analysis in Alzheimer's disease: A population-based study. Neuroimage 36:8–18.

Schouten TM, Loitfelder M, de Vos F, Seiler S, van der Grond J, Lechner A, Hafkemeijer A, Möller C, Schmidt R, de Rooij M,

Rombouts SARB (2016): Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. NeuroImage Clin 11:46–51.

Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD (2009): Neurodegenerative diseases target large-scale human brain networks. Neuron 62:42–52.

Shmueli G (2010): To explain or to predict? Stat Sci 25:289–310.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23:S208–S219.

Teipel SJ, Kurth J, Krause B, Grothe MJ (2015): The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment — Beyond classical regression. NeuroImage Clin 8:583–593.

Thompson PM, Hayashi KM, De Zubicaray GI, Janke AL, Rose SE, Semple J, Hong MS, Herman DH, Gravano D, Doddrell DM, Toga AW (2004): Mapping hippocampal and ventricular change in Alzheimer disease. Neuroimage 22:1754–1766.

Tibshirani R (1996): Regression selection and shrinkage via the Lasso. J R Stat Soc B 58(1):267–288.

Trzepacz PT, Yu P, Sun J, Schuh K, Case M, Witte MM, Hochstetler H, Hake A (2014): Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to alzheimer's dementia. Neurobiol Aging 35:143–151.

Tustison NJ, Cook Pa, Klein A, Song G, Das SR, Duda JT, Kandel BM, van Strien N, Stone JR, Gee JC, Avants BB (2014): Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 99:166–179.

Wee C-Y, Yap P-T, Shen D (2013): Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum Brain Mapp 34:3411–3425.

Westman E, Aguilar C, Muehlboeck JS, Simmons A (2013): Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. Brain Topogr 26:9–23.

Wolpert DH (1992): Stacked generalization. Neural Networks 5: 241–259.

Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, Soininen H, Lötjönen J (2011): Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS One 6:1–9.

Zhang Y, Brady M, Smith S. (2001): Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imag 20(1):45–57.

Zhang Y, Schuff N, Camacho M, Chao LL, Fletcher TP, Yaffe K, Woolley SC, Madison C, Rosen HJ, Miller BL, Weiner MW (2013): MRI markers for mild cognitive impairment: Comparisons between white matter integrity and gray matter volume measurements. PLoS One 8:e66367

Zou H, Hastie T (2005): Regularization and variable selection via the elastic-net. J R Stat Soc 67:301–320.