

# Applying Automated MR-Based Diagnostic Methods to the Memory Clinic: A Prospective Study

Stefan Klöppel<sup>a,b,c,d,\*</sup>, Jessica Peter<sup>b,c,d</sup>, Anna Ludl<sup>a</sup>, Anne Pilatus<sup>a</sup>, Sabrina Maier<sup>a</sup>, Irina Mader<sup>c</sup>, Bernhard Heimbach<sup>a</sup>, Lars Frings<sup>a,f</sup>, Karl Egger<sup>e</sup>, Juergen Dukart<sup>g,h</sup>, Matthias L. Schroeter<sup>h</sup>, Robert Perneczky<sup>i,j,k</sup>, Peter Häussermann<sup>l</sup>, Werner Vach<sup>m</sup>, Horst Urbach<sup>e</sup>, Stefan Teipel<sup>n</sup>, Michael Hüll<sup>a,o</sup>, Ahmed Abdulkadir<sup>b,p</sup> and for the Alzheimer's Disease Neuroimaging Initiative<sup>1</sup>

<sup>a</sup>Center of Geriatrics and Gerontology Freiburg, University Medical Center Freiburg, Freiburg, Germany

<sup>b</sup>Freiburg Brain Imaging, University Medical Center Freiburg, Germany

<sup>c</sup>Departments of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, University Medical Center Freiburg, Freiburg, Germany

<sup>d</sup>Department of Neurology, University Medical Center Freiburg, Freiburg, Germany

<sup>e</sup>Department of Neuroradiology, University Medical Center Freiburg, Freiburg, Germany

<sup>f</sup>Department of Nuclear Medicine, University Medical Center Freiburg, Freiburg, Germany

<sup>g</sup>F. Hoffmann-La Roche, pRED, Pharma Research and Early Development, DTA Neuroscience, Basel, Switzerland

<sup>h</sup>Max Planck Institute for Human Cognitive and Brain Sciences & Clinic for Cognitive Neurology, University of Leipzig, and German Consortium for Frontotemporal Lobar Degeneration, Ulm, Germany

<sup>i</sup>Neuroepidemiology and Ageing Research Unit, School of Public Health, Imperial College of Science, Technology and Medicine London, United Kingdom

<sup>j</sup>Cognitive Impairment and Dementia Services, Lakeside Mental Health Unit, West London Mental Health NHS Trust, London, UK

<sup>k</sup>Departments of Psychiatry and Psychotherapy, Technical University München, Germany

<sup>l</sup>Departments of Gerontopsychiatry und -psychotherapy, LVR Clinic Köln, Germany

<sup>m</sup>Center for Medical Biometry and Medical Informatics, University of Freiburg, Germany

<sup>n</sup>Departments of Psychosomatic Medicine, University of Rostock, and German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany

<sup>o</sup>Clinics for Geronto- and Neuropsychiatry, ZfP Emmendingen, Emmendingen, Germany

<sup>p</sup>Department of Computer Science and BIOS Centre for Biological Signaling Studies, University of Freiburg, Germany

Accepted 12 May 2015

**Abstract.** Several studies have demonstrated that fully automated pattern recognition methods applied to structural magnetic resonance imaging (MRI) aid in the diagnosis of dementia, but these conclusions are based on highly preselected samples that significantly differ from that seen in a dementia clinic. At a single dementia clinic, we evaluated the ability of a linear support

<sup>1</sup>Some data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [http://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf)

<sup>\*</sup>Correspondence to: Stefan Klöppel, MD, Center of Geriatrics and Gerontology Freiburg, Department of Psychiatry and Psychotherapy, University Medical Center Freiburg, Hauptstrasse 5, 79104 Freiburg, Germany. Tel.: +49 761 270 52960; Fax: +49 761 270 70890; E-mail: [stefan.kloeppe@uniklinik-freiburg.de](mailto:stefan.kloeppe@uniklinik-freiburg.de)

vector machine trained with completely unrelated data to differentiate between Alzheimer's disease (AD), frontotemporal dementia (FTD), Lewy body dementia, and healthy aging based on 3D-T1 weighted MRI data sets. Furthermore, we predicted progression to AD in subjects with mild cognitive impairment (MCI) at baseline and automatically quantified white matter hyperintensities from FLAIR-images. Separating additionally recruited healthy elderly from those with dementia was accurate with an area under the curve (AUC) of 0.97 (according to Fig. 4). Multi-class separation of patients with either AD or FTD from other included groups was good on the training set (AUC > 0.9) but substantially less accurate (AUC = 0.76 for AD, AUC = 0.78 for FTD) on 134 cases from the local clinic. Longitudinal data from 28 cases with MCI at baseline and appropriate follow-up data were available. The computer tool discriminated progressive from stable MCI with AUC = 0.73, compared to AUC = 0.80 for the training set. A relatively low accuracy by clinicians (AUC = 0.81) illustrates the difficulties of predicting conversion in this heterogeneous cohort. This first application of a MRI-based pattern recognition method to a routine sample demonstrates feasibility, but also illustrates that automated multi-class differential diagnoses have to be the focus of future methodological developments and application studies.

Keywords: Dementia diagnostics, machine learning, magnetic resonance imaging, prognosis, support vector machine

## INTRODUCTION

An accurate diagnosis of different types of dementia or the prediction of conversion to dementia from a state of mild cognitive impairment (MCI) is important in order to recruit homogeneous populations for diagnostic and treatment studies, clinical trials, and eventually also in the clinical routine. Accurate clinical diagnosis based on medical history, clinical assessment, neuropsychological tests, and laboratory assessment as originally recommended by the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA; [1]) and recently further developed [2] is the gold standard for clinical management. Cognitive testing and clinical examination necessary for differential diagnosing of dementia is, however, labor intensive. As part of the routine workup, imaging initially served to exclude secondary causes such as tumor, abscess, or hydrocephalus [56], but several studies demonstrated information gain from structural magnetic resonance imaging (MRI), positron emission tomography (PET), or computerized tomography scans in the context of dementia diagnostics either in combination with a neuropsychological assessment or in isolation [3–6]. So far, diagnostic use of imaging depends heavily on the quality and experience of the radiologist [7], which may be limited outside specialized centers where the majority of cases is diagnosed [8].

Compared to visual assessment, automated diagnostic methods based on MRI are more reproducible and have demonstrated a high accuracy in detecting a range of pathological conditions [9]. Automated MRI

based volumetry separates Alzheimer's disease (AD) or MCI from healthy aging [10–14], but also, clinically more challenging, separates between different types of dementia [11, 15], or between those with MCI who convert to dementia or remain stable [16–21]. Some studies on structural imaging already considered more than two diagnostic options or used probabilistic rather than categorical diagnostic labels [20, 22–24]. These supervised machine-learning approaches run on a standard PC and rely on a set of labelled training data—for example structural MRI and reliably established diagnostic label for each subject—to diagnose new cases in the absence of expert radiologists.

Although reported performance is encouraging, drawing conclusions on the usefulness of these automated methods in routine application is difficult for several reasons: i) previous studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [25] recruited a sample representative of the population for clinical trials rather than clinical routine [26]. They excluded patients with comorbidities associated with structural brain changes (e.g., stroke, Parkinson's disease, atrophy caused by alcohol abuse) or cognitive impairment unrelated to dementia (e.g., depression); ii) MRI scans with lower data quality were often excluded. While clinicians can read through some image artifacts [27], the effect on automated methods is largely unexplored; iii) With few exceptions [11, 28], studies have either reported cross-validation accuracy or have taken test and training data from the same source. For maximum usefulness, classification methods need to perform well even when training cases differ in characteristics such as co-morbidity, age, or scan-parameters from the clinical set to which the

scanner is applied; iv) Finally, studies on automated diagnosing have not considered the presence of multiple pathologies in the same individual despite their high frequency. White matter hyperintensities (WMH) most often related to vascular pathology, co-occur in a range of neurodegenerative dementias [29]. WMH extent and location affect clinical phenotype [30], but the occurrence of multiple neurodegenerative pathologies is also frequently observed [31, 32].

