Neurobiology of Aging 81 (2019) 58–66

Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging

The value of hippocampal volume, shape, and texture for 11-year prediction of dementia: a population-based study

Hakim C. Achterberg [a,*], Lauge Sørensen [b,c], Frank J. Wolters [d,e], Wiro J. Niessen [a,f], Meike W. Vernooij [d,g], M. Arfan Ikram [d,g], Mads Nielsen [b,c], Marleen de Bruijne [a,b]

[a] Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC, Rotterdam, the Netherlands
[b] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[c] Biomediq A/S, Copenhagen, Denmark
[d] Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands
[e] Department of Neurology, Erasmus MC, Rotterdam, the Netherlands
[f] Imaging Science and Technology, Department of Applied Sciences, Delft University of Technology, Delft, the Netherlands
[g] Department of Radiology, Erasmus MC, Rotterdam, the Netherlands

ARTICLE INFO

Article history:
Received 31 July 2018
Received in revised form 25 April 2019
Accepted 11 May 2019
Available online 28 May 2019

Keywords:
Dementia
Hippocampal shape
Hippocampal texture
Prediction
MRI
Population-based

ABSTRACT

Hippocampal volume and shape are known magnetic resonance imaging biomarkers of neurodegeneration. Recently, hippocampal texture has been shown to improve prediction of dementia in patients with mild cognitive impairment, but it is unknown whether texture adds prognostic information beyond volume and shape and whether the predictive value extends to cognitively healthy individuals. Using 510 subjects from the Rotterdam Study, a prospective, population-based cohort study, we investigated if hippocampal volume, shape, texture, and their combination were predictive of dementia and determined how predictive performance varied with time to diagnosis and presence of early clinical symptoms of dementia. All features showed significant predictive performance with the area under the receiver operating characteristic curve ranging from 0.700 for texture alone to 0.788 for the combination of volume and texture. Although predictive performance extended to those without objective cognitive complaints or mild cognitive impairment, performance decreased with increasing follow-up time. We conclude that a combination of multiple hippocampal features on magnetic resonance imaging performs better in predicting dementia in the general population than any feature by itself.

1. Introduction

Dementia is a neurological syndrome that results from various underlying pathologies, leading to neurodegeneration and cerebral atrophy. Neurodegeneration takes place over the course of many years, and irrevocable degenerative changes in the brain exist by the time clinical symptoms of dementia manifest. Consequently, there is increasing need for tools that identify individuals at high risk of dementia in the population to facilitate development of preventative and curative measures. Magnetic resonance imaging (MRI) allows noninvasive imaging of the brain and is widely used to support dementia diagnosis, and that of its pathological subtypes, of which Alzheimer's disease (AD) is the most common. Of specific interest to dementia, and AD in particular, is the hippocampus that is affected early in the disease process (Braak and Braak, 1997; West et al., 1994, 2004). Because abnormalities of the hippocampus precede the symptoms of cognitive decline, MRI can not only be used to diagnose AD but also to predict its future development.

Hippocampal volume measured in MRI has shown to be predictive of dementia in patients with mild cognitive impairment (MCI) (Devanand et al., 2007; Jack et al., 1999) as well as in community-dwelling individuals (den Heijer et al., 2006). To measure hippocampal atrophy even more specifically, hippocampal shape has also been used to diagnose dementia (Li et al., 2007; Wang et al., 2007) and predict dementia in both subjects with MCI (Costafreda et al., 2011; Ferrarini et al., 2009) and in a sample of the general population (Achterberg et al., 2014). Recently, studies suggested that a novel hippocampal imaging marker, namely hippocampal texture, may further improve prediction of conversion from MCI to AD (Chincarini et al., 2011; Sørensen et al., 2016). Hippocampal texture may be a valuable marker, as it is thought to reflect change in tissue texture as a consequence of characteristic

* Corresponding author at: Departments of Radiology and Medical Informatics, Erasmus MC, Rotterdam 3015GE, the Netherlands. Tel.: +31628144048; fax: +31107034033.
E-mail address: h.achterberg@erasmusmc.nl (H.C. Achterberg).

https://doi.org/10.1016/j.neurobiolaging.2019.05.007

pathological changes of dementia such as neurofibrillary tangles and amyloid-β plaques for AD. However, it is unknown how suitable texture and the combinations of texture with volume and/or shape are for the purpose of dementia prediction in a general population. It is also unclear which marker shows the earliest signs of disease.

We therefore computed volume, shape, and texture of the hippocampi on MRI in nondemented subjects from a population-based study, to determine the predictive value of each MRI imaging biomarker and their combinations for the occurrence of dementia during 11 years of follow-up.

2. Materials and methods

2.1. Study population

This study was embedded in the Rotterdam Study: a prospective, population-based cohort study among inhabitants aged >55 years from the Ommoord area in Rotterdam, The Netherlands. The Rotterdam Study methods have been described previously (Hofman et al., 2015; Ikram et al., 2015). In brief, between 1990 and 1993, 7983 individuals agreed to participate (response figure 78%). Of these, 965 elderly subjects were randomly selected to undergo MRI of the brain in 1995–1996 (den Heijer et al., 2006). As part of the eligibility criteria, we excluded individuals who had dementia, were blind, or had MRI contraindications. This left 832 persons eligible for participation. Among these, 563 persons gave their written informed consent to participate in the present study (response rate, 68%). Of the 563 participants, 52 developed claustrophobia, leaving 511 participants with an MRI scan. For 1 subject, we could not retrieve all data required for the calculation of the texture features, leaving 510 participants with complete data. These individuals constitute the study population for this study. The Rotterdam Study has been approved by the medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. Written informed consent was obtained from all participants.

2.2. Dementia screening and surveillance

Participants were screened for dementia at each center visit using the Mini–Mental State Examination (MMSE) and the Geriatric Mental Schedule organic level (de Bruijn et al., 2015). Those with MMSE <26 or Geriatric Mental Schedule >0 underwent further investigation and informant interview including the Cambridge Examination for Mental Disorders of the Elderly. In addition, the entire cohort was continuously under surveillance for dementia through electronic linkage of the study center with medical records from general practitioners and the regional institute for outpatient mental health care. A consensus panel headed by a consultant neurologist established the final diagnosis according to standard criteria for dementia (DSM-III-R). Follow-up until 1st January 2006 was virtually complete at the time of this study. Within this period, participants were censored at date of dementia diagnosis, death, loss to follow-up, or administrative censoring date, whichever came first. The completeness of follow-up (Clark et al., 2002) was 99.0%, meaning that of all potential person-years of follow-up in the study, only 1.0% was lost due to censoring other than dementia or death.

