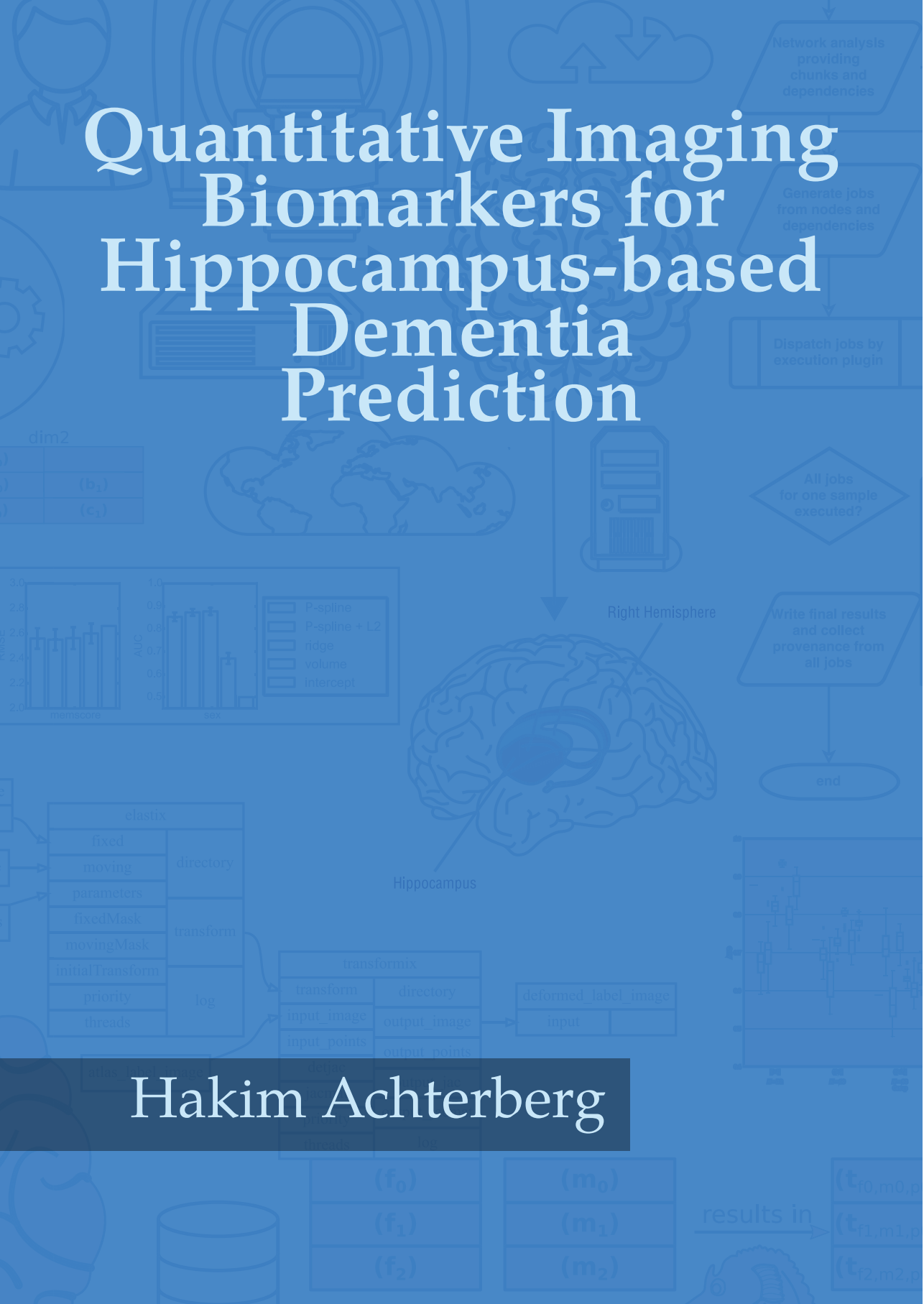
Quantitative Imaging Biomarkers for Hippocampus-based Dementia Prediction



Hakim Achterberg

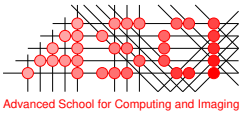
Quantitative Imaging Biomarkers for Hippocampus-based Dementia Prediction

Hakim C. Achterberg

Cover design by Hakim Achterberg & Ton Everaers
Thesis layout by Hakim Achterberg & Ton Everaers

The work in this thesis was conducted at the departments of Radiology and Medical Informatics of the Erasmus MC, Rotterdam, the Netherlands.

This work was carried out in the ASCI graduate school.
ASCI dissertation series number 383.



For financial support for the publication of this thesis, the following organizations are gratefully acknowledged: Alzheimer Nederland, the ASCI graduate school, the Dutch Heart Foundation, and the department of Radiology of the Erasmus MC. I would also like to thank Quantib BV. for their financial support.

ISBN 978-94-6332-268-3
Printed by GVO drukkers & vormgevers

Quantitative Imaging Biomarkers for Hippocampus-based Dementia Prediction

Kwantitatieve beeldmaten van de hippocampus voor het voorspellen van dementie

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 6 december 2017 om 13.30 uur

door

Hakim Christiaan Achterberg
geboren te Tilburg

Erasmus University Rotterdam

Promotiecommissie

Promotor:	Prof.dr. W.J. Niessen
Overige leden:	Prof.dr. P. Golland Prof.dr.ir. B.P.F. Lelieveldt Prof.dr. A. van der Lugt
Copromotor:	Dr. M. de Bruijne

Contents

Chapter 1

General introduction

“It occurred to me that at one point it was like I had two diseases - one was Alzheimer’s, and the other was knowing I had Alzheimer’s."

— Sir Terry Pratchett

Dementia refers to a broad category of neurodegenerative disorders that impair cognition, with Alzheimer's Disease the most prevalent. It was estimated that in 2010 there were 35.6 million people affected by dementia worldwide (Prince et al., 2013). It is expected that due to the growth and increasing age of the population, this number will double every 20 years (Prince et al., 2013). The worldwide dementia-related healthcare costs were estimated at 604 billion US dollars in 2012 and are expected to rise even faster than the prevalence, making dementia a public health priority of the World Health Organization (Organization et al., 2012).

Even though there are currently no cures for dementia, early and differential diagnosis is important as it allows people suffering from dementia to plan their future while they are still able to do so and it allows for non-pharmacological interventions that can help improve cognition and quality of life (Prince et al., 2011). Additionally, early and differential diagnosis of dementia is instrumental for the research of dementia and for the development and use of preventive strategies. Because the underlying pathology of dementia precedes the symptoms of cognitive decline, imaging of the underlying pathology can be a very valuable tool for early diagnosis (Jack et al., 2010; Sperling et al., 2011).

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that provides excellent soft-tissue contrast. This makes it very suitable for brain imaging and it is nowadays commonly used both for dementia research and in clinical practice (McKhann et al., 2011). Currently, MRI is mostly used as a qualitative tool in which radiologists use their knowledge and experience to gauge the disease status of a subject. The qualitative nature of the current MRI practice makes it hard to achieve objective disease status assessment, to analyze large datasets in a consistent way, and to notice subtle effects (e.g differences between subjects, changes over time). With the advance of automated image analysis methods, it has become possible to extract quantitative brain features from MRI scans, leading to the concept of so called quantitative imaging biomarkers.

A quantitative imaging biomarker is defined as "*an objective characteristic derived from an in vivo image measured on a ratio or interval scale as indicators of normal biological processes, pathogenic processes, or a response to a therapeutic intervention*" (Kessler et al., 2015). As the definition already implies, quantitative imaging biomarkers can aid in the understanding of normal biological processes (Casey et al., 2000; Courchesne et al., 2000), diagnosis and diseases prediction (Bron et al., 2015), drug discovery and clinical trials (Pien et al., 2005).

One of the brain areas of interest in dementia is the hippocampus, as it is affected early in the disease process (Braak and Braak, 1997; West et al., 2004). Hence quantitative assessment of the hippocampus may result in potential quantitative imaging biomarkers for the diagnosis and prediction of dementia. However, many challenges need to be addressed to enable routine extraction of hippocampal biomarkers and use these markers in diagnostic and prognostic models. In this thesis, I address a number of these challenges, making contributions to different stages of the process

of biomarker extraction. Here I will first discuss the challenges, prior to describing how these are address in the current thesis.

First, to extract hippocampal quantitative imaging biomarkers from an MRI brain scan, the exact location and outline of the hippocampus in the MRI scan must be known. The process of finding the boundary of an anatomical structure in an image, such as the hippocampus, is called segmentation. There are many methods proposed for segmentation of MR images Balafar et al. (2010). For the segmentation of subcortical brain structures, most methods are atlas-based (Cabezas et al., 2011) or use supervised learning to classify voxels as belonging to a structure or background (Morra et al., 2008, 2010). Both groups of methods rely on training data: an example dataset with known segmentations. Atlas-based methods are very good at finding the general location of a structure, but can have trouble matching the fine details. In contrast, voxel classification approaches are generally very good at modeling local appearance, but may lack contextual information. Considering these different strengths and weaknesses, methods have successfully combined the two approaches, resulting in a very robust and accurate segmentation method (van der Lijn et al., 2012). However, an important limitation of current voxel classification approaches is that the training data needs to be similar to the data to be segmented. However, with exception of some very well-controlled studies, this is generally not the case and thus there is a clear need for supervised segmentation approaches that can cope with differences between train and test data.

Second, once a hippocampus segmentation is obtained, quantitative imaging biomarkers such as the hippocampal volume, hippocampal shape features, and the image texture in the hippocampus, can be computed. These quantitative imaging biomarkers can subsequently be used as part of a diagnosis/prediction model. It is known that hippocampal volume, shape, and texture are all useful for the diagnosis of dementia (Chincarini et al., 2011; Jack et al., 1999; Li et al., 2007). Additionally, hippocampal volume has been shown to be predictive for dementia in a random sample of the general population (den Heijer et al., 2006), making it valuable for early diagnosis. For hippocampal shape and texture it is not yet established whether they provide predictive information for dementia. Additionally, the value of combining hippocampal volume, shape, and texture for the prediction of dementia have not yet been studied.

Third, the use of a quantitative imaging biomarker in a diagnosis/prediction model is straightforward for biomarkers such as volume, which can be expressed as a single value. However, complex biomarkers such as shape or texture may contain many more and up to hundreds of variables, making the analysis and use of such biomarkers more complex. To deal with this complex data, classification or regression models with a regularization added to them can be used. While such methods have been successful at diagnosing and predicting dementia based on the high dimensional data, they generally do not provide insight why a certain classification is achieved. This hampers the understanding and interpretation of the biomarker and results in a

"black box" model which works, but of which it is not really known why or how it works. This may hamper the acceptance of such biomarkers in clinical practice.

Fourth, for large imaging studies, such as population imaging studies, or to pool data from different centers or cohorts, it is very important that all the quantitative imaging biomarkers are derived in exactly the same way. As the calculation and analysis of these biomarkers is often not a single operation, but a workflow of a number of steps, it is of utmost importance that this is performed correctly and consistently. For the integrity of the study, it is important that there is a clear log of all the data processing steps and that the analysis is reproducible. For many studies, workflows are implemented by a script and a log is generated from that script. However, these logs are often not machine readable and not exhaustive, making the reproduction of the results at best difficult and at worst impossible.

In this thesis, I address different aspects of these four aforementioned issues and limitations of quantitative biomarkers. In chapter 2, to address the first challenge, I investigate if the hippocampus segmentation method of van der Lijn et al. (2012) can be made more robust against changes in the acquisition protocol by creating appearance models based on the tissue content instead of the image intensities in the voxels.

To address the second challenge, in chapter 3 I relate hippocampal shape to the future development of dementia in the general population and compare its predictive value to that of hippocampal volume. In chapter 4 I describe a similar analysis that includes hippocampal volume, shape, and texture; I compared the predictive value of the biomarkers and various combinations of the biomarkers to see if they can complement each other.

In chapter 5, the third challenge is addressed by exploring a new way of hippocampal shape analysis that allows for better insight in the relationship between shape and outcome. This method based on spatial regularization using P-spline regression.

Finally, in chapter 6 the fourth challenge is addressed by introducing a framework for creating workflows, called Fastr, which helps formalizing an image analysis workflow and ensures reproducible and traceable analysis.

Chapter 2

Local appearance features for robust MRI brain structure segmentation across scanning protocols

Hakim C. Achterberg
Dirk H.J. Poot
Fedde van der Lijn
Meike W. Vernooij
M. Arfan Ikram
Wiro J. Niessen
Marleen de Bruijne

Local appearance features for robust MRI brain structure segmentation across scanning protocols.
SPIE Medical Imaging, 2013

Segmentation of brain structures in magnetic resonance images is an important task in neuro image analysis. Several papers on this topic have shown the benefit of supervised classification based on local appearance features, often combined with atlas-based approaches. These methods require a representative annotated training set and therefore often do not perform well if the target image is acquired on a different scanner or with a different acquisition protocol than the training images. Assuming that the appearance of the brain is determined by the underlying brain tissue distribution and that brain tissue classification can be performed robustly for images obtained with different protocols, we propose to derive appearance features from brain-tissue density maps instead of directly from the MR images. We evaluated this approach on hippocampus segmentation in two sets of images acquired with substantially different imaging protocols and on different scanners. While a combination of conventional appearance features trained on data from a different scanner with multi-atlas segmentation performed poorly with an average Dice overlap of 0.698, the local appearance model based on the new acquisition-independent features significantly improved (0.783) over atlas-based segmentation alone (0.728).

2.1 Introduction

The volume and shape of brain structures, such as the hippocampus, are being investigated as biomarkers for early detection, differential diagnosis, and monitoring of treatment efficacy of new drugs in neurodegenerative diseases such as Alzheimer's disease. This has caused large interest in automated methods that can segment these structures from MRI scans accurately, robustly, and reproducibly. Atlas-based approaches, sometimes complemented with a global statistical intensity model, have successfully been applied for this purpose (Hammers et al., 2007; Heckemann et al., 2006). The intensity models can correct for registration errors and have shown to increase accuracy compared to segmentations based solely on atlases (Leung et al., 2010a). However, not all brain structures can be accurately modeled by a global intensity model because they have spatially varying intensity distributions or show no contrast with the background along large parts of their boundaries.

One way to tackle this problem is by using voxel classifiers that are based on a large number of local image appearance features. When supplied with representative training data, these supervised methods can provide a much richer description of brain structure appearance than models based on intensity alone and they show a better accuracy (van der Lijn et al., 2012; Morra et al., 2008, 2010). However, their

performance drops when the unlabeled target image is not acquired on the same MRI scanner and with the same protocol as the training images (Morra et al., 2010).

To increase robustness against intensity differences between training and target image, normalization procedures have been proposed. Most of these methods match the histogram of the target image to that of the training image (Han and Fischl, 2007; Nyúl et al., 2000). This strategy works well when differences in intensity distribution are small, but cannot handle images acquired with very different imaging protocols, in which the mapping might be highly non-linear.

We propose a robust way of performing local appearance modeling in supervised MR brain segmentation. We assume that the appearance of the brain in MR images is largely explained by the distribution of gray matter, white matter and cerebrospinal fluid (CSF). We therefore propose to extract local appearance features from tissue maps instead of from MR images. This allows us to build on a large body of work on robust tissue segmentation methods that can handle different imaging protocols (Ashburner and Friston, 2005; Poot et al., 2011; Vrooman et al., 2007; Zhang et al., 2001). Rather than relying only on global image properties like a histogram, these methods incorporate information on a voxel level including regularization and partial volume models.

As far as we know, features derived from tissue maps have not been used before to reduce scanner dependence of appearance models for brain structure segmentation. Morra et al. (Morra et al., 2008, 2010) included features derived from tissue maps in their brain structure classification approach based on AdaBoost. However, they also included many intensity-based features. The resulting feature set was not robust to scanner differences as shown in Morra et al. (2010)

2.2 Materials and Method

The proposed robust features are incorporated in a segmentation method that combines spatial information with local appearance. This segmentation method has been applied successfully for segmentation of the hippocampus and cerebellum (van der Lijn et al., 2012; Morra et al., 2010). In this paper we aim to make such methods also applicable to the situation in which the imaging data to be segmented are acquired under different conditions than the training data.

2.2.1 Imaging data

For our experiments we used two manually annotated datasets selected from the Rotterdam Scan Study (Ikram et al., 2011). Dataset 1 consisted of 8 MR subjects scanned on a 1.5T scanner (Siemens Healthcare). The sequence used was a custom designed, inversion recovery, 3D half-Fourier acquisition single-shot turbo spin echo (HASTE) sequence with an in-plane (sagittal) resolution of 1×1mm and a slice thickness of 1.25mm. The use of two echo times led to two images: HASTE odd

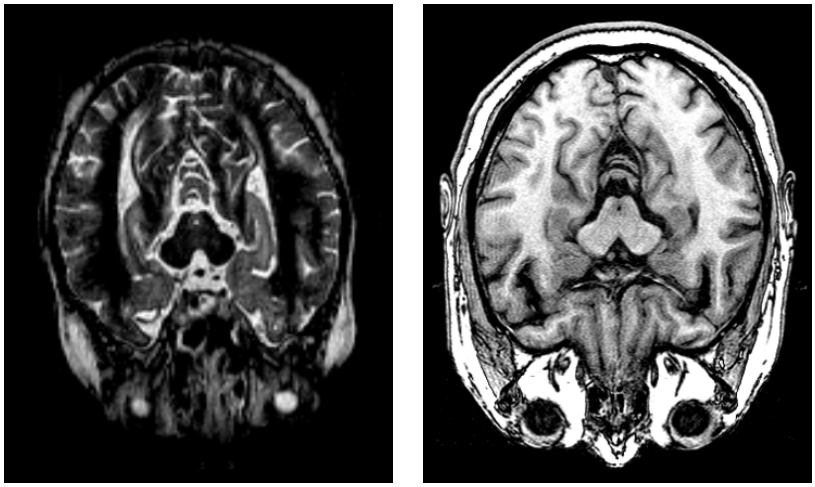


Figure 2.1: *Examples of the MR scans used, left: HASTE Odd weighted scan from dataset 1, right: T1 weighted scan fom dataset 2.*

(HOdd) and HASTE even (HEven) (van der Lijn et al., 2012). Dataset 2 consisted of 18 MR subjects scanned on a 1.5T Scanner (General Electric Healthcare). The sequences used were a T1-weighted 3D Fast RF Spoiled Gradient Recalled Acquisition in Steady State with an inversion recovery pre-pulse (FASTSPGR-IR) sequence with an in-plane resolution (axial) of 0.49×0.49mm and a slice thickness of 0.8mm, a proton density (PD) weighted sequence, and a fluid-attenuated inversion recovery (FLAIR) sequence (Ikram et al., 2011; van der Lijn et al., 2012). All slices were contiguous. For both datasets, manual segmentations were available for all subjects, which were created using the same anatomical definition of the hippocampus, but by different experts. The resulting images have a large difference in appearance, e.g. compare the scans in Figure 2.1. There was no overlap in subjects between the two datasets.

2.2.2 Robust appearance feature extraction

The local appearance model is based on Gaussian scale-space features derived from brain tissue maps. To obtain the gray matter, white matter, and CSF tissue maps, we used a method proposed by Poot et al. (2011). This method uses multispectral MR images to produce a partial volume segmentation of the different tissue types. In the segmentation process, the point spread functions of the different MR sequences, as well as potential bias fields, are taken into account. Furthermore, the different scans are iteratively aligned to correct for inter-scan subject movement. This allows for the optimal use of all image information from scans with different contrast

weightings, orientations, and resolutions. To minimize the influence of the differences in resolution between the train and test images, all tissue maps were estimated on a grid with the same orientation and resolution. A regularization term is introduced which penalizes voxels containing multiple tissue types as well as voxels with a tissue type different from their neighbors.

The tissue segmentation results in three tissue density maps describing gray matter, white matter, and CSF. To reduce the dimensionality of the resulting feature space, also a T1 weighted image was simulated from the tissue density maps using a fixed set of mean tissue intensities obtained from the T1 scans, to potentially replace the three tissue maps.

Similar to van van der Lijn et al. (2012), the appearance features consisted of the original intensities in the three tissue maps and the following Gaussian scale-space features from all three tissue maps: blurred map, first and second order derivatives, gradient magnitude, Laplacian, Gaussian curvature, and the eigenvalues of the Hessian matrix. All scale-space features were extracted at 3 different scales (1mm, 2.2mm, and 5mm), leading to a set of 147 features per input image. A similar feature set was extracted for the simulated T1-weighted image, resulting in 49 features.

2.2.3 Brain structure segmentation

The brain structure segmentation method combines prior spatial and appearance information. Spatial information is derived from multiple atlas images, which are individually registered to the target image.

The appearance probability map is created by taking the posterior probability of a moderated k-nearest neighbour (kNN) voxel classifier (Alkoot and Kittler, 2002), which is based on the features described in the previous subsection. The training data for the appearance model was created by randomly sampling 5 percent of the voxels belonging to the hippocampus and the same number of voxels from a 4mm band around the hippocampus in each training image. To find the optimal set of features for the kNN from all candidate features, a feature selection scheme was used. The feature selection procedure starts with a sequential forward feature selection, in each step adding the feature that increases the area under the ROC curve (AUC) the most when added, until there is no feature left that improves the result. Subsequently, there is a backward search to eliminate features which do not contribute positively to the classification. During the feature selection, the data of half of the subjects in the training set was used for training and the other half for testing of the selected feature set. After the feature selection data from all subjects of the training set was used to train the final kNN classifier.

To create the spatial probability map, all manually segmented atlases from dataset 2 were registered non-rigidly to the target image, except if the atlas and the target were the same image. Registrations were performed with Elastix (Klein et al., 2010b) with mutual information as the similarity metric. A spatial probability map was

created for each subject by averaging the transformed manual segmentations.

The final segmentation was obtained by combining the appearance probability map P_a and the spatial probability map P_s. Assuming P_a and P_s are independent, the hippocampus segmentation can be obtained with $S = P_a P_s > (1 - P_a)(1 - P_s)$ which can be simplified to $S = P_a + P_s > 1$. In the experiments without the appearance model, thresholding the spatial probability map at 0.5 sufficed.

2.2.4 Experiments

Using the two datasets, we conducted two experiments to evaluate the newly proposed features. First, the newly proposed features were assessed when training and testing on data from the same dataset using leave-one-out cross-validation ('intra-scanner'). Subsequently, we tested the performance of the appearance model and the resulting segmentation on one dataset, when the appearance model was trained on the other dataset ('inter-scanner').

In both of these experiments three feature sets were compared. The first set of features (**O**) is derived from the original image intensities of the HOdd in dataset 1 and the T1 in dataset 2. The second set of features is derived from the three tissue maps (**T**), and the third set of features is derived from the simulated T1 map (**S**). In addition, we compare the different feature sets to multi-atlas segmentation without appearance model (atlas only, **A**).

For evaluation, the AUC of the voxel classification was computed from the true and false positives in the same area in which the classifier is trained (a radius of 4 mm around the hippocampus). In addition, the Dice similarity indices (DSI) between the automatic segmentations and the corresponding manual segmentation were computed. DSIs between different feature sets and methods were compared using a paired t-test with a significance level of 0.05.

2.3 Results

The results of the appearance based voxel classification are presented in Table 1. In an intra-scanner setting, all tested feature sets lead to good classification results with an AUC above 0.92. In dataset 1 there is almost no difference between the features used, whereas in dataset 2 **O** results in a slightly higher AUC than **T** or **S**. In an inter-scanner setting, **O** no longer has predictive power, whereas the AUC of **T** and **S** is only slightly lower (around 0.90) than in the intra-scanner setting.

The segmentation results are presented in Table 2.2. In an intra-scanner setting again for dataset 1 all feature sets give similar results, while for dataset 2 **O** results in a slightly higher DSI. These results are consistent with those presented in Table 2.1. In both datasets, the DSI significantly increases when an appearance model is included ($P<0.001$) compared to the DSI of multi-atlas segmentation alone **A**. In the inter-scanner setting, inclusion of the appearance model with features based on the

	Test data	Training data	Original scan (**O**)	Simulated T1 (**S**)	Tissue maps (**T**)
Intra-scanner	set 1	set 1	0.932 (0.010)	0.932 (0.007)	0.939 (0.004)
	set 2	set 2	0.941 (0.012)	0.925 (0.013)	0.924 (0.016)
Inter-scanner	set 1	set 2	0.436 (0.016)	0.911 (0.013)	0.893 (0.013)
	set 2	set 1	0.524 (0.022)	0.889 (0.023)	0.891 (0.019)

Table 2.1: *Performance of the voxel classifier in area under the ROC curve as* mean (standard deviation).

original image deteriorates atlas segmentation results. In contrast, the proposed **T** and **S** significantly contribute to the segmentation ($P<0.001$), with an increase in DSI of 0.06 in dataset 1 when compared to **A**. In dataset 2, the increase in DSI is 0.02, which is significant for **S** ($P<0.001$), but not for **T** ($P=0.094$). Compared to the intra-scanner results, the increase in DSI obtained with **T** and **S** is only slightly lower in the inter-scanner setting.

Figures 2.2 and 2.3 show example scans, tissue maps, simulated T1 images and the corresponding appearance probability maps and segmentations for the two datasets. Note that many false positives occur outside the 4 mm band from which the background samples were extracted in the training set. This does not significantly influence segmentation accuracy, as the atlas-based spatial probability map is (almost) zero outside this band, which ensures correct classification of background voxels.

2.4 Discussion & Conclusion

The combination of atlas-based methods and classifier-based appearance models in brain structure segmentation has shown promising results in the literature. In this work, we evaluate new features for the appearance model, with which these methods can be made more robust against differences in scanner and acquisition protocol between training and testing data.

The intra-scanner results from Table 2.1 and 2.2 show that the proposed features can discriminate between hippocampus and background appearance as good as features derived from the original images. An appearance model based on the original features cannot handle the inter-scanner situation; the DSI drops by 0.12. In contrast, the newly proposed features can – even in the inter-scanner setting – help to discriminate between hippocampus and background.

	Test data	Training data	Atlas (**A**) only	Original (**O**) scan	Simulated (**S**) T1	Tissue (**T**) maps
Intra-scaner	set 1	set 1	0.728 (0.032)	0.815 (0.020)	0.809 (0.023)	0.818 (0.016)
	set 2	set 2	0.790 (0.041)	0.841 (0.029)	0.832 (0.027)	0.830 (0.027)
Inter-scanner	set 1	set 2	0.728 (0.032)	0.698 (0.046)	0.787 (0.017)	0.783 (0.019)
	set 2	set 1	0.790 (0.041)	0.715 (0.035)	0.816 (0.030)	0.812 (0.030)

Table 2.2: *Segmentation results. Dice similarity with the manual segmentation of both hippocampi as* mean (standard deviation).

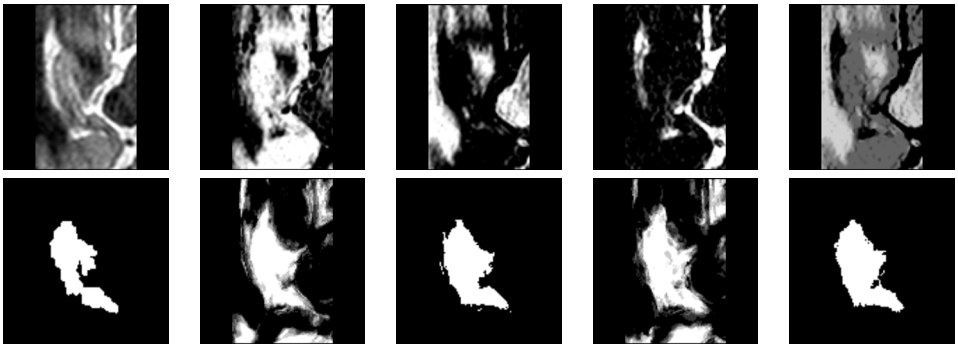


Figure 2.2: *Top row from left to right: crop-out around the hippocampus HOdd image from dataset 1, corresponding gray matter map, white matter map, CSF map, and simulated T1 weighted image. Bottom row: manual segmentation, appearance probability map and final segmentation based on **S** (simulated T1) in intra-scanner and on **S** in inter-scanner setting*

The segmentation accuracy in inter-scan settings might be lower than in the intra-scanner settings, however the contribution of the appearance models is clear when comparing to the atlas-only situation: there is a significant improvement ($P<0.001$) in DSI of up to 0.06 in dataset 1 and up to 0.03 in dataset 2. This shows the proposed method is valuable in situations where data from different scanners or with different acquisition protocols is pooled, e.g. in multi-center studies.

Overall, the DSI are better for dataset 2. The lower performance for dataset 1 may be because the atlas images used are from dataset 2: the difference between atlas and target images slightly deteriorates the registration quality and the datasets

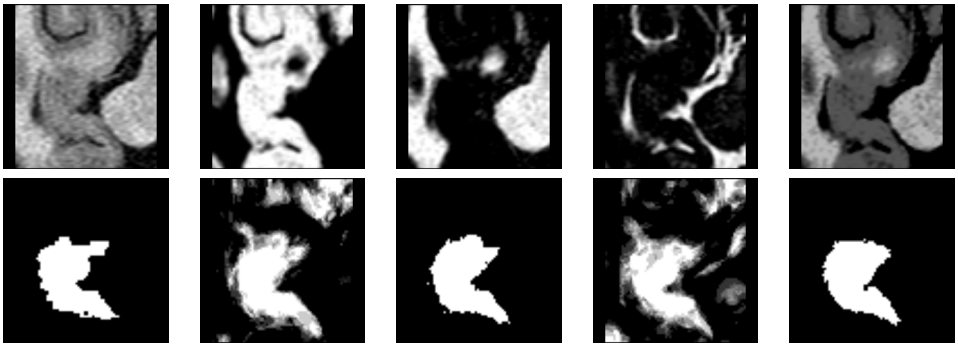


Figure 2.3: *Top row from left to right: crop-out around the hippocampus of one of T1-weighted image from dataset 2, corresponding gray matter map, white matter map, CSF map, and simulated T1-weighted image. Bottom row: manual segmentation, appearance probability map and final segmentation based on* ***S*** *(simulated T1) in intra-scanner and on* ***S*** *in inter-scan setting*

are segmented by different observers. This is reflected by the lower resulting DSI for dataset 1 when only the atlas is used. The appearance model is able to compensate for most of these registration errors.

For dataset 2 the DSIs are comparable to literature (Fischl et al., 2002; Han and Fischl, 2007; Heckemann et al., 2010; van der Lijn et al., 2012). Several authors report better results for hippocampus segmentation (Leung et al., 2010a; Wolz et al., 2010). These studies use manifold learning for finding an optimal segmentation propagation (Wolz et al., 2010) or atlas selection from a large set of atlases (Leung et al., 2010a). In this work, we use a simple averaging of the atlases to create the spatial probability maps. However, these maps can also be created using more advanced atlas combination and selection methods.

An important assumption in this work is that tissue segmentation is robust against changes in scanning protocols. We believe this assumption is valid, as good contrast between tissue types is essential in any sequence designed for structural analysis. Additionally, our result shows that we can use the tissue information from different scanners to build a robust appearance model.

The two datasets used in this study are very different: not only are they acquired on a different MR scanner, they also have a different resolution and use a different MR sequence. Even though this data is acquired within a single longitudinal study, the differences in other studies may be more subtle. In cases where the train and test data are more similar, the benefits from the proposed features and appearance model might be less pronounced. However, this study does serve as an extreme example to show the potential of these features for robust segmentation across scanning protocols.

In conclusion, this work shows that it is possible to improve hippocampus segmen-

tation with an appearance model trained on data acquired with a different imaging protocol and/or MR scanner than the target image. The proposed method can therefore be valuable for the analysis of brain imaging data from multiple sources, such as multi-centre studies or studies re-using data acquired in clinical practice, which requires robustness to many different imaging protocols and scanner types.

Chapter 3

Hippocampal shape is predictive for the development of dementia in a normal, elderly population

Hakim C. Achterberg
Fedde van der Lijn
Tom den Heijer
Meike W. Vernooij
M. Arfan Ikram
Wiro J. Niessen
Marleen de Bruijne

Hippocampal shape is predictive for the development of dementia in a normal, elderly population.
Human Brain Mapping, 2014

Previous studies have shown that hippocampal volume is an early marker for dementia. We investigated whether hippocampal shape characteristics extracted from MRI scans are predictive for the development of dementia during follow up in subjects who were nondemented at baseline. Fur- thermore, we assessed whether hippocampal shape provides additional predictive value independent of hippocampal volume. Five hundred eleven brain MRI scans from elderly nondemented participants of a prospective population-based imaging study were used. During the 10-year follow-up period, 52 of these subjects developed dementia. For training and evaluation independent of age and gender, a subset of 50 cases and 150 matched controls was selected. The hippocampus was segmented using an automated method. From the segmentation, the volume was determined and a statistical shape model was constructed. We trained a classifier to distinguish between subjects who developed dementia and subjects who stayed cognitively healthy. For all subjects the a posteriori probability to develop dementia was estimated using the classifier in a cross-validation experiment. The area under the ROC curve for volume, shape, and the combination of both were, respectively, 0.724, 0.743, and 0.766. A logistic regression model showed that adding shape to a model using volume corrected for age and gender increased the global model-fit significantly ($P = 0.0063$). We conclude that hippocampal shape derived from MRI scans is predictive for dementia before clinical symptoms arise, independent of age and gen- der. Furthermore, the results suggest that hippocampal shape provides additional predictive value over hippocampal volume and that combining shape and volume leads to better prediction.