To overcome these limitations and assess a realistic clinical scenario, we implemented a prospective study in a single memory clinic with liberal inclusion criteria. We applied automatic tools to determine to what extent gold standard clinical diagnosis could be reflected by automated MRI based classification. High level multidisciplinary diagnosis is available in expert centers, such as a dedicated memory clinic, but not part of routine care. Therefore, an automated method that can serve to assign patients to clinical categories with similar accuracy than clinical gold standard diagnosis would be a major step forward in clinical care of dementia. To this end, we extracted volumetric information from 3D-T1 weighted data sets to separate different neurodegenerative conditions. WMH were quantified on Fluid Attenuated Inversion Recovery (FLAIR) images [33–36]. In addition to computing receiver-operator characteristics (ROC) curves, we examined positive (PPV) and negative predictive values (NPV) and therewith the ability to reach a diagnostic decision with high reliability in a substantial fraction of patients or the ability to rule out some possible diagnoses in a reliable manner to reduce the further diagnostic process compared to the current work up. We separately recruited a small number of healthy controls (HC) to evaluate the performance of the classifier more completely.

## MATERIAL AND METHODS

All referrals to the Freiburg dementia clinic with reported progressive decline of cognitive functioning and age above 40 were screened between June 2011 and June 2014 (total 1,303 cases). Inclusion criteria were kept liberal for a good representation of the typical population of a memory clinic. Therefore, the presence of disorders affecting brain structure (e.g., stroke, past alcohol-abuse, depression, and trauma) did not lead to exclusion unless deemed the exclusive cause for present cognitive symptoms. Routine clinical workup did not change for the study and included clinical evaluation (neurological examination, medical

history, etc.) with input from relatives when available, a neuropsychological assessment and structural imaging, sometimes complemented by laboratory tests, such as blood tests. PET, single photon emission computed tomography imaging, and cerebrospinal fluid (CSF) analyses were arranged as part of the routine when diagnostic confidence remained low after assessing medical history, clinical examination, neuropsychological testing, and MRI. Keeping the routine workup unchanged also meant that subjects without cognitive impairment or referrals with a previous CT or MRT examination did not enter the study as no additional MRI was required. Cognitive performance was assessed using the CERAD battery [37], sometimes extended by trail-making tests A and B [38] as well as tests of aphasia. The clinical dementia rating scale (CDR) was used to evaluate severity of impairment [39]. All MCI patients had a CDR score of 0.5 and no impairment in activities of daily living (ADL). For patients with dementia, we focused on mild to moderate dementia stages including those with impairment in ADL and a CDR of  $\geq 0.5$  and  $\leq 2$  as separating different types of dementia becomes difficult in late stages. 18 HC were recruited through advertisements and received the same clinical assessment and imaging protocol. Normal cognitive performance was confirmed using the sensitive Montreal Cognitive Assessment (MOCA) [40]. Informed written consent was obtained from all participants and the study was approved by local ethics committee and internationally registered (ID: DRKS00003199).

### *Gold standard diagnosis*

Gold standard for differential diagnosis of dementia, in case of dementia at baseline, was established based on the baseline assessment in a panel decision and according to established diagnostic criteria including a standard visual evaluation of MRI images and all other available data [2, 41, 42].

For subjects fulfilling criteria for MCI at baseline [43], follow-up clinical assessment was considered the diagnostic gold standard: Prediction of conversion from MCI to dementia based on data from baseline was expected to be easier the earlier after baseline the conversion happened, i.e., the closer to onset the patient was at baseline. On the other hand, the reliability of a diagnosis of a subject with stable MCI based on data from the baseline was assumed to increase with the follow-up interval. To make use of the available follow-up data, we defined a minimal follow-up period of 12 months but used longer follow-up data for the grouping

when available: MCI to dementia converters fulfilled criteria of dementia within two years of the baseline visit. Subjects that converted later than twelve months were excluded. Subjects classified as stable MCI had to remain stable over their full observation period of at least 12 months. Consequently, subjects that were observed over a period of less than twelve months and remained stable during this period were excluded.

#### Questionnaire (pre- and post clinical scoring)

To obtain further insights in the diagnosis process and the potential added value of MRI and thereof derived automated method, the clinicians were asked to complete a questionnaire for each participant. When dementia was determined clinically at baseline, the specific type had to be indicated. For those with MCI at baseline, clinicians had to indicate if they

expected conversion to dementia within 12 months. A three-point scale allowed clinicians to rate their diagnostic confidence (i.e., low: <70%, intermediate: 70–90%, high: >90%). Such a coarse discrete rating was preferred over a continuous rating of certainties as a continuous scale would be pseudo-accurate and because a categorization is clinically useful. To compare the automated method to clinicians in their accuracy of prediction of conversion, we used a six-point scale predicting 12-months conversion with confidence levels ranging from <10% (stability expected with high confidence), 10–30%, 30–50%, 50–70%, 70–90%, up to >90% (conversion expected with high confidence). Clinicians completed the questionnaire twice, first after clinical interview and neuropsychological testing (i.e., pre-MRI clinical scoring), and second after they had access to the MRI scan and the report from an expert neuroradiologist (i.e., post-MRI clinical scoring).

Table 1

Included subjects in training or test set. Please note: Provided references may refer to slightly different samples. \*Mini-Mental State Examination (MMSE) scores were not available for all subjects. Last rows includes relative volumes of white matter hyper intensities (WMH) for the frontal lobe, the temporal lobe, and the whole brain

	HC	AD	FTD	LBD	MCI stable	MCI converter
ADNI [25]						
Age	74.7 ± 5.8	75.2 ± 7.8			73.4 ± 7.8	74.5 ± 7.0
f/m	185/177	129/149			104/170	51/65
MMSE	29.0 ± 1.2	23.1 ± 2.0			27.7 ± 1.8	26.6 ± 1.9
AIBL [44]						
Age	74.5 ± 7.5	72.6 ± 7.9				
f/m	81/60	18/13				
MMSE	N.A.	N.A.				
Train3 [24]						
Age	55.7 ± 9.2	61.6 ± 6.4	58.6 ± 6.4			
f/m	8/15	11/9	5/5			
MMSE	29.0 ± 1.3	22.8 ± 3.8	24.2 ± 4.0			
Train4 [64]						
Age		72.7 ± 7.1	64.9 ± 8.9			
f/m		5/1	8/9			
MMSE		23.8 ± 1.5	24.5 ± 3.8			
Train5 [65]						
Age				73.4 ± 4.6		
f/m				6/10		
MMSE				21.7 ± 4.5		
Train6 [66]						
Age	67.9 ± 7.6	70.4 ± 8.3	62.4 ± 5.4	74.4 ± 4.4		
f/m	25/35	15/11	6/6	1/6		
MMSE*	27.6 ± 6.2	21.8 ± 4.1	24.9 ± 4.6	25.1 ± 1.7		
Test						
Age	70.3 ± 8.3	76.1 ± 7.1	66.0 ± 6.3	72.9 ± 3.3	72.8 ± 6.9	74.3 ± 6.7
f/m	9/9	76/46	7/5	1/3	8/8	6/6
MMSE	29.0 ± 1.7	20.4 ± 4.1	21.8 ± 3.5	21.0 ± 3.6	26.1 ± 2.9	23.5 ± 2.4
FLAIR available	(n = 12, 66%)	(n = 70, 57%)	(n = 3, 25%)	(n = 1, 25%)	(n = 7, 43%)	(n = 6, 50%)
temporal	5.0 ± 2.3	5.9 ± 2.5	5.6 ± 4.7	4.9 ± N.A.	5.8 ± 3.0	5.2 ± 1.7
frontal	6.4 ± 2.1	8.2 ± 4.1	4.3 ± 2.9	6.7 ± N.A.	9.0 ± 6.2	6.2 ± 2.2
#total WM	6.3 ± 1.8	7.3 ± 2.8	4.7 ± 3.3	5.8 ± N.A.	7.4 ± 3.5	6.1 ± 1.8

Table 2

Scanning parameters of training data. TR, repetition time; TE, echo time, FA, flip angle; RES, resolution of volume in pixels; VX, voxel size in mm; FS, magnetic field strength in Tesla

Dataset	Sequence	# scanners	TR [ms]	TE [ms]	FA	RES [ms]	VX	FS
ADNI	MP-RAGE	>50	2,400/2,300 (typical)	~3	8°/9° (typical)	192 × 192 × 160	2400/2300 (typical)	1.5/3.0
AIBL	MP-RAGE	2	2,300	2.98	9°	240 × 256 × 160	1 × 1 × 1.2 mm	3.0T
Train 3	MDEFT or MP-RAGE	2	1,300	3.93	n.a.	256 × 256 × 128	1 × 1 × 1.5 mm	3.0T
Train 4	MP-RAGE	1	2,200	2.15	12°		1 × 1 × 1 mm	3.0 T
Train 5	MP-RAGE	1	1,100	4.3		256 × 256 × 160	1 × 1 × 1 mm	1.5 T
Train 6	MP-RAGE	1	2,500	4.82	7°	256 × 256 × 192	1 × 1 × 1 mm	3.0

### MRI training and testing data

We combined training data from multiple sources to train the predictive computer models. This resulted in a fairly large training sample of more than 1,600 instances, and we assumed that combining data from multiple centers would help to learn models that are more robust to systematic inter-scanner effects. As indicated in Table 1, there was a large imbalance in the number of training instances per class. Subjects with AD and elderly controls were most abundant, while other types of dementia were scarcer. While elderly controls and AD patients were acquired on many scanners, cases diagnosed with non-AD dementia were acquired on very few scanners. The strong correlation of some diagnosis with acquisition scanner/sequence implied potential decrease in classification accuracy as the classifier could learn to separate scanners rather than disease [44].