During the study period, 52 subjects were diagnosed with dementia. The median interval between MRI acquisition and dementia diagnosis was 4.0 years with an interquartile range of 4.8 years (ranging from 0.74 to 11.13).

2.3. Memory assessment and mild cognitive impairment

The memory assessment used in this study was performed at the time of the MRI acquisition. MCI was defined as the combination of subjective cognitive complaints and objective cognitive impairment, in the absence of dementia, following the criteria proposed by Petersen et al. (1997). Subjective cognitive complaints were evaluated by interview, which included 3 questions about memory (feeling forgetful, worry about forgetting things, and word-finding difficulties) and questions on everyday functioning (problems with orientation in familiar places and difficulties getting dressed, using a key, leaving the stove on, or storing things in an unusual place). Subjective cognitive complaints were scored positive when an individual answered positive to at least 1 of these questions. For assessment of objective cognitive impairment, we used a cognitive test battery comprising a letter-digit substitution test, Stroop test, verbal fluency test, and 15-word verbal learning test. To obtain more robust measures, we constructed compound scores by principal component analysis (PCA) for memory function (immediate and delayed recall), information-processing speed (letter-digit substitution test, Stroop reading, and color-naming subtasks), and executive function (Stroop interference subtask, letter-digit substitution test, and verbal fluency). Objective cognitive impairment on each of the domains was defined as a test score below −1.5 standard deviations of the age- and education-adjusted mean of the study population. Given that our sample is a cross-section of the general population, we assume that the means of the study population closely approximate the population means and can be used safely for the above definitions.

2.4. Selection of cases and controls

The entire data set, hereafter referred to as the cohort set, contained 52 subjects who developed dementia and 458 subjects who did not develop dementia within the follow-up period of up to 11 years (for dementia case ascertainment, see Section 2.2). To train and test a model independent of age and gender, an age- and gender-matched subset was identified, hereafter referred to as the matched set. The matching was performed using the following criteria: (1) the gender had to be the same, (2) the follow-up time of the controls should be at least as long as the time to diagnosis of the corresponding case, (3) the age could not differ more than 1.5 years, and (4) controls did not develop dementia during the entire follow-up period. With these criteria, it was possible to select 3 unique age- and gender-matched controls per case for 50 of the cases. The remaining 2 cases were not included in the matched set. Characteristics of the cohort set and matched set are listed in Table 1. The cohort is the same as in Achterberg et al. (2014), except for 1 subject who did not have all required data for the calculation of the texture features.

2.5. MRI scan protocol

All subjects were scanned in the period 1995–1996 on a Siemens 1.5T scanner. The sequence used was a custom-designed inversion recovery, three-dimensional (3D) half-Fourier acquisition single-shot turbo spin echo sequence. This sequence had the following acquisition parameters: inversion time 4.400 ms, repetition time 2.800 ms, effective echo time 29 ms, matrix size 192 × 256, flip angle 180°, slice thickness 1.25 mm, and acquired in sagittal direction. The images were reconstructed to a 128 × 256 × 256 matrix with a voxel dimension of 1.25 × 1.0 × 1.0 mm. The acquired MRI scans of the 510 subjects in this study had a mean signal-to-noise ratio (SNR) (Magnotta et al., 2006) of 27.79 with a standard deviation of 3.82 in the hippocampus.

Table 1
Characteristics of the subjects in various subgroups

Group	N	Women (%)	Age (y) mean (Std)	MMSE median (IQR)	SCC (%)	MCI (%)	TTD (y) median (IQR)
Cohort							
All	510	49.8	73.49 (7.89)	28 (2)	58.0	2.7	4.02 (4.78)
Cases	52	61.5	79.02 (6.37)	27 (2)	76.9	9.6	4.02 (4.78)
Controls	458	48.5	72.86 (7.80)	28 (2)	55.9	2.0	n/a
Matched							
Cases	50	60.0	78.75 (6.34)	27 (3)	78.0	10.0	4.02 (4.88)
Controls	150	60.0	78.72 (6.40)	28 (2)	59.3	2.0	n/a
Time to diagnosis [0–3]							
Cases	18	55.6	80.15 (6.18)	26.5 (4.75)	88.9	22.2	1.11 (0.67)
Controls	54	55.6	80.27 (6.18)	28 (2)	55.6	1.9	n/a
Time to diagnosis (3–6]							
Cases	15	60.0	79.16 (4.96)	27 (2)	73.3	6.7	4.05 (1.67)
Controls	45	60.0	79.21 (4.92)	29 (2)	60.0	0.0	n/a
Time to diagnosis (6–11]							
Cases	17	64.7	76.90 (7.11)	27 (3)	70.6	0.0	6.84 (1.82)
Controls	51	64.7	76.66 (7.20)	28 (2.5)	62.7	3.9	n/a
No MCI							
Cases	38	65.7	79.54 (6.63)	27 (2.75)	68.4	0.0	5.30 (3.73)
Controls	417	48.4	72.78 (7.75)	28 (2)	51.6	0.0	n/a
No objective cognitive complaints							
Cases	37	64.9	79.37 (6.63)	27 (2)	70.3	0.0	5.31 (3.48)
Controls	387	49.4	73.03 (7.68)	28 (2)	55.6	0.0	n/a
No subjective cognitive complaints							
Cases	12	66.7	77.96 (7.83)	27 (3.5)	0.0	0.0	5.51 (3.38)
Controls	202	45.0	72.21 (7.61)	28 (2)	0.0	0.0	n/a

Key: MMSE, Mini–Mental State Examination; IQR, interquartile range; SCC, subjective cognitive complaints; MCI, mild cognitive impairment; TTD, time to diagnosis.

2.6. MRI features

All MRI scans were corrected for bias fields using the nonparametric nonuniform intensity normalization (N3) algorithm (Sled et al., 1998), and the left and right hippocampi were automatically segmented. Based on these segmentations, 3 different types of MRI hippocampus features were computed: bilateral hippocampal volume, bilateral hippocampal shape, and bilateral hippocampal texture.

2.6.1. Hippocampus segmentation

The hippocampi were automatically segmented using a method based on multi-atlas registration, a statistical intensity model, and a regularizer to promote smooth segmentations (van der Lijn et al., 2008). These components were combined in an energy model, which was globally optimized using graph cuts. Atlasses consisted of manually delineated images from 20 participants from the same population as used in this study. The atlas set contains images of 20 subjects selected to cover the population variation in age, sex, and hippocampus size. The atlasses were used lateralized, meaning only hippocampi of the same hemisphere were used as atlasses. Leave-one-out experiments on the atlas images showed mean Dice similarity indices of 0.85 $\pm$ 0.04 and 0.86 $\pm$ 0.02 for the left and right hippocampi, respectively. The segmentation results of all images used in this study were inspected by a trained observer and manually corrected in case of large errors; 2 cases and 69 controls were manually corrected.