3.1 Introduction

There are 5.4 million individuals suffering from Alzheimer's disease (AD) in the USA Thies and Bleiler (2011). Furthermore, it is estimated that on average patients suffering from dementia use three times more medical care compared to people not suffering from dementia in the same age range. Although there are currently no cures or drugs to prevent dementia, recent studies show the promise of better drugs to slow or halt the progression of dementia (Neugroschl and Sano, 2010; Scarpini et al., 2003). Still, all damage suffered by the brain as a result of dementia is irreversible. This makes early detection—preferably before clinical symptoms appear—of great importance.

Experienced radiologists can recognize the typical brain atrophy patterns associated with dementia and can support the traditional diagnosis based on cognitive tests

(Scheltens et al., 1992). In recent years, a large number of articles have explored the possibilities of computer-aided diagnosis and prediction of dementia in MRI.

Chan et al. (2001) showed that atrophy in demented subjects is most pronounced in the hippocampus, amygdala, and entorhinal cortex. Also, it has been shown that the hippocampal subregions are not equally affected by dementia and that localized atrophy within the hippocampus can be linked to dementia (e.g. Apostolova et al., 2006; Csernansky et al., 2005; Scher et al., 2007). Hippocampal shape can be used as a measure for very localized atrophy. Hippocampal shape in relation to dementia has been studied and shown to contain information to distinguish demented subjects from healthy controls (e.g. Ferrarini et al., 2009; Gerardin et al., 2009).

It is well established that gray matter atrophy related to dementia is visible on MRI, even before clinical symptoms become apparent (den Heijer et al., 2006; Jack et al., 1999; Scheltens et al., 2002). Additionally, histology studies show that certain hippocampal subfields are affected stronger and earlier by atrophy (West et al., 2004), suggesting that local analysis is best suited to detect dementia in an early stage. Considering that the hippocampus is one of the brain structures affected earliest and strongest by dementia, in this study we used hippocampal shape to investigate localized hippocampal atrophy.

Many studies investigating dementia using MRI, have focused on distinguishing demented subjects from controls (DeCarli et al., 1995; Li et al., 2007). Other studies compared MR images of cognitively normal subjects to subjects with mild cognitive impairment (MCI), which is considered a precursor of dementia. However, not all MCI subjects develop dementia; some remain MCI for a long time and some even revert to cognitively normal.

To be able to create a model for dementia prediction, longitudinal clinical data is required. This allows researchers to check the cognitive function of subjects over time. When this data is available, new subject groups can be identified: MCI converters (MCI-c), subjects with MCI who develop dementia during a follow-up period, and MCI nonconverters (MCI-nc). Some studies try to distinguish between MCI-c and MCI-nc (Apostolova et al., 2006; Ferrarini et al., 2009; Jack et al., 1999). Other studies try to detect very early dementia by comparing MCI-c subjects to controls (Davatzikos et al., 2008). In contrast to these studies, the subjects in our population were all nondemented at baseline. Furthermore, they form a cross section of middle aged and elderly people from the general population.

This article aims to answer the following questions: Is hippocampal shape extracted from MRI scans predictive for the development of dementia before clinical symptoms? If so, does hippocampal shape provide extra predictive value over hippocampal volume?

A preliminary version of this study has already been made available in a workshop article Achterberg et al. (2010). In the current study, we included 511 instead of 94 subjects and evaluated the models more extensively.

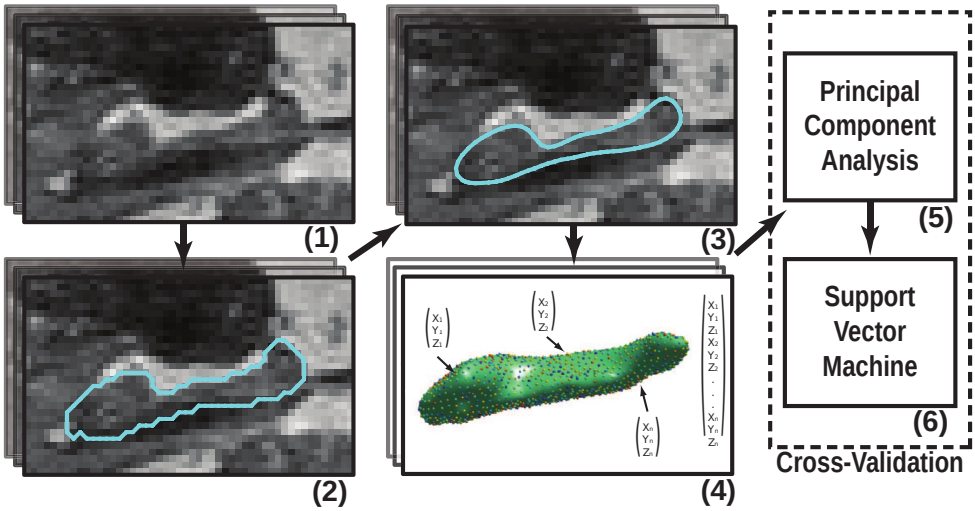


Figure 3.1: *Overview of methods used: (1) MRI scans of the brain were acquired. (2) In each scan, the left and right hippocampus was segmented. (3) The segmentations were postprocessed. (4) Points were distributed over each surface, such that points on a different scans correspond with each other, and were concatenated to create one feature vector per scan. (5) The dimensionality of the feature vectors was reduced using principal component analysis. (6) A Support Vector Machine classifier was used to predict dementia development for each scan. Step (5) and (6) were performed in a cross-validation manner.*

3.2 Materials and Methods

MRI scans of the brain were acquired from 511 nondemented, elderly persons (Data section). During the 10-year clinical follow-up period, 52 subjects (10%) developed dementia (any type). For these subjects, MRI scans taken up to 10 years before the clinical diagnosis were available, allowing the evaluation of the predictive value of imaging biomarkers.

From the MRI scans, the hippocampus was segmented using an automated method (Hippocampus Segmentation section). Subsequently, a statistical shape model was constructed using the segmented data (Shape Representation section). Features from this model were used to train a classifier that distinguishes subjects who will develop dementia from subjects who will stay cognitively healthy (Classification section). The performance of this classifier was evaluated in a cross-validation manner (Evaluation section). A schematic overview of the methods is shown in 3.1. The predictive value of hippocampal volume, shape, and the combination of both shape and volume were compared to each other.

3.2.1 Data

The imaging data used in this study was a subset taken from the Rotterdam Scan Study: a prospective, population-based MRI study on age-related neurological diseases (den Heijer et al., 2003). For 511 nondemented, elderly subjects, MRI scans and the age, gender, dementia diagnosis, and time of followup were available.

All subjects were scanned in 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 characteristics: inversion time 4400 ms, repetition time 2800 ms, effective echo time 29 ms, matrix size 192×256, flip angle 180°, slice thickness 1.25 mm, acquired in sagittal direction. The images were reconstructed to a $128 \times 256 \times 256$ matrix with a voxel dimension of $1.25 \times 1.0 \times 1.0$mm.

Study participants were followed during a 10-year period. During this period, they were invited for four cognitive follow-up tests, and the general practitioners records were tracked for diagnosis of dementia. Dementia screening followed a strict two-step protocol(den Heijer et al., 2006); initially, participants were cognitively screened with the Mini Mental State Examination (MMSE) and the Geriatric Mental Schedule. If the results of this initial screening indicated possible dementia, a more thorough cognitive testing was performed for verification. During the study period, 52 persons 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.

The entire dataset, hereafter referred to as the cohort set, contained 52 prodromal dementia cases and 459 persons who did not develop dementia. To train and test a model independent of age and gender, an age- and gender-matched subset of 50 prodromal dementia subjects and 150 controls was identified, hereafter referred to as the matched set. Characteristics of the cohort set and matched set can be found in 3.1. None of the subjects were demented at the time the MRI scan was taken.

	cohort set		matched set	
	prodromal dementia	controls	prodromal dementia	controls
	$N = 52$	$N = 459$	$N = 50$	$N = 150$
women (%)	61.5	48.6	60.0	60.0
age (years)	79.02 (6.44 std)	72.87 (7.81 std)	78.75 (6.41 std)	78.72 (6.43 std)
	[64.37 – 88.73]	[58.96 – 89.83]	[64.37 – 88.73]	[64.37 – 89.83]
memory complainers (%)	57.7	26.8	58.0	27.3
MMSE	27 (3 iqr)	28 (2 iqr)	27 (3 iqr)	28 (2 iqr)
	[20 – 30]	[14 – 30]	[20 – 30]	[21 – 30]
time to diagnosis (years)	4.0 (4.8 iqr)	n/a	4.0 (4.9 iqr)	n/a
	[0.73 – 10.28]	n/a	[0.73 – 10.28]	n/a

Table 3.1: *Characteristics of the subjects in cohort set and matched set. Values given as* mean (standard deviation) *or* median (interquartile range)*, and range given as* [min-max].

Because memory impairment is the first detectable neuropsychological sign of incipient dementia, we questioned persons on subjective memory complaints. This was done by a single question: "Do you have complaints about your memory performance?". Furthermore, objective memory performance was assessed using a 15-word verbal learning task (den Heijer et al., 2006) resulting in a memory score.

To increase the sample size in the matched set, we selected three unique controls per case; this was possible for 50 cases. The matching was performed using the following criteria: the gender had to be the same, the follow-up time of the controls should be at least as long as the time to diagnosis of the corresponding case, and the age could not differ more than 1.5 years. To avoid significant age differences, the mean age of the controls was kept as close as possible to the age of the case. We verified that the age matching resulted in no significant difference between groups with a paired *t*-test.

3.2.2 Hippocampus segmentation

Hippocampi were automatically segmented using a segmentation method based on multiatlas 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 is globally optimized using graph cuts. As training data, we used manually delineated images from 20 participants from the same population. Leave-one-out experiments on the training images showed mean Dice similarity indices of 0.85±0.04 and 0.86±0.02 for the left and right side. The final segmentation results of the 511 images used in this study were inspected by a trained observer (TdH) and manually corrected in case of large errors; two cases and 69 controls were manually corrected.

Because the creation of the shape model requires one single-body object, we extracted the largest single body from the segmentation and applied a hole-fill. This ensures that the object can be described by one single surface. Furthermore, an antialiasing step was used to smooth the binary segmentation Whitaker (2000).

From the segmentation, the hippocampal volume was calculated, which was subsequently corrected for intracranial volume. The intracranial volume was calculated by registration of a single brain mask to the target image and calculating the volume inside the brain mask Ikram et al. (2008).

3.2.3 Shape representation

The hippocampal shapes were described by corresponding points on the surfaces using the entropy-based particle system as presented before Cates et al. (2006, 2007)[1]. This method aims at finding a uniform sampling of the shapes while minimizing the

[1]We used the software provided by the authors. ShapeWorks: An open-source tool for constructing compact statistical point-based models of ensembles of similar shapes. Scientific Computing and Imaging Institute (SCI). Can be found at http://www.sci.utah.edu/software.html

information content of the resulting shape model, leading to a compact model with optimal point correspondences. By describing both these criteria as entropies, they can be combined in a single model in a natural way.

When describing shapes by a set of N points on their surface, shapes can be seen as points in a $3N$-dimensional space and a collection of shapes forms a distribution in this space. The variance in this distribution can be caused by real shape differences or by errors in the point correspondence between the shapes. By moving the points over the surface for the individual shapes, the point correspondence error can be reduced. However, when the sampling is approximately uniform, the real shape differences will not change. While minimizing the variance in the distribution of shapes will thus change the description of the shape, in our case the point sampling, in such a way the correspondence is optimal without losing the real variation between shapes. This optimization is performed using a gradient descent algorithm. During gradient descent optimization, the points are constrained to lie on the surfaces of original segmentations.

All hippocampus segmentations were isotropically scaled to have equal volumes before the creation of the shape model, to exclude any volume information from the model. We created a shape model for both the left and the right hippocampus separately. The number of points (N) to represent each shape was set to 1,024. The shapes were aligned rigidly using Procrustes analysis (Goodall, 1991) at regular intervals of 25 iterations. To determine the number of iterations required for convergence, one optimization was run for 1,600 iterations, saving intermediate results every 10 iterations. For every intermediate output, the point displacements in the last 10 iterations were calculated. The optimization converged after 150–200 iterations. In all our following experiments, we ran the optimization for 200 iterations.

3.2.4 Classification

We trained statistical pattern classifiers to discriminate between subjects who developed dementia and those who remained cognitively healthy during the follow up period. This resulted in an estimated class for each subject, allowing us to evaluate how well we can predict dementia development in our dataset.

For classification three different feature sets were used: volume (normalized for intracranial volume), shape and a combination of both (hereafter referenced to as shape+volume). Volume and shape were combined by scaling the volume and shape feature vectors to have equal total variance and then concatenating them. All feature sets were created by concatenating the feature vectors derived from left and right hippocampus.

The dense sampling of points on a shape leads to a high dimensional feature space: two shapes with 1,024 points in a 3D space results in a 6,144 dimensional feature space. To reduce the dimensionality of the feature space a principal component analysis (PCA) retaining 99% of the variance was applied. After PCA, the number of

dimensions was reduced to around 175.

A Support Vector Machine (SVM) classifier was used in all experiments. For completeness, we tested other classifiers, but none outperformed the SVM. For the shape and shape+volume features, the SVM classifier used a radial basis kernel. For the volume features, the SVM classifier used a linear kernel; the dimensionality was only two (right and left hippocampal volume) and a radial basis kernel did not improve classification performance.

The slack parameter (controlling the trade-off between a large margin and small error on the training data) and the scale parameter of the radial basis function were estimated automatically by a grid search using leave-group-out cross-validation on the training data. Each fold of the cross-validation contained one case and its corresponding controls, thus preserving the age- and gender-matching. An SVM results in a signed distance to the decision boundary for each subject. To convert this distance to the posterior probability that a subject belongs to a class, we use the inverse logit function

$$P(d) = \frac{1}{1 + \exp(-d)}, \tag{3.1}$$

with d the distance to the decision boundary. We used these posterior probabilities both to compute receiver operating characteristic (ROC) curves and for regression analysis. All classification tests were performed using the PRTools (Duin et al., 2004) Matlab toolbox and libsvm (Chang and Lin, 2001).

3.2.5 Evaluation

All classification tests were performed in a leave-group-out cross-validation loop. We trained (including the dimension reduction by PCA and parameter estimation of the SVM) on the matched dataset, except for one case and its matching controls, and estimated classification rate on the left out subjects. This was repeated for every case in the set. This keeps the age-and gender-matching intact, but still allows for as many cross-validation folds as possible.

The results were stratified for time to diagnosis; using only results of cases (and the corresponding controls) who developed dementia within a certain time interval. The following time intervals were used: less than 3 years ($N = 72$), 3–6 years ($N = 60$), and more than 6 years ($N = 68$) until diagnosis.

To evaluate the predictive value of hippocampal volume and shape in a population setting, we added the remaining 311 subjects who were not in the matched set. The class label for these subjects was estimated using a classifier trained on the entire matched set. These results combined with the cross-validation results on the matched set, constituted a predicted class label and posterior probability for all 511 subjects in the cohort. Even though the vast majority of the additional 311 subjects were controls, we add these subjects so we can evaluate a cross section of a normal, elderly population.

For all classification experiments, ROC curves were computed. We used the area under the ROC curve (AUC) as a measure of predictive value. Furthermore, we calculated the sensitivity, specificity, and risk ratio for specific points on the ROC curve. These points were chosen so that they had either a sensitivity or specificity of 0.8.

In practice, any imaging biomarker for dementia will be used in combination with other variables. To investigate the predictive value of the posterior probabilities in such a situation, we used SAS 9.2 to fit a logistic regression model with the posterior probabilities of the cohort set as an independent variable and the future development of dementia as the dependent variable. We considered three models: the classical model with only volume posterior as an independent variable, the classical model extended with shape information, and finally, the new model using the shape+volume posterior. All models were corrected for age and gender by including these as covariates in the model. These regressions provide insight in the relation between posterior probabilities and actual development of dementia. In addition, the significance level can be computed for every term in every model, indicating the value of a posterior as an imaging biomarker. Finally, we investigated the overall fit of the model by comparing the log-likelihood of the models; a higher log-likelihood means a better model fit. A likelihood-ratio test was used to estimate the significance of model fit improvements. The likelihood-ratio test is based on the fact that the log-likelihood ratio of two comparable models follows a χ_1^2 distribution.

Our data were taken from a population-based study and includes subjects who have varying cognitive abilities; there were no demented subjects present at inclusion, but some subjects might have memory complaints or lesser cognitive impairment. To investigate the predictive value of hippocampal shape and volume before any symptoms arise, we stratified the data into groups that exclude different forms of decreased cognitive function. We created four subgroups by excluding groups with various levels of cognitive impairment: (1) subjective memory complaints, (2) a memory score lower than the average memory score minus one standard deviation, (3) a memory score lower than the average memory score minus one and a half standard deviations, and (4) considered having MCI 3.4. We defined subjects to belong to the MCI group if they have both a memory complaint and a memory score lower than the average minus one and a half standard deviation.

The experiments described above do not make a distinction between dementia subtypes. Our database does not have a complete subtype differential diagnosis, but we know which subjects were clinically diagnosed with AD. To see if predictive value is different for the more homogenous group of patients with AD, we repeated the classifier training and testing for a subset of the data, including only the subjects with AD.

To investigate the left-right hippocampal asymmetry, we performed the classification using only one of the two hippocampi. We did this for the left and the right hippocampus and with each feature set (volume, shape and the combination

shape+volume).

Finally, we visualized the areas Figure 3.7 that contribute most to the classifier by calculating the discriminative direction of the classifier. We used the method introduced by Golland (2002) to locally approximate the discriminative direction at a representative point. As a representative point, we selected the point where the line between the class means intersects the decision boundary.

We standardized the elements of the discriminative direction by multiplying the coefficients with the standard deviation of the feature. These standardized coefficients indicate how much each feature contributes to the final classification results. The strength and sign of the contribution for each surface-point on the classifier is color-coded Figure 3.7. The color contains the contribution of the point when moving in the direction of the surface normal.

3.3 Results

The ROC curve for both the matched set and the cohort set using volume, shape, and the shape1volume features is shown in 3.2. Shape performs best performance when the specificity is high, whereas volume performs best when the sensitivity is around 80%. Overall, the combination shape1volume outperforms both volume and shape individually. On the matched set, the AUC for volume, shape, and shape+volume posteriors was, respectively, 0.734, 0.715, and 0.769. On the cohort set, the AUC was, respectively, 0.724, 0.743, and 0.766. 3.3 shows the AUC for the matched set stratified by time to diagnosis. It can be seen that the predictive value of the volume features is very good for a short time to diagnosis, whereas more than 3 years before diagnosis the predictive value of volume decreases. For shape and especially the shape1volume features, the predictive value does not decrease when the time to diagnosis increases.

The sensitivity and specificity for the specified points on the ROC curve are presented in Table 3.2. In most cases, volume has a better specificity, sensitivity, and risk ratio than shape at the two selected points on the ROC curve. Combining volume and shape improves the sensitivity when the specificity is fixed at 80%, but only has values in between shape and volume when fixing the sensitivity at 80%. Figure 2 however shows that although at 80% sensitivity volume shows a high specificity, overall shape+volume will still outperform volume only.

For the logistic regression, we present three models in Table 3.3. In the first model, we show the results of a regression of volume, corrected for age and gender. In this model, volume is highly significant ($P < 0.0001$). In the second model, shape is added as an extra term to the regression. In the new model, volume posteriors ($P = 0.0012$) and shape posteriors ($P = 0.0071$) both are significant. In the last model, we show a model with the combined shape1volume feature posterior, corrected for age and gender. The shape+volume posterior is highly significant ($P < 0.0001$).

For the model with the volume, the log likelihood was -139.1; for the model with volume and shape, it was -135.4; and for the model with shape+volume features, it was -132.0. A likelihood-ratio test revealed that the model improvement when adding shape to the first model is significant ($P = 0.0063$). When replacing the separate shape and volume posteriors with the shape+volume posterior, the model again improves significantly ($P = 0.0098$).

In Figure 3.4, it can be seen that excluding the subjects considered to have MCI or with a memory score of lower than the mean minus 1.5 standard deviations has little effect on the results; for shape+volume features AUC decreases by 0.007 and 0.006, respectively. When excluding subjects with subjective memory complaints or a memory score lower than the mean minus 1.0 standard deviation, the AUC decreases with 0.057 and 0.072, respectively, for shape+volume features. However, even with a loss of 0.072, there is still significant predictive value with an AUC of 0.69.

In 3.5, we show a figure similar to 3.3, but using only the cases that were diagnosed with dementia of Alzheimer's type. Cases with other types of dementia, and their controls, were excluded from the dataset. This resulted in a dataset with 41 cases in the cohort set and 39 cases in the matched set. On that matched set, the AUC was 0.690, 0.662, and 0.722 for volume, shape, and shape+volume, respectively. This is lower in than in the original dataset.

Figure 3.6 shows the classification results when only using the left or right hippocampus, as well as the combination of both.

Finally, the discriminative direction of the classifier is presented in Figure 3.7. For the left hippocampus, the most influential points appear to be in the CA1 and subiculum subfields. Also, the tip of the hippocampal tail contains points with very high coefficients. For the right hippocampus, the pattern is different; there are influential points located on the inferior side of the subiculum and CA2 subfield. Additionally, there is a group of points on the head of the right hippocampus, which appears to be on the interface of the CA1 and subiculum. Lastly, the right hippocampus also shows very high coefficients at the tip of the tail.

3.4 Discussion

Hippocampal shape extracted from MRI scans is predictive for dementia before clinical symptoms arise, independent of age, gender, and hippocampal volume: we clearly show that shape contains predictive information with an AUC for shape of 0.715 on the matched set and 0.743 on the cohort set.

Stratification of the results on the time to dementia diagnosis shows that the predictive value of hippocampal volume is largest for the subjects who developed dementia soon after scan time, while with an increasing time between scan and diagnosis, the predictive value decreases. In contrast, for shape, there is no clear relation between the predictive value and time to diagnosis; this might indicate

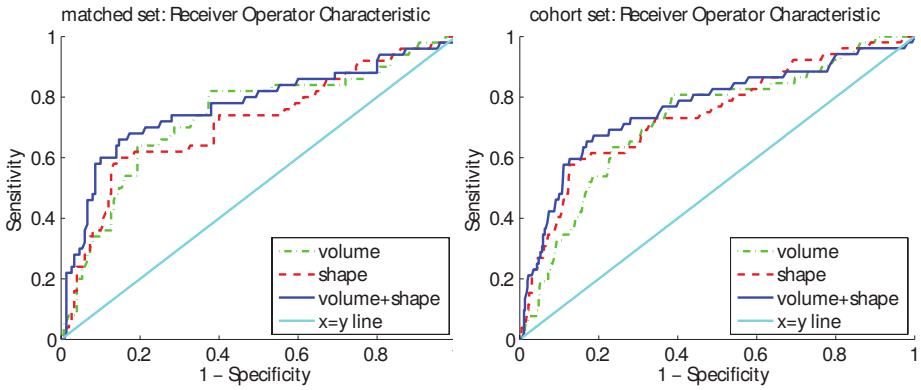


Figure 3.2: *The receiver operating characteristic (ROC) curve. Left: the matched set. right: the cohort set.*

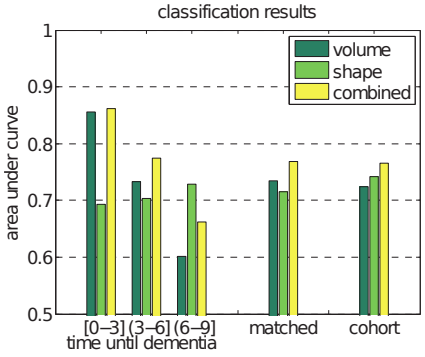
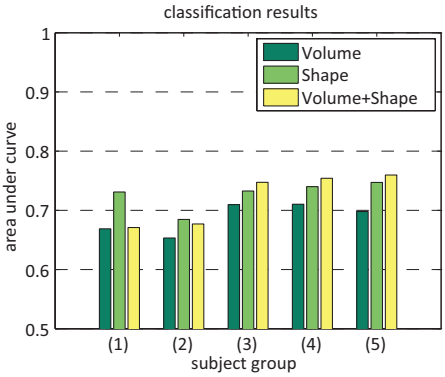


Figure 3.3: *The area under the ROC curve. The results on the matched set stratified by time to diagnosis in years, and the results on the complete matched and cohort sets. Time until dementia diagnosis is a continuous variable, and the interval is defined in interval notation. In interval notation, a bracket means inclusive and a parentheses means exclusive.*

that shape features are less dependent on the time to diagnosis. Therefore, shape seems to provide valuable information for early detection. This is in line with the findings on histology West et al. (2004); the fact that the early neuronal loss caused by AD is localized in the CA1 and subiculum subfields of the hippocampus, indicates that hippocampal shape might be more suited to detect dementia in a prodromal phase than hippocampal volume. In Figure 3.7, it appears indeed that the CA1 and subiculum subfields play a role in the early prediction of dementia.

It is not only important to identify subjects who will develop dementia early, but



group	exclusion criteria	N	N $_{cases}$
1	subjective memory complaint	358	22
2	$S_{mem} < \bar{S}_{mem} - s$	417	28
3	$S_{mem} < \bar{S}_{mem} - 1.5s$	471	43
4	MCI: subjective memory complaints and $S_{mem} < \bar{S}_{mem} - 1.5s$	497	47
5	none	511	52

Figure 3.4: *Left: The AUC results on subsets of the cohort. Right: information about subsets used. The subsets were created by excluding: (1) subjects with memory complaints, (2) subjects with a S_{mem} (memory score) lower than the population mean minus a standard deviation, (3) subjects with a S_{mem} 1.5 standard deviations lower than the population mean, (4) subjects considered MCI (memory complaints and S_{mem} 1.5 standard deviations lower than the population mean) and finally (5) the entire cohort without exclusions. N is the number of subjects left in a group after exclusion and N_{cases} is the number of cases left.*

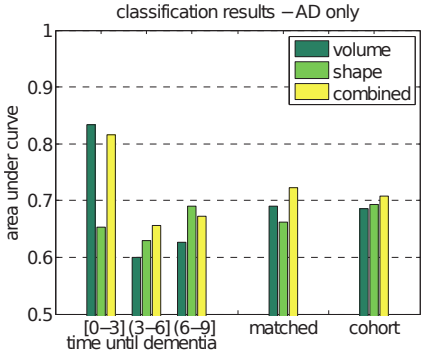


Figure 3.5: *The area under the ROC curve. The dataset only contained cases that have AD, other types of dementia were excluded. The results on the matched set stratified by time to diagnosis in years, and the results on the complete matched and cohort sets. Time until dementia diagnosis is a continuous variable and the interval is defined in interval notation. In interval notation, a bracket means inclusive and a parentheses means exclusive.*

Cohort set						
feature set	Volume		Shape		Volume + shape	
sensitivity	53.8	80.0	61.5	80.0	67.3	80.0
specicifity	80.0	61.4	80.0	45.8	80.0	56.0
risk ratio	3.92	5.60	5.05	3.17	6.29	4.60
Matched set						
feature set	Volume		Shape		Volume + shape	
sensitivity	64.0	80.0	62.0	80.0	68.0	80.0
specicifity	80.0	62.7	80.0	38.7	80.0	53.3
risk ratio	4.05	4.33	3.72	2.06	4.52	3.27

Table 3.2: *Senstivity, specificity and risk ratio for fixed locations on the ROC curve. The points on the ROC used were those that had either a sensitivity or specificity of 0.8*

	Parameter	Estimate	Wald 95% Confidence Limits	*P* value
Model 1: volume posterior (Log Likelihood: -139.0860)				
	volume	0.0575	(0.0342 – 0.0809)	< .0001
Model 2: volume and shape posterior (Log Likelihood: -135.3668)				
	volume	0.0423	(0.0167 – 0.0678)	0.0012
	shape	0.0283	(0.0077 – 0.0489)	0.0071
Model 3: shape+volume posterior (Log Likelihood: -132.0318)				
	shape+volume	0.0478	(0.0322 – 0.0633)	< .0001

Table 3.3: *Logistic regression models fitted on the cohort set. All models are corrected for age and gender; in all cases age and gender was not significant ($P > 0.5$).*

also before any symptoms arise; subjects might have memory complaints or MCIs years before dementia diagnosis, and this damage is irreversible. Excluding subjects with MCI or a very low memory score (more than 1.5 standard deviations lower than the mean) does not change the results considerably, indicating that hippocampal shape is also predictive for the development of dementia in subjects who did not yet have MCI. When we exclude subjects with subjective memory complaints or with a memory score lower than the mean minus a standard deviation, the predictive value is not as high as in the complete dataset, but we can still predict dementia with an AUC of over 0.67. This shows that hippocampal volume and shape are predictive for

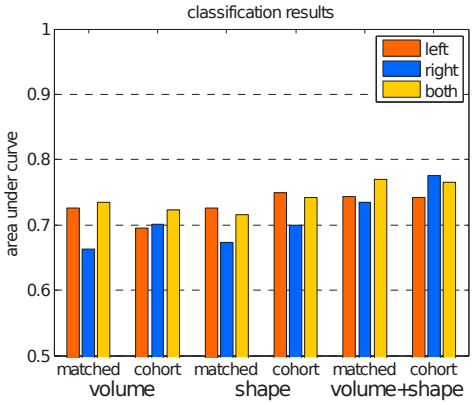


Figure 3.6: *The area under the ROC curve for left and right hippocampus separate and combined.*

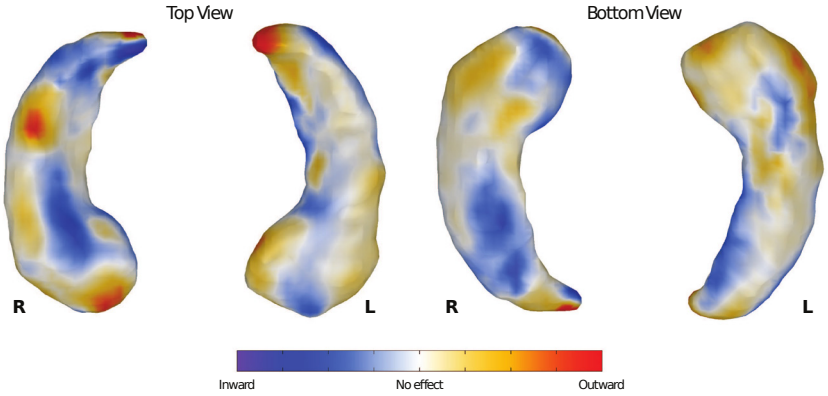


Figure 3.7: *The discriminative direction of the classifier. The colors represent coefficients of the classifier localized on the hippocampal surface. The posterior probability of developing dementia increases if the points move in the direction indicated by the colors: blue points further inward and red/yellow points further outward indicate a higher chance of developing dementia.*

dementia development in cognitively normal subjects.

Hippocampal shape provides additional predictive value over hippocampal volume: Figure 3.2 shows that classification based on shape has similar predictive value as volume and that combining shape and volume increases predictive value. When looking at the AUCs this is even more clear: shape has an AUC of 0.715 on the matched set and 0.743 on the cohort set, and volume has 0.734 on the matched set and 0.724 on the cohort set. This suggests that on the matched set volume has more predictive value than shape, while on the cohort set, shape has more predictive value

than volume. However, when shape and volume features are combined in a single classifier, in both datasets the AUC increases. These observations are reinforced when taking into account the ROC curves and the sensitivity, specificity, and risk ratio for predefined points on the ROC curve.

If hippocampal volume and shape would be used in the prediction of development of dementia, it would be combined with other predictors (e.g., genetic biomarkers) which might increase the predictive value of a model. We created simple models, corrected for age and gender, to estimate the added value of using both volume and shape over just volume (see Table 3.3). In the first model, the volume posterior probability was a highly significant term. When we added the shape posterior probability to this model, the model improved significantly and both volume and shape were significant. In the third model, the separate volume and shape posterior probabilities where substituted by the shape+volume posterior probability, which resulted in a model with a better global fit.

When considering only cases who develop AD instead of cases who develop any type of dementia, the trends in the results appear to remain the same, but the absolute performance seems to be worse. This may be explained by the smaller number of shapes available for training in these experiments. For a reliable subgroup analysis, more data will be required.

The regression analysis indicates that shape is more predictive than volume, just as the classification results show in the cohort set. However, in all cases the shape1volume features resulted in a better predictive value than volume alone; in the regression using both volume and shape posterior probabilities even increased the model fit significantly.