The largest number of publicly available scans was from the ADNI-study (<http://www.adni-info.org>). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a \$60 million, 5-year public-private partnership. The public data base provides extended longitudinal data including neuropsychological tests, genetic data, imaging, and CSF data. The second publicly available data set was obtained from the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) [45]. The AIBL MRI data was acquired at only two sites, with a similar protocol as the ADNI. Further, we included data from four memory clinics to study the ability of separation of different types of dementia as detailed in Tables 1 and 2. We included AD, frontotemporal dementia (FTD, without distinguishing subtypes) and Lewy body dementia (LBD) as well as healthy elderly.

Of note, diagnoses in the training set typically relied on clinical criteria with biomarker confirmation (e.g., PET, CSF) only in a subset.

Training data for the predicting of conversion from MCI was derived from ADNI only. The respective subjects were classified with MCI at baseline and were sub-divided into two groups of patients. As with the test data from the local clinic, we considered the available follow-up intervals which were longer than those from the local set: one group was formed by subjects who remained MCI in all follow-ups and their last follow-up was at least 360 days after the baseline check, and the other converting to AD within 720 days after the baseline examination. Individuals whose disease status changed more than once or fell in neither category were excluded from the training set.

Prospective test data were acquired locally on a 3T Siemens system using a 32-channel head coil. 3D-FLAIR and T1-weighted images were acquired with a 1 mm isotropic resolution. Parameters for the MPRAGE sequence included: TR 1440 ms and TE = 2.15 ms with a parallel imaging protocol (<4 min). FLAIR was performed in just under 8 min with TR = 5000 ms and TE = 388 ms. Scan quality in the test set was graded as good quality (no artifacts), intermediate quality (some movement artifacts and/or reduced contrast without affecting visual assessment of scans) and low quality (substantial artifacts reducing visual interpretability of scans). None of the training data were acquired on the same scanner as the test data.

### Image pre-processing and adjustment for covariates

The same images that were inspected by the neuroradiologists were used for the automated diagnosis pipeline. To this end, all images were pre-processed as detailed below in order to obtain data suitable for the automated algorithm. We employed pre-processing

as in previous work [46] as it performed among the best in terms of area under the curve [28]. Effects of age, gender, and total intracranial volume were removed. Instead of adjusting for covariates by kernel linear regression, we adjusted using Gaussian process regression with a non-linear covariance function as in previous work [44].

T1-weighted images of all data sets were pre-processed using the VBM8 Toolbox (<http://dbm.neuro.uni-jena.de/vbm/>). This essentially performs tissue segmentation into gray matter (GM), white matter, and CSF, and a subsequent normalization to a standard stereotactic Montreal Neurological Institute (MNI) space formed by 550 healthy subjects. Modulation was applied after affine spatial transformation in order to correct local volume changes by the non-linear spatial transformation procedure and to preserve the amount of GM. This processing step computed a map of GM tissue probability for every subject that was spatially aligned across subjects. Based on these maps, we computed three dot-product matrices that captured the similarities between pairs of GM maps. The first was computed from the raw modulated GM maps, the second from the smoothed (8 mm FWHM) modulated GM maps, and the third of the region-wise weighted averages of the modulated GM maps. The weight for the last dot-product matrix were taken from the LONI probabilistic brain atlas [47], as the latter was reported to perform best in prediction of the conversion from MCI to AD compared to two other atlases [48]. To reduce confounding effects of total intracranial volume, age, and gender, we adjusted each dot-product kernel using Gaussian process regression as in previous work [44]. The methodological details are outlined in the Supplementary Methods section. For the classification, the three dot-product kernels were normalized such that the mean of each diagonal was one and then the kernels were summed up. Data from the local memory clinic (i.e., the test data), from which the performance is reported, neither entered the estimation of regression parameters nor the training of the SVM.

#### *Quantification of white matter changes*

We rated total volume of WMH in temporal and frontal lobe using FLAIR images from scans of the local memory clinic. WMH lesions load in these two lobes has been shown to affect cognitive speed and memory [49]. To meet the requirements of the clinical setting for fast and automated detection of WMH, we used a simpler version of our previously proposed processing pipeline [36] and applied a threshold to the

standardized FLAIR image [50]. We first generated a native space white matter mask using the unified segmentation algorithm [51] as implemented in SPM8. Second, FLAIR images were normalized to identical interquartile ranges. Voxel with values  $>0.5$  in white matter mask and FLAIR image were considered as WMH in a binary image. This step identified bright regions in the FLAIR image but reduced the number of false positives by a restriction to voxels with an at least 50% probability of belonging to the white matter segment. We co-registered the Talairach Daemon Lobe atlas [52] to each subjects structural images (i.e., T1 and FLAIR) and computed the relative amount of detected WMH with respect to the total size of temporal and frontal lobe, respectively. As a plausibility check, WMH were correlated with age across all prospective subjects of the local memory clinic with available FLAIR image as a strong positive correlation could be expected from the literature [53].

#### *SVM-based predicting and results reporting*

For all analyses, the same adjusted kernel was used for a soft-margin *C*-Support Vector Machine (SVM) [54] classification. All models were trained using *libsvm* ([55], <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>; version 3.18). For all learning tasks, the cost parameter *C* was selected among the candidates  $C = \{\exp(-4), \exp(-2), \dots, \exp(2)\}$  such that the cross-validation performance on the training set was highest.

SVMs are supervised binary classifiers, meaning that they learn a decision function based on training examples that belong to two classes. The grouping is based on a decision value, which codes the distance to a separating plane. Values close to zero indicate uncertain decisions. Binary classification is appropriate for the classification of stable MCI versus progressive MCI and, e.g., the classification of AD versus FTD. In contrast, a four class SVM (i.e., HC; AD; FTD; LBD) was used for the differential diagnosis of dementia. The multi-class SVM implementation in *libsvm* employs a one-versus-one classification as the algorithm internally computes  $(m \cdot (m-1))/2$ , where *m* is the number of classes. In case of four classes, this results in six decision values. To simplify the performance analyses, these six values were combined into (pseudo-) probabilities according to [56]. This results in one probability value per class. For example, a subject could get 78% AD, 8% FTD, 10% HC, and 4% LBD, in which case the probability for AD was highest and picked as discrete prediction. Note, that the probabilities add up to 100%. Posterior probabilities for the four possible

diagnostic labels predicted by the multi-class classifier were used for the ROC curves (Fig. 4) to plot the index class against the combination of all other classes and radar plots to illustrate the distribution of class specific probabilities. Thus, although the ROC analysis is intrinsically binary, the data presented in Fig. 4 were obtained from the outputs from a multi-class classifier. The proportion of WMH relative to the volume of frontal or temporal lobe formed two extra axes in the radar plot with the case with the maximum volume of WMH in the whole sample scaled to the maximum of the axis. Following discussions with clinicians we choose to use radar plots to display the probability profile for each individual visually. These plots provide visual patterns for the typical cases although the ordering of diagnostic entities could misleadingly indicate that some diagnosis would be more similar to each other. In addition, as probabilities sum to one (i.e., 100%), the presence of two pathologies would lead to relatively low probabilities for both while high probabilities for both would be more straightforward to interpret. The analysis of the differential diagnosis using a multi-class classifier is complemented by a binary distinction between AD and FTD, as the distinction between different types of dementia is often difficult and because relatively many cases per group were available in the test set. For that classifier, we report separate ROC when excluding cases with low scan quality ( $n=24$ ) or coexisting cerebral pathology ( $n=34$ ).