2.6.2. Hippocampal volume

The left and right hippocampal volumes were calculated directly from the segmentation and were subsequently normalized by dividing by the intracranial volume. The left and right normalized hippocampal volumes were concatenated into a 2-dimensional volume feature vector. The intracranial volume was computed by deformable registration of a single brain mask to the target image and calculating the volume inside the brain mask (Ikram et al., 2010).

2.6.3. Hippocampal shape

The shape features used are the same as in the study by Achterberg et al. (2014). Because the hippocampus segmentation can have single voxel holes or contain small non-connected regions and creation of the shape model requires a single-body object, we extracted the largest single body from the segmentation for each of the hippocampi and applied a hole-filling operation before computing shape features. Furthermore, an antialiasing step was used to smooth the binary segmentation (Whitaker, 2000).

First, the 3D shape model was fitted to the left and right preprocessed hippocampus segmentations separately. Shapes were described in 3D by 1024 corresponding points per hippocampal surface (left and right separately). These points were defined using the entropy-based particle system as presented before (Cates et al., 2006, 2007). This method aims at finding a uniform sampling of the shapes while minimizing the information content of the resulting shape model, leading to a compact model with optimal point correspondences. The resulting hippocampal shapes were then corrected for the hippocampal volume by scaling the point coordinates.

The point coordinates of the left and right hippocampi were concatenated into a single feature vector describing both shapes jointly. To reduce the number of features, a PCA was applied that retained 99% of the variance. In our experiments, the transformation of the PCA was estimated only on the design set and then applied to the corresponding test set (see Fig. 1). For each outer cross-validation training fold, the PCA resulted in either 116 or 117 components, that is, the shape feature vector was either 116- or 117-dimensional.

2.6.4. Hippocampal texture

Texture features were calculated using a previously published method that was developed on a different cohort (Sørensen et al., 2016). In brief, the MRI scan was filtered using a Gaussian derivative-based, multiscale, rotation-invariant filter bank comprising 28 filters (7 base filters [the 3 eigenvalues of the hessian, gradient magnitude, Laplacian of the Gaussian, Gaussian curvature, and the Frobenius norm of the Hessian] at scales 0.6, 0.85, 1.2, and 1.7 mm), and 28 corresponding filter response histograms were computed using the filter responses in both the left and the right hippocampus jointly. Each histogram was estimated using adaptive binning with 9 bins and normalized to sum to one. The final texture descriptor comprised the concatenated histograms and was of dimensionality 252. The exact same settings as specified in Sørensen et al. (2016) were used.

2.6.5. Combination of features

Feature types (i.e., volume, shape, and texture) were combined into new feature sets (e.g., volume + shape, volume + shape + texture) by concatenation. To ensure equal influence, each feature type was normalized for the sum of the eigenvalues of its covariance matrix in the design set (see Fig. 1 or Section 2.8). This ensured that each feature type contained the same amount of variance in the combined feature vector.

2.7. Classification for dementia prediction

We train classifiers to discriminate between subjects who develop dementia during follow-up and subjects who remain cognitively intact, based on different MRI feature configurations (volume, shape, and texture in isolation and all possible combinations). For all configurations except volume, we used a soft-margin support vector machine (SVM) classifier (Cortes and

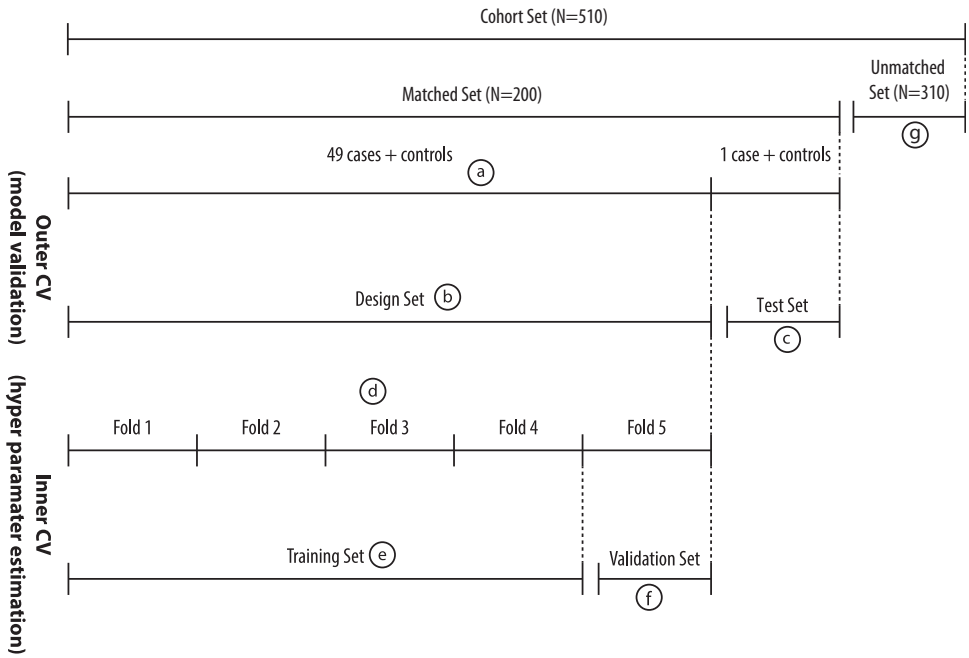

Fig. 1. Overview of the cross-validation setup. Two nested cross-validation loops are used. The outer cross-validations loop is a leave-case-out cross-validation on the matched data. The entire matched set (a) is split in a design set (b) consisting of 49 cases and their matching controls and a test set (c) of one case and the matching controls. For estimating the hyperparameters of the SVM, a 5-fold cross-validation on the design set (d) is used. In this cross-validation, the design set (b) is split in a training set (e) consisting of 80% of the design set and validation set (f) consisting of the remaining 20% of the design set. The unmatched set (g) was scored by training and SVM on the entire matched set and applying this to the unmatched set. In this procedure, the SVM hyperparameters were determined using 5-fold cross-validation on the matched set. Abbreviation: SVM, support vector machine.

Vapnik, 1995) with a radial basis function (RBF) kernel. A linear SVM was used for hippocampal volume. In all cases, the training samples were weighted with a 3 (for cases) to 1 (for controls) ratio to compensate for the matching of 3 controls per case, meaning that the case class was regularized more than the control class in the SVM.