Relation to literature

In this article, we predicted dementia development with pattern recognition methods using hippocampal shape and volume, in a normal, elderly population. Dementia classification based on hippocampal shape has been investigated in various diagnostic studies before. Li et al. (2007) have performed classification to distinguish AD subjects from controls with up to 94.9% accuracy. Gerardin et al. (2009) reported accuracies of up to 83%, sensitivity 83%, and specificity 84% for MCI versus control classification. Ferrarini et al. (2009) reported accuracies of up to 90%, sensitivity 88%, specificity 92% for AD versus control classification and accuracies of up to 80%, sensitivity 80%, specificity 80% on MCI-converter (MCI-c) versus MCI-non converter (MCI-nc). Leung et al. (2010b) reported an AUC of up 0.67 on MCI-c versus MCI-nc subjects.

These studies use similar methods to our study, but given their case-control design were diagnostic of nature. Therefore, our work cannot be compared to the mentioned studies. Diagnostic accuracy in extreme groups (dementia or MCI patients vs. healthy controls) is very different from prediction of disease development in an asymptomatic population, as differences between subjects in the latter group are much more subtle.

In our study, all persons were nondemented at scan time and developed dementia only later. Therefore, our results support the use of shape as a predictive marker.

The most notable other imaging methods used for extracting features for dementia classification are based on voxel-based morphometry (VBM) (e.g. Fan et al., 2007; Klöppel et al., 2008) and cortical thickness (e.g. Desikan et al., 2009; Querbes et al., 2009). Cuingnet et al. (2010) compared these methods of dementia classification, using a large dataset from the AD Neuroimaging Initiative database. They compared three groups of methods: VBM, cortical thickness measurements and hippocampus volume/shape based methods. They found that for AD versus control classification the whole brain methods outperformed the hippocampus-based methods. However, for MCI-c versus control classification the hippocampal methods were competitive with the whole-brain methods. This result confirms that the hippocampus is one of the regions in the brain where atrophy is noticeable first in subjects with dementia.

We are not aware of any work using pattern recognition techniques to evaluate predictive value of hippocampal shape on similar data used in our study. There are studies which use statistical methods (e.g., regression or analysis of variance) to evaluate predictive value. Csernansky et al. (2005) and Apostolova et al. (2010) studied hippocampal shape using comparable subject groups. Their studies were more descriptive of nature making it impossible to quantitatively compare their results to our study. We can, however, qualitatively compare the discriminative direction obtained in our study to the maps obtained by Csernansky et al. Csernansky et al. (2005) (Figure. 3) and Apostolova et al. (2010) (Figure 1). For the left hippocampus, the discriminative direction maps presented in Figure 7 appears to match the atrophy and significance maps presented by Csernansky and Apostolova respectively: most influential points are found in the CA1 and Subiculum subfields. Csernansky also provides the direction of change, which corresponds with our results. For the right hippocampus, the similarity between the studies is lower: there are areas which contribute to our classification in the CA2 subfield that Csernansky or Apostolova do not find. This may be partly due to the fact that the discriminative direction in our work is based on the classifier that uses all points jointly, rather than the group differences per point as used by Csernasky and Apostolova. Also, 3.6 6 shows asymmetry in the classification performance of the left and right hippocampus, indicating that the right hippocampus might not contribute much discriminative information to the classifier.

Many studies have shown asymmetry in hippocampal volume (Karas et al., 2004; Morra et al., 2009b; Scher et al., 2011), atrophy rates (Morra et al., 2009a; Zhou et al., 2009b), or report differences in the diagnostic value of the left and right hippocampus (Csernansky et al., 2005; Tepest et al., 2008). However, the asymmetry and the direction of asymmetry are not consistent across studies. It has been suggested that the asymmetry depends on the stage of dementia; the left hippocampus is affected first by dementia related atrophy and the right hippocampus follows with a time lag (Karas et al., 2004; Morra et al., 2009b; Zhou et al., 2009b). In our data, the left

hippocampus was found to be more predictive for dementia, which fits the suggested pattern for asymmetry; in our subjects, the disease is in a very early stage, and it is possible that the left hippocampus is already affected, while the right hippocampus is still unaffected. The combination of the left and the right hippocampal features in one classifier generally improves the classification result, indicating that asymmetry might be relevant for the prediction of dementia.

The quality of the segmentation method obviously has an important influence on the accuracy of the predictions. The automatic method used in this work has been shown to produce accurate segmentations; in a leave-one-out experiment on a manually labeled subset of the dataset used in our experiments a mean SI of 0.86 was obtained (van der Lijn et al., 2008). The technique did occasionally label parts of the parahippocampal gyrus or the entorhinal cortex as foreground. However, these errors were manually corrected so we expect their influence on the classification results to be negligible. Furthermore, the leave one out experiments showed that as an atlas based method, the segmentation method has a tendency to underestimate the volume of large hippocampi and to overestimate the volume of small hippocampi. This bias toward the population mean is also likely to affect the shape of the segmentations, and could therefore negatively influence the classification accuracy. This effect may be reduced by selecting a more representative subset of atlases from a larger library. Segmentation methods based on this strategy tend to yield a higher overall accuracy than using all atlases (Aljabar et al., 2009; Barnes et al., 2008; Collins and Pruessner, 2010; Leung et al., 2010a). Unfortunately, we could not experimentally verify this, since we did not have access to a larger template library.

In our work, we represented hippocampal shape by points on the surfaces of the hippocampus and optimized point correspondence using an entropy-based particle method (Cates et al., 2007). Ferrarini et al. (2009) and Leung et al. (2010b) also use point clouds to represent hippocampal shape but optimize correspondence by using, respectively, adaptive mesh optimization (Ferrarini et al., 2007) and minimal description length (Davies et al., 2002).

Besides other correspondence optimization methods, also different shape descriptions have been used. Brechbühler et al. (1995) described shapes using spherical harmonics, which was the representation used in Gerardin et al. (2009) to describe hippocampi. Both Morra et al. (2009a) and Qiu et al. (2009) use a map-based representation for hippocampal shape, creating a deformation map to a template surface for each subject. Morra et al. (2009a) use a medial axis method for creating correspondence, where Qiu et al. (2009) uses a large diffeomorphic deformation metric matching method (LDDMM). In Qiu et al. (2008), shape differences derived by LDDMM are described directly by the deformation field, via the Jacobian of the transformation.

We decided to use the entropy-based particles method for shape correspondence modeling in our study, as it allows explicit optimization of correspondence. Moreover, an implementation of this method is publicly available and has been successfully

used to study the shape of the hippocampus in patients with schizophrenia (Cates et al., 2007)

3.5 Conclusion

In this article, we have shown that hippocampal shape extracted from MRI scans is predictive for dementia before clinical symptoms arise, independent of age and gender. Furthermore, hippocampal shape provides additional predictive value over hippocampal volume.

Shape seems to have more predictive value for subjects who develop dementia more than 6 years after scan time or exhibit less symptoms, while volume is more predictive for subjects who develop dementia shortly after scan time or exhibit more symptoms. Therefore, shape can be of great value for prodromal prediction of dementia onset.

Future treatment trials in AD will include persons with the most early and maybe even presymptomatic stage of the disease. It will therefore be increasingly important to detect persons in these stages. We show in this article that hippocampal shape may be used to partially identify these stages. Combination of hippocampal shape with other early biomarkers such as PET imaging and cerebrospinal fluid markers will be ultimately necessary for most optimal prodromal diagnosis.

Chapter 4

The Value of Hippocampal Volume, Shape and Texture for 11-year Prediction of Dementia: a population-based study

Hakim C. Achterberg
Lauge Sørensen
Frank J. Wolters
Wiro J. Niessen
M. Arfan Ikram
Meike W. Vernooij
M. Nielsen
M. de Bruijne

The Value of Hippocampal Volume, Shape and Texture for 11-year Prediction of Dementia: a population-based study.
Submitted

Hippocampal volume and shape are known MRI imaging biomarkers of dementia. 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. We investigated if hippocampal volume, shape, texture, and their combination were predictive of dementia in the general population, and determined how predictive performance varied with time to diagnosis and presence of early clinical symptoms of dementia. In 1995-1996, a random selection of 510 non-demented subjects from the population-based Rotterdam Study underwent brain MRI of whom 52 developed dementia during an 11-year clinical follow-up period. Left and right hippocampi were segmented using an automated method. From the segmentations, the volume was determined and a statistical shape model was fitted resulting in shape features, and texture features within the segmentations were computed from the corresponding brain MRI. Using these features, we trained classifiers to distinguish between subjects who developed dementia and subjects who stayed cognitively healthy. For all subjects, the probability to develop dementia was estimated using the classifiers in a cross-validation experiment. 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 (e.g., 0.935 for < 3 years, 0.809 for $3 - 6$ years, and 0.632 for > 6 years, for the combination of volume and texture). We conclude that a combination of multiple hippocampal features on MRI performs better in predicting dementia in the general population than any feature by itself. Best prediction was achieved with the combination of hippocampal volume and texture.

4.1 Introduction

Dementia is a neurological syndrome that can have varying underlying pathologies leading to neurodegeneration and cerebral atrophy. Dementia develops over the course of many years and has led to irrevocable degenerative changes in the brain by the time clinical symptoms manifest. Consequently, there is increasing need for tools to identify individuals at high risk 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.

Of specific interest is the hippocampus that is affected early in the disease process (Braak and Braak, 1997; West et al., 1994, 2004). Because the underlying pathology precedes the symptoms of cognitive decline, MRI can not only be used to diagnose dementia, but also to predict future development of dementia.

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 Alzheimer's disease (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 pathological changes of dementia such as neurofibrillary tangles (NFTs) and amyloid-β (Aβ) 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 non-demented 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.

4.2 Materials and Methods

4.2.1 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, 563 elderly subjects were randomly selected to undergo MRI of the brain in 1995-1996. 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.

Group		N	Women	Age		MMSE		SCC	MCI	Time to diagnosis	
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

Table 4.1: *Characteristics of the subjects in various subgroups.*

4.2.2 Selection of cases and controls

The entire dataset, hereafter referred to as the cohort set, contained 52 subject who developed dementia and 458 persons who did not develop dementia within the follow-up period of up to 11 years (for dementia case ascertainment, see paragraph on dementia screening below). 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: the gender had to be the same, the follow-up time of the controls should be at least as long as the time to diagnosis of the corresponding case, the age could not differ more than 1.5 years, and controls did not develop dementia during the entire follow-up period. With these criteria, it was possible to select three 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 4.1. The cohort is the same as in (Achterberg et al., 2014), except for one subject that did not have all required data for the calculation of the texture features.

4.2.3 Dementia screening and surveillance

Participants were screened for dementia at each centre visit using the Mini-Mental State Examination (MMSE) and the Geriatric Mental Schedule (GMS) organic level (de Bruijn et al., 2015). Those with MMSE$<$26 or GMS$>$0 underwent further investigation and informant interview including the Cambridge Examination for Mental Disorders of the Elderly (CAMDEX). Additionally, the entire cohort was continuously under surveillance for dementia through electronic linkage of the study centre with medical records from general practitioners and the regional institute for outpatient mental healthcare. 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.

During the study period, 52 persons 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.

4.2.4 Memory assessment and mild cognitive impairment (MCI)

MCI was defined as the combination of subjective cognitive complaints and objective cognitive impairment, in the absence of dementia. Subjective cognitive complaints were evaluated by interview, which included three 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 one 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 for memory function (immediate and delayed recall), information-processing speed (letter-digit substitution test, Stroop reading and colour-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.

4.2.5 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 x 256, flip angle 180 degrees, slice thickness 1.25 mm, acquired in sagittal direction. The images were reconstructed to a 128 x 256 x 256 matrix with a voxel dimension of 1.25 x 1.0 x 1.0 mm.

4.2.6 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 hippocampus were automatically segmented. Based on these segmentations three different types of MRI hippocampus features were computed: volume, shape, and texture.

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. Atlases consisted of manually delineated images from 20 participants from the same population as used in this study. Leave-one-out experiments on the atlas images showed mean Dice similarity indices of 0.85 +/- 0.04 and 0.86 +/- 0.02 for the left and right hippocampus 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; two cases and 69 controls were manually corrected.

Hippocampal volume

The hippocampal volume was calculated directly from the segmentation and was subsequently normalized by dividing by the intracranial volume. 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).

Hippocampal shape

The shape features used are the same as in (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 prior to computing shape features. Furthermore, an anti-aliasing step was used to smooth the binary segmentation (Whitaker, 2000).

First, the shape model was fitted to the left and right pre-processed hippocampus segmentations separately. Shapes were described by 1024 corresponding points on the surfaces. 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 hippocampus were combined into a single feature vector describing both shapes jointly. To reduce the number of features, a principal component analysis (PCA) was applied that retained 99% of the variance. In our experiments, the transformation of the PCA was estimated only on the training data and then applied to the corresponding test samples. For each cross-validation training fold, the PCA resulted in either 116 or 117 components.

Hippocampal texture

Texture features were calculated using a previously published method (Sørensen et al., 2016). In brief, the MRI scan was filtered using a Gaussian derivative-based, multi-scale, rotation-invariant filter bank comprising 28 filters (7 base filters [the three 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 histograms of each of the filter responses within both hippocampi were computed. 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.

Combination of features

Feature types were combined by concatenation. To ensure equal influence, each feature type was normalized for the sum of the eigenvalues of its covariance matrix in the training set. This ensured that each feature type contained the same amount of variance in the combined feature vector.

4.2.7 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 Vapnik, 1995) with a radial basis function kernel (RBF). 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 cases class was regularized more than the controls class in the SVM.

The hyper-parameters of the SVM were determined by optimizing the area under the receiver operating characteristic (ROC) curve (AUC) in cross-validation on the training data. 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-hyper parameters: 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 hyper-parameters 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 hyper-parameter combination. The search range for C was $e^0, e^1, e^2, ..., e^9, e^{10}$ and the search range for γ was $e^{\log\gamma(init)-4}, e^{\log\gamma(init)-3}, ..., e^{\log\gamma(init)+4}$. The linear SVM only has one hyper parameter 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.

4.2.8 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 three matching controls and trained on all remaining data in the matched set. 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 hyper-parameter optimization can be sensitive to the fold layout of the 5-fold cross-validation on the training set. We therefore report results of the experiment repeated 25 times with a different random ordering of the data, which leads to different fold layouts of the internal cross-validation. Different configurations were compared by comparing the ROC-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 three timeframes (i.e., diagnosis within 3 years after MRI (N=72); diagnosis between 3-6 years (N=60); and diagnosis more than 6 years after MRI (N=68)).

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 over all 25 runs was used. The various logistic regression models all included age, sex 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.

4.3 Results

4.3.1 Prediction of conversion to dementia

The AUCs for dementia prediction in the matched set and cohort set are provided in Figure 4.1. 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 +

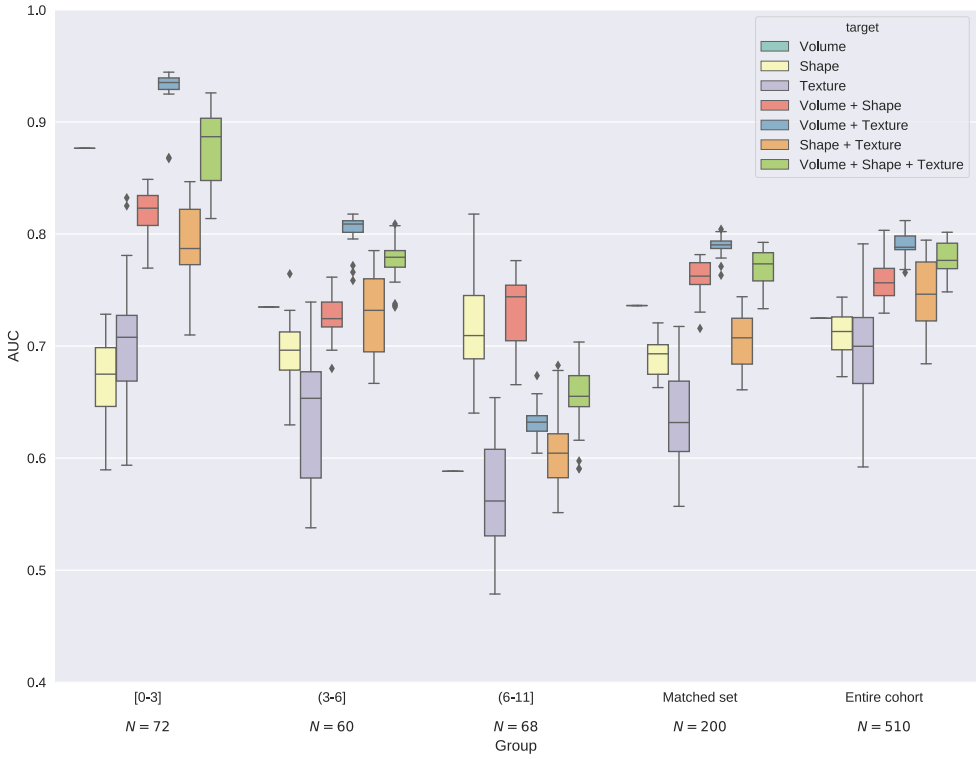


Figure 4.1: *Classification result for different feature configurations for the matched and cohort sets, and for the matched set stratificatied by time to diagnosis. The box plot shows the AUCs for the 25 random permutations of the data. The first three columns are the matched set stratified in different time intervals (in years) between scanning and follow-up dementia diagnosis.*

texture (0.707 and 0.746). In Table 4.2 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 pair-wise combinations, only volume + texture was significantly better than each single feature. The combination of all three feature types was significantly better than either shape or texture alone, but did not significantly outperform any of the pair-wise 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 Figure 4.1. Strikingly, volume has an interquartile

Matched Set		Random	Volume	Shape	Texture	V + S	V + T	S + T	All
	AUC	0.50	0.74	0.71	0.69	0.77	0.79	0.74	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	

Table 4.2: *Area under the curve for the mean posterior probability for each feature configuration and P-values for pairwise comparisons of feature configurations. P-values are obtained using the DeLong test. The colors indicate significance level. Green to yellow indicates significant, yellow to red indicates trending towards significance, and red indicates clearly non-significant differences.*

		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

Table 4.3: *Significance of feature configurations in the different subgroups.*

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.

4.3.2 Stratification by time-to-conversion

Figure 4.1 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 4.3, 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. Even though volume and

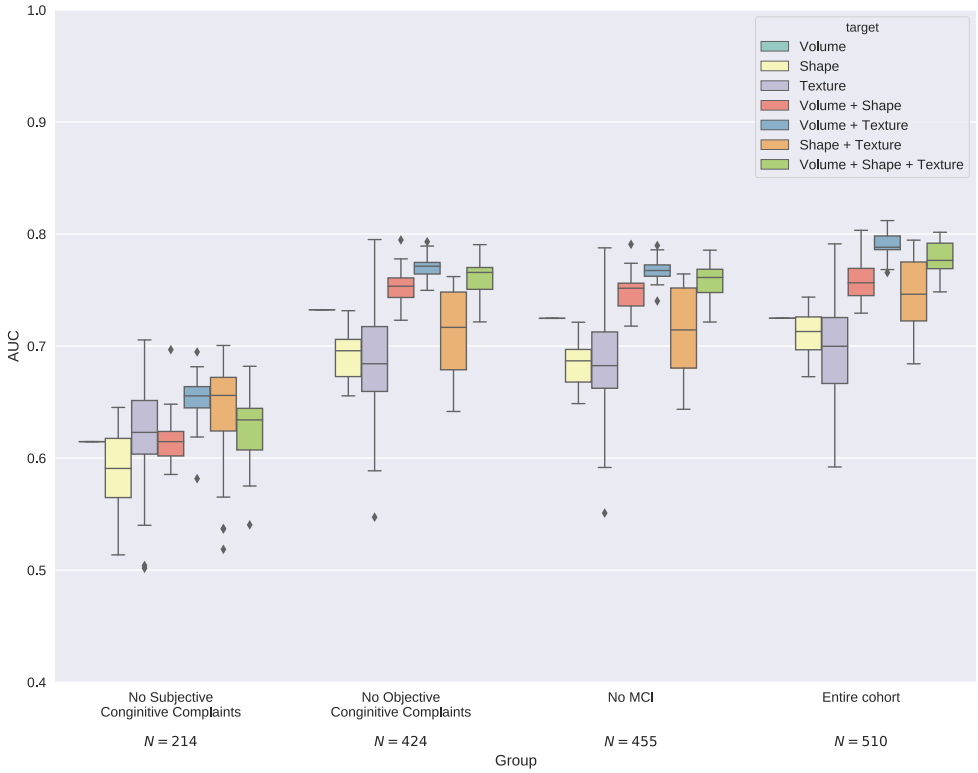


Figure 4.2: *Classification result for different feature configurations in subgroups of the cohort set where subjects with different symptoms of early dementia has been excluded. The box plot shows the AUCs for the 25 random permutations of the data. The entire cohort represents all subjects in the study, other groups are, from right to left with increasing number of subjects excluded.*

volume + shape were trending towards 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 towards significance.

4.3.3 Dementia prediction in the cognitively healthy

Figure 4.2 and Table 4.3 show the prediction performance in subgroups of the cohort set where different levels of cognitive impairment were excluded. Note that the

entire cohort in Figure 4.1 and Figure 4.2 is the same and can be used as a reference point. Excluding subjects that 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 that 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) to 0.160 (shape). None of the feature configurations reached significant predictive value. It should be noted that this subgroup only contained 12 cases.

4.3.4 Assessing dementia risk using MRI, age and sex

In Table 4.4, logistic regression models to assess dementia risk based on MRI imaging biomarkers, age and sex are presented. All models include age and sex as independent variables and have dementia as dependent (outcome) variable. All models were fitted on the entire cohort set. The initial model used only the volume posterior probability. When adding the shape posterior probability as an independent variable, both volume and shape were highly significant ($P<0.0001$). A log-likelihood ratio test showed that the improvement of the full model (with volume and shape) compared to the nested model (with only volume) was significant ($P<0.0001$). For texture, the results were the same, in the model combining volume posterior and texture posterior probabilities, both were highly significant ($P<0.0001$), and the difference between the nested and full model was also significant ($P=0.0001$). Finally, in the model including all three MRI imaging biomarkers, volume was still highly significant ($P<0.0001$), as were shape and texture ($P=0.0017$ and $P=0.0065$, respectively). The full model was significantly better than both the model with shape and volume ($P=0.006$) and the model with texture and volume ($P=0.0017$).

4.4 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.,

Parameter	Estimate	Wald 95% Confidence Limits	*P* value
Model 0: volume posterior (Log Likelihood: -151.921)			
Volume (linear)	−38.6691	(−50.6602 – −26.6779)	< 0.0001
Model 1: volume and shape posteriors (Log Likelihood: -142.790)			
Volume (linear)	−41.4203	(−53.7803 – −29.0603)	< 0.0001
Shape	6.2396	(3.3201 – 9.1591)	< 0.0001
Model 2: volume and texture posteriors (Log Likelihood: -143.964)			
Volume (linear)	−36.1053	(−48.4266 – −23.7839)	< 0.0001
Texture	5.5204	(2.7420 – 8.2988)	< 0.0001
Model 3: volume, shape and texture posteriors (Log Likelihood: -139.020)			
Volume (linear)	−38.7279	(−51.3049 – −26.1510)	< 0.0001
Shape	4.9976	(1.8741 – 8.1211)	0.0017
Texture	4.0817	(1.1438 – 7.0195)	0.0065

Table 4.4: *Logistic regression models with the MRI imaging biomarkers (posterior probabilities) as independent variables and development of dementia as the dependent variable. All models also include age and sex. Age was significant in all models, sex was not significant in any model.*

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 two feature types performed better than the best of the two individual features. However, combining all three types of features did not improve the predictive performance further. A possible explanation for this could be that shape and texture carry similar information. It could also be due to the substantial increase in dimensionality when combining shape and texture. Few studies have combined volume with shape or texture for dementia prediction, and no previous studies have combined all three MRI imaging 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 prior to non-linear classification which allowed for learning non-linear relations between the two 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.

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 imaging 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 two separate covariates.

The effect of developing future dementia on the MRI imaging 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 to 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 three time-to-diagnosis stratifications (72, 60, and 68 subjects, with only 17, 15, and 16 cases) and the trends visible in Figure 4.1 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, even though 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 non-linear 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 two 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 imaging 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 imaging 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 are features (Achterberg et al., 2014).

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 imaging biomarkers showed significant predictive performance for subjects who were cognitively healthy, up to 3 years prior to dementia diagnosis. Combining the different MRI imaging biomarkers improved the prediction; all combinations, except volume and shape, showed predictive performance in subjects without objective cognitive complaints and up to 6 years prior to dementia diagnosis.

Chapter 5

Spatially regularized shape analysis of the hippocampus using P-spline based shape regression

H.C. Achterberg
Johan J. de Rooi
Meike W. Vernooij
M. Arfan Ikram
Wiro J. Niessen
Paul H.C. Eilers
M. de Bruijne

Spatially regularized shape analysis of the hippocampus using P-spline based shape regression.
Submitted

Shape analysis is increasingly important to study changes in brain structures in relation to clinical neurological outcomes. This is a challenging task due to the high dimensionality of shape representations and the often limited number of available shapes. Current techniques counter the poor ratio between dimensions and sample size by using regularization in shape space, but do not take into account the spatial relations within the shapes. This can lead to models which are biologically implausible and difficult to interpret. We propose to use *P*-spline based regression, which combines a generalized linear model (GLM) with the coefficients described as *B*-splines and a penalty term that constrains the regression coefficients to be spatially smooth. Owing to the GLM, this method can naturally predict both continuous and discrete outcomes and can include non-spatial covariates without penalization. We evaluated our method on hippocampus shapes extracted from MR images of 510 elderly people. We related the hippocampal shape to age, memory score, and sex. The proposed method retained the good performance of ridge regression, but produced smoother coefficient fields that are easier to interpret.

5.1 Introduction

Quantitative structural neuroimaging is playing an increasingly important role in research on neurological diseases. For example, the brain tissue volumes such as grey or white matter and, brain structure volume, such as the hippocampus have been described as imaging biomarkers of disease risk and potential targets to unravel the pathophysiological pathways of disease (den Heijer et al., 2006; Ikram et al., 2010). Yet, it is increasingly understood that volume may not capture all changes in brain structure, and shape has been proposed as a potentially more sensitive marker (Achterberg et al., 2014). The relation between the shape of the hippocampus and dementia has been extensively researched. (Apostolova et al., 2010; Csernansky et al., 2005; Ferrarini et al., 2009; Gerardin et al., 2009; Li et al., 2007; Morra et al., 2009a)

Some shape analysis methods focus on dementia prediction or diagnosis based on hippocampal shape. These methods generally use a support vector machine (SVM) trained on a high-dimensional representation of the shape (Achterberg et al., 2014; Apostolova et al., 2010; Costafreda et al., 2011; Ferrarini et al., 2009; Gerardin et al., 2009; Li et al., 2007). Other methods aim to localize the areas where shape changes occur in dementia (Achterberg et al., 2014; Csernansky et al., 2005; Morra et al., 2009a). There are a number of methods that can achieve the latter: (1) aggregating shape differences in predefined regions of interest (e.g. hippocampal subfields) and subsequent analysis using standard linear multi-variate regression (Csernansky et al.,

2005), (2) point-wise statistical tests on the surface of the shapes to find points that are significantly related to the outcome (Apostolova et al., 2006, 2010; Ferrarini et al., 2009; Morra et al., 2009a,b), (3) visualizing the discriminative direction of the SVM (Achterberg et al., 2014; Zhou et al., 2009a), and (4) visualizing local atrophy maps (Csernansky et al., 2005).

As anatomical structures are generally smooth, neighbouring points are expected to behave in a similar way, whereas distant points may not influence each other. This correlation between points will lead to an ill-posed regression problem if no regularization is used. None of the aforementioned shape analysis methods explicitly model the spatial correlation between points in the shapes. Point-wise statistical testing ignores relations between points altogether, while machine learning approaches—though they can model point interactions—ignore the spatial relation between points.

By using spatial information for regularization, the resulting regression problem becomes well-posed and the resulting models will be more realistic, reflected in better interpretable coefficient maps (Cuingnet et al., 2012; Sabuncu et al., 2011). Also, it could potentially improve the model fit and generalizability of the model. To perform shape analysis explicitly incorporating spatial information, we use a P-spline based regression model (Marx and Eilers, 1999). This approach is based on a generalized linear model (GLM) (McCullagh and Nelder, 1989) and uses B-splines to model a smooth coefficient field with an extra spatial regularization term to further control smoothness. The use of the GLM ensures a strong basis in traditional statistics. The spline representation of the regression coefficients and additional spatial smoothing regularization make it suitable for the problem of shape analysis. The GLM uses a link function that allows for multiple outcome types (e.g. linear, classes, counts) and can include multiple covariates from different sources (with different penalties) in a natural way.

We applied our method on both synthetic as well as real hippocampus shapes. The shapes of the left hippocamus of 510 elderly subjects were extracted and related to age, memory score, and sex, demonstrating the ability to use the model for both linear (age and memory score) and logistic (sex) regression. Also, we visualized the regression coefficients on the hippocampal surface to investigate their interpretability.

5.2 Methods

In section 5.2.1 we explain the P-spline based regression model. For brevity we will refer to it as P-spline regression in the remainder of the paper. Then in section 5.2.2 we describe the shape representation.

In this paper we use a consistent notation where a bold upper case letter denotes a matrix (e.g. $\mathbf{M}$), a bold lower case letter denotes a vector (e.g. $\mathbf{y}$), a normal upper case letter denotes a function (e.g. $F(x)$) and a normal lower case letter denotes a scalar (e.g. c). In some cases subscripts indicate the size of the vector/matrix, in

which case the subscript will be in the form $m \times n$ for matrices or $n \times 1$ for vectors (e.g. $\mathbf{M}_{m\times n}$ or $\mathbf{y}_{n\times 1}$).

5.2.1 Shape regression using P-splines

We first describe linear P-spline regression in 1D to avoid some complexities with notation. Subsequently, we extend the method to the GLM, which allows for different types of outcome variables. Considering that the hippocampus can be described by a 2D surface living in a 3D space, we briefly describe how the method is extended to 2D. Finally, we describe the model selection used for estimating the hyper-parameters.

5.2.1.1 P-spline regression in 1D

In regression, we try to find the relation between an observed response vector $\mathbf{y}_{m\times 1}$ and design matrix $\mathbf{X}_{m\times p}$, where m is the number of samples and p the number of independent variables. In this section we restrict ourselves to the situation where $\mathbf{y}_{m\times 1}$ is a continuous variable, in the next section this will be generalized for other types of responses. We aim to minimize $Q(\boldsymbol{\alpha}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2$, where $\boldsymbol{\alpha}_{p\times 1}$ are the regression coefficients. In shape regression, typically p is much larger than m, leading to an ill-posed problem. To obtain a realistic estimate, we need to constrain $\boldsymbol{\alpha}$.

In medical image analysis the shapes encountered are generally smooth (e.g. organs, brain structures). Accordingly, we expect the differences between shapes to be smooth. Hence, given the problem of shape analysis, we constrain $\boldsymbol{\alpha}$ to be spatially smooth.

To achieve this, we assume that $\boldsymbol{\alpha}$ is sampled from a smooth, continuous coefficient field. We describe $\boldsymbol{\alpha}$ by a linear combination of smooth B-splines, $\boldsymbol{\alpha}_{p\times 1} = \mathbf{B}_{p\times n}\boldsymbol{\beta}_{n\times 1}$, where $\mathbf{B}$ is a matrix encoding the B-spline basis functions, $\boldsymbol{\beta}$ is a new set of regression coefficients and n is the number of B-spline basis functions used. This leads to

$$Q(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\mathbf{B}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{U}\boldsymbol{\beta}\|^2, \tag{5.1}$$

where $\mathbf{U}_{m\times n} = \mathbf{X}\mathbf{B}$. Depending on the number of used B-spline basis functions, the matrix $\mathbf{U}$ can become full column rank ($m > \mathrm{rank}(\mathbf{U})$). This can be seen as a dimension reduction by using $\mathbf{U}$ as a new matrix of observations to estimate the unknown regression coefficients $\boldsymbol{\beta}$. If the number of used B-spline basis functions is chosen considerably lower than m, the linear regression problem becomes well-posed.

Note that since $\mathbf{X}$ is captured in $\mathbf{U}$, we do not require $\mathbf{X}$ any more in subsequent equations. This is advantageous, as the number of columns in $\mathbf{U}$ is much smaller than in $\mathbf{X}$.

As in Eilers and Marx (1996), we place a generous amount of equidistant knots that will make $\boldsymbol{\alpha}$ overly flexible. Subsequently, $\boldsymbol{\alpha}$ is constrained to be smooth by

adding a difference penalty on the B-spline coefficient vector $\boldsymbol{\beta}$:

$$Q^* = Q + P = \|\mathbf{y} - \mathbf{U}\boldsymbol{\beta}\|^2 + \lambda \sum_{k=d+1}^{n} (\Delta_k^d \beta)^2. \tag{5.2}$$

With P the difference penalty, Δ_k^d the kth finite difference operator of order d and λ a nonnegative regularization parameter. λ Influences the amount of penalization: $\lambda = 0$ leads to an unpenalized solution, whereas a very large λ will make the dth derivative of $\boldsymbol{\beta}$ zero, such that the coefficients are described by a polynomial of degree $d - 1$. When $\lambda = 0$ this is the normal solution for linear regression. Increasing λ will make $\boldsymbol{\beta}$ smoother. The optimal λ can be selected using cross-validation.