#### Diagnostic usefulness and comparison to clinicians

We were interested in the ability of the classifier to assign a significant fraction of the population with a sufficiently large confidence to the correct class. To rate the usefulness of the computer tool, we aimed to identify the cut point for the probability output which

results in a PPV above 0.9 or in a NPV above 0.95. For computing PPV and NPV, we used the observed prevalence in the test data sets. Based on internal *a priori* panel discussions, we regarded the computer tool as clinical useful, if at least one of the potential diagnoses could be given above the PPV threshold for 15% of all patients. Alternatively, the tool was deemed useful if at least one of the four possible diagnoses could be excluded for 25% of patients. To assess the usefulness, we plotted NPV and PPV as function of the probabilistic classifier output.

To assess the necessity for an additional diagnostic tool, we report the proportion of cases for which clinicians rated their own diagnostic certainty as low (i.e., <70%) at pre- and post-MRI clinical scoring. We compared the performance achieved by clinicians at post-MRI to the predictions of the SVM in respect to the clinical gold standard defined by follow-up examinations. In addition, we computed ROC and the area under the ROC curve (AUC) for the correct identification of MCI converters by comparing the probabilistic outcome of the SVM [57] against the diagnostic certainty of clinicians on the six-point scale. No comparisons between clinicians and the computer tool was attempted for differential diagnosis of dementia as clinician's rating at post-MRI was no longer independent from the gold standard diagnoses while a comparison with pre-MRI clinical scoring was deemed uninformative as both decisions would be based on completely different information.

## RESULTS

### Screening

A total of 1,303 were screened for this study over an interval of three years. 961 cases did not enter this study because clinical interview and neuropsycholog-

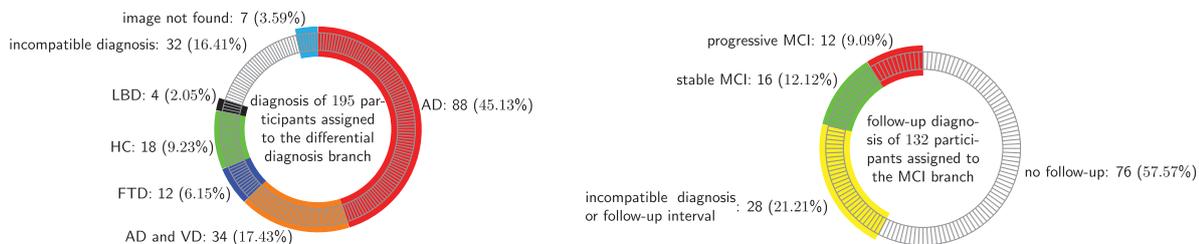


Fig. 1. Distribution of patients entering the memory clinic and their inclusion in the differential diagnosing of dementia (left panel) and predicting of MCI conversion (right panel). LBD, Lewy body dementia; FTD, frontotemporal dementia; AD, Alzheimer's disease; HC, healthy controls; VaD, vascular dementia; MCI, mild cognitive impairment.

ical testing indicated no need for MRI and in 10 cases, no image was taken (e.g., patient did not show up). Imaging was deemed unnecessary when it had already been done recently, when subjects were cognitively normal, or when impairment could be fully attributed to a non-neurodegenerative cause (e.g., stroke, severe depression). Cases with very typical medical history and cognitive profile tended to receive a CT rather than MRI scan and were thus unavailable to this study. In line with the German guidelines for diagnostics in MCI and dementia [58], individuals with MCI did not always receive imaging. Three hundred and forty-two (26.8% of 1303) subjects were referred to the Freiburg dementia clinic with reported progressive decline of cognitive function, age above 40 and the request for MRI. Of those, 327 entered one of the two arms of the study (Fig. 1). Therefore, fifteen were excluded from the study after being referred to the dementia clinic. Ten subjects were scheduled for a scan but were not scanned (e.g., subjects did not show up, or could not be scanned because of a metallic implant), and five subjects were not assigned to one of the study arms because of other reasons (e.g., depression). Visual inspection indicated an accurate detection of WMH. Across all included subjects who also had a usable FLAIR image, the percentage of WMH in the frontal lobe ( $p < 0.001$ ) and temporal lobe ( $p = 0.019$ ) was positively correlated with age. On average, approximately 5% of the voxels of the temporal and frontal lobe volumes were classified as hyperintense (see Table 1, bottom row, for statistics grouped by diagnosis).

#### Prediction of conversion to dementia

Of all 132 scans that entered the MCI study branch, a total of 28 cases (16 stable) with 12-months follow-up gold standard diagnosis were available for final analyses (Table 1 and Fig. 1, lower panel). A large proportion of individuals had to be excluded because insufficient follow-up data were available ( $n = 76$ ). Two cases converted after 720 days, 8 cases converted within less than 90 days, and four cases had incompatible diagnosis, such as dementia of unknown origin or normal pressure hydrocephalus. Thirteen cases remained stable but were observed for less than 360 days and therefore excluded. Of note, one subject converted from MCI to FTD and was not included in order to focus on conversion from MCI to dementia due to AD. In only 10% of the cases, clinicians were very certain (>90% certainty) at pre-MRI clinical scoring whether a subject would progress to dementia or remain stable. That figure rose to 29% of cases after MRI (Fig. 2).

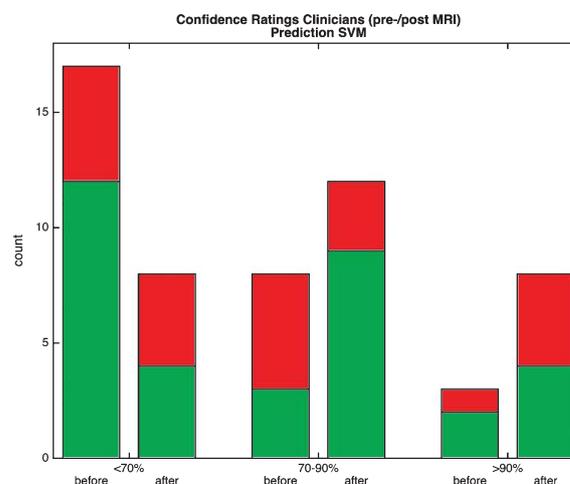


Fig. 2. Histograms showing increased levels of diagnostic confidence for the prediction of conversion from MCI (x-axis) by clinicians after learning about the MRI-results. In addition, a separation into cases correctly (green) or incorrectly (red) predicted by the SVM indicates no association between the diagnostic confidence of clinicians and the accuracy of the SVM.

Figure 2 indicates that there was no association between the confidence rating of clinicians and the correctness of the SVM. The predictive ratings by clinicians were most accurate when they rated their diagnostic confidence as high irrespective of pre- or post-MRI clinical scoring. The AUC was highest for the clinicians (0.81), followed by 0.80 for the training set using SVM. The AUC of the test set based on the SVM decision value was 0.73. If the default threshold from the SVM was taken to perform a classification, sensitivity of detecting progressive MCI was 0.25 and specificity was 0.93. ROC-curves as well as plots of PPV and NPV for the SVM are depicted in Fig. 3. PPV and NPV were constantly below 90% and 95% respectively and thus did not indicate a clinically useful performance.

#### Differential diagnostics of dementia

A total of 195 subjects entered the study branch dementia (Fig. 1, upper panel). Thirty-two subjects were excluded because no or no compatible gold standard diagnosis was available (e.g., dementia of unknown origin, primary progressive aphasia without further differentiation, progressive supranuclear palsy, Balint-Holmes-Syndrom in posterior cortical atrophy). A demographic and clinical evaluation of test subjects with dementia is reported in Table 1. Clinicians were certain with their diagnosing in 42% of cases at pre-MRI clinical scoring. The figure rose to 80%