The hyperparameters of the SVM were determined by optimizing the area under the receiver operating characteristic curve (AUC) in cross-validation on the design set (see Fig. 1). A special 5-fold cross-validation was used in which a case and its matching controls were kept together, ensuring every fold was properly age- and gender-matched. An RBF SVM has two hyperparameters: C which controls the amount of regularization used in the soft-margin and γ which controls the scale of the RBF kernel. The estimation of the hyperparameters was performed in a similar fashion to Sørensen et al. (2016). An initial estimate for the γ parameter was generated using the Jaakkola Heuristic (Jaakkola et al., 1999). Then a grid around this estimated γ_{init} and spanning a large range of C was searched to find the optimal hyperparameter combination. The search range for C was $e^0, e^1, e^2, \ldots, e^9, e^{10}$ and the search range for γ was $e^{\log(\gamma_{init})-4}, e^{\log(\gamma_{init})-3}, \ldots, e^{\log(\gamma_{init})+4}$. The linear SVM only has 1 hyperparameter C. It was optimized in the same way as for the RBF SVM but using a one-dimensional grid search. The parameter range searched was identical to the RBF SVM.

The posterior probability for a subject to develop dementia, given the observed feature configuration, was computed from the SVM discrimination function d as $P(d) = 1/(1 + e^{-d})$.

For the implementation of the classification experiments, we used the Python library scikit-learn (Pedregosa et al., 2011). The SVM implementation used by scikit-learn is libsvm (Chang and Lin, 2001). The Jaakkola index we implemented ourselves using the Shark C++ machine learning library (Igel et al., 2008) as a reference implementation.

2.8. Analyses

Within the matched set of 50 cases and 150 controls, a leave-case-out cross-validation scheme was used, where in each fold, the classifiers were evaluated on a case and its 3 matching controls and created using all remaining data in the matched set as the design set. As noted in the previous section, the hyperparameters of the SVM were selected using a 5-fold cross-validation on the design set. The model training, validation, testing procedure is schematically drawn in Fig. 1. For the remaining 310 subjects in the cohort set, a model was trained on the entire matched set and applied to all remaining subjects. Classification performance was evaluated using AUC.

The hyperparameter optimization can be sensitive to the fold layout in the 5-fold cross-validation on the design set. We therefore report results of the experiment repeated 25 times with a different random ordering of the data in the design set, which leads to different fold layouts of the internal cross-validation. Different configurations were compared by comparing the receiver operating characteristic curves of the mean posterior (over the 25 repetitions) using the DeLong test (DeLong et al., 1988). We use a p-value of 0.05 as the threshold for significance.

To investigate the performance of different features by time between the MRI scan and dementia diagnosis, we then performed analyses stratifying the matched set into 3 timeframes (i.e., diagnosis within 3 years after MRI [N = 72]; diagnosis between 3 and 6 years [N = 60]; and diagnosis more than 6 years after MRI [N = 68]). This is a stratification of the already computed classifier outputs, and no additional training/testing procedures were performed.

To assess performance of feature types independent of early clinical symptoms of cognitive decline, we performed several analyses on subsets of the cohort set, excluding individuals with (1) subjective cognitive complaints, (2) objective cognitive complaints, and (3) MCI.

Considering that a good prediction model likely will need non-imaging features in addition to the MRI features, we evaluated the added value of combining age and sex with probabilities obtained from the classification models using a logistic regression model with dementia development as outcome. Like with the DeLong tests, the mean posterior probability for each subject overall 25 runs was used. The various logistic regression models all included age, sex, and MMSE score in addition to the SVM probabilities for the MRI features. For nested models, we used a log likelihood ratio test to validate if the nested model was significantly worse than the full model.

3. Results

3.1. Prediction of conversion to dementia

The AUCs for dementia prediction in the matched set and cohort set are provided in Fig. 2. Of the individual feature types, volume performed best (median AUC of 0.736 on the matched set and 0.725 on the cohort set), followed by shape (0.693 and 0.713), and finally texture (0.632 and 0.700). Of the combined features sets, volume + texture performed best (median AUC 0.790 on the matched set and 0.788 on the cohort set), followed by the combination of all features, volume + shape + texture (0.773 and 0.776), shape + volume (0.762 and 0.756), and finally shape + texture (0.707 and 0.746). In Table 2, the resulting *p*-values of the DeLong tests on the mean posterior of the 25 repetitions are given. It can be seen that on the matched set, all feature configurations showed significant predictive value. Most differences between feature configurations were not significant. None of the individual features performed significantly different from each other. Of the pairwise combinations, only volume + texture was significantly better than each single feature. The combination of all 3 feature types was significantly better than either shape or texture alone but did not significantly outperform any of the pairwise combinations (volume + texture actually scored better on average, but this difference was also not statistically significant).

The variation in the results over the different feature configurations with respect to the data permutations is depicted in Fig. 2. Strikingly, volume has an interquartile range (IQR) of zero, i.e., no variation. This is probably because it used a linear SVM and therefore only 1 parameter was optimized in the cross-validation on the training set or because the resulting classifier was less sensitive to the C parameter due to the less flexible decision boundary of a linear SVM. For the remaining configurations that all used an RBM SVM, texture was the least stable with an IQR of 0.063 on the matched set and 0.059 on the cohort set, followed by shape + texture, shape, volume + shape, volume + shape + texture, and finally by volume + texture with the smallest IQR of 0.007 for the matched set and 0.013 for the cohort set.

3.2. Stratification by time-to-conversion

Fig. 2 shows results of the cohort set, matched set, and the matched set stratified by time to dementia diagnosis. For the shortest interval between scan and dementia diagnosis (up to 3 years), volume + texture performed best with median AUC of 0.935, followed by volume + shape + texture, volume, volume + shape, shape + texture, texture, and finally shape with 0.675. As can been seen in Table 3, all feature configurations were significantly predictive. In the interval from 3 to 6 years, the order was almost the same, but differences were smaller; the best was volume + texture with 0.809, followed by volume + shape + texture, volume, shape + texture, volume + shape, shape, and finally texture with 0.562. Although volume and volume + shape were trending toward