The penalty term P can be written as a multiplication of $\boldsymbol{\beta}$ and a banded matrix $\mathbf{D}_d$ representing the difference operator Δ_k^d. $\mathbf{D}_d$ can be calculated recursively with $\mathbf{D}_0 = I_{n\times n}$ and the elements of $\mathbf{D}_d$ being the column wise difference of $\mathbf{D}_{d-1}$. We can now write the penalty as:

$$P = \lambda \boldsymbol{\beta}^T \mathbf{D}_d^T \mathbf{D}_d \boldsymbol{\beta}. \tag{5.3}$$

The minimization of Q^* leads to the following system of equations for $\boldsymbol{\beta}$:

$$(\mathbf{U}^T\mathbf{U} + \lambda \mathbf{D}_d^T \mathbf{D}_d)\boldsymbol{\beta} = \mathbf{U}^T\mathbf{y}. \tag{5.4}$$

Note that for $d = 0$, Eq. 5.4 simplifies to ridge regression (Hoerl and Kennard, 1970).

5.2.1.2 The generalized linear model for P-spline regression

The GLM is a generalization of the ordinary linear regression that suits response distributions from the exponential family (e.g. the normal, binomial and Poisson distribution). In Marx and Eilers (1999) it is shown that P-spline regression can also be used with the GLM. The representation of the regression coefficients by B-splines can be achieved by simply substituting $\mathbf{X}\boldsymbol{\alpha}$ with $\mathbf{XB}\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\beta}$ in the link function, which relates the linear model with the response variable.

Adding the penalty term to a GLM is slightly different from adding it to an ordinary linear regression; rather than minimizing the sum of squared differences, the GLM maximizes the log-likelihood of $\boldsymbol{\beta}$ given the observations, $l(\boldsymbol{\beta}; \mathbf{U}, \mathbf{y})$. Using the method of scoring iterative equations this simplifies to

$$\hat{\boldsymbol{\beta}}_t = (\mathbf{U}^T\hat{\mathbf{V}}_{t-1}\mathbf{U})^{-1}\mathbf{U}^T\hat{\mathbf{V}}_{t-1}\hat{z}_{t-1}, \tag{5.5}$$

where $\hat{\mathbf{V}} = \mathrm{diag}(\hat{v}_{ii}) = \mathrm{diag}\{[h'(\hat{\eta}_i)]^2/\mathrm{var}(\mathbf{Y}_i)\}$, η is the linear predictor, and h' is the derivative of the inverse link function. The working vector $\hat{z}$ has entries $\hat{z}_i = (y_i - \hat{\mu}_i)/(h'(\hat{\eta}_i) + \hat{\eta}_i)$. Convergence of Eq. 5.5 gives the (unpenalized) maximum likelihood parameter estimates.

To get the penalized maximum likelihood, the original log-likelihood function is modified by subtracting the penalty term from the original log-likelihood function, which becomes

$$l(\boldsymbol{\beta}; \mathbf{U}, \mathbf{y})^* = l(\boldsymbol{\beta}; \mathbf{U}, \mathbf{y}) - \frac{1}{2}\lambda \boldsymbol{\beta}^T \mathbf{D}_d^T \mathbf{D}_d \boldsymbol{\beta}, \tag{5.6}$$

which is analog to Eq. 5.2. The factor $\frac{1}{2}$ is required to cancel out the factor of 2 that appears when differentiating the penalty.

Maximization of the penalized log-likelihood in Eq. 5.6 leads to a small change in the scoring algorithm in Eq. 5.5,

$$\hat{\boldsymbol{\beta}}_{\lambda,t} = (\mathbf{U}^T \hat{\mathbf{V}}_{t-1} \mathbf{U} + \lambda \mathbf{D}_d^T \mathbf{D}_d)^{-1} \mathbf{U}^T \hat{\mathbf{V}}_{t-1} \hat{z}_{t-1}, \tag{5.7}$$

Eq. 5.7 can be viewed as a penalized form of an iterative weighted regression of the working vector on $\mathbf{U}$, where $\hat{\mathbf{V}}$ and $\hat{z}$ depend on the choice of λ. On convergence with λ fixed, we obtain the estimated smooth coefficient vector, $\hat{\boldsymbol{\alpha}} = \mathbf{B}\hat{\boldsymbol{\beta}}_\lambda$.

5.2.1.3 2D *P*-spline regression

In our application, the hippocampal surface is described as a 2D signal, as explained in Subsection 5.2.2. Therefore, we require a version of *P*-spline regression which works on 2D surfaces. For this, we describe the shape as a combination of tensor product *B*-splines as proposed in Marx and Eilers (2005).

An introduction to tensor product *B*-splines is provided in Eilers and Marx (2003) and a more complete description is given by Dierckx (1993). Tensor product *B*-spines exist in the $\nu \times \breve{\nu}$ plane and are the product of two univariate *B*-splines, B_r and $\breve{B}_s$. In our case, the 2D *B*-spline basis is defined by a regular grid of $n \times \breve{n}$ *B*-splines with an unknown coefficient matrix $\boldsymbol{\Gamma}_{n\times\breve{n}} = [\gamma_{rs}]$. For a given *B*-spline knot grid, a flexible surface can be approximated at, for example, the digitized coordinates. For $j = 1, ..., p$ and $k = 1, ..., \breve{p}$,

$$\alpha(\nu_j^*, \breve{\nu}_k^*) = \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} B_r(\nu_j^*) \breve{B}_s(\breve{\nu}_k^*) \gamma_{rs}. \tag{5.8}$$

Using an "unfolded" matrix notation, Eq. 5.8 can be written as

$$\mathrm{vec}\{\boldsymbol{\alpha}\} = \mathbf{T}^* \boldsymbol{\gamma}, \tag{5.9}$$

where vec is a vectorization operator that creates a vector containing all elements of a given matrix, $\boldsymbol{\gamma} = \mathrm{vec}(\boldsymbol{\Gamma})$ and the matrix $\boldsymbol{\alpha}_{p\times\breve{p}} = [\alpha(\nu_j^*, \breve{\nu}_k^*)]$. $\mathbf{T}^*{}_{p\breve{p}\times n\breve{n}}$ is a matrix containing the tensor product of the *B*-spline basis functions. We follow the definition in Marx and Eilers (2005), section 2.1.

Given an $m \times p\breve{p}$ regressor matrix $\mathbf{X}$ with vectorized 2D observations, we can use *B*-splines to model the regression coefficients. Similar to Eq. 5.1, we use $\boldsymbol{\alpha}_{p\breve{p}} = \mathbf{T}^*_{p\breve{p}\times n\breve{n}} \boldsymbol{\gamma}_{n\breve{n}}$ to obtain,

$$Q(\boldsymbol{\gamma}) = \|\mathbf{y} - \mathbf{X}\mathbf{T}^*\boldsymbol{\gamma}\|^2 = \|\mathbf{y} - \mathbf{M}\boldsymbol{\gamma}\|^2, \tag{5.10}$$

with $\mathbf{M} = \mathbf{X}\mathbf{T}^*$. Even though using tensor product B-splines, the dimension of the estimation is reduced from $p\breve{p}$ to $n\breve{n}$, the estimation can still be ill-posed for moderately complex surfaces.

In the spirit of 1D P-spline regression, a penalty term is added to the differences of the B-spline coefficients. Because the knot spacing in both dimensions might not be equal, separate row (P_{row}) and column penalties (P_{column}) can be added. Considering that in 2D the dimensionality is very high, we also add a penalty similar to the one used in ridge regression: a penalty on the L2-norm of the B-spline coefficients (P_{ridge}). This leads to

$$\begin{aligned} Q^*(\boldsymbol{\gamma}) &= Q(\boldsymbol{\gamma}) + P_{\text{row}}(\boldsymbol{\gamma}) + P_{\text{column}}(\boldsymbol{\gamma}) + P_{\text{ridge}}(\boldsymbol{\gamma}) \\ &= \|\mathbf{y} - \mathbf{M}\boldsymbol{\gamma}\|^2 + \lambda_1\|\mathbf{P}_1\boldsymbol{\gamma}\|^2 + \lambda_2\|\mathbf{P}_2\boldsymbol{\gamma}\|^2 + \lambda_3\|\boldsymbol{\gamma}\|^2, \end{aligned} \tag{5.11}$$

where λ_1, λ_2 and λ_3 are the penalty parameters for the rows, the columns and a ridge term respectively, and $\mathbf{P}_1$ and $\mathbf{P}_2$ are the appropriate versions of $\mathbf{D}$ used in the 1D case for the row and column penalty. More details about how these matrices are constructed can be found in Marx and Eilers (2005).

Similar to the 1D P-splines the method can also be applied to the GLM. The link function then uses $\mathbf{M}\boldsymbol{\gamma}$ instead of $\mathbf{X}\boldsymbol{\alpha}$, and the GLM for 2D observations becomes:

$$l(\boldsymbol{\gamma}; \mathbf{M}, \mathbf{y})^* = l(\boldsymbol{\gamma}; \mathbf{M}, \mathbf{y}) - \lambda_1\|\mathbf{P}_1\boldsymbol{\gamma}\|^2 - \lambda_2\|\mathbf{P}_2\boldsymbol{\gamma}\|^2 - \lambda_3\|\boldsymbol{\gamma}\|^2. \tag{5.12}$$

This function can be maximized by iteratively weighted least-squares using in the iterative function

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{\lambda_1,\lambda_2,t} =& (\mathbf{M}^T\hat{\mathbf{V}}_{t-1}\mathbf{M} + \lambda_1\mathbf{P}_1^T\mathbf{P}_1 + \lambda_2\mathbf{P}_2^T\mathbf{P}_2 + \lambda_3\mathbf{I})^{-1} \\ & \times \mathbf{W}^T\hat{\mathbf{V}}_{t-1}\hat{\mathbf{z}}_{t-1}, \end{aligned} \tag{5.13}$$

where the weight matrix $\mathbf{W} = \text{diag}\{h(\hat{\eta}_i)\}/var(Y)$.

For a given λ_1, λ_2, and λ_3, the sandwich operator of the regression coefficients is given by:

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\gamma}}) \approx& (\mathbf{M}^T\mathbf{M} + \lambda_1\mathbf{P}_1^T\mathbf{P}_1 + \lambda_2\mathbf{P}_2^T\mathbf{P}_2 + \lambda_3\mathbf{I})^{-1} \\ & \times \mathbf{M}^T\mathbf{M} \\ & \times (\mathbf{M}^T\mathbf{M} + \lambda_1\mathbf{P}_1^T\mathbf{P}_1 + \lambda_2\mathbf{P}_2^T\mathbf{P}_2 + \lambda_3\mathbf{I})^{-1}. \end{aligned} \tag{5.14}$$

This can be translated back to points on the surface by $\text{var}(\hat{\alpha}) = \boldsymbol{\Gamma}^* \text{var}(\hat{\boldsymbol{\gamma}})\boldsymbol{\Gamma}^{*T}$. The standard error can be computed by taking the square root of the diagonals of $\text{var}(\hat{\alpha})$.

5.2.2 Corresponding shape description

P-spline regression requires the shapes to be represented as a function in the form $f : \mathbb{R}^2 \to \mathbb{R}$ that can be regularly sampled in points on a rectangular grid. Considering

that the hippocampus shape is roughly cylindrical, we describe it by the radius of the hippocampus as a function of the position along the principal curve s and the angle φ, leading to $\rho(s, \varphi)$. We use deformable image registration to create a mean space and define correspondences between individual shapes using this mean space.

To create an unbiased mean shape template we use a method similar to Seghers et al. (2004). Given a set of K binary masks representing shapes, S_k with $k = 1, ..., K$, the goal is to create a mean shape $\overline{S}$ defined as the average in a common space Ω_c:

$$\overline{S}(\mathbf{x}) = \frac{1}{K} \sum_{i=0}^{K} S_i(T_i(\mathbf{x})), \tag{5.15}$$

with $T_i(\mathbf{x}) : \Omega_c \rightarrow \Omega_{S_i}$ the point transformation from the common space to the subject space Ω_{S_i} and $S_i(T_i(\mathbf{x}))$ representing the deformed shape.

The transformations $T_i(\mathbf{x})$ are derived from pairwise image registrations. To compute T_i, shape S_i is registered to all shapes S_j, resulting in a set of transformations $R_{i,j}(\mathbf{x})$. When image S_i is used as fixed image we define the transformations $R_{i,j}(\mathbf{x}) : \Omega_{S_i} \rightarrow \Omega_{S_j}$. By averaging the transformations $R_{i,j}$, the mean transformation of image S_i, $R_i(\mathbf{x}) : \Omega_{S_i} \rightarrow \Omega_c$, is obtained:

$$R_i(\mathbf{x}) = \frac{1}{K} \sum_{j=1}^{K} R_{i,j}(\mathbf{x}), \tag{5.16}$$

The transformation $T_i(\mathbf{x})$ is then calculated by inverting R_i, $T_i(\mathbf{x}) = R_i^{-1}(\mathbf{x})$. Note that the identity transformation $R_{i,i}$ is also taken into account in equation (5.16).

From the mean shape $\bar{S}$ a principal curve $C_c(s) : (I) \rightarrow \Omega_c$ is derived using the method presented by Hastie and Stuetzle (1989)[1]. A principal curve is a one-dimensional curve that passes through the middle of a p-dimensional dataset. In our case the curve passes through the middle of the hippocampus.

At equidistant points along the principal curve we define a normal plane spanned by a normal vector $\boldsymbol{n}$ and binormal vector $\boldsymbol{b}$ of the principal curve. We calculate the tangent $\boldsymbol{t} = \frac{C_s}{ds}$ by central finite differences of the curve (the normal using $\boldsymbol{n} = \boldsymbol{t} \times \boldsymbol{f}$, with $\boldsymbol{f}$ a fixed vector, and the binormal by $\boldsymbol{b} = \boldsymbol{t} \times \boldsymbol{n}$). In every normal plane we extra a number of equiangular vectors from the principal curve to the surface and place points on the hippocampal surface by finding the intersection of the vectors and $\overline{S}$.

Using the point transforms $\mathbf{T}_i(\mathbf{x})$ the points on the principal curve and the corresponding surface points are transformed back to the individual shape spaces Ω_i. This ensures correspondence between the points on the curve and the directions in which we sample.

To sample each shape S_i, the vectors originating from the principal curve C_i pointing towards the surface point are again followed to the point where the vector

[1] We used the the R package `princurve` which is an R port of the original S code from Trevor Hastie

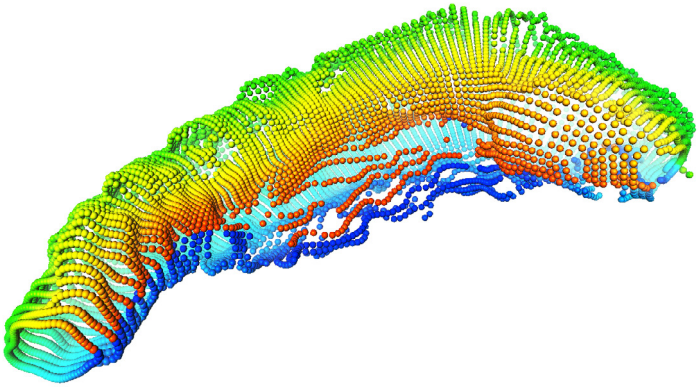


Figure 5.1: *An example of the sample of the hippocampus. The color of the points depicts the angle component of the coordinate φ.*

crosses the boundary of the shape in S_i. We then define $\rho_i(s,\varphi)$ as the distance between the principal curve C_i and the surface point. The design matrix $\mathbf{X}$ has one column for each combination of s and φ and one row for each shape. The entries of $\mathbf{X}$ are the corresponding $\rho_i(s,\varphi)$.

5.2.3 Coefficient field description

In P-spline regression, the regression coefficients are described by a B-spline grid (e.g. $\mathbf{B}\boldsymbol{\beta}$). For this we put a two dimensional B-spline grid on the shape parametrization $\rho_i(s,\varphi)$. The B-spline grid is normal in the direction of s, and cyclic in the direction of φ. This ensures a cylindrical B-spline grid to match the topology of the hippocampus. To create a cyclic grid, the matrix $\mathbf{D}$ in Eq. 5.3 is supplemented with extra rows. For a first order difference matrix $\mathbf{D}_1$ will be of the form:

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{5.17}$$

Note that the last row wraps around, and would not be present in the non-cyclic B-spline basis. The number of rows added will be equal to the order d of $\mathbf{D}_d$.

5.2.4 Coefficient field and shape effect field

The regression we propose in this work predicts an outcome variable in a subject (e.g. age) based on the shape of a brain structure. As we assume that the causality

is reversed, the shape is affected by the outcome, these models are called backward models. For interpretation it is more relevant to visualize the original direction, i.e. how the shape changes given changes in the outcome variable (the corresponding forward model). Haufe et al. (2014) proposed a method to turn a backward model into the corresponding forward model using:

$$\mathbf{A} = \mathbf{\Sigma}_{\mathbf{x}} \gamma \mathbf{\Sigma}_{\hat{\mathbf{y}}}^{-1}. \tag{5.18}$$

This turns a coefficient field γ into what we call a shape effect field ($\mathbf{A}$).

5.3 Experiments and Results

To test the proposed *P*-spline regression method, we perform a number of experiments both on real and synthetic data. This section will first describe both the real data and the synthetic data, followed by a short description of the evaluation. Finally, the experiments and their results are presented.

5.3.1 Data

5.3.1.1 MRI data

The imaging data used in this study was obtained from the Rotterdam Scan Study, a longitudinal MRI study on age-related diseases (den Heijer et al., 2003). In the period 1995-1996, 510 non-demented elderly subjects (223 female) of 55 years and older were scanned on a Siemens 1.5T scanner. The sequence used was an inversion recovery, 3D half-Fourier acquisition single-shot turbo spin echo sequence (HASTE) with inversion time 4400 ms, repetition time 2800 ms, effective echo time 29 ms, matrix size 192×256, flip angle 180 degrees, slice thickness 1.25 mm, acquired in sagittal direction. The images were reconstructed to a $128 \times 256 \times 256$ matrix with a voxel dimension of $1.25 \times 1.0 \times 1.0$mm.

The average age was 73.5 years, with a standard deviation of 7.9 years. The youngest subjects was 59.0 years old, and the oldest subject was 89.8 years old. The memory score of the subjects was assessed using a 15 word delayed (15 minutes) recall task. The average memory score was 5.82, with a standard deviation of 2.66, and ranged from 0 to 15.

5.3.1.2 Hippocampus segmentation

Hippocampi were automatically segmented using a segmentation 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. As training data we used manually delineated images from 20 participants of the same population. In

leave-one-out experiments on the training images a mean Dice similarity index of 0.85 ± 0.04 (van der Lijn et al., 2008) was achieved. The final segmentation results of the 510 images used in this study were inspected by a trained observer and manually corrected in case of large errors. To save computational time and to for visualization purposes, we only considered the left hippocampus in this work.

5.3.1.3 Synthetic data

To test the performance of the model under different conditions, we created a number of synthetic datasets which are referred to as phantoms. A single phantom is composed of the entire design matrix for an experiment $\mathbf{X}^{\text{phantom}}$ and corresponding outcome vector $\mathbf{y}^{\text{phantom}}$. The separate shape samples $\mathbf{X}_i^{\text{phantom}}$ are the rows in the design matrix.

Each sample $\mathbf{X}_i^{\text{phantom}}$ in a phantom $\mathbf{X}^{\text{phantom}}$ is composed of three different contributions:

1. a fixed signal encoding the outcome of the regression (signal, $\mathbf{X}_i^{\text{signal}}$).
2. a random variable representing a plausible hippocampus shape (base shape, $\mathbf{X}_i^{\text{baseshape}}$).
3. a random variable representing noise on the individual surfaces landmarks ($\mathbf{X}_i^{\text{noise}}$).

The final sample $\mathbf{X}_i^{\text{phantom}}$ is a weighted combination of the three contributions

$$\mathbf{X}_i^{\text{phantom}} = w^s \mathbf{X}_i^{\text{signal}} + w^b \mathbf{X}_i^{\text{baseshape}} + w^n \mathbf{X}_i^{\text{noise}}. \tag{5.19}$$

The outcomes $\mathbf{y}_i^{\text{phantom}}$ are random variables drawn from a normal distribution. The relation between the samples and the outcome of a phantom is given by $\mathbf{y}_i^{\text{phantom}} \propto \mathbf{X}_i^{\text{signal}} \boldsymbol{\beta}^{\text{phantom}}$, with $\boldsymbol{\beta}^{\text{phantom}}$ being the ground truth regression coefficients. Figure 5.2 shows the components of the phantom. The signal $\mathbf{X}_i^{\text{signal}}$ is a Gaussian blob in the 2D shape representation space, and therefore the ground truth shape effect field is also a Gaussian blob.

The base shape signal contributions $\mathbf{X}_i^{\text{baseshape}}$ are random instantiations from a shape model built from the real imaging data. The shape model was created by a principal component analysis (PCA) on the segmentations using the representation described in section 5.2.2. The shape model consists of shape components which are weighted by a set of coefficients. Assuming that these coefficients follow a Gaussian distribution, we can use the shape model to generate new shape samples.

We created sets of phantoms with a varying signal-to-noise ratio (SNR). To this end, the noise contributions ($\mathbf{X}_i^{\text{baseshape}} + \mathbf{X}_i^{\text{noise}}$) were fixed by setting w^b and w^n, and then w^s was chosen so that the SNR of the phantom had the desired value. For

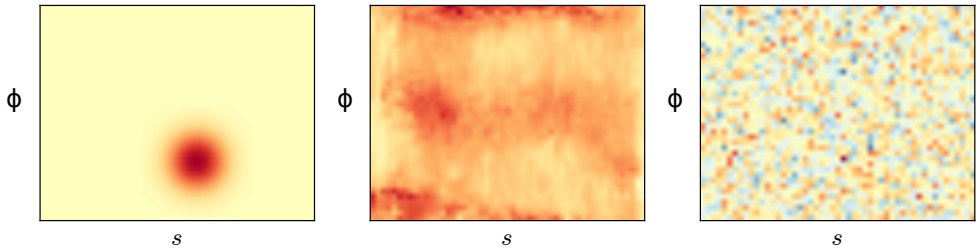


Figure 5.2: *Examples of the different components of the phantoms. From left to right: the ground truth signal, the base shape signal, and the Gaussian noise component.*

these experiments we defined the SNR of the phantom as the maximum SNR over the entire shape surface:

$$\mathrm{SNR}(\mathbf{X}^{\mathrm{phantom}}) = \max_{s,\varphi} \frac{\sigma(\mathbf{X}^{\mathrm{signal}})}{\sigma(\mathbf{X}^{\mathrm{baseshape}} + \mathbf{X}^{\mathrm{noise}})}, \tag{5.20}$$

with σ the sample standard deviation of $\rho_i(s, \varphi)$ over the different phantom samples.

We created two series of phantoms: (1) a series of phantoms describing a mean hippocampus shape with only a signal and Gaussian noise added, which will refer to as *mean shape phantoms* ($w^b = 0, w^n = c$), and (2) a series of phantoms following the shape distribution found in real data with only the signal added, which we will refer to as *hippocampus shape phantoms* ($w^b = 1, w^n = 0$). Both series contained phantoms for ten different SNRs, ranging from 0 to 16, and 1000 shape samples per phantom that were split in half to create a training and a test set. The constant c was chosen in such a way that the amount of Gaussian noise is on average the same as the natural shape distribution found in the real data.

5.3.2 Evaluation

5.3.2.1 Regression models

We evaluate three different methods: *P-spline* regression, *P-spline + L2* regression and *ridge* regression.

Ridge regression is a simple, yet effective, regularized regression method. It adds a L2-norm penalty term to the standard linear regression, leading to $Q(\boldsymbol{\alpha}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2 + \|\boldsymbol{\alpha}\|^2$. For the GLM the likelihood function can be defined as $l(\boldsymbol{\beta}; \mathbf{U}, \mathbf{y})^* = l(\boldsymbol{\beta}; \mathbf{U}, \mathbf{y}) - \frac{1}{2}\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$ and solved using iterative weighted least squares.

For *P*-spline regression we use the formulation as in Eq. 5.12. We use two different versions of the model: (1) *P*-spline regression with $\lambda_3 = 0$, disabling the ridge term, and (2) *P*-spline regression with the ridge term enabled, but with $\lambda_1 = \lambda_2$ to avoid an excessive amount of hyper-parameters. In the remainder of the paper we refer to the first model as *P-spline* and to the second model as *P-spline + L2*.

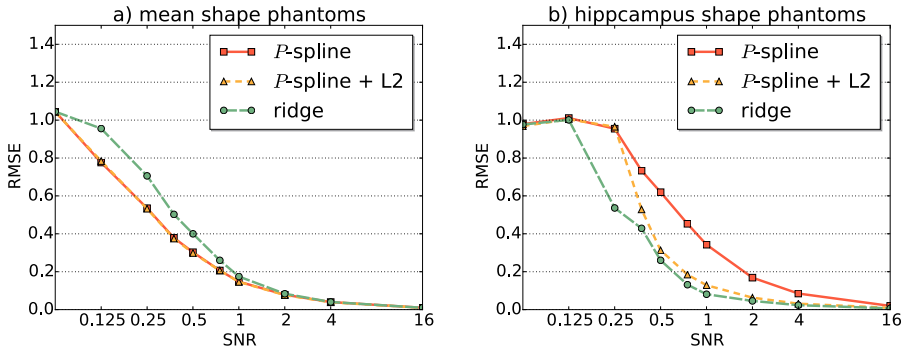


Figure 5.3: *a) result of the* mean shape phantoms*, b) result of* hippocampus shape phantoms.

5.3.2.2 Model selection

The used regression models have a number of hyper-parameters that need to be optimized. These model parameters are estimated using a 5-fold cross-validation procedure on the training data. In the cross-validation procedure we use a grid-search over a range of parameters. The optimal parameters are chosen by maximizing the cross-validated log-likelihood.

5.3.2.3 Performance metrics

For linear and logistic regression two different performance measures were used to compare the three regression models. For linear regression, root mean square error (RMSE) was used. For logistic regression, the results were evaluated with the area under the receiver operator characteristic curve (AUC). To obtain meaningful performance metrics, leave-one-out cross-validation was used for the real data and independent train and test sets were created for the phantom data. This ensured that all the test samples were unseen by the tested model.

5.3.3 Phantom experiments

The performance of the different regression models for *mean shape phantoms* is depicted in Figure 5.3a. For a low signal to noise ratio, *ridge* regression has a higher error rate, but when the SNR increases to 2 (meaning the maximum SNR over the surface is 2, see Eq. 5.20), all methods seem to converge to the same performance. In Figure 5.4 we show some example coefficient and shape effect fields. From these fields it is clear that the smoothing of the *P-spline + L2* is very efficient in removing noise and has a positive effect on the interpretability of the coefficient and shape effect fields.

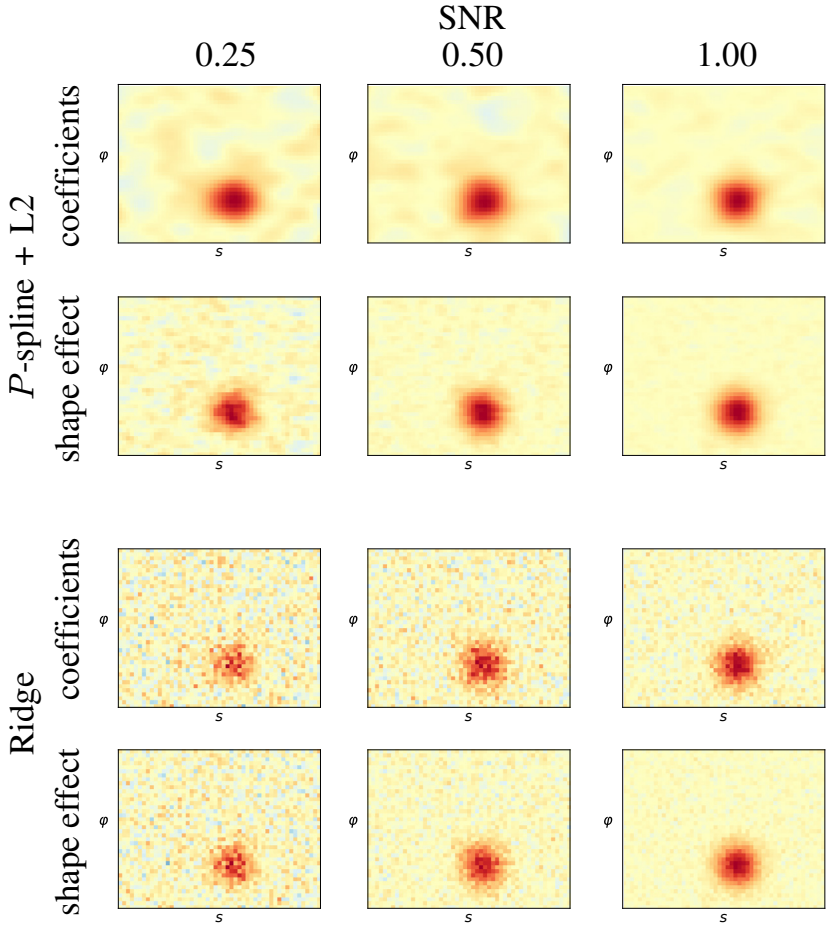


Figure 5.4: *Resulting coefficient and shape effect fields for the* mean shape phantoms*. The axes φ and s give the location on the hippocampus surface (see section 5.2.2). The ground truth shape effect is shown on the left side of Figure 5.2. The scale of the figures is not scaled equally because the scale depends on the amount of regularization added.*

The prediction errors for *hippocampus shape phantoms* can be found in Figure 5.3b. In this case, the *ridge* regression performs better than *P-spline* regression over the entire SNR range. *P-spline* regression without an L2 term takes until an SNR of over 16 before it converges to the performance of *ridge* regression. When a L2 term is added, *P-spline + L2* regression performs worse than *ridge* regression for very low SNRs, but converges quickly to the performance of *ridge* regression. However between 0.125 SNR and 0.5 SNR there is a substantial difference in performance

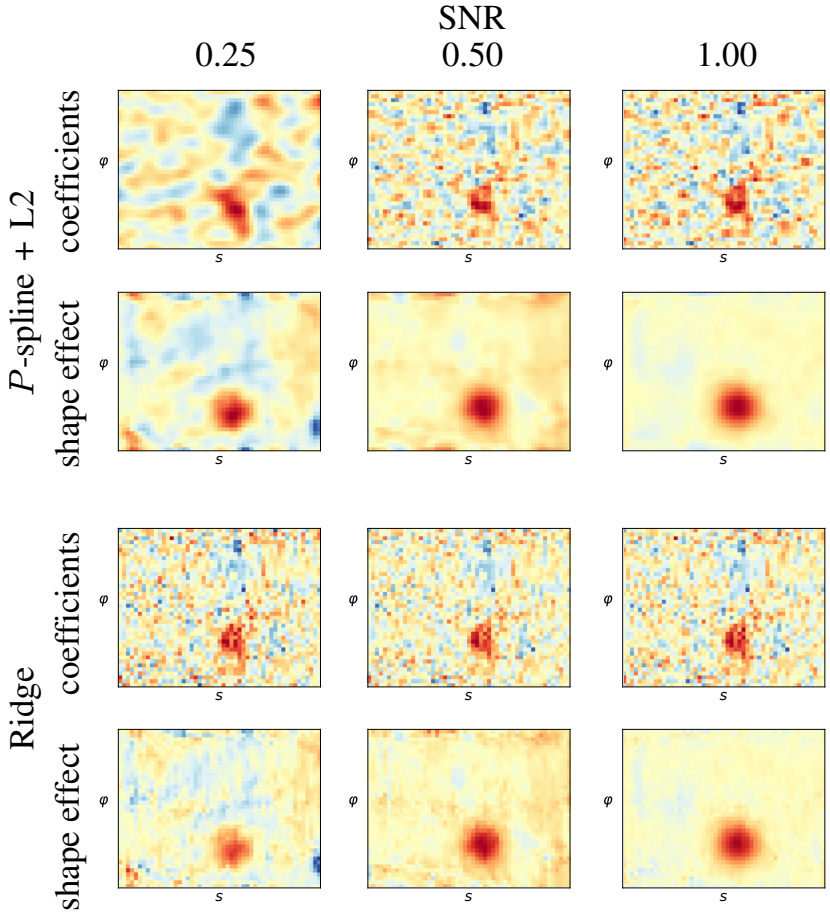


Figure 5.5: *Resulting coefficient and shape effect fields for the* hippocampus shape phantoms. *The axes φ and s give the location on the hippocampus surface (see section 5.2.2). The ground truth shape effect is shown on the left side of Figure 5.2. The scale of the figures is not scaled equally because the scale depends on the amount of regularization added.*

between the methods.