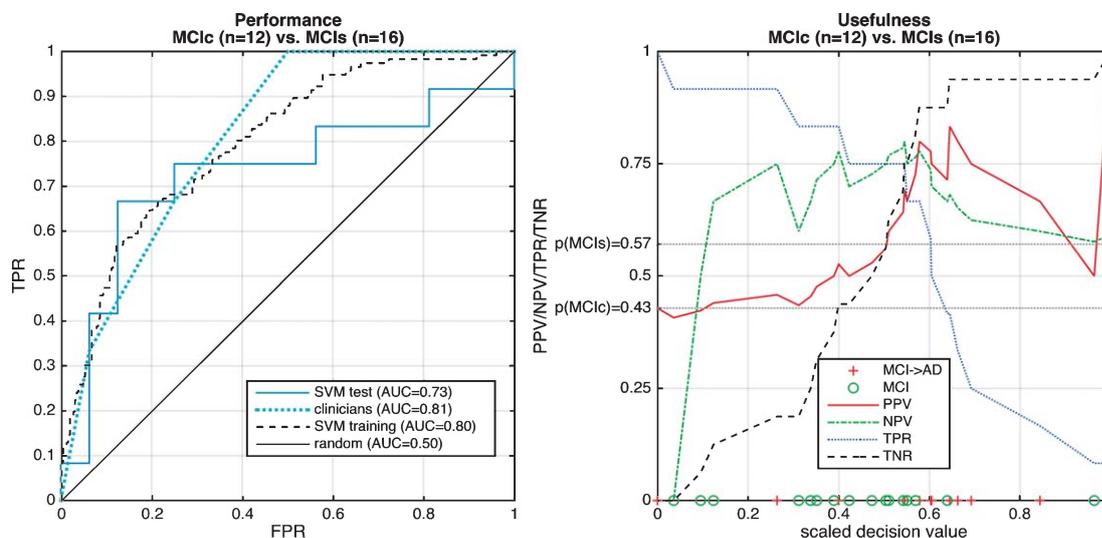


Fig. 3. Separating stable (MCIs) from those converting to dementia (MCIc). Left: ROC curve for different levels of diagnostic confidence (clinicians) and decision values (SVM). The cross-validated SVM performance on train set (dashed black line), test set (solid blue line) and performance by clinicians at post-MRI (dotted light blue) is shown. FPR, false positive rate; TPR, true positive rate. Right: True positive (TPR) and negative rate (TNR) together with positive (PPV) and negative predictive value (NPV). Markers on the x-axis indicate individual cases: green circles: MCIs; red crosses: MCIc.  $p(\text{MCIs})$  and  $p(\text{MCIc})$  indicate the fraction of stable and progressive MCI subjects in the sample, respectively.

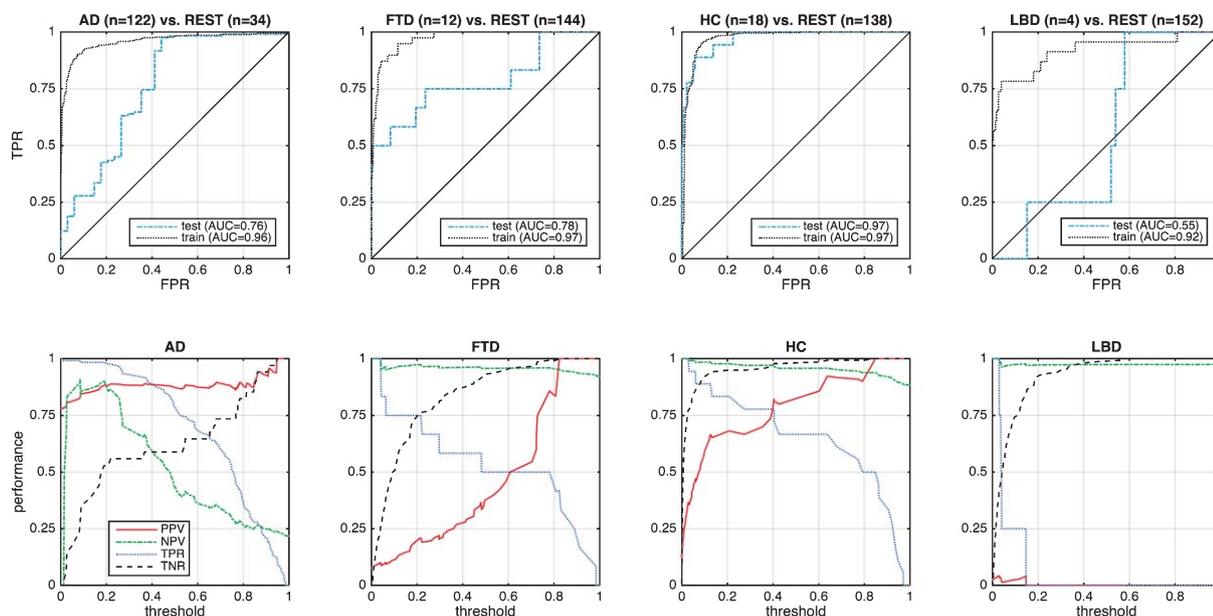


Fig. 4. Performance of multiclass differential diagnosis of dementia. The top row displays the ROC curve of each class versus the rest. Dotted black line and solid blue line indicate cross-validated training using cross-validation and test performance respectively. The bottom row shows several performance measure such as positive predictive value (PPV), negative predictive value (NPV), true positive rate (TPR), and true negative rate (TNR). See main text for AUCs of the training set.

at post-MRI clinical scoring. Performance of differential diagnosis is shown in Figs. 4–6. Separation of controls from rest was achieved with high accuracy

(AUC = 0.97). Classification of AD and FTD against all other entities were above chance level (AUC = 0.76 for AD and AUC = 0.78 for FTD; Fig. 4)

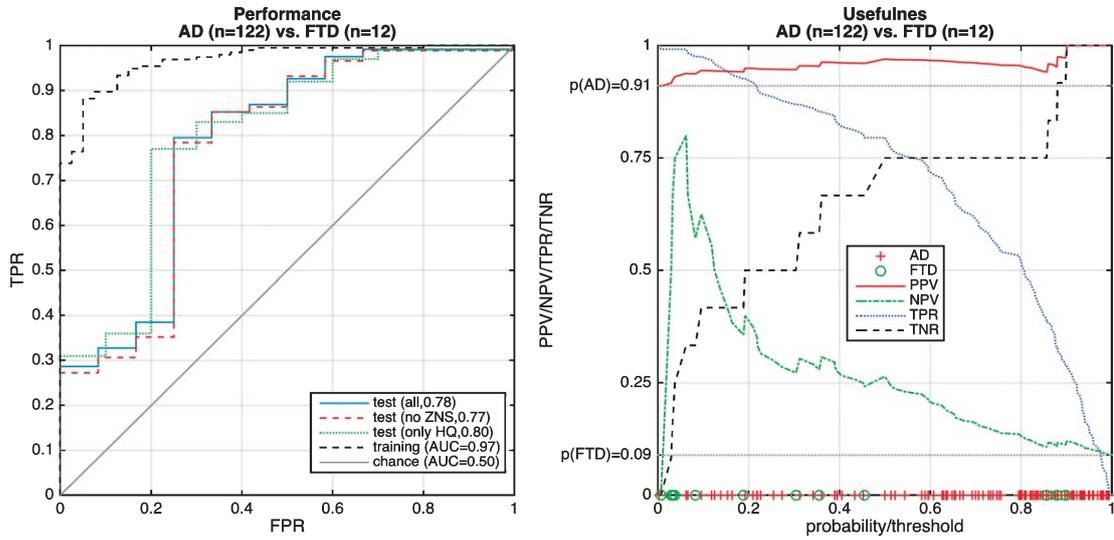


Fig. 5. Performance of differential diagnosis of FTD versus AD. Left: ROC curve, where the dashed black line indicates the cross-validated result from the training data and the solid blue line the test result. Red and green dashed lines illustrate the performance when cases are restricted to those with high quality (HQ) or cases without comorbid brain disorders (no comorb). Right: Indication of usefulness in terms of PPV and NPV. Markers on the x-axis indicate FTD (green circles) and AD (red crosses).