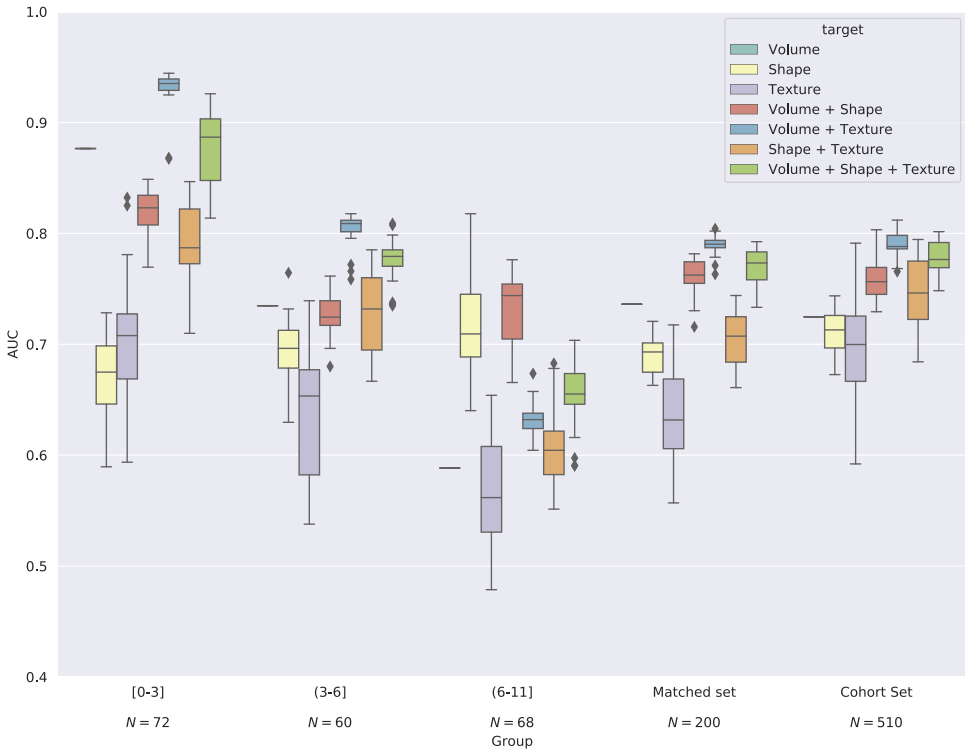

Fig. 2. Classification result for different feature configurations for the matched and cohort sets, and for the matched set stratified by time to diagnosis. The box plot shows the AUCs for the 25 random permutations of the data. The first 3 columns are the matched set stratified in different time intervals (in years) between scanning and follow-up dementia diagnosis. ◆ represent samples that are considered outliers (based on a function of the interquartile range). Abbreviation: AUC, area under the receiver operating characteristic curve.

Table 2
Area under the receiver operating characteristic curve for the mean posterior probability for each feature configuration (in gray) and *p*-values of differences in ROC curve between feature configurations

Matched Set	AUC	Random 0.50	Volume 0.74	Shape 0.71	Texture 0.69	V + S 0.77	V + T 0.79	S + T 0.74	All 0.78
Random	0.50		*0.0004*	*0.0009*	*0.0021*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
Volume	0.74	*0.0004*		0.5685	0.3776	0.1945	*0.0212*	0.9354	**0.0732**
Shape	0.71	*0.0009*	0.5685		0.6693	**0.0533**	*0.0444*	0.3883	*0.0455*
Texture	0.69	*0.0021*	0.3776	0.6693		**0.0963**	*0.0058*	**0.0762**	*0.0162*
V + S	0.77	*0.0000*	0.1945	**0.0533**	**0.0963**		0.4888	0.3789	0.7529
V + T	0.79	*0.0000*	*0.0212*	*0.0444*	*0.0058*	0.4888		**0.0631**	0.2232
S + T	0.74	*0.0000*	0.9354	0.3883	**0.0762**	0.3789	**0.0631**		0.1133
All	0.78	*0.0000*	**0.0732**	*0.0455*	*0.0162*	0.7529	0.2232	0.1133	

Key: AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic; S + T, shape + texture; V + S, volume + shape; V + T, volume + texture. *p*-values are obtained using the DeLong test. Bold and italics indicate significance level. Italics indicates significance ($p < 0.05$), and bold indicates trending toward significance ($0.05 \leq p < 0.10$).

significance, only the combinations that include texture were significantly predictive. For more than 6 years between scan and diagnosis, the order was very different; the best performing configuration was volume + shape with median AUC of 0.744, followed by shape, volume + shape + texture, volume + texture, shape + texture, volume, and finally texture with 0.562. None of the feature configurations showed a significant predictability; only shape and volume + shape were trending toward significance.

3.3. Dementia prediction in the cognitively healthy

Fig. 3 and Table 3 show the prediction performance in subgroups of the cohort set where different levels of cognitive impairment were excluded. Note that the entire cohort in Figs. 2 and 3 are the same and can be used as a reference point. Excluding subjects who already had MCI at scan time barely influenced the results; the decrease in median AUC was between 0.000 (volume) and 0.039 (shape + texture). All feature configurations still had a significant predictability. Excluding the subjects with objective cognitive complaints also did not influence the result much; the median change in AUC was between a 0.008 increase (volume) and a 0.045 decrease (volume + texture). Again, all feature configurations showed significant predictability. Finally when all subjects who had subjective cognitive complaints at scan time were excluded, the performance was also lower than in the cohort. The median drop in AUC varied between 0.079 (texture) and 0.160 (shape). None of the feature configurations reached significant predictive value. It should be noted that this subgroup contained only 12 cases.

3.4. Assessing dementia risk using MRI, age, sex, and MMSE score

In Table 4, logistic regression models to assess dementia risk based on MRI imaging biomarkers, age, and sex are presented. All models include age, sex, and MMSE score as independent variables and have development of dementia as dependent (outcome) variable. All models were fitted on the entire cohort set. The initial model used only the volume posterior probability, which when combined with the other variable was not significant ($p = 0.114$). When adding the shape posterior probability as an independent variable, both volume and shape were significant (shape $p < 0.001$; volume $p = 0.038$). A log likelihood ratio test showed that the improvement of the full model (with volume and shape) compared with the nested model (with only volume) was significant ($p < 0.001$). When combining volume posterior and texture posterior probabilities, texture was highly significant ($p < 0.001$), whereas volume was only trending toward significance ($p = 0.087$), and the difference between the nested and full model was significant ($p < 0.001$). Finally, in the model including all 3 MRI imaging biomarkers, volume was again significant ($p = 0.042$), as were shape and texture ($p = 0.014$ and $p = 0.004$, respectively). The full model was significantly better than both the model with shape and volume ($p = 0.013$) and the model with texture and volume ($p = 0.004$).