In Figure 5.5 the coefficient and shape effect fields of the regression with the shape based phantoms are shown. This experiment shows the importance of the shape effect fields. In all cases the localization of the coefficients is not very good, as only part of the ground-truth signal is found. The ridge regression suffers from more pixelated noise, whereas the *P-spline + L2* regression has a smooth coefficient field that suffers from false positive coefficients clusters, especially at lower SNR.

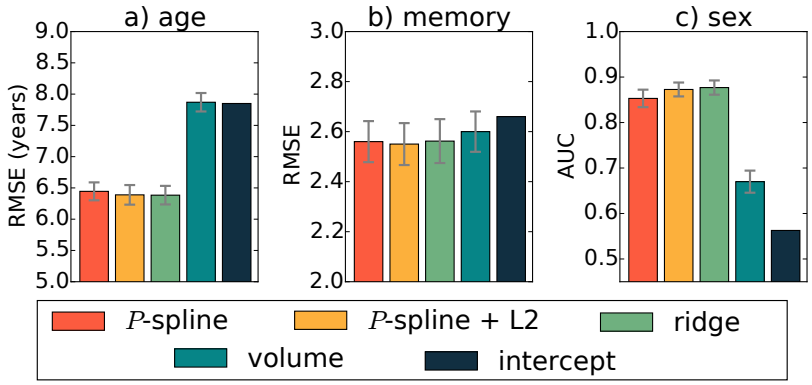


Figure 5.6: *Cross-validated regression results on the hippocampus segmentation of 510 subjects obtained in a cross-validation experiment: a) linear regression with age as dependent variable, b) linear regression with memory score as dependent variable, c) logistic regression with sex as outcome. Note that a low RMSE is better, while a high AUC is better. The first three bars show* P-spline*,* P-spline + L2 *and* ridge *regression with shape as the independent variable. The volume bar shows the regression where instead of shape, volume was used as the independent variable. The intercept bar shows the expected result when fitting an intercept to the data (the standard deviation for linear regression and fraction of the largest class for logistic regression)*

By changing the coefficients into shape effects, for both methods the localization improves a lot and again the *P-spline + L2* produces a smoother result.

5.3.4 Real data experiments

For the experiments on the real data we performed regression on an MR brain dataset of 510 subjects. We performed two linear regressions with shape as the independent variables and either age or memory score as the dependent variable. Additionally, we performed a logistic regression with shape as the independent variable and sex as the dependent variable.

To show that our models are naturally extended to include non-penalized variables, we additionally performed regressions with either age or memory score correcting for the other variable and sex.

The results of the three experiments on the real data are presented in Figure 5.6. We benchmarked these regressions on shape against the same regression but then with volume as the independent variable. For the linear regressions we also plotted the standard deviation in the data, to show the performance of fitting only an intercept, whereas for logistic regression the intercept fit was assumed to be the fraction of the largest class.

From Figure 5.6a, it can be observed that a regression with age and volume yield

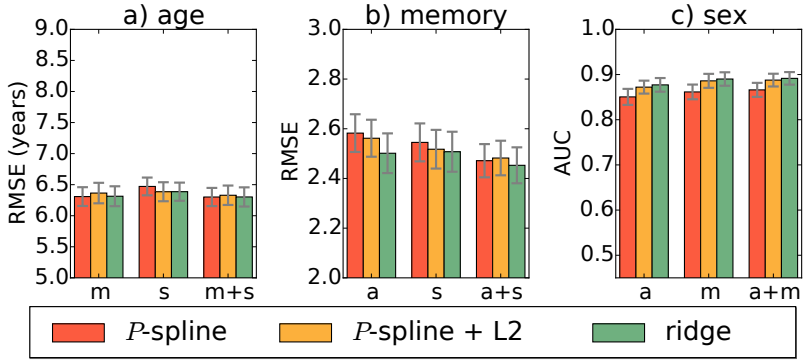


Figure 5.7: *Cross-validated result of the regression with clinical covariates added: a) linear regression with age as dependent variable, b) linear regression with memory score as dependent variable, c) logistic regression with sex as dependent variable. The models are corrected for age (a), memory score (m), sex (s) or a combination (e.g. a+s for age and sex). Note that a low RMSE is better, while a high AUC is better.*

the same error (7.87) as fitting only an intercept (7.87). When using shape data, the error drops to 6.4. The *ridge* regression (6.38) and *P-spline + L2* (6.39) have a slightly lower RMSE than the *P-spline* (6.45), but a paired t-test showed that there were no statistically significant differences ($P < 0.05$) between the three regression methods.

In 5.6b it can be seen that for a regression with memory as an outcome, the difference between the methods are smaller. For memory score fitting volume (2.60) is slightly better than fitting only an intercept (2.66). There were no significant differences between regression using volume as the independent variable and any of methods using shape as the independent variable.

Figure 5.6c shows that volume is not a very good predictor for sex (AUC of 0.67). However, shape can predict sex surprisingly well in our dataset. With an AUC of 0.85, 0.87, and 0.88 for *P-spline*, *P-spline + L2* and *ridge* regression. Both the *P-spline + L2* and *ridge* performance significantly better than the pure *P-spline* regression ($P < 0.01$).

Figure 5.8 shows the coefficient and shape effect fields of the regression of shape with age, memory score or sex as outcome. The coefficient and shape effect fields derived from *P-spline + L2* regression are smoother and look more pronounced than those obtained with *ridge* regression.

The hippocampus is not a solid structure but more like a Swiss roll with subfields that are bands from head to tail on the rolled up sheet. The *P-spline + L2* regression coefficients projected onto the mean hippocampal surface seem to roughly follow these subfields. Furthermore, for both methods the coefficients of age and memory

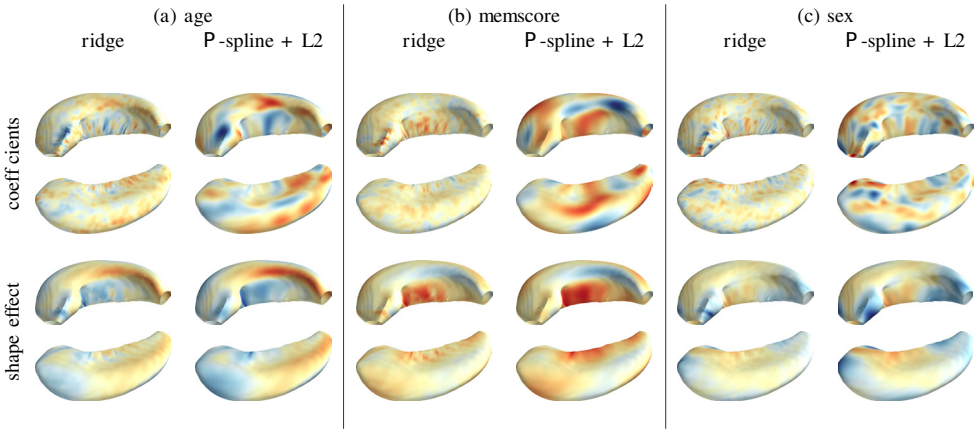


Figure 5.8: *The regression coefficient fields (top row) and shape effect fields (bottom row) for regression with the outcomes with outcomes (a) age, (b) memscore, and (c) sex shown on the mean left hippocampus surface. For each map there are two view of the hippocampus given one from the top, with the head on the left and the tail on the right and one from the bottom. The color range is scaled per sub image, as we are interested in localization and relative volumes, and not the absolute ranges of the shape effects.*

score appear to be negatively correlated. This is expected as older people generally have lower memory scores.

5.3.4.1 Clinical covariates

We repeated the regression experiments on the MR brain data and corrected for covariates. We used age and memory score as outcomes, and age, memory score, and sex as covariates. We tested with just a single covariate or both remaining variables as covariates in the model. The results are presented in Figure 5.7. For age the addition of covariates has little influence on the results. In contrast, adding both age and sex as covariates for predicting memory score improves the predictive value of the model.

The resulting shape effect maps of the regressions are shown in Figure 5.9. It is apparent that for age, adjusting for clinical covariates has no visible effect. However, for the memory score the shape effect fields change slightly when adding age, sex, or both age and sex. This is most visible in the red areas which are more pronounced when adding covariates. This seems to be more pronounced with *P-spline + L2* than with *ridge* regression.

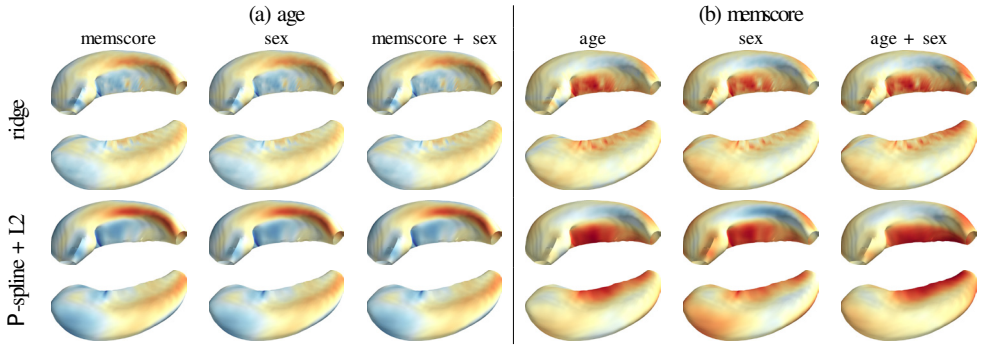


Figure 5.9: *The shape effect fields of the regression with outcomes (a) age and (b) memscore that include clinical covariates shown on the mean hippocampus surface. The included covariate(s) are given at the top of the corresponding column. The color range is scaled per sub image, as we are interested in the localization and not the absolute ranges of the coefficients.*

5.4 Discussion

The spatial regularization of *P-spline* regression leads to spatially smooth coefficient fields and better prediction than *ridge* regression in the presence of uncorrelated Gaussian noise. The resulting shape effect fields are also smoother and appear to be slightly more pronounced.

However, pure *P-spline* regression does not perform well in a situation where the noise is spatially smooth. In Figure 5.3b it can be seen that *P-spline* regression performs worse than *ridge* regression on the phantoms with spatially smooth noise. To counter this, we introduced a hybrid method that has both a spatial smoothness term and a L2 term, the *P-spline + L2* regression. The *P-spline + L2* regression outperforms the *P-spline* regression and performs similar to *ridge* regression except for very low SNR. On real MRI data the *P-spline + L2* regression yielded similar prediction accuracy (no significant differences) as the *ridge* regression.

The spatial regularization used in the *P-spline* and *P-spline + L2* regression methods yields smoother coefficient and shape effect fields which are easier to interpret and seem biologically plausible. For example, in Figure 5.8 for the memory score, the results from *P-spline + L2* regression show a clear cluster of negative coefficients in the area where the CA2-3 hippocampal subfield is located and positve coefficients in the Subiculum. For the regressions with age and especially with memory score as dependent variable, the shape effect fields appear to reflect a structure that matches the structure of the hippocampal subfields. However, to verify this a hippocampal subfield segmentation would be required, which is not feasible as the subfields are not discernible on our 1.5T MR data.

We demonstrated that the presented regression methods can be easily extended

to include clinical covariates. From Figure 5.9 it can be seen that in some cases the shape effect field changes when a covariate is added. It is thus important in brain shape analysis studies to include the relevant covariates. Previous approaches to shape analysis for diagnosis or prediction of dementia are mostly based on machine learning methods such as the non-linear SVM (Achterberg et al., 2014; Gerardin et al., 2009; Lao et al., 2004). With these methods, adding clinical covariates is not trivial because it is important to scale the dimensions properly to determine the importance of different types of information in the classifier. Our approach based on the GLM does not have this problem because it is based on the log-likelihood and not on distances in the feature space. Also, the aforementioned methods are aimed at classification, making them unsuitable for other types of outcomes (e.g. continuous variables). In constrast, the GLM allows for different outcomes by using different link functions. Finally, the non-linear SVM is very good at prediction when provided with sufficient training data, but in contrast to the proposed approach it is difficult to visualize what information the resulting classifier uses (Achterberg et al., 2014; Zhou et al., 2009a). This makes it hard to understand how the SVM makes a decision and to find what shape changes are related to the outcome variable.

Shape analysis methods for visualizing the differences between groups are often based on point-wise testing (Apostolova et al., 2010; Ferrarini et al., 2009; Morra et al., 2009a). For each point in a shape, a statistical test is performed to see if it significantly relates to an outcome. To correct for multiple comparisons, the null-distribution is typically generated by performing the analysis on multiple permutations of the data. These methods help to localize the areas of interest in shapes, but cannot be used to predict outcomes for unseen samples. Additionally, because all points are tested separately, there is no interaction between the different points in the model.

The *P-spline + L2* method combines the favourable qualities of good diagnosis/prediction performance with the ability to visualize the shape differences that are related to the outcome. It both has the ability to apply the model on unseen samples as in classification approaches and to create a shape effect field that can be easily interpreted as in point-wise testing. Because it is based on the GLM, it is easy to adjust for covariates and use both categorized and continuous outcomes. Contrary to other shape analysis techniques, our method explicitly incorporates spatial information to create coherent coefficient and shape effect fields. Though the spatial information does not improve prediction on our real shape data, it does make the shape effect fields more coherent without a loss of predictive value.

There have been few papers that applied spatial regularization for medical shape analysis. Gutman et al. (2013) used Tikhonov regularization to perform a spatially regularized linear discriminant analysis (LDA) on hippocampus shapes. This resulted in smoother coefficient maps while having a similar power as a combination of principal component analysis (PCA) and LDA. However, the LDA can only be used for classification, whereas a GLM can be used for a number of different types of

outcomes. Another use of spatial regularization for shape analysis, is the localized component analysis (LoCA) (Alcantara et al., 2007). LoCA is similar to a PCA, but does not only create components based purely on the variance across the training set, but also encourages components that describe more localized shape variations. It does so by adding a new energy term that summarizes the spatial locality of each component. The resulting components appear more interpretable than PCA components. This decomposition, like the PCA, is unsupervised and does not guarantee that the components are most suitable for regression or classification.

While little prior work is available on spatial regularization in shape analysis, spatially regularized methods have been used more frequently in voxel based analysis and functional MRI (fMRI). For voxel-based morphometry, Sabuncu et al. (2011) introduced the Relevance Voxel Machine and Cuingnet et al. (2012) introduced a spatially regularized SVM. These methods yield smooth and plausible coefficient maps while retaining competitive classification accuracy, but do not naturally include different covariates and outcomes.

For fMRI data often a spatial regularization is used (e.g. Penny et al., 2005; Purdon et al., 2001) and models have been introduced based on Total Variation (TV). TV was initially introduced for image denoising (Rudin et al., 1992) and penalizes the $L1$-norm of the image gradient to perserve edges. This method has been adapted and used for fMRI models (Baldassarre et al., 2012; Gramfort et al., 2013; Michel et al., 2011). The TV used is similar to the regularization terms used in this paper, however the regularization terms in this paper are based on the $L2$-norm. The $L2$-norm seems more appropriate to us because we do not expect changes with sharp edges in the shapes investigated, but rather spatially smooth changes.

The methods described in this paper have a few limitations. First, our shape representation is created to match a 2D B-spline grid, and is in its current form only suitable for cylinder-like structures. To support other topologies, an alternative parametrization is needed, e.g a type of unorganized splines. Secondly, we observed in phantom experiments that in the presence of correlated, spatially smooth noise the P-splines may fit to this noise, reducing the performance in cases where the SNR is low. When the signal is sparse, adding a sparsity term ($L1$ on the coefficients) might help to avoid fitting on the noise and to focus on the real signal. However, in real data this problem was minimal and the addition of a $L2$ penalty terms solved these problems. Finally, the current implementation of the method requires the inverse of a matrix with dimensions equal to the number of features. Changing the optimization method used could help solve the problem more efficient, allowing for larger scale problems.

In conclusion, we presented P-spline regression for shape analysis of brain structures such as the hippocampus. The spatial regularization combines good predictive value with interpretable shape effect fields. As it is based on the GLM, it can naturally handle different types of outcomes and include non-penalized clinical covariates. P-spline regression can help to relate shape to clinical covariates in an understandable

way with the potential of application in clinical routine, e.g. in decision rules. It can both help predict values for unseen cases and provide insight into the regions of the shape related to clinical outcomes.

Chapter 6

Fastr: a workflow engine for advanced data flows in medical image analysis

Hakim C. Achterberg
Marcel Koek
Wiro J. Niessen

Fastr: a workflow engine for advanced data flows in medical image analysis.
Frontiers in ICT, 2016

With the increasing number of datasets encountered in imaging studies, the increasing complexity of processing workflows, and a growing awareness for data stewardship, there is a need for managed, automated workflows. In this paper we introduce Fastr, an automated workflow engine with support for advanced data flows. Fastr has built-in data provenance for recording processing trails and ensuring reproducible results. The extensible plugin-based design allows the system to interface with virtually any image archive and processing infrastructure. This workflow engine is designed to consolidate quantitative imaging biomarker pipelines in order to enable easy application to new data.

6.1 Introduction

In medical image analysis, most methods are no longer implemented as a single executable, but as a workflow composed of multiple programs that are run in a specific order. Each program is executed with inputs that are predetermined or resulting from the previous steps. With increasing complexity of the methods, the workflows become more convoluted and encompass more steps. This makes execution of such a method by hand tedious and error-prone, and makes reproducing the exact chain of processing steps in subsequent studies challenging. Therefore, solutions have been created that are based on scripts that perform all the steps in the correct order.

In population imaging, data collections are typically very large and are often acquired over prolonged periods of time. As data collection is going on continuously, the concept of a 'final' dataset is either non-existent or defined after a very long follow up time. Commonly, analyses on population imaging datasets therefore define intermediate cohorts or time points. To be able to compare intermediate cohorts, all image analysis methods need to produce consistent results over time and should be able to cope with the ever growing size of the population imaging. Therefore the process of running analysis pipelines on population imaging data needs to be automated to ensure consistency and minimize errors.

When different population imaging cohorts are combined in multi-center imaging studies or imaging biobanks (e.g. ADNI (Mueller et al., 2005), OASIS (Marcus et al., 2007b), The Heart-Brain Connection (van Buchem et al., 2014) and BBMRI-NL2.0[1]) where data is often acquired from different scanners, the challenge of ensuring consistency and reliability of the processing results also calls for automated processing workflows.

[1] http://www.bbmri.nl

Traditionally, this is accomplished by writing scripts created specifically for one processing workflow. This can work well, but generally the solutions are tailor-made for a specific study and software environment. This makes it difficult to apply such a method to different data or on a different infrastructure than originally intended. With evolving computational resources, in practice this approach is therefore not reproducible and difficult to maintain. Additionally, for transparency and reproducibility of the results it is very important to know exactly how the data was processed. To accomplish this, an comprehensive data provenance system is required.

Writing a script that takes care of all the aforementioned issues is a challenging and time consuming task. However, many of the components are generic for any type of workflow and do not have to be created separately for each workflow. Workflow management systems can be used to address these issues. These systems help formalize the workflow and can provide features such as provenance as part of the framework, removing the need to address these for every seperate workflow.

For our use cases we desire a workflow management system that works with the tools found in the domain of image analysis, can handle advanced data flows (explained more in detail in section 6.2.3), has strong provenance handling, can handle multiple version of tools, flexible execution backend, can be embedded in our infrastructure. There are already a number of workflow systems available, but none of them fit all our criteria (see Table 6.1).

The most notable open-source, domain-specific workflow system that we are aware of is Nipype (Gorgolewski et al., 2011), which is aimed at creating a common interface for a variety of neuroimaging tools. It also features a system for creating workflows. The tool interfaces of Nipype are elaborate, but Nipype only tracks the version of tools, but does not manage it. This means the system is only aware of the currently installed version of the tool, and cannot offer multiple versions simultaneously.

LONI pipeline (Dinov et al., 2010; Rex et al., 2003) and CBrain (Sherif et al., 2014) also have been developed for the domain of medical image analysis. They include workflow engines, but these systems are part of larger environments which includes data management and processing backends. This makes it difficult to integrate in our infrastructure. Furthermore, LONI is closed-source, which make it even more difficult to integrate it.

The XNAT storage system also has a related workflow system called XNAT pipeline engine (Marcus et al., 2007a). The pipeline engine is integrated nicely with the XNAT storage system and works with simple data flows. However, it does not handle advanced data flows and does not provide tool versioning.

Besides the workflow systems specific for the domain of medical image analysis, there are a number of other notable workflow systems that are either domain-independent or have been created for a different domain. Taverna (Oinn et al., 2006) and KNIME (Berthold et al., 2008, 2009) are well-known and mature workflow management systems. These systems are domain-independent, but mostly used in the

bioinformatics field. Their support for local binary targets is limited, and therefore not suitable for using most medical imaging analysis tools. KNIME needs tools to be create with their API and Taverna is mostly focussed on web services.

Finally, Galaxy (Goecks et al., 2010) is a web-based workflow system for bioinformatics. It is mainly focussed on next-generation sequencing (NGS). It has a large repository of tools, web interface and large support in their domain. However, the system is not designed for batch processing and it does not support complex data-flows.

We developed an image processing workflow framework for creating and managing processing pipelines: Fastr. The framework is designed to build workflows that are agnostic to where the input data is stored, where the resulting output data should be stored, where the steps in the workflow will be executed, and what information about the data and processing needs to be logged for data provenance. To allow for flexible data handling, the input and output of data is managed by a plugin-based system. The execution of the workflow is managed by a pluggable system as well. The provenance system is a built-in feature that ensures a complete log of all processing steps that led to the final result.

In the following section we discuss the design of Fastr. In Section 6.3 we present the resulting software. Finally, we discuss related work and future directions in section 6.4.

Table 6.1: *An overview of workflow systems and the important features of each. The column* Data Flow *can have the value* simple *or* advanced. Simple *means the workflow system supports only sequential data flows whereas* advanced *indicates support for more complex data flows (e.g. the data flows in Section 6.2.3).*

Workflow software						
Name	*Open-Source*	*Language*	*Data flow*	*Tools*	*Versioning of tools*	*Citation*
CBrain	yes	Ruby	simple	binaries	yes	(Sherif et al., 2014)
Fastr	yes	Python	advanced	binaries	yes	
Galaxy	yes	Python	simple	binaries	yes	(Goecks et al., 2010)
KNIME	yes	Java	advanced	Wrappers for Java, Python, Perl code	no	(Berthold et al., 2008, 2009)
LONI pipeline	no	Java	advanced	binaries	yes	(Dinov et al., 2010; Rex et al., 2003)
Nipype	yes	Python	advanced	binaries	no	(Gorgolewski et al., 2011)
Taverna	yes	Java	advanced	webservices	no	(Oinn et al., 2006)
XNAT pipeline engine	yes	Java	simple	binaries	no	(Marcus et al., 2007a)

6.2 Design

The Fastr workflow design follows similar principles as flow-based programming (Morrison, 2010). This paradigm defines applications as a network of black boxes, with predefined connections between the black boxes that indicate the data flow. The black boxes can be reordered and reconnected to create different workflows. However, it should be noted that other aspects of the paradigm are not met, so our design can at most be considered to have flow-based programming aspects.

In Fastr, the workflow is described as a `Network`, which is a directional acyclic graph. The `Nodes` of this `Network` are based on templates that we call `Tools`. These `Nodes` can be interpreted as the black boxes from the flow-based programming paradigm. In the next subsection we will discuss the `Tools` in more detail. After that, we will describe the `Network` and its components in more detail using an example from medical image analysis.

6.2.1 Tools

In Fastr the `Tools` are the *blueprints* for the `Nodes`: they describe the input, output, and behaviour of the `Node`. The `Tools` are composed of three main parts: general metadata, a target, and an interface. The `Tools` are stored as XML or JSON files. An example of a simple `Tool` that adds two list of integers element-wise is given in Listing 6.1. The general metadata contains information about the `Tool` such as id, version, author, license. The target describes how to set the execution environment properly, e.g. by setting the correct search path to use a specific version of the software. The interface describes the inputs and outputs of a `Tool` and how the `Tool` executes given a set of inputs and outputs.

The tools are specified in a schema. This schema validates the internal python data structures (after conversion from XML or JSON) and is specified as a JSON schema. The schemas are located in the source code. There is a schema for the general `Tool`[2] and a schema for the `FastrInterface`[3]. Other types of `Interfaces` can also defined by their own data schema files.

The content of the interface tag depends on the class of `Interface` used. The default `Interface` class in Fastr creates a call to a command-line program given the set of Inputs and Outputs. In the example there are two inputs and one output. In Fastr, the minimal information required for an `Interfaces` to function is the id, cardinality and data type for each Input and Output. The cardinality is the number of values a sample contains (e.g. an argument requiring a point in 3D space, represented by three float values, would have a cardinality of 3).

[2] https://bitbucket.org/bigr_erasmusmc/fastr/src/default/fastr/resources/schemas/Tool.schema.json

[3] https://bitbucket.org/bigr_erasmusmc/fastr/src/default/fastr/resources/schemas/FastrInterface.schema.json

Listing 6.1: *The XML code that defines the AddInt `Tool`. Note that though it might seem the two author entries are redudant or conflicting, the first one states the author of the `Tool` description file, whereas the second states the author of the underlying command (addint.py in this case).*

```
<tool id="AddInt" name="Add two integers" version="1.0">
<description>Add two integers together.</description>
<authors>
  <author name="Hakim Achterberg" email="h.achterberg@erasmusmc.nl"
      ↪ url="http://www.bigr.nl/people/HakimAchterberg" />
</authors>
<command version="0.1" url="">
  <targets>
    <target os="*" arch="*" interpreter="python" paths="./" bin="
        ↪ addint.py" />
  </targets>
  <description>
    addint.py value1 value2
    output = value1 + value2
  </description>
  <authors>
    <author name="Marcel Koek" email="m.koek@erasmusmc.nl" url="http
        ↪ ://www.bigr.nl/people/MarcelKoek" />
  </authors>
</command>
<repository />
<interface>
  <inputs>
    <input id="left_hand" name="left hand value" datatype="Int"
        ↪ prefix="--in1" cardinality="1-*" repeat_prefix="false"
        ↪ required="true" />
    <input id="right_hand" name="right hand value" datatype="Int"
        ↪ prefix="--in2" cardinality="as:left_hand" repeat_prefix="
        ↪ false" required="true" />
  </inputs>
  <outputs>
    <output id="result" name="Resulting value" datatype="Int"
        ↪ automatic="True" cardinality="as:left_hand" method="json"
        ↪ location="^RESULT=(.*)$">
      <description>The summation of left_hand and right_hand.</
          ↪ description>
    </output>
  </outputs>
</interface>
</tool>
```

In Fastr there is a notion of datatypes: each input and output has a (set of) data types it accepts or produces. The datatypes in Fastr are plugins that, in the simplest form, only need to expose their id, but can be extended to include functionality like validators and handlers for multi-file data formats. Data types can be simple values or point to files.

Fastr checks if the datatypes of a linked input and output are (or at least can be) compatible. In addition, data types can be grouped, which is useful for groups of programs using a common (io) library (for example, programs created with The Insight Segmentation and Registration Toolkit[4] (Yoo et al., 2002) can read/write a number of images formats which we grouped together in a pseudo-datatype).

6.2.2 Networks

After `Tools` are defined, a workflow can be created by linking a set of `Tools` which results in a `Network`. Once a `Network` is defined it can be executed. Figure 6.1 shows a graphic representation of an atlas-based segmentation workflow, using the image registration software Elastix (Klein et al., 2010a). Elastix can register two images by optimizing the transformation applied to a moving image to match it to a fixed reference image.

There are different classes of `Nodes`: normal `Nodes` (gray blocks in Figure 6.1), `Source Nodes` (green), `Constant Nodes` (purple) and `Sink Nodes` (blue). Data enters the `Network` through a `Source Node` and leaves the `Network` through a `Sink Node`. A `Constant Node` is similar to the `Source Nodes`, but has its data defined as part of the `Network`. When a `Network` is executed, the data for the `Source Nodes` and `Sink Nodes` has to be supplied. The specifics of the `Source Nodes` and `Sink Nodes` will be discussed in section 6.2.4. The normal `Nodes` process the data as specified by the `Tool`.

The data flow in the `Network` is defined by links (the arrows in Figure 6.1). A link is a connection between the output of a `Node` and the input of another `Node`. A link can manipulate the flow of the data, which will be discussed in section 6.2.3.

The `Nodes` and links in the `Network` form a graph from which the dependencies can be determined for the execution order. Since all `Nodes` are black-boxes that can operate independently of each other, this allows for `Nodes` to be executed in parallel as long as the input dependencies are met.

6.2.3 Data Flow

In Fastr, a sample is defined as the unit of data that is presented to an input of a `Node` for a single job. It can be a simple scalar value, a string, a file, or a list of the aforementioned types. For example, in the *addint* `Tool` presented in Listing 6.1, the *left_hand* and *right_hand* inputs of the `Tool` are required to be (lists of) integers. The

[4] www.itk.org

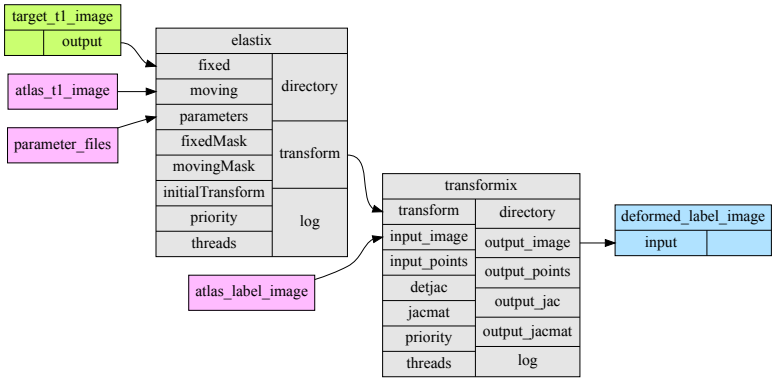


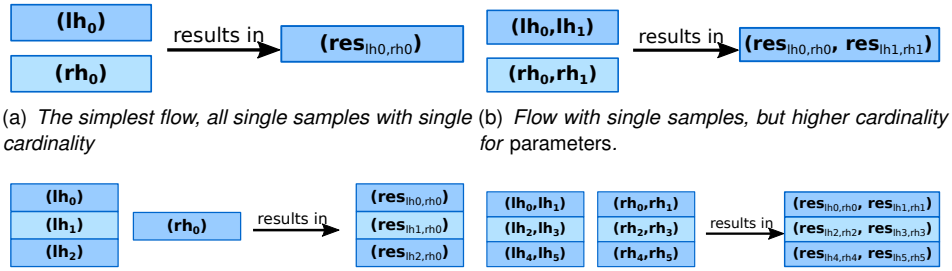
Figure 6.1: *Example `Network` representing a single atlas-based segmentation workflow implemented using the open source Elastix image registration software. Green boxes are `Source Nodes`, purple `Constant Nodes`, gray normal `Nodes`, and blue `Sink Nodes`. Each `Node` contains two columns: the left column represents the inputs,the right column represents the outputs of the `Node`. The arrows indicate links between the inputs and outputs. This image was generated automatically from the source code.*

result output will generate a sample that contains a list of integers. As the cardinality of *right_hand* and *result* are defined to be the same as the *left_hand*, they will all have to same length.

Fastr can handle multiple samples on a specific input. Figure 6.2 shows multiple examples to explain how Fastr handles multiple samples as input (lh for *left_hand* and rh for *right_hand*, and in which output samples (res for *result*) this results. In Figure 6.2(a) we present the simplest situation, in which one sample with one value is offered to each input and one sample with one value is generated. In Figure 6.2(b), the *left_hand* and *right_hand* inputs have one sample with two values. The result is a sample with two values, as one result value is created per input value.

To facilitate batch processing a `Node` can be presented with a collection of samples. These collections are multi-dimensional arrays of samples. In Figure 6.2(c), we depict a situation where three additions are performed. Three samples are offered to the *left_hand* input and one sample is offered to the *right_hand* input. This results in three samples: each sample of the *left_hand* input was used in turn, whereas the samples for the *right_hand* were considered constant. In Figure 6.2(d), there are three samples for the *left_hand* and *right_hand* inputs. The result is again three samples, as now each pair of samples from *left_hand* and *right_hand* inputs was taken.

This is useful for simple batch processing where a task should be repeated a number of times for different input values. However, in certain situation (e.g. multi-atlas segmentation) it is required to register every fixed image to every moving image. To simplify this procedure Fastr can switch from pairwise behaviour to cross product



(a) *The simplest flow, all single samples with single cardinality*

(b) *Flow with single samples, but higher cardinality for* parameters.

(c) *A flow where only one input (*fixed*) is offered multiple samples.*

(d) *A flow where two inputs (*fixed *and* moving*) are offered multiple samples.*

Figure 6.2: *Illustration of the data flows in a* `Node`*. Each rectangle is a sample, and a block of rectangles represents a sample collection. The value is printed in each rectangle, where the commas separate multiple values. The samples lh are offered to the* left_hand *input, the sample rh to* right_hand *input. The sample* res *is generated for the* result *output. The subscript of sample* res *indicates which input samples were used to generate the result.*

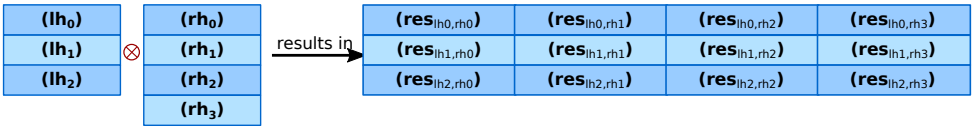


Figure 6.3: *Illustration of the data flows in a* `Node` *that has multiple input groups. The default operator creates a new sample for each combination of input groups.*

behavior. In Figure 6.3, this is depicted graphically. Every combination of *left_hand* and *right_hand* sample is used for registration and the result is a two-dimensional array of transformation samples that in turn contain two transformations each.