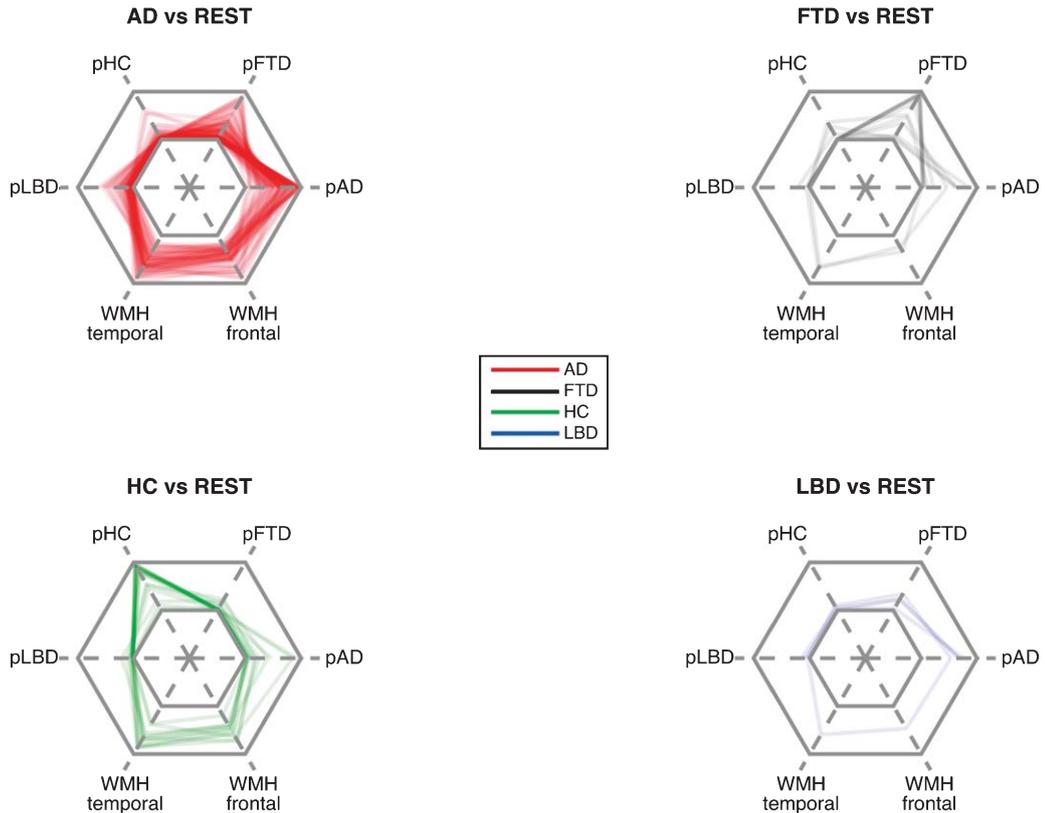


Fig. 6. Radar plot illustrating the posterior probability for each diagnostic class and the proportion of white matter hyperintensities (WMH). Center hexagon indicates minimum diagnostic probability or WMH load. Each line represents one case. All cases with dementia are shown in each subplot but with intransparent lines for a different class to aid visualization. Lines representing cases without available FLAIR imaging do not show values for the WMH. pHC, probability of healthy controls; pFTD, probability of FTD; pAD, probability of AD; pLBD, probability of LBD; WMH temporal, proportion of WMH in temporal lobe; WMH frontal, proportion of WMH in frontal lobe.

the posterior probability for AD and FTD, respectively. Performances on the training sets were all  $>0.9$ . Specifically AUCs were 0.96, 0.97, 0.97, and 0.92 for AD, FTD, HC, and LBD versus REST, respectively. As an example, a binary classifier separating AD from FTD (Fig. 5) revealed PPV constantly above 90% due to the highly unbalanced groups. Of note, the PPV increased with the decision value indicating additional diagnostic certainty from the classifier output. In contrast, NPV was always low. Excluding cases with low scan quality, additional CNS pathology did not substantially alter the results. Only four cases with LBD were included in the study, which challenges a meaningful interpretation of the classifiers ability to separate LBD from other types of dementia.

Radar plots (Fig. 6) show the combination of diagnostic probabilities for each alternative diagnosis together with the volume of WMH. Cases with AD frequently showed a high probability for AD and low probabilities for the other groups. Volumes of WMH were often relatively high. A subset of AD cases showed high probabilities for FTD and another for LBD. Of note, as the sum of probability values for each case always adds to one, a high probability for AD automatically means a low probability for the other groups. The majority of HC were correctly assigned.

## DISCUSSION

We set out to explore the clinical usefulness of automated diagnosing in a university based memory clinic with liberal inclusion criteria and when either MCI converters or the type of current dementia had to be identified.

Clinicians and SVM achieved a similar performance for the detection of MCI converters (Fig. 3). Encouragingly, performance of the SVM did not differ substantially between test and train set arguing against severe overfitting. Overall, the achieved prediction accuracy of 65% on the test set was low but a balanced accuracy of 77% ( $AUC = 0.8$ ) on the train set was comparable to existing studies typically reporting values close to or above 80% [59–61]. On the test set PPV and NPV were never above the predefined thresholds ( $PPV >0.9$ ;  $NPV >0.95$ ). As detailed below, it is likely that the MCI cases recruited into the study were particularly difficult to diagnose which may explain this relatively low performance on the test set.

In the multi-class differential-diagnosis model, HC were well separated from all neurodegenerative diseases. The multi-class classifier achieved AUCs

between 0.92 and 0.97 on the training set when separating one class from the rest, despite combining data from different sources and not explicitly modeling the systematic variance between the sites. This confirms the validity of the concept of using computer-aided diagnosis tools. However, this result was expected, since all MRI data sets were known to contain significant and detectable differences between diagnostic groups. Interestingly, performance separating HC from cases with neurodegeneration was equally good for training and test set which underlines the robustness of this separation. On the other hand, separating AD from all other groups was relatively inaccurate and this was true even for the binary separation between AD and FTD (Fig. 5). Fig. 5 also indicates that eliminating cases with low image quality or with additional structural brain changes did not further increase classification accuracy. At first sight, the observed drop between train and test set could indicate overfitting. While this possibility cannot be fully excluded, we consider it relatively unlikely given the heterogeneity of the training data. Systematic differences between training and test set in terms of recruitment strategy or the use of parallel imaging sequences for the test data only are likely explanations. We did not use WMH lesion load to define a separate diagnostic category (i.e., vascular dementia) but quantified it in addition to atrophy for each case. Visual inspection revealed good detection accuracy with around 5% of the volume of temporal and frontal lobe identified to contain WMH. A direct comparison to existing studies is difficult as those typically report WMH load across the whole brain. For that, Burton and colleagues suggested to define a percentage over 4.5% as an indication of severe lesion load [49].

Radar plots (Fig. 6) indicate that among cases with AD, subsets showed a high likelihood of belonging to the group of FTD or LBD cases. Relatively high probabilities for both could indicate the presence of two different pathologies in the same individual or alternatively, a misdiagnosing. We have no means of separating both possibilities with certainty. The co-occurrence of FTD and AD in the same individual is relatively rare except for the logopenic subtype but high for AD and LBD [31, 62].

Besides reporting ROC curves and radar plots, we also aimed to evaluate the need for an additional diagnostic method and its ability to make decisions with a sufficient NPV and PPV. Confidence ratings by clinicians indicated the need for a further diagnostic tool primarily for the detection of MCI converters (Fig. 2) in line with previous work [4]. This need is likely

to be higher outside dementia centers where clinicians and radiologists typically have less experience with dementia diagnostics and where less detailed neuropsychological test results are available.

As an additional aspect of clinical usefulness, we examined PPV and NPV. For the binary separation of AD from FTD (Fig. 5), PPV was constantly above 90%. This is most likely due to the highly imbalanced groups (AD:  $n = 122$ , 91%; FTD:  $n = 12$ , 9%) but a further improvement of the PPV with the SVM-based decision value indicates some added value. Looking at the multi-class classification, rare conditions such as FTD and LBD showed the expected high NPV. On the other hand NPV and PPV for the separation of HC from all other groups were high but are clinically less difficult.

As we did not change routine clinical work in the memory clinic, the final sample consisted of less than 30% of cases that were initially referred. Most importantly, subjects could enter the study only when clinicians deemed the acquisition of an MRI scan as clinically warranted. Therefore, and in line with the German guidelines for diagnostics in MCI [58], imaging was not done routinely for subjects with MCI. Finally, cases with already manifest dementia and a very typical medical history and neuropsychological profile tended to receive a CT rather than MRI as this is available during the same visit. These factors made it more likely, that the cases finally recruited into the study were in fact more difficult to diagnose than the full sample initially referred to the clinic.

In this study, training and test data were strictly separated and differed in several characteristics which likely reduced diagnostic accuracy: Training data were acquired for research studies, typically applying substantially more exclusion and sometimes different inclusion criteria. In addition, accelerated imaging protocols which can systematically influence volumetry [63] were used for the test set only. Training data from the different diagnostic groupings were not equally balanced across scanners and differed substantially in number for each entity. Methods we recently presented to adjust for between scanner differences [44] for binary classification problems could not be applied to the current data as they require data from HC from all scanners. Using the set of HC in the test set would have challenged the interpretation of the classifier performance on HC. Another limitation of the training set is the low number of cases with FTD and particularly LBD which contributed to relatively low classification accuracy of LBD already for the training set. Applying feature selection might further improve the

performance. However, these issues are not straightforward to implement in a multi-class setting.