4. Discussion

We found that hippocampal volume, shape, and texture individually were all significant predictors of development of dementia. This is in line with previous results; MRI hippocampal volume is predictive of the development of AD in MCI subjects (Devanand et al., 2007; Jack et al., 1999) and in the same sample of the general population as used in this work (den Heijer et al., 2006). Hippocampal shape is also predictive of development of dementia in both MCI subjects (Costafreda et al., 2011; Ferrarini et al., 2009) and in the same sample of the general population as used in this work (Achterberg et al., 2014). Hippocampus texture has only been shown to be predictive of the development of dementia for subjects with MCI (Chincarini et al., 2011; Sørensen et al., 2016). Our results demonstrate that texture is also predictive at an earlier stage in a sample of the general population with subjects scanned up to 11 years before clinical dementia diagnosis.

Combination of MRI features performed in general better than individual features. All pairwise combinations of 2 feature types

Table 3
Significance of feature configurations in the different subgroups

	p-Value vs random						
	N	Volume	Shape	Texture	V + S	V + T	S + T	All
x < 3 years	72	*0.0001*	*0.0244*	*0.0147*	*0.0005*	*0.0000*	*0.0004*	*0.0000*
3 <= x < 6 years	60	**0.0519**	0.1205	0.1645	**0.0616**	*0.0087*	*0.0290*	*0.0199*
6 < x years	68	0.5332	**0.0825**	0.2233	**0.0687**	0.3015	0.1960	0.2266
Entire cohort	510	*0.0004*	*0.0001*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
No MCI	455	*0.0021*	*0.0036*	*0.0001*	*0.0003*	*0.0001*	*0.0001*	*0.0002*
No object cognitive complaints	424	*0.0023*	*0.0035*	*0.0003*	*0.0004*	*0.0002*	*0.0003*	*0.0004*
No subjective cognitive complaints	214	0.3858	0.4906	0.1654	0.3624	0.1268	**0.0569**	0.2090

Bold and italics indicate significance level. Italics indicates significance, and bold indicates trending toward significance.

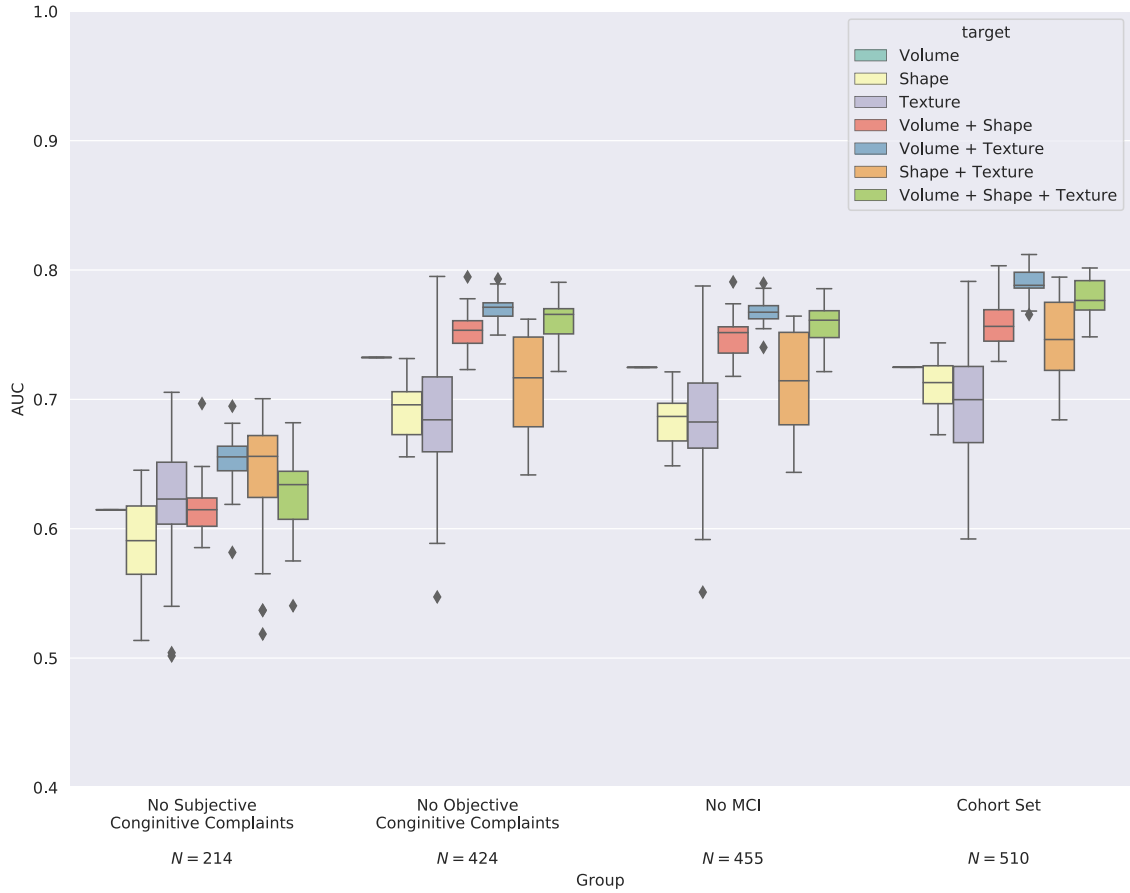

Fig. 3. Classification result for different feature configurations in subgroups of the cohort set where subjects with different symptoms of early dementia have been excluded. The box plot shows the AUCs for the 25 random permutations of the data. The entire cohort represents all subjects in the study and other groups are, from right to left, with increasing number of subjects excluded. ◆ represent samples that are considered outliers (based on a function of the interquartile range). Abbreviation: AUC, area under the receiver operating characteristic curve.

performed better than the best of the 2 individual features. However, combining all 3 types of features did not improve the predictive performance further. A possible explanation for this could be the substantial increase in dimensionality when combining shape and texture, which outweighs the extra information gained. Few studies have combined volume with shape or texture for dementia prediction, and no previous studies have combined all 3 MRI biomarkers for this purpose. In this work, we reproduced the results of Achterberg et al. (2014), which found that the combination of shape and volume led to improved prediction of conversion to dementia in the general, elderly population. Contrary to our results, Sørensen et al. (2016) found that the combination of texture and volume performed not significantly different from texture alone for prediction of MCI-to-AD conversion. This could be caused by a number of differences: in this work, we do not consider a specific dementia type (such as AD), the segmentation of the hippocampi were obtained using different methods, and we combined volume and texture by concatenation of feature vectors before nonlinear classification, which allowed for learning nonlinear relations between the 2 feature types and between individual texture features and volume. In Sørensen et al. (2016), an overall texture score from a classifier was linearly combined with volume.