Sometimes a `Tool` outputs a sample with a higher cardinality which should be treated as separate samples for further processing, or conversely a number of samples should be offered as a single sample to an input (e.g. for taking an average). For this, Fastr offers two flow directives in data links. The first directive is expand, which indicates that the cardinality is to be transformed into a new dimension. This is illustrated in the left side of Figure 6.4. The second directive is collapse, which indicates one or more dimensions in the sample array should be collapsed and combined into the cardinality. This process is illustrated in the right side of Figure 6.4. These flow directives allow for more complex dataflows in a simple fashion and enable users to implement MapReduce type of workflows.

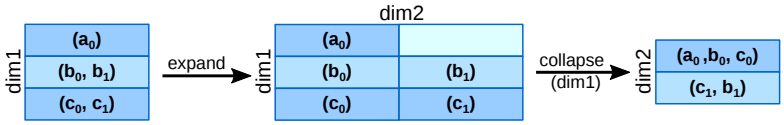


Figure 6.4: *Collapsing and expanding flows. The start situation on the left expands to the situation in the middle after which data collapses the first dimension. Note that in the middle situation there is an empty place in the sample collection (top right). This is possible due to a sparse array representation of the sample collections. This results in two samples with different cardinality in the right-most situation.*

6.2.4 Data input and output

The starting points of every workflow are `Source Nodes`, in which the data is imported into the `Networks`. Similarly the endpoints of every workflow are the `Sink Nodes`, which export the data to the desired location. When a `Network` is constructed only the data type for the `Source Nodes` and `Sink Nodes` needs to be defined. The actual definition of the data is done at runtime using uniform resource identifiers (URI).

Based on the URI scheme, the retrieval and storage of the data will be performed by a plugin. Consider, the following two example URIs:

```
vfs://mount/some/path/file1.txt
xnat://xnat.example.com/data/archive/projects/sandbox/subj...
```

The schemes (in red) of these URI indicate by which plugin the retrieval or storage of the data is handled. For the first URI, vfs indicates that the URI will be handled by the Virtual File System plugin. For the second URI, xnat indicates that the URI will be handled by the XNAT storage plugin. These plugins implement the methods to actually retrieve and store the data. The remainder of the URI is handled by the plugin, so the format of the schemeâĂŹs URI format is defined by the plugin developer.

Plugins can also implement a method to expand a single URI into multiple URIs based on wildcards or searches. In the following example URIs we use wildcards (shown in blue) to retrieve multiple datasets in one go:

```
xnat://xnat.example.com/search?projects=test&subjects=s[0-9]...
vfsregexp://tmp/network_dir/.*/.*/__fastr_result__.pickle.gz
```

The XNAT storage plugin has a direct storage as well as search URI scheme defined. The VFS regular expression plugin, uses the regexp filter to generate a list

of matching vfs URIs. This illustrates that a plugin can expand an url into urls of a different type, and the newly generated urls will be handled by the appropriate plugin.

The use of URIs makes the `Network` agnostic to the location and storage method of the source and target data. Also it allows easy loading of large amounts of resources using wildcards, csv files or search queries.

Currently, Fastr includes plugins for input/output from the (virtual) filesystem, csv files and XNAT. New plugins can be created easily as there are only a few methods that need overwriting. It is also possible to make plugins which can only read data, only write data, or only perform search queries. This allows users to create plugins purely for reading or writing.

Fastr does not include a credential store or other solution for authentication. For all `Network` based input/output plugins (e.g. the XNAT plugin) a netrc file stored in the user's home directory is used for authentication. However, for running Fastr on a grid without a shared network drive this might lead to problems.

6.2.5 Execution

The Fastr framework is designed to offer flexible execution of jobs. The framework analyzes the workflow and creates a list of jobs, including dependencies, that need to be executed. Then it dispatches the jobs to an execution plugin. The plugins can run jobs locally or dispatch them to an execution system such as a cluster, grid, or cloud. A different plugin can be selected for each run allowing for easy switching of the execution backend.

The Fastr execution system consists of a number of components that work together in a layered fashion (see Figure 6.5). The execution starts when the `Network` execute method is called. We will call the machine on which the `Network` execution is started the Submit Host.

Fastr analyzes the `Network` and divides it in chunks that can be processed further. For each chunk the `Network` determines in what order the `Nodes` have to be processed and then executes the `Nodes` in the correct order. When a `Node` is executed, it analyzes the samples on each input and creates a job for each combination input (as specified by the data flow directives).

Jobs contain all information needed to run a single task (e.g. input/output arguments, `Tool` used, etc). The jobs are then dispatched by an execution plugin. The plugin can run the job remotely (e.g on a compute cluster or cloud) or locally (in which case the Submit Host and Execution Host are the same).

Jobs are executed on the Execution Host, and during this step the arguments are translated from urls to actual paths/values. Subsequently, the `Tool` sets the environment for execution according to the target specification and invokes the interface. The interface executes the actual `Tool` commands. Once the interface returns its results, they are validated and the paths in the results are translated back

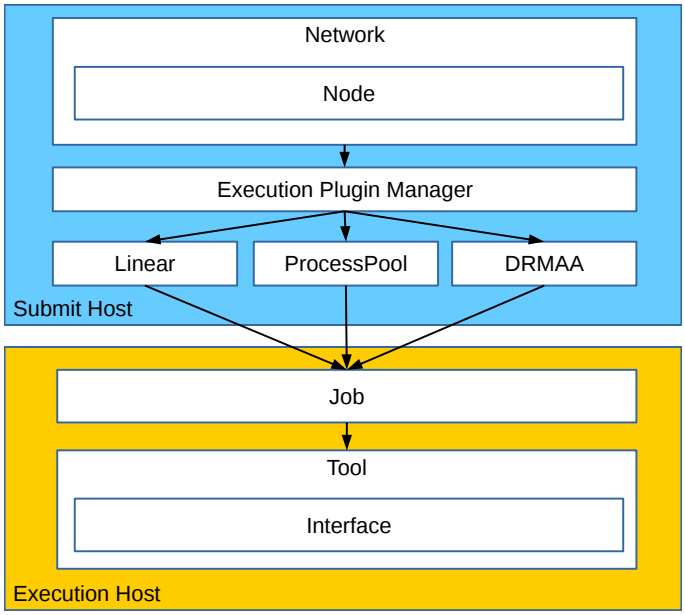


Figure 6.5: *An overview of the execution components in Fastr. The* `Network` *controls the main execution, it sorts the* `Nodes` *required and executes those, resulting in a list of jobs to be run. The jobs are dispatched via an execution plugin. The job is then executed. On execution all arguments are translated to values and paths which the* `Tool` *can use. The* `Tool` *then sets the environment and finally calls the Interface for the actual running of the underlying task.*

into urls.

Once the job execution is finished, the execution plugin will trigger a callback on the Submit Host that reads the job result and updates the `Network` accordingly. If a chunk is finished, the `Network` will process the next chunk, using the updated information. If all chunks are finished, the `Network` execution is done.

Currently, Fastr supports functional plugins for processing locally and on a cluster (using the DRMAAv1 API[5]). Future plugins will focus on flexible middleware for grid/cluster/cloud, like Dirac[6], that offer support for a wide range of systems. For creating a new plugin, five methods need to be implemented: an initialization and a cleanup method as well as methods for queueing, releasing and cancelling a job.

[5] http://www.drmaa.org
[6] http://diracgrid.org

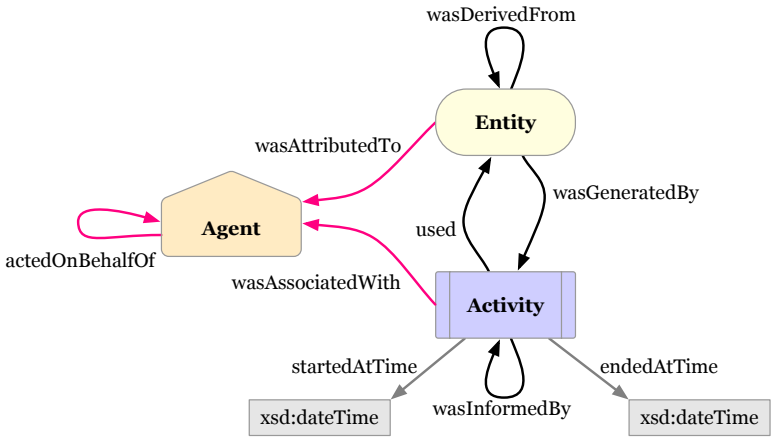


Figure 6.6: *The three base classes of the provenance data model with their relating properties. The agents are orange pentagons, the entities are yellow ovals and the activities are depicted as blue squares. This image is copied from PROV-O: The PROV Ontology. Copyright ©2015 W3C® (MIT, ERCIM, Keio, Beihang). http://www.w3.org/Consortium/Legal/2015/doc-license*

6.2.6 Provenance

Data provenance is a built-in feature of Fastr and is based on an implementation of the *W3C PROV-DM: Prov Data model recommendation* (Belhajjame et al., 2013). Fastr records all relevant data during execution and ensures that for every resulting file a complete data provenance document is included. The standard format of a provenance document is PROV-N, which can be serialized to PROV-JSON or PROV-XML.

In Figure 6.6 the three base classes and the properties of how they relate to each other are illustrated. For Fastr, `Networks`, `Tools` and `Nodes` are modelled as agents, jobs as activities and data objects as entities. The relating properties are naturally valid for our workflow application. The hierarchy and topology of the `Network` follows automatically from the relating properties between the classes, but in order to make the provenance document usable for reproducibility, extra information is stored as attributes on the classes and properties. For every `Tool`, the version is stored. For every data sample, the value or file path and a checksum is stored. For every job the start and end time of execution, the stdout and stderr logs are stored, the end status (success, success with warnings, failed, etc), and an exhaustive description of the execution environment.

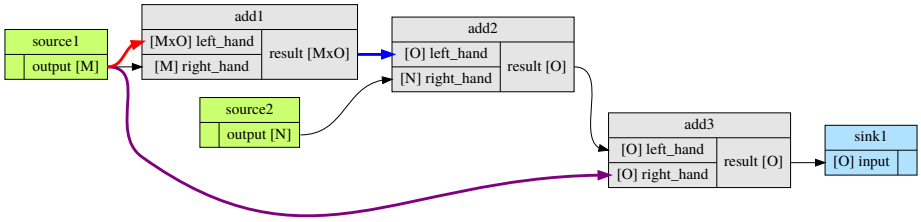


Figure 6.7: *A example of flow visualization. The colored arrows indicate the flow directive in the link: red for expand, blue for collapse and purple for a combination of both. After each input and output the dimensions are printed in square bracket. In this workflow the dimensions N and O should match, but the system can only validate this at runtime.*

6.2.7 Visualization

To give the user insight in the data flow through the `Network`, it is possible to visualize the `Network` using graphviz (Gansner and North, 2000). The figures in this paper that show examples of `Networks` (Figures 6.1 and 6.7), are generated automatically by Fastr. Fastr plots the `Tool` as a collection of inputs and outputs and draws the links between them.

Because Fastr allows for more advanced data flows, there is a few visualization options that can aid users in validating the data flow. First, the color of a link changes if the flow in the Link is different. Second, there is an option to draw the dimension sizes in a `Network`. This shows the number of dimension and the expected size (as symbols). A simple example of the visualization of a more advanced dataflow is given in Figure 6.7.

6.3 Evaluation

A functional version of Fastr is available from https://bitbucket.org/bigr_erasmusmc/fastr. Fastr is open-source and free to use (under the Apache license 2.0). The framework is written in Python and easy to install using the python package index (`pip install`) [7] or using the included setuptools from the source distribution. Fastr is platform independent and runs on Linux, Mac and Windows environments. However, Linux support is much more stable, since that is the platform used in most processing environments.

Documentation is available at http://fastr.readthedocs.io; it includes a quick start tutorial, a user manual and a developer reference of the code. The documentation is built using Sphinx.

The Fastr software is composed of core modules and plugins. The core modules

[7] https://pypi.python.org/pypi/fastr

implement the networking, data flow and interfacing with the plugins. The plugins provide the data input/output, and execution functionality. Fastr is tested for code quality using both unit tests and functional tests. The unit tests are limited to the core modules and ensure the integrity of the core on a fine grained level. The functional testing covers the building and execution of small `Networks`. The functional tests validate the functional requirements of Fastr. Both the unit and the functional tests are performed continuously using the continuous integration framework Jenkins[8].

Currently we are using Fastr for a number of workflows for several single-center and multi-center studies. For example, the Rotterdam Scan Study (Ikram et al., 2011), containing over 12.000 scan sessions, uses a analysis pipeline implemented in Fastr for the preprocessing, tissue type segmentation, white matter lesion segmentation and lobes segmentation of brain MR images (see Figure 6.8). The data is fetched from the archive and is processed in a cluster environment. The resulting data is stored in an image archive.

[8] https://jenkins.io

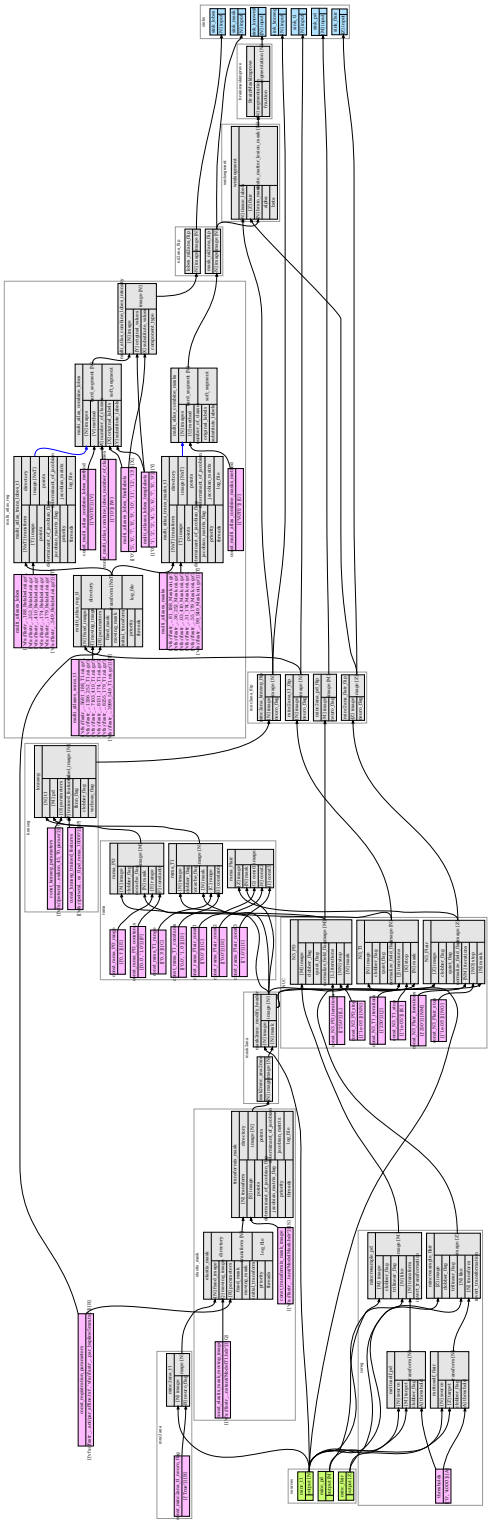


Figure 6.8: *The graphical overview of the processing pipeline used for the Rotterdam Scan Study. It performs brain masking, bias field correction, segments the brain tissues, white matter lessions and different lobels of brain.*

Fastr has been used to run this workflow on new batches of subjects since mid 2015. It performance has proven to be very stable as the workflow always succeeded. The overhead is limited as the Fastr workflow engine uses only a fraction of the resource compared the underlying `Tools`.

6.4 Discussion

With Fastr we created a workflow system that allows users to rapidly create workflows. The simple access to advanced features makes Fastr suitable for both simple and complex workflows. Workflows created with Fastr will automatically get data provenance, support for execution on various computational resources, and support for multiple storage systems. Therefore, Fastr speeds up the development cycle for creating workflows and minimizes the introduction of errors.

Fastr offers a workflow system that works with tools that can really be black boxes, they do not need to implement a specific API as long as their inputs and outputs can be defined. Fastr can manage multiple versions of tools, as we believe it is important to be able to keep an environment where all the old versions of tools are available for future reproducibility of the results. Additionally, it provides provenance records for every result for reproducibility of the experiments. Batch processing and advanced data flows are at the core of Fastrs design. Fastr communicates with processing backends and data providers via plugins allowing interoperability with other components of research infrastructures.

6.4.1 Workflow languages

Most workflows systems and languages are simpler with respect to data flow. However, there are two languages that have features similar to that of Fastr. Taverna, using the SCUFL2 language, has a concept of a dot product or cross product for input ports. This is equivalent to the use of input groups in Fastr. Also the MOTEUR (Glatard et al., 2008) system, using the GWENDIA (Montagnat et al., 2009) language, has the same cross product and dot product concepts.

A main difference between Fastr and the other two languages is that Fastr describes the data as N-D arrays, and a cross product increases the number of dimensions, whereas GWENDIA and SCUFL2 follow the list (of lists) principle. Of course a list of lists can be seen as a 2D array, but that is not used by the aforementioned languages.

There is also the recent effort of the Common Workflow Language, CWL, (Amstutz et al., 2016). The CWL includes a specification for tools and workflows. The CWL has a support for an optional *scatter* directive. This allows a cross product type of behaviour. However, this is not part of main specification, but rather an optional feature.

6.4.2 Limitations

The Fastr workflow system has been created with some clear goals, but there are also some limitations in the design. First of all, our design is created with automated processing workflows in mind and there is no support for interactive steps in the workflow. This is a design choice and there are no plans to address this issue.

Maybe the largest drawback of Fastr is, that as a new system the amount of `Tools` available is limited. The `Tool` wrappers and interfaces are very flexible, but compared to systems as LONI pipeline and Nipype there is a lack of resources. This is a problem any new system faces and we believe that in time this issue will be resolved.

A similar issue is the limited number of execution backend plugins. The system is plugin based and has the potential to support almost any computational resource, but currently only supports local execution and cluster environments. We will add new plugins whenever a project requires one, but do not aim to create many additional plugins on the short term. For grid execution, this could be more challenging due to the lack of credentials management in Fastr. Currently we do not facilitate advanced credential storage, which is often an important requirement in grid computing.

The system is currently completely command-line based and offers no graphical user interface (GUI). Since the focus of Fastr is batch processing the target environments are mostly headless. It is good practice to completely decouple core functionality from the user interface, especially when running in headless environments. Therefore we decided to spend our time on creating a solid workflow engine before creating a GUI. We believe that the tooling can always be added and improved later, but that the core design limitations are generally harder to solve in the future. We plan on adding more (graphical) tools which provide more convenient user interaction in the future.

And finally, we are not satisfied with our current test code coverage. We have test for some core functionally, but the code coverage of the unittests on the low side. This is partially offset by the functional testing, but we feel we should improve the test code coverage to avoid technical debt.

6.4.3 Future directions

Because of the differences in design philosophy, Fastr and Nipype are complementary in focus: Fastr is created for managed workflows and has tools and interfaces as a necessity, whereas the interfaces are the primary focus of Nipype. Considering that there are many interfaces available for Nipype, we created a prototype NipypeInterface in Fastr, that allows `Tools` in Fastr to use Nipype for the interface. This is still experimental and there are still some limitations because Nipype and Fastr have incompatible data type systems.

Another option to increase the amount of tools available is to start supporting

Boutiques[9]. Boutiques are an application repository with a standard packaging of tools, so that they can be used on multiple platforms. The boutiques applications are somewhat similar to Fastr `Tools`, as they describe the inputs and output in a JSON file. Additionally, the underlying binaries, scripts, and data are all packaged, versioned, and distributed using Docker[10] containers. It would require to either rewrite the boutique inputs/outputs into a Fastr interface or to create a new interface class for Boutiques.

Although the CWL at the moment is as far as we know not used in the medical imaging domain, we think that support for the CWL is an important future feature for Fastr as we fully support the idea to have a common standard language. Support for CWL tools in Fastr could possible using a new interface class, but the support for workflows would probably need to be an import/export that transcribes workflows from CWL to Fastr and back.

For reproducibility it is important to be able to re-run analyses in exactly the same conditions. Currently Fastr supports environment modules to keep multiple versions of software available at the same time. However, the same version of the software can still be different based on underlying libraries, compiler used and the OS. Virtual Machines or Linux Containers offer a solution to this problem. Linux containers, such as Docker and LXC, are often seen as a light-weight alternative to Virtual Machines. They ensure that the binaries and underlying libraries are all managed, but they use the kernel of the host OS. We plan to add support for Docker containers to make it easier to share tools and improve reproducibility further.

For continuous integration, we have a Jenkins[11] continious integration server that runs our tests nightly. Additionally we use SonarQube[12] for inspecting code quality, technical debt, and code coverage. We are aiming for each release to increase the code coverage and to decrease the technical debt.

Finally, we are working on more (web-based) tooling around Fastr to make it easier to visualize, develop and debug `Networks` and to inspect the results of a run (including provenance information).

[9] http://boutiques.github.io/
[10] https://www.docker.com
[11] https://jenkins.io/
[12] http://www.sonarqube.org/

Chapter 7

General discussion

“It seems that when you have cancer you are a brave battler against the disease, but when you have Alzheimer’s you are an old fart. That’s how people see you. It makes you feel quite alone."

— Sir Terry Pratchett

7.1 Findings and implications

In this thesis I developed and evaluated methods for the extraction and application of quantitative imaging biomarkers for dementia diagnosis and prognosis. Various aspects of quantitative biomarkers development were considered: the extraction of the imaging biomarkers, the interpretability and predictive value of the imaging biomarkers and the IT infrastructure required to extract imaging biomarkers in a standardized fashion, and on a large scale.

7.1.1 Imaging biomarker extraction

The work in this thesis focused on the hippocampus, a structure known to be of importance in dementia. The first step was to obtain a segmentation of the hippocampus. We used a method based on a combination of machine learning and atlas registration for hippocampus segmentation. This method was originally developed for imaging data acquired in the population based Rotterdam Scan Study

A challenge in machine learning methods is the need for training data that is representative for the data to which the algorithm will be applied. Our segmentation method (based on atlas registration and an appearance model in the form of voxel classification) was developed and worked very well for the homogeneous data of the Rotterdam Scan Study, but could not straightforwardly be applied in a settings with more heterogeneous data. To address this issue, I investigated if this problem could be circumvented by using the tissue content of the voxels rather than the intensity values of the voxels as basis for the appearance models. We showed that this led to competitive performance when segmenting across different scan protocols, while retaining a good performance when segmenting scans acquired with the same imaging protocol. This indicates that, at least for the hippocampus, the good performance appearance model is mostly driven by the tissue distributions and not by the more specific imaging acquisition dependent intensity information in the image.

I thus demonstrated that it is possible to adapt segmentation methods to be more robust to differences in the acquisition protocol. This could help greatly with the pooling of data from different sources. Currently, image segmentation methods are most widely applied in controlled settings, such as large population or clinical studies, where the acquisition protocol has been fixed and great care has been taken to create uniform data. However the application of such segmentation methods on data pooled from different studies or in clinical practice has not taken off on a large scale, in part due to the difficulties related to differences in imaging hardware and acquisition settings. Methods that are robust to these differences have the potential to facilitate reuse of data, pooling of existing data, and translation to clinical practice. This could especially have impact for studies into rare diseases (e.g. Lewy Body Dementia), as the creation of a large prospective dataset is challenging in such cases.

7.1.2 Imaging biomarkers for prediction

Once the region of interest has been segmented, quantitative descriptors can be derived to describe the region. For the hippocampus, the simplest, most frequently used descriptor is its volume. In this thesis I investigated less common descriptors such as the shape and texture of the hippocampus in relation to dementia. These quantitative descriptors have been used before, but not in the context of the prediction of dementia in cognitively normal subjects in a population study. In this thesis I show that both shape and texture are predictive of the development of dementia. Additionally, I show that the various combinations of volume, shape and texture generally improve the predictive power over using just a single biomarker.

The predictive power of volume, shape, and texture is at a level where, on its own, it is not good enough for meaningful personalized predictions. The sensitivity and specificity have to be substantially better to be able to predict dementia on the individual subject level. To improve predictive power, the prediction model should include other features including demographic data, cognitive test data, and genetic data. Also, in such studies, we should evaluate the additional value of the imaging biomarkers over other markers.

Even though on a personal level, the prediction is not yet usable, the features I found can be used for risk stratification. A normal population could be split into a high and low risk group, which could be very useful for research into preventive treatments where it is important to find high risk subjects. In a group of high risk subjects it would be easier to measure the effect of a treatment, lowering the required number of inclusions. This could make it easier to run the study and it would lower the cost of the study.

In the investigations of the predictive power of texture and shape I used a non-linear classification approach. While this approach achieved good results, interpreting the results and translating them back to physiological changes turned out to be rather difficult. This was partly due to the nonlinearity of the model, leading to complex relations, and partly because the model of a classifier is not trying to describe the process that connects the features and outcome, but to optimize the ability to separate groups. In an attempt to create models that can both yield good predictive performance and provide insight into the underlying pathology, I investigated the use of spatially penalized regression models using P-spline regression as a basis. These models were based on a general linear model (GLM), making it possible to model different types of outcomes following a distribution from the exponential family (e.g. linear, counts, binary choice). Furthermore, the penalization used allows for different covariates to have a different penalization (or no penalization at all), making it possible to combine different types of features in a single model. I showed that the P-spline based regression models yielded smoother coefficient and shape effect fields than ridge regression, while maintaining a competitive performance. These smoother fields can make it easier to interpret the biological changes related to the outcome

variable. However, it appeared that for dementia prediction the linear models do not perform very well; ridge regression, P-spline regression and linear SVMs all performed poorly compared to a non-linear SVM.

In chapter 3 we visualized the decision boundary of a non-linear SVM, which was very challenging due to the nonlinear nature. For every point in the feature space the separating direction is different, so we tried to find a sensible point in feature space to visualize.

7.1.3 IT infrastructure for imaging biomarker extraction

To extract quantitative biomarkers from large volumes of imaging data, to pool data from different centers or cohorts, or to establish a quantitative imaging biomarker in clinical practice, it is important that they are extracted in a consistent, reproducible way. Many methods used for the extraction of quantitative biomarkers in fact do not consist of a single step, but are rather multiple steps that are combined in a workflow. These workflows are traditionally scripted by individual researchers. In order to make the creation of formalized workflows easier and to add strong data provenance to workflows, in this thesis I developed the workflow management system Fastr. Using Fastr for creating image processing workflows leads to a formalized description of all the steps in a workflow making it easier to reuse parts of the workflow. Also any workflow created with Fastr automatically can make use of various execution plugins (to be executed e.g. locally, on a cluster, or on a cloud) and input/output plugins (to load/save data from e.g. the file system or from imaging databases). The system is currently in use in our department for the processing of multiple large population studies as well as as in an experimental system to process clinical data. This helps greatly in the systematic extraction of biomarkers from thousands of scans.

7.2 Limitations and future directions

In this thesis I made contributions to various aspects of the extraction and use of quantitative imaging biomarkers. Of course, there are many challenges that I could not address and there are limitations to the work in this thesis. I will discuss some of these limitations and the possible continuation of the work.

7.2.1 Robustness to changes in acquisition protocol

Though I demonstrated in Chapter 2 that using tissue distribution maps instead of voxel intensities can help to make appearance models more robust against acquisition protocols, there are some clear limitations to this approach. By using tissue maps, I effectively shifted the problem of differences in scanner protocols from hippocampus segmentation to tissue segmentation. The assumption I made is that it is easier to create a robust tissue segmentation (there are already examples of such methods)

than to change the appearance model. Also, the proposed approach does not easily generalize to other situations. It is effective in the brain, where we can segment the scan in a limited number of tissues (i.e. gray matter, white matter and cerebrospinal fluid). In other organs, where such a segmentation into a limited set of tissues is not possible, this method will not work.

In these situations other methods have to be employed. The most straightforward method would be image normalization, which often is achieved by normalizing the histogram of images in some way (Christensen, 2003; Jager and Hornegger, 2009). There are also patch based methods which can synthesize one modality from others that can be used to synthesize standardized scans (Roy et al., 2011, 2013; Ye et al., 2013). Also, a method based on registration might be possible. First the target image would be registered to a number of templates and subsequently an intensity mapping could be estimated through the joint histogram and used to map the target data and the training data to each other.

In the field of MR acquisition techniques, quantitative acquisition protocols are being developed that measure physical properties, such as the T1 and T2 relaxation time (Tofts, 2005). Such quantitative MRI protocols might make image normalization obsolete, since the image intensities resulting from these acquisitions are supposed to be calibrated, and therefore less likely to vary across different scan sessions and scanners. As these protocols do not directly measure, but rather estimate the physical properties, there still will be some differences in values between the protocols. Furthermore, reliable quantitative MRI protocols tend to cost more scan time and can only be implemented for new studies. There is a large amount of legacy data that could be very useful, but is not scanned with these protocols.

It is possible to use the normalization methods described previously to synthesize a quantitative MRI based on a normal MRI scan, e.g. a quantitative T1 map based on a simple T1 scan. This would always be an estimation based on templates and learned data, but it might be good enough for making robust appearance models or texture models.

Finally there is also the possibility to solve the problem of data heterogeneity by machine learning methods. In the field of transfer learning, classifiers and methods are being developed that are robust against certain differences in the distribution of the training and target data (van Opbroek et al., 2015b; Pan and Yang, 2010). The differences that can be modeled and strategies to do so are dependent on the individual method. For example, by weighting the training data from multiple sources, a classifier could be made that better fits the target data, as long as some of the training data is similar to the target data (Cheplygina et al., 2016; van Opbroek et al., 2015c). Also it is possible to find a feature space transformation that estimates how the data from training sources and target data could be mapped to a shared feature space (van Opbroek et al., 2015a).

7.2.2 Imaging biomarkers for prediction

The quantitative biomarkers of hippocampus volume, shape, and texture have shown to contain predictive information about dementia development. In my analyses I corrected for age and gender, but I did not correct for cognitive markers such as memory score and memory complaints. A limitation of these studies is that I did not evaluate how much current dementia prediction models could be improved by integrating imaging based markers as volume, shape and texture.

For determining the predictive value of hippocampal features, I mostly looked at dementia of any type as an outcome. However, dementia is a disorder that can be caused by a number of different pathologies. To optimize treatment and to support dementia research it is also important to know the subtype of dementia. Owing to the limited number of cases, I did not investigate differential diagnosis in this thesis. Only for Alzheimer's Disease the number of cases was sufficient for analysis. The analysis in a case-control settings with Alzheimer's Disease instead of dementia as the outcome showed similar trends as the analysis with dementia, although due to the limited number of cases the predictive power was lower. It would be of large interest to investigate the use of the proposed quantitative biomarkers for differential diagnosis, but this would require a larger dataset with long term follow-up. Such dataset will become available in view of the population studies that are currently running.

In this thesis a cohort with just over 500 subjects from the Rotterdam Scan Study was used for the evaluation of the quantitative biomarkers in a cross-validation design. This cohort was used because data was prospectively collected, and the study has had over ten years of follow up time. Though this has not been shown in this work, I expect that the methods should be generalizable to different populations and small differences in acquisition protocol. As shown in Chapter 2, in case of larger differences in acquisition protocols, there will be challenges in segmentation and texture analysis that are not yet fully solved by our approach. Of course it is possible to redo the entire analysis, from the segmentation to the prediction in cross-validation, on a new cohort. However, this would only reveal if shape/texture descriptors are also predictive in different settings, but it is not likely that the descriptors can be shared straightforwardly between cohorts. To introduce a quantitative imaging biomarker into clinical practice, a single biomarker definition that can be used on the heterogeneous data being acquired in daily clinical practice would be instrumental.

The analysis of hippocampus shape with P-spline based regression has shown to lead to smoother coefficient maps and shape effect maps and to allow for the combination of features and covariates with different penalizations. The possibility to mix and match features with penalization strategies is interesting, as it allows combining hippocampal volume, shape and texture in one model but with different penalization strategies. This could be an alternative of the combination method used in Chapters 3 and 4. However, there are some important drawbacks to the presented

method. First of all, though the performance of P-spline regression is similar to ridge regression, for both methods the performance for dementia prediction is not so good. From my experiments it was clear that a nonlinear SVM performs better for dementia prediction, which means that gaining interpretability by using a linear method seems infeasible.

Another limitation is the shape representation that is based on a cylindrical coordinate system. It both limits the types of shape that can be described and it introduces a disconnection between the distance (and hence strength of spatial regularization) in the representation and the image space. The first issue could be solved with using a different (less organized) basis for the splines. The method presented in this thesis uses tensor product B-splines which require a regular two dimensional basis. The second issue could be solved by correcting for the physical distance between points in the calculation of the penalty matrix.