A further limitation concerns the definition of the gold standard. Cases used for the training and test set had clinical diagnoses, only sometimes supported by biomarker evidence. As a consequence, the level of diagnostic certainty was limited. Ideally, all training cases should have clear clinical signs and/or corresponding biomarker changes. For the test set, several scenarios would be interesting. For example, a clinical diagnosis could be based only on data available from routine clinical workup and biomarker evidence could be added for validation. While we relied on clinical diagnoses which currently represent the basis of clinical management, in an alternative setting, MRI could be left out completely from the definition of the gold standard and thereby avoid any circularity. Limited access, particularly to cases with FTD and LBD with a high diagnostic certainty, made such approaches infeasible. Restricting our analyses only to cases with PET validation did not change AUCs (Supplementary Figure 2), most likely because of the sometimes inaccurate diagnoses of the training cases. We neither attempted to define subtypes such as posterior cortical atrophy in AD or the language or behavioral subtypes in FTD nor did our gold standard definition consider the presence of more than one neurodegenerative disease. We considered three different types of neurodegenerative dementias in addition to an evaluation of vascular pathology. Although they may cover 90% of dementia cases it is far from complete. In the MCI study arm, stable MCI was defined by stability over 12 months which is comparable short. On the other hand, longer follow-up intervals were available for the majority of subjects. Our study included imaging to reach diagnostic decisions and used this to validate an imaging based method. The two approaches are therefore not fully independent, but this should primarily affect cases with a very clear and highly localized atrophy pattern. When considering the clinical routine outside a specialized memory clinic, expert radiologists are often unavailable and an automated method that can extract the same information as an expert radiologist would have substantial clinical value.

In summary, we report the application of automated classification methods to a routine sample from a memory clinic and thereby present a potential routine application of these methods. We found that reported performances from clean research data—usually cross-validated and regarded as good estimates of the potential performance in practice—could not be consistently achieved on a test set acquired in the clinical

routine. Nevertheless, the presented design could be extended to multicenter studies and applied by radiologists outside specialized memory clinics where the full benefit of the automated method can be played out. In addition, future studies should systematically evaluate the effect of the SVM classification on the diagnostic process in respect of diagnostic accuracy as well as perceived diagnostic confidence if made available to clinicians. At the same time, such studies should evaluate, if clinicians consider radar plots a helpful way of presenting results.

## ACKNOWLEDGMENTS

We would like to thank the team of the Freiburg memory clinic for participating in the study. We would also like to thank Elisabeth Kasper for coordinating data exchange with Rostock. The project was partly funded by a grant from the Deutsche Forschungsgemeinschaft (KL 2415/2-1) and the Federal Ministry for Economic Affairs and Energy (KF3223201LW3). The center Leipzig was supported by the German Consortium for Frontotemporal Lobar Degeneration, funded by the German Federal Ministry of Education and Research (MLS), and by the Parkinson's disease Foundation (MLS & KM; Grant No. PDF-IRG-1307).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by

the Foundation for the National Institutes of Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Authors' disclosures available online (<http://j-alz.com/manuscript-disclosures/15-0334>).

## SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: <http://dx.doi.org/10.3233/JAD-150334>.

## REFERENCES

- [1] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (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 on Alzheimer's Disease. *Neurology* **34**, 939-944.
- [2] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, 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.
- [3] Condefer KA, Haworth J, Wilcock GK (2004) Clinical utility of computed tomography in the assessment of dementia: A memory clinic study. *Int J Geriatr Psychiatry* **19**, 414-421.
- [4] Hentschel F, Kreis M, Damian M, Krumm B, Frölich L (2005) The clinical utility of structural neuroimaging with MRI for diagnosis and differential diagnosis of dementia: A memory clinic study. *Int J Geriatr Psychiatry* **20**, 645-650.
- [5] Geroldi C, Canu E, Bruni AC, Dal Forno G, Ferri R, Gabelli C, Perri R, Iapaolo D, Scarpino O, Sinforiani E, Zanetti O, Frisoni GB (2008) The added value of neuropsychologic tests and structural imaging for the etiologic diagnosis of dementia in Italian expert centers. *Alzheimer Dis Assoc Disord* **22**, 309-320.
- [6] Fox NC, Schott JM (2004) Imaging cerebral atrophy: Normal ageing to Alzheimer's disease. *Lancet* **363**, 392-394.
- [7] Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, Mader I, Mitchell LA, Patel AC, Roberts CC, Fox NC, Jack CR, Ashburner J, Frackowiak RS (2008) Accuracy of dementia diagnosis: A direct comparison between radiologists and a computerized method. *Brain* **131**, 2969-74.
- [8] Prävalenz und Inzidenz von Demenz in Deutschland, Last updated 2009, Accessed on 2009.

- [9] Klöppel S, Abdulkadir A, Jack CR Jr, Koutsouleris N, Mourão-Miranda J, Vemuri P (2012) Diagnostic neuroimaging across diseases. *Neuroimage* **61**, 457-463.
- [10] Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O; Alzheimer's Disease Neuroimaging Initiative (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.
- [11] Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RS (2008) Automatic classification of MR scans in Alzheimer's disease. *Brain* **131**, 681-9.
- [12] Vemuri P, Whitwell JL, Kantarci K, Josephs KA, Parisi JE, Shiung MS, Knopman DS, Boeve BF, Petersen RC, Dickson DW, Jack CR Jr (2008) Antemortem MRI based STructural Abnormality iNdex (STAND)-scores correlate with post-mortem Braak neurofibrillary tangle stage. *Neuroimage* **42**, 559-567.
- [13] Adaszewski S, Dukart J, Kherif F, Frackowiak R, Draganski B, Alzheimer's Disease Neuroimaging Initiative (2013) How early can we predict Alzheimer's disease using computational anatomy? *Neurobiol Aging* **34**, 2815-2826.
- [14] Dukart J, Kherif F, Mueller K, Adaszewski S, Schroeter ML, Frackowiak RSJ, Draganski B, Alzheimer's Disease Neuroimaging Initiative (2013) Generative FDG-PET and MRI model of aging and disease progression in Alzheimer's disease. *PLoS Comput Biol* **9**, e1002987.
- [15] Davatzikos C, Resnick SM, Wu X, Parnpi P, Clark CM (2008) Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuroimage* **41**, 1220-1227.
- [16] Teipel SJ, Born C, Ewers M, Bokde AL, Reiser MF, Moller HJ, Hampel H (2007) Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *Neuroimage* **38**, 13-24.
- [17] Fan Y, Batmanghelich N, Clark CM, Davatzikos C (2008) Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage* **39**, 1731-1743.
- [18] Misra C, Fan Y, Davatzikos C (2009) Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *Neuroimage* **44**, 1415-1422.
- [19] Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, Weiner MW, Knopman DS, Petersen RC, Jack CR (2009) MRI and CSF biomarkers in normal, MCI, and AD subjects: Predicting future clinical change. *Neurology* **73**, 294-301.
- [20] 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.e19-27.
- [21] Heister D, Brewer JB, Magda S, Blennow K, McEvoy LK (2011) Predicting MCI outcome with clinically available MRI and CSF biomarkers. *Neurology* **77**, 1619-1628.
- [22] Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, Boeve BF, Petersen RC, Jack CR (2008) Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *Neuroimage* **39**, 1186-1197.
- [23] Franke K, Ziegler G, Klöppel S, Gaser C (2010) Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *Neuroimage* **50**, 883-892.
- [24] Dukart J, Mueller K, Horstmann A, Barthel H, Möller HE, Villringer A, Sabri O, Schroeter ML (2011) Combined evaluation of FDG-PET and MRI improves detection and differentiation of dementia. *PLoS One* **6**, e18111.
- [25] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005) The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am* **15**, 869-877, xi-xii.
- [26] Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology* **74**, 201-209.
- [27] Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, J LW, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW (2008) The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J Magn Reson Imaging* **27**, 685-91.
- [28] Bron EE, Smits M, van der Flier WM, Vrenken H, Barkhof F, Scheltens P, Papma JM, Steketee RME, Méndez Orellana C, 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, Iftekharrudin 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, Alzheimer's Disease Neuroimaging Initiative (2015) Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *Neuroimage* **111**, 562-579.
- [29] Toledo JB, Arnold SE, Raible K, Brettschneider J, Xie SX, Grossman M, Monsell SE, Kukull WA, Trojanowski JQ (2013) Contribution of cerebrovascular disease in autopsy confirmed neurodegenerative disease cases in the National Alzheimer's Coordinating Centre. *Brain* **136**, 2697-2706.
- [30] Zekry D, Duyckaerts C, Belmin J, Geoffre C, Herrmann F, Moulins R, Hauw J-J (2003) The vascular lesions in vascular and mixed dementia: The weight of functional neuroanatomy. *Neurobiol Aging* **24**, 213-219.
- [31] Lim A, Tsuang D, Kukull W, Nochlin D, Leverenz J, McCormick W, Bowen J, Teri L, Thompson J, Peskind ER, Raskind M, Larson EB (1999) Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series. *J Am Geriatr Soc* **47**, 564-569.
- [32] Toledo JB, Toledo E, Weiner MW, Jack CR, Jagust W, Lee VMY, Shaw LM, Trojanowski JQ, Alzheimer's Disease Neuroimaging Initiative (2012) Cardiovascular risk factors, cortisol, and amyloid- $\beta$  deposition in Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement* **8**, 483-489.
- [33] Anbeek P, Vincken KL, van Osch MJ, Bisschops RH, van der Grond J (2004) Probabilistic segmentation of white matter lesions in MR imaging. *Neuroimage* **21**, 1037-44.
- [34] Qudus A, Fieguth P, Basir O (2005) Adaboost and support vector machines for white matter lesion segmentation in MR images. *Conf Proc IEEE Eng Med Biol Soc* **1**, 463-466.