Table 4
Logistic regression models with the MRI imaging biomarkers (posterior probabilities) as independent variables and development of dementia as the dependent variable

Parameter	Wald 95% confidence		
	Estimate	Limits	*p* value
Model 0: volume posterior (Log Likelihood: −146.79)			
Volume (linear)	−14.9472	(−33.468 to 3.574)	0.114
Model 1: volume and shape posteriors (Log Likelihood: −139.39)			
Volume (linear)	−20.6916	(−40.192 to −1.192)	0.038
Shape	5.7460	(2.776 to 8.716)	<0.001
Model 2: volume and texture posteriors (Log Likelihood: −140.45)			
Volume (linear)	−16.4103	(−35.229 to 2.408)	0.087
Texture	5.0624	(2.227 to 7.898)	<0.001
Model 3: volume, shape, and texture posteriors (Log Likelihood: −136.32)			
Volume (linear)	−20.3747	(−39.970 to −0.780)	0.042
Shape	4.6205	(1.462 to 7.779)	0.004
Texture	3.7617	(0.771 to 6.752)	0.014

Key: MRI, magnetic resonance imaging; MMSE, Mini–Mental State Examination. All models also include age, sex, and MMSE. Age and MMSE were significant in all models, and sex was not significant in any model.

Besides combining the different MRI feature types using an SVM classifier, we also combined them using a logistic regression model on the posterior probabilities obtained from the individual SVM classifiers. It is important to note that while the AUC of the SVM in cross-validation is a direct measure of predictive value, the regression model only indicates how well the posterior probabilities explain variations in the outcome. The regression analysis showed that adding shape or texture to volume improved the model significantly and that combining all features resulted in the best model. This is contrary to the direct combination of MRI features in the SVM classifier, where volume and texture combined performed similarly to all features combined. We believe the difference is in part because the SVM classifier can model nonlinear

relations between the features, whereas the regression only models the linear relation between the aggregated MRI biomarkers. This is supported by the result in Achterberg et al. (2014), where a regression model using the SVM posterior probability of volume and shape features combined resulted in a better model fit than a model that included the posterior probabilities for volume and shape as 2 separate covariates.

As our method uses texture measures for dementia prediction, the question might arise if the changes in texture are due to an increase in head motion. It might be possible that subjects close to dementia conversion exhibit more head motion than cognitively intact persons. To inspect this, we used the MRI SNR in the hippocampus as a proxy measure for head motion and corrected the regression models (as given in Table 4) for this. This showed that there was no relationship between dementia development and SNR (p-values ranging from 0.86 to 0.97) and that correcting for the SNR did not affect the performance of the texture or shape features. Only the p-values for volume increased slightly. This indicates that texture measures are not just an elaborate description of SNR but really contain different information.

The effect of developing future dementia on the MRI biomarkers was noticeable before the MCI stage, in a subgroup without any objective cognitive complaints, and—depending on the feature type—up to 6 years before dementia diagnosis. Some feature configurations even showed borderline significance in the group of 6–11 years before diagnosis. Performance generally decreased with increasing time to diagnosis, but it appeared that shape in isolation and in combinations with other feature types were less sensitive to this. It should be noted, however, that the subsets were very small for the 3 time-to-diagnosis stratifications (72, 60, and 68 subjects, with only 17, 15, and 16 cases) and the trends visible in Fig. 2 were not all significant. In the group without any subjective cognitive complaints, we did not find any significant predictive performance, but we would like to point out that the group, although the total size was 214, only contained 12 cases. Larger studies are needed to see if the lack of significance is because of the small sample size or the lack of signal.

A drawback of the proposed high dimensional classification approach is that it requires more training data for reliable estimation of parameters than a classification approach based on lower dimensional features such as hippocampal volumes. To gauge the reliability of our classifiers, we repeated the main experiment 25 times with different random permutations of the data, which led to a different data fold layout in the SVM classifiers' hyperparameter optimization. Some feature configurations resulted in less-stable classifiers than others. Texture, the individual feature type of highest dimensionality, performed most unstable of the individual features. We tested whether a simpler classifier would improve stability, and indeed a linear SVM resulted in more consistent performance across the 25 permutations. It seemed that the inclusion of volume in the nonlinear SVMs also had a stabilizing effect. This was probably because the features were scaled to obtain equal variance within the feature group, and as there are only 2 volumetric features, they are individually stronger than the individual shape and texture features.

A potential limitation of this study is that we considered all types of dementia jointly. Because dementia can have many different causes with corresponding different changes in the brain, it could prove more difficult to find and describe all these changes. Moreover, the biomarkers relating to the hippocampus may be less relevant for some dementia types (e.g., frontotemporal dementia). Many other MRI biomarker studies have therefore focused on dementia of the AD type. However, the differential diagnosis between different types of dementias is challenging and less reliable in the population study setting where we have no access to additional information such as positron emission tomography amyloid imaging or cerebrospinal fluid markers of abnormal protein buildup. We have therefore chosen to assess the value of the different MRI biomarkers directly for all-cause dementia prediction in the general population. In a previous study on the same cohort, restricting the analysis to only cases with clinical diagnosis of AD resulted in similar trends but a lower overall prediction performance when using volume and shape as features (Achterberg et al., 2014).

The cohort used in this study is a subset of the larger Rotterdam Study (Hofman et al., 2015; Ikram et al., 2015). Although the subjects invited were randomly sampled, the sample is not completely representative of the larger cohort. This is due to the selection criteria (no dementia, not blind, and not MRI contraindications) and the response rate. de Groot et al. (2000) investigated this (on a combination of this cohort and another cohort) and found "Compared with nonparticipants, the participants of the study were younger (mean age difference, 3.8 years; $p < 0.001$) and more educated (5% more subjects with university-level education; $p = 0.05$). Baseline MMSE scores were available for subjects originally invited from the Rotterdam Study and were higher in participants compared with nonparticipants (age- and sex-adjusted mean difference, 0.4 points; $p < 0.001$)." This shows that the cohort of the present study could be considered cognitively healthier than the entire Rotterdam Study. Although not completely representative, we still believe that the results on the cohort set give a good indication of what to expect in a general elderly population.

In conclusion, we have shown that hippocampal volume, texture, and shape are all predictive of future development of dementia in the general population. All MRI biomarkers showed significant predictive performance for subjects who were cognitively healthy, up to 3 years before dementia diagnosis. Combining the different MRI biomarkers improved the prediction; all combinations, except volume and shape, showed predictive performance in subjects without objective cognitive complaints and up to 6 years before dementia diagnosis.