Another issue is how to best establish correspondence between the different shapes in a dataset. In Chapters 3 and 4 I used a method based on entropy introduced by Cates et al. (2007), whereas in Chapter 5 I used a scheme based on registration. The entropy-based approach optimizes the correspondence between all shapes simultaneously. The registration-based scheme uses pair-wise registration to a number of templates, which has the advantage that it is easier to add new samples to the model. Due to the simultaneous optimization used, applying the entropy method to unseen samples is not straightforward. During the hippocampus segmentation itself, also a number of registrations are performed and it should be possible to use these for establishing correspondence. This would actually merge the problem of segmentation and finding correspondence between shapes and solve these problems jointly. The correspondence would then also be based on the image rather than only the binary masks, allowing the context of the shape to be used in establishing correspondence.

7.2.3 IT infrastructure for imaging biomarkers extraction

In this thesis I also contributed towards enabling IT infrastructure for more standardized imaging biomarker extraction: The Fastr framework helps to create formalized, flexible, and consistent workflows. However, this comes at the cost of a learning curve and it forces developers to work in a systematic way. Every step of the workflow needs to be compartmentalized and the flow between the steps should be formally defined. Every step is based on a tool definition that specifies the inputs and outputs of the step and how execution off the step will be performed. This creates overhead as even a simple step will need to be formally described. However, when creating a new workflow using mostly existing tool definitions, the process can actually be faster than traditional scripting. In my personal experience, the formalization of the workflow also tends to give provide insight in the workflow itself and helps to find mistakes.

A drawback of Fastr compared to more established workflow systems (e.g. Nipype

(Gorgolewski et al., 2011), LONI pipeline (Dinov et al., 2010; Rex et al., 2003)) is that the number of tool descriptions available is not as extensive yet. With every new workflow we implement, new tool descriptions are added. The user base of Fastr is still relatively small, but with the increase in users and workflows the ecosystem will grow. Currently Fastr is used in a number of multi-center projects which ensures that the ecosystem will grow and that the most used tools will be added to repository of tool descriptions. However, it will be important to give tutorials about Fastr to reach more users and show them that Fastr can benefit the creation of the workflows.

For the future it is planned that more plugins will be added to the Fastr framework. Using these plugins more interaction with other systems can easily be achieved. There are already plugins for communication with XNAT and execution on a cluster. Currently, a workflow progress visualization web service is being created to which Fastr can broadcast the progress of a workflow run. This provides a visual overview of the status of a workflow, making it possible to track the progress of large scale workflow execution. Also, there is work on a resource manager in the cloud that can schedule jobs and scale the cloud to the requirements of the jobs. This allows users to leverage the computation potential of cloud platforms (e.g. Surf HPC cloud or Amazon Web Services) in a simple and efficient way. These additions make it easier to perform large scale analysis and to process larger cohorts. Because of the design of the framework, each new plugin can immediately benefit all workflows created using Fastr.

References

Achterberg, H.C., van der Lijn, F., den Heijer, T., van der Lugt, A., Breteler, M.M.B., Niessen, W.J., de Bruijne, M.; Prediction of Dementia by Hippocampal Shape Analysis. In: Wang, F., Yan, P., Suzuki, K., Shen, D., editors. Machine Learning in Medical Imaging. Springer Berlin / Heidelberg; volume 6357 of *Lecture Notes in Computer Science*; 2010. p. 42–49.

Achterberg, H.C., van der Lijn, F., den Heijer, T., Vernooij, M.W., Ikram, M.A., Niessen, W.J., de Bruijne, M.; Hippocampal shape is predictive for the development of dementia in a normal, elderly population. Human Brain Mapping 2014;35(5):2359–2371.

Alcantara, D., Carmichael, O., Delson, E., Harcourt-Smith, W., Sterner, K., Frost, S., Dutton, R., Thompson, P., Aizenstein, H., Lopez, O., Becker, J., Amenta, N.; Localized components analysis. Information processing in medical imaging : proceedings of the conference 2007;20:519–31.

Aljabar, P., Heckemann, R.a., Hammers, A., Hajnal, J.V., Rueckert, D.; Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 2009;46(3):726–38.

Alkoot, F.M., Kittler, J.; Moderating k-NN Classifiers. Pattern Analysis & Applications 2002;5(3):326–332.

Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S., Stojanovic, L.. Common Workflow Language, v1.0. 2016.

Apostolova, L.G., Dutton, R.A., Dinov, I.D., Hayashi, K.M., Toga, A.W., Cummings, J.L., Thompson, P.M.; Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. Archives of neurology 2006;63(5):693–9.

Apostolova, L.G., Mosconi, L., Thompson, P.M., Green, A.E., Hwang, K.S., Ramirez, A., Mistur, R., Tsui, W.H., de Leon, M.J.; Subregional hippocampal atrophy predicts Alzheimer's dementia in the cognitively normal. Neurobiology of aging 2010;31(7):1077–1088.

Ashburner, J., Friston, K.J.; Unified segmentation. NeuroImage 2005;26(3):839–51.

Balafar, M.A., Ramli, A.R., Saripan, M.I., Mashohor, S.; Review of brain mri image segmentation methods. Artificial Intelligence Review 2010;33(3):261–274.

Baldassarre, L., Mourão-Miranda, J., Pontil, M.; Structured sparsity models for brain decoding from fMRI data. In: Proceedings - 2012 2nd International Workshop on Pattern Recognition in NeuroImaging, PRNI 2012. IEEE; 2012. p. 5–8.

Barnes, J., Foster, J., Boyes, R.G., Pepple, T., Moore, E.K., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C.; A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. NeuroImage 2008;40(4):1655 – 1671.

Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.. PROV-DM: The PROV Data Model. Recommendation;

W3C; 2013. http://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.; KNIME: The Konstanz information miner. Springer, 2008.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B.; Knime-the konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter 2009;11(1):26–31.

Braak, H., Braak, E.; Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiology of Aging 1997;18(4):351–357.

Brechbühler, C., Gerig, G., Kübler, O.; Parametrization of Closed Surfaces for {3-D} Shape Description. Computer Vision and Image Understanding 1995;61(2):154–170.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., na, D.C.P., Álvarez Meza, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Fatta, G.D., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S.; Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The CADDementia challenge. NeuroImage 2015;111:562 – 579.

de Bruijn, R.F., Bos, M.J., Portegies, M.L., Hofman, A., Franco, O.H., Koudstaal, P.J., Ikram, M.A.; The potential for prevention of dementia across two decades: the prospective, population-based Rotterdam Study. BMC Medicine 2015;13(1):132.

van Buchem, M.A., Biessels, G.J., Brunner la Rocca, H.P., de Craen, A.J., van der Flier, W.M., Ikram, M.A., Kappelle, L.J., Koudstaal, P.J., Mooijaart, S.P., Niessen, W., et al.; The Heart-Brain Connection: A multidisciplinary approach targeting a missing link in the pathophysiology of vascular cognitive impairment. Journal of Alzheimer's Disease 2014;42(s4).

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Bach Cuadra, M.; A review of atlas-based segmentation for magnetic resonance brain images. Computer methods and programs in biomedicine 2011;.

Casey, B., Giedd, J.N., Thomas, K.M.; Structural and functional brain development and its relation to cognitive development. Biological Psychology 2000;54(1-3):241–257.

Cates, J., Fletcher, P.T., Whitaker, R.; Entropy-Based Particle Systems for Shape Correspondence. In: Mathematical Foundations of Computational Anatomy Workshop, MICCAI 2006. 2006. p. 90–99.

Cates, J.E., Fletcher, P.T., Styner, M.A., Shenton, M.E., Whitaker, R.T.; Shape Modeling and Analysis with Entropy-Based Particle Systems. In: Karssemeijer, N., Lelieveldt, B.P.F., editors. IPMI. Springer; volume 4584 of *Lecture Notes in Computer Science*; 2007. p. 333–345.

Chan, D., Fox, N.C., Scahill, R.I., Crum, W.R., Whitwell, J.L., Leschziner, G., Rossor, A.M., Stevens, J.M., Cipolotti, L., Rossor, M.N.; Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. Ann Neurol 2001;49(4):433–442.

Chang, C.C., Lin, C.J.. LIBSVM: a library

for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Cheplygina, V., van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M.; Asymmetric similarity-weighted ensembles for image segmentation. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). 2016. p. 273–277.

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.; Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. NeuroImage 2011;58(2):469–480.

Christensen, J.D.; Normalization of brain magnetic resonance images using histogram even-order derivative analysis. Magnetic Resonance Imaging 2003;21(7):817 – 820.

Collins, D.L., Pruessner, J.C.; Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. NeuroImage 2010;52(4):1355–66.

Cortes, C., Vapnik, V.; Support-vector networks. Machine Learning 1995;20(3):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.; Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. NeuroImage 2011;56(1):212–219.

Courchesne, E., Chisum, H.J., Townsend, J., Cowles, A., Covington, J., Egaas, B., Harwood, M., Hinds, S., Press, G.A.; Normal brain development and aging: Quantitative analysis at in vivo mr imaging in healthy volunteers. Radiology 2000;216(3):672–682. PMID: 10966694.

Csernansky, J.G., Wang, L., Swank, J., Miller, J.P., Gado, M., McKeel, D., Miller, M.I., Morris, J.C.; Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. Neuroimage 2005;25(3):783–792.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., The Alzheimer's Disease Neuroimaging Initiative, ; Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage 2010;.

Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., Colliot, O.; Spatial and anatomical regularization of SVM: A general framework for neuroimaging data. IEEE transactions on pattern analysis and machine intelligence 2012;35(3):682–696.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M.; Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiology of agingAging 2008;29(4):514–523.

Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.; A minimum description length approach to statistical shape modeling. IEEE Trans Med Imaging 2002;21(5):525–537.

DeCarli, C., Murphy, D.G., McIntosh, A.R., Teichberg, D., Schapiro, M.B., Horwitz, B.; Discriminant analysis of MRI measures as a method to determine the presence of dementia of the Alzheimer type. Psychiatry Res 1995;57(2):119–130.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.; Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;:837–845.

Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W., Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., Fischl, B., Initiative, A.D.N.; Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 2009;132(Pt 8):2048–2057.

Devanand, D., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G., Honig, L., Mayeux, R., et al.; Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of Alzheimer disease. Neurology 2007;68(11):828–836.

Dierckx, P.; Curve and surface fitting with splines. Monographs on numerical analysis. Clarendon Press, 1993.

Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., Zamanyan, A., Chakrapani, S., Van Horn, J., Parker, D.S., et al.; Neuroimaging study designs, computational analyses and data provenance using the loni pipeline. PloS one 2010;5(9):e13070.

Duin, R.P.W., Juszczak, P., de Ridder, D., Paclík, P., Pekalska, E., D.M.J.Tax, . PRTools. 2004.

Eilers, P.H., Marx, B.D.; Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemometrics and Intelligent Laboratory Systems 2003;66(2):159–174.

Eilers, P.H.C., Marx, B.D.; Flexible smoothing with B-splines and penalties. Statistical Science 1996;11(2):89–121.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.; COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans Med Imaging 2007;26(1):93–105.

Ferrarini, L., Frisoni, G.B., Pievani, M., Reiber, J.H.C., Ganzola, R., Milles, J.; Morphological hippocampal markers for automated detection of Alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. Journal of Alzheimer's Disease 2009;17(3):643–659.

Ferrarini, L., Olofsen, H., Palm, W.M., van Buchem, M.A., Reiber, J.H.C., Admiraal-Behloul, F.; GAMEs: growing and adaptive meshes for fully automatic shape modeling and analysis. Med Image Anal 2007;11(3):302–314.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.; Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33(3):341–55.

Gansner, E.R., North, S.C.; An open graph visualization system and its applications to software engineering. Software Practice and Experience 2000;30(11):1203–1233.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, ; Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 2009;47(4):1476–1486.

Glatard, T., Montagnat, J., Lingrand, D., Pennec, X.; Flexible and efficient workflow deployment of data-intensive applications on grids with MOTEUR. International Journal of High Performance Computing Applications 2008;22(3):347–360.

Goecks, J., Nekrutenko, A., Taylor, J., et al.; Galaxy: a comprehensive approach for supporting accessible, reproducible, and trans-

parent computational research in the life sciences. Genome Biol 2010;11(8):R86.

Golland, P.; Discriminative direction for kernel classifiers. In: Advances in neural information processing systems. volume 14; 2002. p. 745–752.

Goodall, C.; Procrustes Methods in the Statistical Analysis of Shape. Journal of the Royal Statistical Society Series B (Methodological) 1991;53(2):285–339.

Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S.; Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Frontiers in neuroinformatics 2011;5.

Gramfort, A., Thirion, B., Varoquaux, G.; Identifying predictive regions from fMRI with TV-L1 prior. In: Proceedings - 2013 3rd International Workshop on Pattern Recognition in Neuroimaging, PRNI 2013. 2013. p. 17–20.

Gutman, B.A., Hua, X., Rajagopalan, P., Chou, Y.Y., Wang, Y., Yanovsky, I., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M., Alzheimer's Disease Neuroimaging Initiative, ; Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features. NeuroImage 2013;70:386–401.

Hammers, A., Heckemann, R., Koepp, M.J., Duncan, J.S., Hajnal, J.V., Rueckert, D., Aljabar, P.; Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: a proof-of-principle study. NeuroImage 2007;36(1):38–47.

Han, X., Fischl, B.; Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE transactions on medical imaging 2007;26(4):479–86.

Hastie, T., Stuetzle, W.; Principal curves. Journal of the American Statistical Association 1989;84(406):502–516.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F.; On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 2014;87:96–110.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.; Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 2006;33(1):115–26.

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A.; Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. NeuroImage 2010;51(1):221–7.

den Heijer, T., Geerlings, M.I., Hoebeek, F.E., Hofman, A., Koudstaal, P.J., Breteler, M.M.B.; Use of hippocampal and amygdalar volumes on magnetic resonance imaging to predict dementia in cognitively intact elderly people. Archives of general psychiatry 2006;63(1):57–62.

den Heijer, T., Vermeer, S.E., Clarke, R., Oudkerk, M., Koudstaal, P.J., Hofman, A., Breteler, M.M.B.; Homocysteine and brain atrophy on MRI of non-demented elderly. Brain; a journal of neurology 2003;126:170–175.

Hoerl, A.E., Kennard, R.W.; Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55–67.

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.; The Rotterdam Study: 2016 objectives and design update. European Journal of Epidemiology 2015;30(8):661–

708.

Igel, C., Heidrich-Meisner, V., Glasmachers, T.; Shark. Journal of machine learning research 2008;9(Jun):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.; The rotterdam scan study: design update 2016 and main findings. European Journal of Epidemiology 2015;30(12):1299–1315.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Krestin, G.P., Koudstaal, P.J., Hofman, A., Breteler, M.M.B., Vernooij, M.W.; The Rotterdam Scan Study: design and update up to 2012. European journal of epidemiology 2011;26(10):811–824.

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.; Brain tissue volumes in relation to cognitive function and risk of dementia. Neurobiology of aging 2010;31(3):378–86.

Ikram, M.A., Vrooman, H.A., Vernooij, M.W., van der Lijn, F., Hofman, A., van der Lugt, A., Niessen, W.J., Breteler, M.M.; Brain tissue volumes in the general elderly population: The rotterdam scan study. Neurobiology of Aging 2008;29(6):882 – 890.

Jaakkola, T., Diekhans, M., Haussler, D.; Using the fisher kernel method to detect remote protein homologies. In: In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. AAAI Press; 1999. p. 149–158.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.; Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. The Lancet Neurology 2010;9(1):119 – 128.

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.; Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 1999;52(7):1397–1403.

Jager, F., Hornegger, J.; Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. IEEE Transactions on Medical Imaging 2009;28(1):137–150.

Karas, G.B., Scheltens, P., Rombouts, S.a.R.B., Visser, P.J., van Schijndel, R.a., Fox, N.C., Barkhof, F.; Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. NeuroImage 2004;23(2):708–16.

Kessler, L.G., Barnhart, H.X., Buckler, A.J., Choudhury, K.R., Kondratovich, M.V., Toledano, A., Guimaraes, A.R., Filice, R., Zhang, Z., Sullivan, D.C.; The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Statistical Methods in Medical Research 2015;24(1):9–26. PMID: 24919826.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.; elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 2010a;29(1):196 – 205.

Klein, S., Staring, M., Murphy, K., Viergever, M.a., Pluim, J.P.W.; Elastix: a Toolbox for Intensity-Based Medical Image Registration. IEEE transactions on medical imaging 2010b;29(1):196–205.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J.; Automatic classification of MR scans in Alzheimer's disease. Brain 2008;131(Pt 3):681–689.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S.M., Davatzikos, C.; Morphological classification of brains via high-

dimensional shape transformations and machine learning methods. NeuroImage 2004;21(1):46–57.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S.; Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage 2010a;51(4):1345–59.

Leung, K.K., Shen, K.K., Barnes, J., Ridgway, G.R., Clarkson, M.J., Fripp, J., Salvado, O., Meriaudeau, F., Fox, N.C., Bourgeat, P., Ourselin, S.; Increasing power to predict mild cognitive impairment conversion to Alzheimer's disease using hippocampal atrophy rate and statistical shape models. Med Image Comput Comput Assist Interv 2010b;13(Pt 2):125–132.

Li, S., Shi, F., Pu, F., Li, X., Jiang, T., Xie, S., Wang, Y.; Hippocampal shape analysis of Alzheimer disease based on machine learning methods. AJNR Am J Neuroradiol 2007;28(7):1339–1345.

van der Lijn, F., de Bruijne, M., Klein, S., den Heijer, T., Hoogendam, Y.Y., van der Lugt, A., Breteler, M.M.B., Niessen, W.J.; Automated brain structure segmentation based on atlas registration and appearance models. IEEE transactions on medical imaging 2012;31(2):276–86.

van der Lijn, F., den Heijer, T., Breteler, M.M.B., Niessen, W.J.; Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 2008;43(4):708–720.

Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L.; The extensible neuroimaging archive toolkit. Neuroinformatics 2007a;5(1):11–33.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.; Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience 2007b;19(9):1498–1507.

Marx, B.D., Eilers, P.H.; Generalized linear regression on sampled signals and curves: a p-spline approach. Technometrics 1999;41(1):1–13.

Marx, B.D., Eilers, P.H.; Multidimensional penalized signal regression. Technometrics 2005;47(1):13–22.

McCullagh, P., Nelder, J.A.; Generalized linear models. volume 37. CRC press, 1989.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps, C.H.; The diagnosis of dementia due to alzheimerâĂŹs disease: Recommendations from the national institute on aging-alzheimerâĂŹs association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & Dementia 2011;7(3):263 – 269.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.; Total variation regularization for fMRI-based prediction of behavior. IEEE Transactions on Medical Imaging 2011;30(7):1328–1340.

Montagnat, J., Isnard, B., Glatard, T., Maheshwari, K., Fornarino, M.B.; A data-driven workflow language for grids based on array programming principles. In: Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science. New York, NY, USA: ACM; WORKS '09; 2009. p. 7:1–7:10.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Avedissian, C., Madsen, S.K., Parikshak, N., Hua, X., Toga, A.W., Jack, C.R., Schuff, N., Weiner, M.W., Thompson, P.M.,

Alzheimer's Disease Neuroimaging Initiative, ; Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Human Brain Mapping 2009a;30(9):2766–2788.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Avedissian, C., Madsen, S.K., Parikshak, N., Hua, X., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M.; Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. NeuroImage 2008;43(1):59–68.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Avedissian, C., Madsen, S.K., Parikshak, N., Toga, A.W., Jack, C.R., Schuff, N., Weiner, M.W., Thompson, P.M.; Automated mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. NeuroImage 2009b;45(1 Suppl):S3–15.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M.; Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE transactions on medical imaging 2010;29(1):30–43.

Morrison, J.P.; Flow-Based Programming, 2nd Edition: A New Approach to Application Development. Paramount, CA: CreateSpace, 2010.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.; Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia 2005;1(1):55–66.

Neugroschl, J., Sano, M.; Current treatment and recent clinical research in Alzheimer's disease. Mt Sinai J Med 2010;77(1):3–16.

Nyúl, L.G., Udupa, J.K., Zhang, X.; New variants of a method of MRI scale standardization. IEEE transactions on medical imaging 2000;19(2):143–50.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A., Wroe, C.; Taverna: lessons in creating a workflow environment for the life sciences. Concurrency and Computation: Practice and Experience 2006;18(10):1067–1100.

van Opbroek, A., Achterberg, H.C., de Bruijne, M.; Feature-Space Transformation Improves Supervised Segmentation Across Scanners; Cham: Springer International Publishing. p. 85–93.

van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M.; Transfer learning improves supervised image segmentation across imaging protocols. IEEE Transactions on Medical Imaging 2015b;34(5):1018–1030.

van Opbroek, A., Vernooij, M.W., Ikram, M.A., de Bruijne, M.; Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. Medical image analysis 2015c;24(1):245–254.

Organization, W.H., et al.; Dementia: a public health priority. World Health Organization, 2012.

Pan, S.J., Yang, Q.; A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 2010;22(10):1345–1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg,

V., et al.; Scikit-learn: Machine learning in python. Journal of Machine Learning Research 2011;12(Oct):2825–2830.

Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J.; Bayesian fMRI time series analysis with spatial priors. NeuroImage 2005;24(2):350–362.

Pien, H.H., Fischman, A.J., Thrall, J.H., Sorensen, A.; Using imaging biomarkers to accelerate drug development and clinical trials. Drug Discovery Today 2005;10(4):259 – 266.

Poot, D., Ikram, M., Vernooij, M., de Bruijne, M., Niessen, W.; Improved tissue segmentation by including an MR acquisition model. In: Liu, T., Shen, D., Ibanez, L., Tao, X., editors. Multimodal Brain Image Analysis, LNCS 7012. Toronto, Canada: Multimodal Brain Image Analysis, LNCS 7012; 2011. p. 152–159.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P.; The global prevalence of dementia: A systematic review and metaanalysis. Alzheimer's & Dementia 2013;9(1):63 – 75.e2.

Prince, M., Bryce, R., Ferri, C.; World Alzheimer Report 2011: The benefits of early diagnosis and intervention. Alzheimer's Disease International, 2011.

Purdon, P.L., Solo, V., Weisskoff, R.M., Brown, E.N.; Locally Regularized Spatiotemporal Modeling and Model Comparison for Functional MRI. NeuroImage 2001;14(4):912–923.

Qiu, A., Fennema-Notestine, C., Dale, A.M., Miller, M.I., Alzheimer's Disease Neuroimaging Initiative, ; Regional shape abnormalities in mild cognitive impairment and Alzheimer's disease. Neuroimage 2009;45(3):656–661.

Qiu, A., Younes, L., Miller, M.I., Csernansky, J.G.; Parallel transport in diffeomorphisms distinguishes the time-dependent pattern of hippocampal surface deformation due to healthy aging and the dementia of the Alzheimer's type. Neuroimage 2008;40(1):68–76.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.A., Démonet, J.F., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P., Alzheimer's Disease Neuroimaging Initiative, ; Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 2009;132(Pt 8):2036–2047.

Rex, D.E., Ma, J.Q., Toga, A.W.; The loni pipeline processing environment. Neuroimage 2003;19(3):1033–1048.

Roy, S., Carass, A., Prince, J.; A Compressed Sensing Approach for MR Tissue Contrast Synthesis; Berlin, Heidelberg: Springer Berlin Heidelberg. p. 371–383.

Roy, S., Carass, A., Prince, J.L.; Magnetic resonance image example-based contrast synthesis. IEEE Transactions on Medical Imaging 2013;32(12):2348–2363.

Rudin, L.I., Osher, S., Fatemi, E.; Nonlinear total variation noise removal algorithm. Physica D: Nonlinear Phenomena 1992;60(1-4):259–268.

Sabuncu, M.R., Leemput, K.V., Van Leemput, K.; The relevance voxel machine (RVoxM): A Bayesian method for image-based prediction. Medical Image Computing and Computer-Assisted InterventionâĂŞMICCAI 2011 2011;:99–106.

Scarpini, E., Scheltens, P., Feldman, H.; Treatment of Alzheimer's disease: current status and new perspectives. Lancet Neurol 2003;2(9):539–547.

Scheltens, P., Fox, N., Barkhof, F., Carli, C.D.; Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. Lancet Neurol 2002;1(1):13–21.

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E.C., Valk, J.; Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. Journal of Neurology, Neurosurgery & Psychiatry 1992;55(10):967–972.

Scher, A.I., Xu, Y., Korf, E.S.C., Hartley, S.W., Witter, M.P., Scheltens, P., White, L.R., Thompson, P.M., Toga, A.W., Valentino, D.J., Launer, L.J.; Hippocampal morphometry in population-based incident Alzheimer's disease and vascular dementia: the HAAS. Journal of Neurology, Neurosurgery & Psychiatry 2011;82(4):373–6.

Scher, A.I., Xu, Y., Korf, E.S.C., White, L.R., Scheltens, P., Toga, A.W., Thompson, P.M., Hartley, S.W., Witter, M.P., Valentino, D.J., Launer, L.J.; Hippocampal shape analysis in Alzheimer's disease: a population-based study. Neuroimage 2007;36(1):8–18.

Seghers, D., Agostino, E.D., Maes, F., Vandermeulen, D.; Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques. In: Barillot, C., Haynor, D.R., Hellier, P., editors. Brain. Springer; volume 3216 of *Lecture Notes in Computer Science*; 2004. p. 696–703.

Sherif, T., Rioux, P., Rousseau, M.E., Kassis, N., Beck, N., Adalat, R., Das, S., Glatard, T., Evans, A.C.; CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. Frontiers in Neuroinformatics 2014;8:54.

Sled, J.G., Zijdenbos, A.P., Evans, A.C.; A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 1998;17(1):87–97.

Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative, , the Australian Imaging Biomarkers, , of Ageing, L.F.S.; Early detection of Alzheimer's disease using MRI hippocampal texture. Human Brain Mapping 2016;37(3):1148–1161.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H.; Toward defining the preclinical stages of alzheimerâĂŹs disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & Dementia 2011;7(3):280 – 292.

Tepest, R., Wang, L., Csernansky, J.G., Neubert, P., Heun, R., Scheef, L., Jessen, F.; Hippocampal surface analysis in subjective memory impairment, mild cognitive impairment and Alzheimer's dementia. Dement Geriatr Cogn Disord 2008;26(4):323–329.

Thies, W., Bleiler, L.; 2011 Alzheimer's disease facts and figures. Alzheimer's & dementia : the journal of the Alzheimer's Association 2011;7(2):208–44.

Tofts, P.; Quantitative MRI of the brain: measuring changes caused by disease. John Wiley & Sons, 2005.

Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., Breteler, M.M.B., Niessen, W.J.; Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. NeuroImage 2007;37(1):71–81.

Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J.C., Csernansky, J.G., Miller, M.I.; Large deformation diffeo-

morphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE transactions on medical imaging 2007;26(4):462–70.

West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C.; Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. Lancet 1994;344(8925):769–772.

West, M.J., Kawas, C.H., Stewart, W.F., Rudow, G.L., Troncoso, J.C.; Hippocampal neurons in pre-clinical Alzheimer's disease. Neurobiology of aging 2004;25(9):1205–1212.

Whitaker, R.T.; Reducing Aliasing Artifacts in Iso-Surfaces of Binary Volumes. In: IEEE Symposium on Volume Visualization and Graphics. Los Alamitos, CA, USA: IEEE Computer Society; 2000. p. 23–32.

Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D.; LEAP: learning embeddings for atlas propagation. NeuroImage 2010;49(2):1316–25.

Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E.; Modality Propagation: Coherent Synthesis of Subject-Specific Scans with Data-Driven Regularization; Berlin, Heidelberg: Springer Berlin Heidelberg. p. 606–613.

Yoo, T., Ackerman, M., Lorensen, W., Schroeder, W., Chalana, V., Aylward, S., Metaxas, D., Whitaker, R.; Engineering and algorithm design for an image processing API: A technical report on ITK - The Insight Toolkit; volume 85. p. 586–592.

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

Zhou, L., Hartley, R., Wang, L., Lieby, P., Barnes, N.; Identifying anatomical shape difference by regularized discriminative direction. IEEE transactions on medical imaging 2009a;28(6):937–50.

Zhou, L., Lieby, P., Barnes, N., Régladе-Meslin, C., Walker, J., Cherbuin, N., Hartley, R.; Hippocampal shape analysis for Alzheimer's disease using an efficient hypothesis test and regularized discriminative deformation. Hippocampus 2009b;19(6):533–40.

Summary

Deriving quantitative features from images has the ability to aid in the understanding of normal biological processes, diagnosis and diseases prediction, drug discovery and clinical trials. The requires the development of so-called quantitative imaging biomarkers. In this thesis I addressed different aspects of the extraction of quantitative imaging biomarkers from MR brain imaging data and their use for prediction of dementia. As the hippocampus is affected by dementia in an early stage, I focussed on quantitative MR imaging biomarkers related to the hippocampus: hippocampus volume, shape and texture.

Robust segmentation

Before quantitative features of the hippocampus can be derived, the hippocampus needs to be segmented in the brain MRI data. Several papers on this topic have shown the benefit of supervised classification based on local appearance features, often combined with atlas-based approaches. These methods require a representative annotated training set and therefore often do not perform well if the target image is acquired on a different scanner or with a different acquisition protocol than the training images. Assuming that the appearance of the brain is determined by the underlying brain tissue distribution and that brain tissue classification can be performed robustly for images obtained with different protocols, in chapter 2 I propose to derive appearance features from brain-tissue density maps instead of directly from the MR intensities. I evaluated this approach on hippocampus segmentation in two sets of images acquired with substantially different imaging protocols and on different scanners. While a combination of multi-atlas segmentation and conventional appearance features trained on data from a different scanner performed poorly with an average Dice overlap of 0.698, the local appearance model based on the new acquisition-independent features significantly improved (Dice of 0.783) over atlas-based segmentation alone (Dice of 0.728).

Predictive value of hippocampal shape and texture

Previous studies have shown that hippocampal volume is an early marker for dementia. Considering that hippocampal shape also has been shown to be a biomarker

for dementia, in chapter 3 I investigated whether hippocampal shape characteristics extracted from MRI scans are predictive for the development of dementia during follow up in subjects who were nondemented at baseline. Furthermore, I assessed whether hippocampal shape provides additional predictive value independent of hippocampal volume. Five hundred eleven brain MRI scans from elderly nondemented participants of a prospective population-based imaging study were used. During the 10-year follow-up period, 52 of these subjects developed dementia. For training and evaluation independent of age and gender, a subset of 50 cases and 150 matched controls was selected. The hippocampus was segmented using an automated method. From the segmentation, the volume was determined and a statistical shape model was constructed. I trained a classifier to distinguish between subjects who developed dementia and subjects who stayed cognitively healthy. For all subjects the a posteriori probability to develop dementia was estimated using the classifier in a cross-validation experiment. The area under the ROC curve for volume, shape, and the combination of both were, respectively, 0.724, 0.743, and 0.766.

Recently, hippocampal texture has also 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 predictive value extends to cognitively healthy individuals. In chapter 4 I investigated if hippocampal volume, shape, texture, and their combination were predictive of dementia in the general population, and determined how predictive performance varied with time to diagnosis and presence of early clinical symptoms. The same data as in chapter 3 was used, except for one subject for which the matching scan to the hippocampus mask was not available. The same volume and shape features were used as in chapter 3 and additionally the texture features were computed. A similar cross-validation experiment was performed as in chapter 3. All features showed significant predictive performance with the area under the ROC 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 subjective memory complaints or MCI, performance decreased over follow-up (e.g. 0.935 for <3 years, 0.809 for 3–6 years, and 0.632 for >6 years, for the combination of volume and texture). I conclude that a combination of multiple hippocampal features on MRI performs better in predicting dementia in the general population than any feature by itself. The best prediction was achieved with the combination of volume and texture.

Spatially regularized shape regression

Shape analysis is a challenging task due to the high dimensionality of shape representations and the often limited number of available shapes. Current point distribution model techniques counter the poor ratio between dimensions and sample size by using regularization in feature space, but do not take into account the spatial relations

within the shapes. This can lead to models which are biologically implausible and are difficult to interpret. In chapter 5 I propose to use P -spline regression, which combines a generalized linear model (GLM) with the coefficients described as B-splines and a penalty term that constrains the regression coefficients to be spatially smooth, for shape analysis. Owing to the GLM, this method can naturally predict both continuous and discrete outcomes and can include non-spatial covariates without penalization. I evaluated the method on the same hippocampus shapes of 510 elderly people as used in chapter 4. I related the hippocampal shape to age, memory score, and sex. The proposed method retained the performance of ridge regression, but produced smoother coefficient fields and shape effect maps that are easier to interpret.

Fastr workflow framework for biomarker extraction

For large imaging studies, possibly in a multi-center setting, including population and clinical imaging studies, it is very important that all the quantitative imaging biomarkers are derived in exactly the same way to be able to pool the data. With the increasing number of datasets encountered in imaging studies, the increasing complexity of processing workflows, and a growing awareness for data stewardship, there is a need for managed, automated workflows. In chapter 6 I introduce Fastr, an automated workflow engine with support for advanced data flows. Fastr has built-in data provenance for recording processing trails and ensuring reproducible results. The extensible plugin-based design allows the system to interface with virtually any image archive and processing infrastructure. This workflow engine is designed to consolidate quantitative imaging biomarker pipelines in order to enable easy application to new data.