- [35] Lao Z, Shen D, Liu D, Jawad AF, Melhem ER, Launer LJ, Bryan RN, Davatzikos C (2008) Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Acad Radiol* **15**, 300-313.
- [36] Klöppel S, Abdulkadir A, Hadjidemetriou S, Issleib S, Frings L, Thanh TN, Mader I, Teipel SJ, Hüll M, Ronneberger O (2011) A comparison of different automated methods for the detection of white matter lesions in MRI data. *Neuroimage* **57**, 416-422.
- [37] Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C (1989) The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* **39**, 1159-1165.
- [38] Reitan RM (1955) The relation of the trail making test to organic brain damage. *J Consult Psychol* **19**, 393-394.
- [39] Morris JC (1993) The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412-2414.
- [40] Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* **53**, 695-699.
- [41] Neary D, Snowden JS, Gustafson L, Passant U, Stuss D, Black S, Freedman M, Kertesz A, Robert PH, Albert M, Boone K, Miller BL, Cummings J, Benson DF (1998) Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria. *Neurology* **51**, 1546-1554.
- [42] McKeith IG, Dickson DW, Lowe J, Emre M, O'Brien JT, Feldman H, Cummings J, Duda JE, Lippa C, Perry EK, Aarsland D, Arai H, Ballard CG, Boeve B, Burn DJ, Costa D, Del Ser T, Dubois B, Galasko D, Gauthier S, Goetz CG, Gomez-Tortosa E, Halliday G, Hansen LA, Hardy J, Iwatsubo T, Kalaria RN, Kaufer D, Kenny RA, Korczyn A, Kosaka K, Lee VMY, Lees A, Litvan I, Lodos E, Lopez OL, Minoshima S, Mizuno Y, Molina JA, Mukaetova-Ladinska EB, Pasquier F, Perry RH, Schulz JB, Trojanowski JQ, Yamada M, Consortium on DLB (2005) Diagnosis and management of dementia with Lewy bodies: Third report of the DLB Consortium. *Neurology* **65**, 1863-1872.
- [43] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270-279.
- [44] Kostro D, Abdulkadir A, Durr A, Roos R, Leavitt BR, Johnson H, Cash D, Tabrizi SJ, Scihill RI, Ronneberger O, Klöppel S, Investigators Track-HD (2014) Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *Neuroimage* **98**, 405-415.
- [45] Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo N, Martins RN, Maruff P, Masters C, Milner A, Pike K, Rowe C, Savage G, Szoek C, Taddei K, Villemagne V, Woodward M, Ames D, AIBL Research Group (2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr* **21**, 672-687.
- [46] Abdulkadir A, Peter J, Ronneberger O, Brox T, Klöppel S (2014) Voxel-based multi-class classification of AD, MCI, and elderly controls. *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2014 CADDe-mentia Challenge* **2014**, 1-8.
- [47] Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, Poldrack RA, Bilder RM, Toga AW (2008) Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* **39**, 1064-1080.
- [48] Ota K, Oishi N, Ito K, Fukuyama H, Study SEAD-J, Group (2014) A comparison of three brain atlases for MCI prediction. *J Neurosci Methods* **221**, 139-150.
- [49] Burton EJ, Kenny RA, O'Brien J, Stephens S, Bradbury M, Rowan E, Kalaria R, Firbank M, Wesnes K, Ballard C (2004) White matter hyperintensities are associated with impairment of memory, attention, and global cognitive performance in older stroke patients. *Stroke* **35**, 1270-1275.
- [50] Hirono N, Kitagaki H, Kazui H, Hashimoto M, Mori E (2000) Impact of white matter changes on clinical manifestation of Alzheimer's disease: A quantitative study. *Stroke* **31**, 2182-2188.
- [51] Ashburner J, Friston KJ (2005) Unified segmentation. *Neuroimage* **26**, 839-51.
- [52] Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT (2000) Automated Talairach Atlas labels for functional brain mapping. *Hum Brain Mapp* **10**, 120-131.
- [53] Schmahmann JD, Smith EE, Eichler FS, Filley CM (2008) Cerebral white matter: Neuroanatomy, clinical neurology, and neurobehavioral correlates. *Ann N Y Acad Sci* **1142**, 266-309.
- [54] Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* **20**, 273-297.
- [55] Chang C-C, Lin C-J (2001) LIBSVM: A Library for Support Vector Machines.
- [56] Wu TF, Lin CJ, Weng RC (2004) Probability estimates for multi-class classification by pairwise coupling. *J Mach Learn Res* **5**, 975-1005.
- [57] Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, eds. MIT Press, Cambridge, MA, pp. 61-74.
- [58] Frölich PDL (2010) S3-Leitlinie "Demenzen". *Nervenarzt* **81**, 796-806.
- [59] Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Alzheimer's Disease Neuroimaging Initiative (2015) Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* **104**, 398-412.
- [60] Doyle OM, Westman E, Marquand AF, Mecocci P, Velas B, Tsolaki M, Kloszewska I, Soininen H, Lovestone S, Williams SCR, Simmons A (2014) Predicting progression of Alzheimer's disease using ordinal regression. *PLoS One* **9**, e105542.
- [61] Tang X, Holland D, Dale AM, Younes L, Miller MI (2015) Baseline shape diffeomorphometry patterns of subcortical and ventricular structures in predicting conversion of mild cognitive impairment to Alzheimer's disease. *J Alzheimers Dis* **44**, 599-611.
- [62] Piguet O, Halliday GM, Creasey H, Broe GA, Kril JJ (2009) Frontotemporal dementia and dementia with Lewy bodies in a case-control study of Alzheimer's disease. *Int Psychogeriatr* **21**, 688-695.
- [63] Krueger G, Granziera C, Jack CR, Gunter JL, Littmann A, Mortamet B, Kannengiesser S, Sorensen AG, Ward CP, Reyes DA, Britson PJ, Fischer H, Bernstein MA (2012) Effects of

- MRI scan acceleration on brain volume measurement consistency. *J Magn Reson Imaging* **36**, 1234-1240.
- [64] Frings L, Dressel K, Abel S, Saur D, Kümmerer D, Mader I, Weiller C, Hüll M (2010) Reduced precuneus deactivation during object naming in patients with mild cognitive impairment, Alzheimer's disease, and frontotemporal lobar degeneration. *Dement Geriatr Cogn Disord* **30**, 334-343.
- [65] Perneczky R, Haussermann P, Diehl-Schmid J, Boecker H, Forstl H, Drzezga A, Kurz A (2007) Metabolic correlates of brain reserve in dementia with Lewy bodies: An FDG PET study. *Dement Geriatr Cogn Disord* **23**, 416-422.
- [66] Teipel SJ, Grothe MJ, Filippi M, Fellgiebel A, Dyrba M, Frisoni GB, Meindl T, Bokde ALW, Hampel H, Klöppel S, Hauenstein K, EDSO study group (2014) Fractional anisotropy changes in Alzheimer's disease depend on the underlying fiber tract architecture: A multiparametric DTI study using joint independent component analysis. *J Alzheimers Dis* **41**, 69-83.
- [67] Coleman T, Li Y (1996) An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J Optim* **6**, 418-445.