Disclosure statement

The authors have to disclose that Wiro Niessen is cofounder, part-time Chief Scientific Officer, and stockholder of Quantib BV, Lauge Sørensen is an employee at Biomediq A/S, and Mads Nielsen is an employee at Biomediq A/S and a shareholder of Biomediq A/S. The remaining authors have no commercial or financial relationships that could be seen as potential conflicts of interest.

Acknowledgements

Funding for this research has been provided by the Netherlands Organisation for Scientific Research NWO, the Danish National Advanced Technology Foundation (project 034-2011-5, Early MRI diagnosis of Alzheimers Disease), and Eurostars (project 8234, MR Brain Image Quantification in Dementia).

References

Achterberg, H.C., van der Lijn, F., den Heijer, T., Vernooij, M.W., Ikram, M.A., Niessen, W.J., de Bruijne, M., 2014. Hippocampal shape is predictive for the development of dementia in a normal, elderly population. Hum. Brain Mapp. 35, 2359–2371.

Braak, H., Braak, E., 1997. Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiol. Aging 18, 351–357.

Cates, J., Fletcher, P.T., Whitaker, R., 2006. Entropy-based particle systems for shape correspondence. In: Mathematical Foundations of Computational Anatomy Workshop, MICCAI 2006, pp. 90–99.

Cates, J.E., Fletcher, P.T., Styner, M.A., Shenton, M.E., Whitaker, R.T., 2007. Shape modeling and analysis with entropy-based particle systems. In: Karssemeijer, N., Lelieveldt, B.P.F. (Eds.), IPMI. Vol. 4584 of Lecture Notes in Computer Science. Springer, Heidelberg, pp. 333–345.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: A Library for Support Vector Machines. Software available at. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A., Nobili, F., 2011. Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. NeuroImage 58, 469–480.
Clark, T.G., Altman, D.G., Stavola, B.L.D., 2002. Quantification of the completeness of follow-up. Lancet 359, 1309–1310.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learn. 20, 273–297.
Costafreda, S.G., Dinov, I.D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Wahlund, L.-O., Spenger, C., Toga, A.W., Lovestone, S., Simmons, A., 2011. Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. NeuroImage 56, 212–219.
de Bruijn, R.F., Bos, M.J., Portegies, M.L., Hofman, A., Franco, O.H., Koudstaal, P.J., Ikram, M.A., 2015. The potential for prevention of dementia across two decades: the prospective, population-based Rotterdam Study. BMC Med. 13, 132.
de Groot, J.C., de Leeuw, F.-E., Oudkerk, M., Van Gijn, J., Hofman, A., Jolles, J., Breteler, M.M., 2000. Cerebral white matter lesions and cognitive function: the rotterdam scan study. Ann. Neurol. 47, 145–151.
DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 837–845.
den Heijer, T., Geerlings, M.I., Hoebeek, F.E., Hofman, A., Koudstaal, P.J., Breteler, M.M.B., 2006. Use of hippocampal and amygdalar volumes on magnetic resonance imaging to predict dementia in cognitively intact elderly people. Arch. Gen. Psychiatry 63, 57–62.
Devanand, D., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G., Honig, L., Mayeux, R., Stern, Y., Tabert, M.H., de Leon, M.J., 2007. Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of Alzheimer disease. Neurology 68, 828–836.
Ferrarini, L., Frisoni, G.B., Pievani, M., Reiber, J.H.C., Ganzola, R., Milles, J., 2009. Morphological hippocampal markers for automated detection of Alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. J. Alzheimer's Dis. 17, 643–659.
Hofman, A., Brusselle, G.G.O., Murad, S.D., van Duijn, C.M., Franco, O.H., Goedegebure, A., Ikram, M.A., Klaver, C.C.W., Nijsten, T.E.C., Peeters, R.P., Stricker, B.H.C., Tiemeier, H.W., Uitterlinden, A.G., Vernooij, M.W., 2015. The Rotterdam Study: 2016 objectives and design update. Eur. J. Epidemiol. 30, 661–708.
Igel, C., Heidrich-Meisner, V., Glasmachers, T., 2008. Shark. J. Mach. Learn. Res. 9, 993–996.
Ikram, M.A., van der Lugt, A., Niessen, W.J., Koudstaal, P.J., Krestin, G.P., Hofman, A., Bos, D., Vernooij, M.W., 2015. The rotterdam scan study: design update 2016 and main findings. Eur. J. Epidemiol. 30, 1299–1315.
Ikram, M.A., Vrooman, H. a., Vernooij, M.W., den Heijer, T., Hofman, A., Niessen, W.J., van der Lugt, A., Koudstaal, P.J., Breteler, M.M.B., 2010. Brain tissue volumes in relation to cognitive function and risk of dementia. Neurobiol. Aging 31, 378–386.
Jaakkola, T., Diekhans, M., Haussler, D., 1999. Using the Fisher kernel method to detect remote protein homologies. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. AAAI Press, pp. 149–158.
Jack, C.R., Petersen, R.C., Xu, Y.C., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Waring, S.C., Tangalos, E.G., Kokmen, E., 1999. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 52, 1397–1403.
Li, S., Shi, F., Pu, F., Li, X., Jiang, T., Xie, S., Wang, Y., 2007. Hippocampal shape analysis of Alzheimer disease based on machine learning methods. AJNR Am. J. Neuroradiol 28, 1339–1345.
Magnotta, V.A., Friedman, L., BIRN, F., 2006. Measurement of signal-to-noise and contrast-to-noise in the fbirn multicenter imaging study. J. Digital Imaging 19, 140–147.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Machine Learn. Res. 12, 2825–2830.
Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Kokmen, E., Tangelos, E.G., 1997. Aging, memory, and mild cognitive impairment. Int. Psychogeriatrics 9, 65–69.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.
Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2016. Early detection of Alzheimer's disease using MRI hippocampal texture. Hum. Brain Mapp. 37, 1148–1161.
van der Lijn, F., den Heijer, T., Breteler, M.M.B., Niessen, W.J., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 43, 708–720.
Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J.C., Csernansky, J.G., Miller, M.I., 2007. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE Trans. Med. Imaging 26, 462–470.
West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C., 1994. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. Lancet 344, 769–772.
West, M.J., Kawas, C.H., Stewart, W.F., Rudow, G.L., Troncoso, J.C., 2004. Hippocampal neurons in pre-clinical Alzheimer's disease. Neurobiol. Aging 25, 1205–1212.
Whitaker, R.T., 2000. Reducing aliasing artifacts in iso-surfaces of binary volumes. In: IEEE Symposium on Volume Visualization and Graphics. IEEE Computer Society, Los Alamitos, CA, USA, pp. 23–32.