In summary, in this thesis I addressed a number of challenges relating to the extraction of quantitative imaging biomarkers: the segmentation of images acquired with different protocols/scanners, the validation of hippocampus shape and texture descriptions as quantitative imaging biomarkers for dementia prediction, the development of statistical methods for interpretable shape analysis, and a workflow framework for large-scale, reproducible biomarker extraction. While the implementation and validation of this thesis focussed on the hippocampus extracted from MRI brain data, some of the solutions and concepts are more generally applicable.

Samenvatting

Dementie is een verzamelnaam voor een categorie van neurodegeneratieve ziektes die het denkvermogen, communicatie en geheugen aantasten, waarvan de ziekte van Alzheimer de meest voorkomende is. In 2010 waren er naar schatting 35.6 miljoen mensen wereldwijd die leden aan de ziekte van Alzheimer. Naar verwachting zal door de bevolkingstoename en een langere levensverwachting dit aantal elke 20 jaar verdubbelen. De wereldwijde zorgkosten gerelateerd aan dementie werden geschat op 604 miljard dollar voor het jaar 2012 en zullen naar verwachting nog sneller stijgen dan het aantal mensen dat dementie krijgt.

Ondanks dat er nog geen genezing voor dementie mogelijk is, is een vroege diagnose belangrijk om mensen met dementie en hun omgeving de mogelijkheid te bieden om zich te kunnen voorbereiden op deze ziekte. Bovendien zijn er therapieën gericht op de verbetering van het geheugen en denkvermogen die weliswaar geen genezing bieden maar wel de kwaliteit van het leven kunnen verbeteren. Ten slotte is de vroege diagnose van dementie belangrijk voor het onderzoek naar de preventie en behandeling hiervan. Gezien het feit dat de onderliggende ziekteprocessen vaak langere tijd aan de gang zijn voordat de symptomen zich manifesteren, kan beeldvorming van het brein een zeer waardevolle bijdrage leveren aan een vroege diagnose.

MRI is een niet-invasieve beeldvormende techniek die een zeer goed contrast laat zien tussen zachte weefsels. Dit maakt MRI uitermate geschikt voor de beeldvorming van de hersenen en MRI wordt tegenwoordig dan ook zowel in de klinische praktijk als in het medisch (dementie) onderzoek veelvuldig gebruikt. Momenteel worden MRI beelden veelal kwalitatief beoordeeld; een radioloog interpreteert de beelden en probeert een inschatting te maken van de ziektestatus. Deze kwalitatieve beoordeling heeft een aantal nadelen: 1) het is moeilijk om tot een objectief resultaat te komen (elke radioloog kan het anders beoordelen), 2) grote datasets consistent te beoordelen is moeilijk en arbeidsintensief, en 3) subtiele veranderingen zijn soms moeilijk waar te nemen. Door de opmars van automatische beeldanalyse technieken is het nu mogelijk om kwantitatieve maten van de hersenen te bepalen aan de hand van MRI scans, bijvoorbeeld het volume van de hersenen of van een deel van de hersenen. Deze ontwikkeling leidt tot het concept van een kwantitatieve beeld biomarker.

Een kwantitatieve beeldvormende biomarker wordt gedefinieerd als: *ëen objectieve eigenschap van een in vivo beeld, gemeten op ratio of interval schaal, als indicator van normale biologische processen, pathogene processen, of een resultaat van een thera-*

peutische interventie." Zoals de definitie aangeeft kunnen kwantitatieve beeldmaten nuttig zijn voor het begrijpen van normale biologische processen, voor de diagnose en voorspelling van ziektes, en voor het evalueren van het effect van medicijnen of therapieën.

In deze thesis ga ik in op een aantal aspecten van het bepalen van kwantitatieve biomarkers aan de hand van hersen MRI scans en het gebruik ervan voor het voorspellen van dementie. Doordat MRI een zeer veelzijdige techniek is, die naast anatomie bijvoorbeeld ook functie af kan beelden, zijn er veel verschillende potentiële kwantitatieve biomarkers die uit MRI hersenbeelden geëxtraheerd zouden kunnen worden. Omdat de hippocampus een hersenstructuur is waarvan het bekend is dat deze in een vroeg stadium door dementie wordt beïnvloed, richt ik me in dit proefschrift specifiek op beeldmaten gerelateerd aan de hippocampus: het volume, de vorm en de textuur van de hippocampus.

Robuuste beeldsegmentatie

Voordat we een kwantitatieve beschrijving van de hippocampus kunnen verkrijgen, moeten we eerst bepalen waar in de MRI scan de hippocampus zich precies bevind. Het omlijnen van een structuur in een beeld heet ook wel segmentatie. Er zijn meerdere publicaties over dit onderwerp die het nut aantonen van machinaal leren, een techniek waarbij de computer op basis van voorbeelden leert hoe lokale beeldkarakteristieken gebruikt kunnen worden om een object te segmenteren. Vaak wordt deze methode dan gecombineerd met een atlas voor ruimtelijke informatie. Deze methode heeft echter als nadeel dat er representatieve, reeds gesegmenteerde voorbeelden moeten zijn, en zal dan ook niet werken als de nieuwe beelden niet genoeg op de voorbeeld beelden lijken. Als de beelden die geanalyseerd moeten worden met een andere MRI scanner of een ander acquisitie protocol gemaakt zijn dan de voorbeeld data waarop geleerd is, zal de segmentatie methode niet goed werken.

Als we aannemen dat het lokale structuur van een MRI scan van de hersenen grotendeels bepaald wordt door de verhouding van de witte stof, grijze stof en het liquor (hersenvocht) in het brein, kunnen we de MRI scan beschrijven aan de hand van deze verhouding in plaats van de originele beeldintensiteit. Hiermee introduceren we in feite een tussenstap waarin de MRI beelden van verschillende oorsprong geharmoniseerd worden. In hoofdstuk 2 van dit proefschrift evalueer ik deze aanpak voor hippocampus segmentatie door het toepassen ervan op twee verzamelingen beelden die met een substantieel ander acquisitie protocol en op verschillende scanners zijn gemaakt. Terwijl de originele combinatie van een atlas met beeldkarakteristieken niet goed werkte (slechter dan enkel de atlassen gebruiken), werkte de nieuwe variant met beeldkarakteristieken gebaseerd op de hersenweefsels beter dan alleen het gebruik van de atlassen.

De voorspellende waarde van hippocampus vorm en textuur voor dementie

Eerdere studies hebben aangetoond dat het volume van de hippocampus een vroege indicator is voor dementie. Ook is aangetoond dat de vorm van de hippocampus bruikbaar kan zijn voor de diagnose van dementie. Daarom heb ik in hoofdstuk 3 van dit proefschrift onderzocht of de vorm van de hippocampus zoals gevonden in MRI beelden nuttig is voor de voorspelling van dementie bij mensen, die op dat moment nog geen dementie hebben. Aangezien de vorm en het volume van de hippocampus enigszins gerelateerd zullen zijn, heb ik gekeken of de vorm van de hippocampus voorspellende waarde toevoegt onafhankelijk van het hippocampus volume.

Voor dit onderzoek gebruikte ik 511 MRI hersenscans van niet-demente oudere deelnemers aan een groot prospectief bevolkingsonderzoek. Deze mensen werden na de scan meer dan 11 jaar gevolgd en in die tijd ontwikkelden 52 deelnemers dementie. Om de modellen onafhankelijk van leeftijd en geslacht te kunnen maken, gebruikte ik voor de training van de modellen 50 mensen met dementie en 150 controles die qua leeftijd en geslacht overeenkwamen met de mensen met dementie.

De linker en rechter hippocampus van iedere deelnemer werden automatisch gesegmenteerd. Van deze segmentatie werd het volume berekend en een beschrijving van de vorm gemaakt. Een classificatiesysteem dat een onderscheid maakt tussen deelnemers die dement worden tijdens de studie en deelnemers die cognitief gezond blijven, werd getraind op deze data. Voor elke deelnemer schatte ik met het classificatiesysteem de kans dat deze deelnemer dementie ontwikkelt. De voorspellende waarde van de schattingen van de verschillende classificatiesystemen werd gemeten met de oppervlakte onder de ROC curve. Dit was voor volume, vorm en de combinatie van vorm respectievelijk 0.724, 0.743, en 0.766. Dat geeft gaan dat alle classificatiesystemen voor rond de driekwart van de gevallen een goede voorspelling kunnen geven en dat volume het minst goed presteerde en de combinatie van volume en vorm het beste.

Recent ook aangetoond dat de textuur van de hippocampus in MRI scans kan helpen bij het voorspellen of mensen met lichte cognitieve problemen ook daadwerkelijk dementie ontwikkelen. Het is echter nog niet bekend of de hippocampus textuur ook verbetering brengt in de voorspelling van dementie bij (cognitief) gezonde mensen. Daarom heb ik in hoofdstuk 4 het eerdere onderzoek met hippocampus volume en vorm uitgebreid met textuur. Ik onderzocht of hippocampus volume, vorm, textuur en alle mogelijke combinaties daarvan voorspellende informatie voor dementie bevatten. Ook bekeek ik hoe de voorspellende waarde varieerde met de tijd tussen de hersen MRI scan en de dementie diagnose, en met de mogelijke cognitieve klachten die deelnemers al hadden aan het begin van de studie.

Voor dit onderzoek werden bijna dezelfde data gebruikt als eerder, echter was er voor één deelnemer niet genoeg informatie beschikbaar om de textuurmaten te

bereken. Dezelfde volumes en vormmaten als eerder werden gebruikt en nieuwe textuurmaten werden toegevoegd. Een soortgelijk experiment met een classificatiesysteem werd gebruikt om te schatten wat de voorspellende waarde van de hippocampus maten voor dementie is.

Alle maten (hippocampus volume, vorm, textuur en de combinaties) bevatten significante voorspellende waarde, variërend van een oppervlakte onder ROC curve van 0.700 voor alleen textuur tot 0.788 voor de combinatie van volume en textuur. Als we in de analyse mensen die aan het begin van de studie al subjectieve geheugenklachten hadden buiten beschouwing laten, blijft de voorspellende waarde significant. Wel was het duidelijk dat naarmate er meer tijd zat tussen de MRI scan en de ontwikkeling van dementie, de voorspelling moeilijker werd: voor het voorspellen van dementie binnen een tijdsbestek van 3 jaar was het 0.935, voor 3 tot 6 jaar was het 0.809, en voor meer dan 6 jaar was het 0.632 (wat niet significant meer is). Ik concludeer dan ook dat de combinatie van maten beter werkt dan een enkele maat, en dat de combinatie van volume en textuur het beste werkt, waarbij we voorspellende waarde aan konden tonen tot 6 jaar voor het optreden van de dementie.

Ruimtelijke informatie en vormregressie

Vormanalyse in medische beelden is lastig door de hoge dimensionaliteit van vormbeschrijvingen en de gewoonlijk beperkte hoeveelheid voorbeeld vormen die beschikbaar zijn om een model te bouwen. Huidige methoden lossen dit probleem op door het toepassen van regularisatie in de vormenruimte, maar deze regularisatie houdt geen rekening met de ruimtelijke relaties binnen vormen. In de vormenruimte worden alle punten op een vorm als onafhankelijk van de rest gezien, maar het is zeer aannemelijk dat als twee punten op een vorm zich dicht bij elkaar bevinden ze een soortgelijk gedrag zullen vertonen. Het negeren van zulke ruimtelijke relaties kan leiden tot modellen die biologisch onwaarschijnlijk en moeilijk te interpreteren zijn. In hoofdstuk 5 stel ik voor om P-spline regressie te gebruiken voor medische vormanalyse. Omdat P-spline regressie is gebaseerd op een gegeneraliseerd linear model kan deze methode verschillende soorten uitkomsten voorspellen (bijvoorbeeld continue maten of klassen) en kan deze ook andere (niet-vorm) variabelen met geen of andere regularisatie bevatten.

Ik evalueerde de methode voor hippocampi vormanalyse op dezelfde 510 MRI scans als voor het onderzoek naar de hippocampus textuur. Daarvoor bekeek ik de relatie tussen de hippocampus vorm en leeftijd, geheugen score en geslacht. De voorgestelde methode behield dezelfde prestatie als een huidige methode, maar deed dit met realistischere, gladdere coëfficiënten velden en vorm effect velden die gemakkelijker te interpreteren zijn.

Fastr workflow software voor het bepalen van biomarkers

Voor grote medische beeldstudies, waaraan mogelijk meerdere medische centra deelnemen, zoals bevolkingsbeeldonderzoek en klinische studies waarin beelden verzameld worden, is het zeer belangrijk dat alle kwantitatieve beeldmaten op exact dezelfde manier afgeleid zijn omdat ze anders niet goed vergeleken kunnen worden. Doordat beeldstudies steeds omvangrijker worden, steeds meer data bevatten, en er de steeds complexere beeldanalyse workflows, is er een groeiend besef dat de data goed en veilig beheerd moeten worden. Dit kan gerealiseerd worden door hiervoor een gestandaardiseerde workflow benadering te kiezen.

In hoofdstuk 6 van dit proefschrift introduceer ik hiervoor Fastr, een softwarepakket dat geavanceerde workflows beheert en automatiseert. Fastr formaliseert de manier waarop analyse workflows worden geïmplementeerd en heeft een ingebouwd herkomst traceer mechanisme (provenance) voor het registreren van alles bewerkingen die worden uitgevoerd op de data en kan zodoende reproduceerbaarheid garanderen. Het provenance document wat bij elk gegenereerd resultaat wordt afgegeven is, als het ware, een certificaat voor betrouwbaarheid. Het modulaire ontwerp van Fastr staat toe om door middel van plugins samen te werken met bijna elk soort beeldarchief en rekenomgeving. Fastr is ontworpen om geconsolideerde workflows voor kwantitatieve beeldmaten goed en gemakkelijk te kunnen implementeren en beheren, zodat het snel op nieuwe data toegepast kan worden.

Samenvattend heb ik in dit proefschrift een aantal facetten met betrekking tot de extractie en het gebruik van kwantitatieve beeldmaten in medische beelden bekeken: de segmentatie van beelden die verkregen zijn met verschillende acquisitie protocollen, de validatie van hippocampus vorm en textuur als kwantitatieve beeldmaten voor de voorspelling van dementie, een statistische methode voor het interpreteerbaar analyseren van vormen, en een workflow framework voor de grootschalige, reproduceerbare extractie van beeldmaten. De methodes zijn geïmplementeerd en gevalideerd in de context van de kwantitatieve analyse van de hippocampus in het kader van dementie diagnostiek en prognostiek. Veel van de aspecten van de oplossingen en concepten die gepresenteerd zijn, zijn echter breder toepasbaar.

Dankwoord

> *“A friend is someone who knows all about you and still loves you.*"
>
> — **Elbert Hubbard**

Als promovendus word je geacht zelfstandig te werken, maar in de praktijk werk je toch veel met anderen samen en hebben allerlei mensen invloed op je werk. De ene keer is dat door directe samenwerking, de andere keer is het door goede feedback te krijgen. Soms krijg je net dat zetje dat je nodig hebt als je vast zit of die positieve energie van een goede brainstormsessie. Tenslotte is er ook gewoon de broodnodige afleiding van het werk om je hoofd leeg te maken. Ik wil graag iedereen bedanken die me heeft geholpen tijdens mijn promotietraject.

Ten eerste wil ik graag professor Wiro Niessen bedanken omdat hij mij de mogelijkheid heeft gegeven dit te doen. Wiro, je hebt me aangenomen als promovendus en me later ook gesteund toen ik meer begon te werken aan de infrastructuur. Soms was het lastig te combineren, maar ik ben erg blij dat ik de kans heb gekregen om mijn werkzaamheden te verleggen en toch mijn thesis af te kunnen maken. Ook voor het feit dat je zoveel hebt kunnen investeren in de infrastructuur projecten waarin ik zo geloof, ben ik je zeer dankbaar. Ik weet dat het moeilijk is geld voor dergelijke projecten te vinden en dat het soms zwemmen tegen de stroom in is. Het feit dat je een heel infrastructuur team hebt weten te creëren, is zeer bijzonder in de academische wereld. Ik heb er vertrouwen in dat we dat avontuur ook tot een goed einde zullen brengen.

Als mijn dagelijks begeleider is Marleen waarschijnlijk de persoon die de meeste invloed heeft gehad op dit proefschrift. Marleen, je was er altijd om feedback te geven wanneer ik dat nodig had, maar misschien nog wel belangrijker, je hebt me altijd gesteund en gezorgd dat ik steeds aan mijn thesis ben blijven werken. Dank voor alle steun en geduld door de jaren heen.

I would also like to thank all members of the committee to whom I will be defending my thesis. Professor Aad van der Lugt, professor Boudewijn Lelieveldt, and professor

Polina Golland thank you for your time and effort it costs to take part in the inner committee. Especially I would like to thank professor Polina Golland for coming all the way from Boston. I was very happy and mildly surprised to hear that you accepted the invitation, because it felt like more than I could ask for. Professor Meike Vernooij, professor Paul Eilers and Janne Papma thank you for your interest and willingness to take part in the defense.

Natuurlijk zou ik graag mijn paranimfen willen bedanken voor hun steun. Annegreet, jij begon met je master thesis vlak nadat ik zelf met mijn promotie ben begonnen. Daarna begon jij ook aan een promotie en later kwamen we samen in de infrastructuur groep. Je bent altijd actief geweest in de groep en ik heb altijd erg fijn met je samen gewerkt. Dank je wel voor de leuke tijd. Dennis, wij kennen elkaar al vanaf de middelbare school en je bent onderdeel van dat groepje vrienden die er altijd voor elkaar zijn. Het is grappig dat we ondanks dat we na de middelbare school andere studies zijn gaan doen, nu allebei promotieonderzoek doen in de beeldverwerking.

Toen ik net begon aan mijn promotie kwam ik terecht op een kamer met Fedde, Renske en Marius. Het heeft niet lang geduurd voordat ik erachter kwam hoeveel geluk ik daarmee had. Het was een gezellige kamer waar ik veel heb kunnen leren. Fedde, jij was altijd een groot stuk positieve energie, het was heerlijk om met jou te brainstormen en ik heb je input in mijn project erg gewaardeerd. Renske, jij was een geweldige collega en je regelde altijd van alles. Soms leek het wel of je het sociale hart van BIGR was. Ik denk dat iedereen uit die tijd zeer blij was met jou in de groep. Marius, wij hadden vaak veel leuke (al dan niet werk gerelateerde) discussies. Ook hebben we samen de prachtige nieuwbouw discovery show kunnen bekijken door ons raam.

Nu deel ik een kamer met het infrastructuur team. Na jaren van een relatief eenzaam PhD project was het een verademing om echt in een team te kunnen werken. Marcel, wij zijn samen het infrastructuur avontuur begonnen en het is prachtig te zien hoe het nu echt uitgroeit tot een mooi geheel. Graag wil ik je ook bedanken voor alle sociale randzaken en het introduceren van het mountainbiken. Esther en Annegreet, jullie hebben ervoor gezorgd dat er structuur kwam in het infrastructuur team. Jullie planning en orde heeft de efficiëntie van het team vergroot op een manier die heel prettig is. Marcel, Adriaan, Mattias and Thomas, it is great to have more technical colleagues to build the infrastructure together. Although we all have slightly different backgrounds (which I think helps complementing each other) we all like the good life (e.g. beers, food, (board) games).

I would like to thank all my (ex-)colleagues who made BIGR such a pleasant place to work. Especially I would like to thank those who made the group great when I arrived and the time after: Andres, Annegreet, Arna, Coert, Carolyn, Dirk, Emily, Erik, Esben,

Esther, Eugene, Fedde, Gennady, Gerardo, Ghassan, Gijs, Gokhan, Henri, Hortense, Hui, Ihor, Jifke, Marcel, Marius, Marleen, Mart, Michiel, Nora, Petra, Pierre, Rahil, Reinhard, Stefan, Theo, Veronika and Wyke. We had a good time together with great activities such as the outings, the BIGR-BAKR competition, and the movie nights. It made working at BIGR so much more fun. I also would like the thank the newest generation to keep the tradition of social activities alive, I really like the board game evenings. Thank you Florian, Gerda, Martijn, Sebastian, Willem, Zahra.

In the model-based medical image analysis subgroup we present our work to each other, gave each other feedback and we proofread (parts of) each others papers. Thank you for the feedback and all the things I learned from those presentations.

Ik wil graag Désirée en Petra bedanken voor alle hulp en steun voor de (sociale) activiteiten in de groep. Jullie zijn de stille krachten die veel mogelijk maken. Désirée, dank je wel dat je mij hebt willen begeleiden bij de procedures voor de afronding van de thesis.

In de nieuwbouw kregen we nieuwe buren, dit bleken zeer goede en gezellig buren. Rebecca, Rozanna, Carolina, Renske, Taihra en Anouk, door het goede contact met jullie heb ik eindelijk het gevoel gekregen echt deel te zijn van radiologie.

Ton, dank je wel voor de hulp met de opmaak van mijn thesis. Ik kwam je elke keer storen en je was altijd aardig en behulpzaam. Ik ben erg blij met het resultaat dat toch echt wel beter is geworden door je hulp.

Thomas, Baldur en Marcel, jullie enthousiasme maken dat ik BBMRI een zeer leuk project vind. Dat we met groepen van andere instituten zo kunnen samenwerken en echt iets samen bouwen is speciaal. Het is ook leuk om te zien dat we vanuit verschillende perspectieven toch goed een common ground kunnen vinden.

Marcel, Rebecca, Rozanna, Renske, Mattias, Ewoud, Gabriella, and Joost thank you for making the research drinks a success. It is great to have a social gathering and to get to know the colleagues better.

I would like to thank Adrian Dalca, Kristin McLeod, Maxime Taquet, Sinara Vijayan, and Katherine Gray for starting the MICCAI student board with me. It was a great experience to create something together and you were a great team. Furthermore I would like to thank Lena Filatova, Danielle Pace, Mathias Unberath, Bernhard Fuerst, Duygu Sarikaya, Hanne Klause for taking over the project and making sure it continues.

Johan en Paul, jullie hebben me geholpen met het hoofdstuk over de statistiek. Jullie

hebben de initiële implementatie van de methode gemaakt en het mij uitgelegd zodat ik er verder mee aan de slag kon. Het was niet het makkelijkste werk, maar ik heb er veel van geleerd en vind dat we er een mooi stuk over geschreven hebben.

Lauge and Mads, I would like to thank you for your cooperation on the hippocampus texture analysis. Lauge, because I conducted the research while I was working primarily as a scientific programmer, the project took longer than we hoped, but in the end we wrote a good paper together.

Meike en Arfan, jullie hebben altijd al mijn papers gelezen en er voor gezorgd dat de epidemiologische achtergrond en het data gebruik goed was. Frank, jij hebt mijn paper over de hippocampus textuur flink verbeterd en gezorgd dat de epidemiologische achtergrond en beschrijving veel correcter werd. Dank je wel voor alle goede inbreng.

Astrid, wij hebben heel wat uren samen in de trein doorgebracht. Daar bespraken allerlei zaken en hielden we elkaar op de hoogte van ons lief en leed. Dank je wel voor het aangenaam maken van het forenzen.

Dennis, Carolien, Dave, Sophie, Maarten, Toine, en Ana, het is fijn om zulke vrienden te hebben. We kennen elkaar al lang en hebben al veel meegemaakt. Het is zo prettig te weten dat jullie, ondanks dat we elkaar niet altijd even regelmatig zien, er altijd zijn.

Natuurlijk wil ik graag mijn familie bedanken. Lieve papa en mama, jullie zijn er altijd en hebben me altijd gesteund. Misschien zeg ik het niet vaak, maar ik weet dat ik echt geluk heb met zulke ouders. Carol, ik ben me er van bewust dat het promotietraject ook voor jou niet altijd even makkelijk was, vooral omdat jij het druk had met je inburgering in Nederland en het moederschap naast een universitaire studie. Dank je wel voor alles wat we samen hebben meegemaakt. Als laatste wil ik Yoyo en Yawen bedanken, jullie hebben het promotietraject er niet altijd makkelijker op gemaakt, maar jullie hebben me laten zien dat er meer in het leven is. Ik houd zielsveel van jullie en zou het voor geen goud anders gedaan hebben.

Publications

Journal Papers

- **H.C. Achterberg**, L. Sørensen, F.J. Wolters, W.J. Niessen, M.A. Ikram, M.W. Vernooij, M. Nielsen and M. de Bruijne, The Value of Hippocampal Volume, Shape and Texture for 11-year Prediction of Dementia: a population-based study, *submitted*.

- **H.C. Achterberg**, J.J. de Rooi, M.W. Vernooij, M.A. Ikram, W.J. Niessen, P.H.C. Eilers and M. de Bruijne, Spatially regularized shape analysis of the hippocampus using *P*-spline based shape regression, *submitted*.

- E.E. Bron, R.M.E. Steketee, G.C. Houston, R.A. Oliver, **H.C. Achterberg**, M. Loog, J.C. van Swieten, A. Hammers, W.J. Niessen, M. Smits and S. Klein, Diagnostic classification of arterial spin labeling and structural MRI in presenile early-stage dementia, *Human Brain Mapping*, 2014; 35(9):4916-4931.

- **H.C. Achterberg**, M. Koek and W.J. Niessen, Fastr: a workflow engine for advanced data flows in medical image analysis, *Frontiers in ICT*, 2016; 3(15)

- **H.C. Achterberg**, F. van der Lijn, den Heijer, M.W. Vernooij, M.A. Ikram, W.J. Niessen and M. de Bruijne, Hippocampal shape is predictive for the development of dementia in a normal, elderly population, *Human Brain Mapping*, 2014; 35(5), 2359-2371

- V. Prčkovska, **H.C. Achterberg**, M. Bastiani, P. Pullens, E. Balmashnova, B. M. ter Haar Romeny, A. Vilanova and A. Roebroeck, Optimal Short-Time Acquisition Schemes in High Angular Resolution Diffusion-Weighted Imaging, *International Journal of Biomedical Imaging*, 2013

Conference Papers

- **H.C. Achterberg**, D.H.J. Poot, F. van der Lijn, M.W. Vernooij, M.A. Ikram, W.J. Niessen and M. de Bruijne, Local appearance features for robust MRI brain structure segmentation across scanning protocols, *SPIE Medical Imaging*, 2013

Workshop Papers

- M. Koek, **H.C. Achterberg**, M. de Groot, E. Vast, S. Klein and W.J. Niessen, Population Imaging Study IT Infrastructure: An Automated Continuous Workflow Approach, *1 st MICCAI Workshop on MAnagement and Processing of images for Population ImagiNG*, 2015

- **H.C. Achterberg**, M. Koek and W.J. Niessen, Fastr: a workflow engine for advanced data flows, *1 st MICCAI Workshop on MAnagement and Processing of images for Population ImagiNG*, 2015

- A.G. van Opbroek, **H.C. Achterberg** and M. de Bruijne, Feature-Space Transformation Improves Supervised Segmentation Across Scanners, *Machine Learning Meets Medical Imaging*, 2015

- V. Terzopoulos, **H.C. Achterberg**, A. Plaisier, A.M. Heemskerk, M. de Groot, D.H.J. Poot, W.J. Niessen, J. Dudink and S. Klein, A 3D atlas of MR diffusion parameters in the neonatal brain, *MICCAI Workshop on Perinatal and Paediatric Imaging: PaPI 2012*, 2012

- **H.C. Achterberg**, F. van der Lijn, den Heijer, A van der Lugt, M.M.B. Breteler, W.J. Niessen and M. de Bruijne, Prediction of Dementia by Hippocampal Shape Analysis, *Machine Learning in Medical Imaging*, 2010

Conference Abstracts

- **H.C. Achterberg**, A.G. Van opbroek, M.Koek, D.Bos M.W. Vernooij, M.A. Ikram, H. Hulshoff pol, W.J. Niessen, and A. Van der lugt, Quantitative Imaging Biomarkers for Biobanking, Global Biobank Week, 2017

- T. Kroes, **H.C. Achterberg**, B. van Lew, M. Koek, A.G. van Opbroek, A. Versteeg, and B. Lelieveldt, Pipeline Inspection and Monitoring, Global Biobank Week, 2017

- M. Koek, **H.C. Achterberg**, A.G. van Opbroek, A. Versteeg, D. Bos, B. van Lew, T. Kroes, M. Zwiers, Y. Caspi, H. Hulshoff Pol, A. van der Lugt, and W.J. Niessen, Imaging Biomarker Infrastructure, Global Biobank Week, 2017

- M. Hansson, **H.C. Achterberg**, E.H.G. Oei and S. Klein, Evaluation of two multi-atlas cartilage segmentation models for knee MRI: data from the Osteoarthritis Initiative, 9th International Workshop on Osteoarthritis Imaging , in press

- E. Vast, M. Koek, M. de Groot, **H.C. Achterberg**, A. Dekker, J. van Soest, W.J. Niessen and S. Klein, Sharing medical imaging data for multi-center studies, EUDAT 3rd Conference , 2014

- **H.C. Achterberg**, M. de Bruijne, F. van der Lijn, den Heijer, M.W. Vernooij, M.A. Ikram and W.J. Niessen, Hippocampal Shape Predicts Development of Dementia in the General Population, RSNA 2011: 97nd Scientific Assembly and Annual Meeting of the Radiological Society of North America , 2011
- H.A. Vrooman, M.M.S. Jasperse, M. Koek, F. van der Lijn, **H.C. Achterberg**, M. de Bruijne, A. van der Lugt and W.J. Niessen, A computer-aided diagnosis system for the early and differential diagnosis of neurodegenerative disease, ECR Conference Proceedings 2011 , 2011
- **H.C. Achterberg**, F. van der Lijn, den Heijer, A van der Lugt, M.M.B. Breteler, W.J. Niessen and M. de Bruijne, Prediction of Dementia by Hippocampal Shape Analysis, Dutch BME conference 2011 , 2011

PhD Portfolio

PhD period 2009-2017
Departments Radiology & Medical Informatics
Research School ASCI

In-depth courses

Bayesian Methods and Bias Analysis (Erasmus MC)	2010
Advanced Pattern Recognition (ASCI)	2010
Int. Summer School on Biomedical Imaging, Berder Island, France (IEE EMBS)	2010
Modern Statistical Methods (NIHES)	2010
Scientific Writing in English for Publication (Erasmus MC)	2011
Principles of Research in Medicine and Epidemiology (NIHES)	2011
Summer School on Imaging in Neurology, Dubrovnik, Croatia (EIBIR)	2011
Training Course Speaking in Public (Erasmus MC)	2011
Survival Analysis (NIHES)	2012
English Biomedical Writing and Communication (Erasmus MC)	2012
Knowledge driven Image Segmentation (ASCI)	2012
Biomedical Image Analysis Summer School, Paris, France (MICCAI)	2012
C++ course (BIGR)	2013-2014
Software and Data Carpentry Instructor Training, Manchester, UK (Elixir)	2015

Seminars and workshops

Dutch Bio-Medical Engineering Conference - BME, Egmond aan Zee (oral presentation)	2011
Medical Imaging Symposium for PhD students - MISP, Utrecht (attendance)	2013
Medical Imaging Symposium for PhD students - MISP, Leiden (attendance)	2014
XNAT workshop - Saint Louis, USA (presentation and attendance)	2016
XNAT developer workshop - Rotterdam, the Netherlands (organization and attendance)	2017

Research seminar series

Biomedical Imaging Group Rotterdam Seminars, bi-weekly (4 presentations)	2009-2016
Medical Informatics Research Lunch, bi-weekly (2 presentations)	2009-2016

International conferences

Medical Image Computing and Computer-Assisted Intervention -
MICCAI, Beijing, China (poster at workshop) 2010
Medical Image Computing and Computer-Assisted Intervention -
MICCAI, Toronto, Canada (attendance) 2011
Radiological Society of North America -
RSNA, Chicago, USA (oral presentation) 2011
Medical Image Computing and Computer-Assisted Intervention -
MICCAI, Nice, France (attendance) 2012
SPIE: Medical Imaging -
SPIE, Orlando, USA (oral presentation) 2013
SPIE: Medical Imaging -
SPIE, San Diego, USA (attendance) 2014
Medical Image Computing and Computer-Assisted Intervention -
MICCAI, Munich, Germany (attendance) 2015
Global Biobank Week -
Stockholm, Sweden (2 oral presentations and 1 poster) 2017

Awards, nominations and grants

2nd best oral presentation award, Dutch Bio-Medical Engineering Conference 2011

Teaching experience

Supervision master thesis project - Lin Zhu,
Project: Optimal Multi-atlas Segmentation Strategies for Brain Structures 2012
Supervision student project - Vasilis Terzopoulos,
Project: A 3D Atlas Of MR Diffusion Parameters in the Preterm Nionatal Brain 2012
Teaching Introduction to Image Processing to medical students 2011-2013

Open-Source Software

Fastr: Workflow engine https://fastr.readthedocs.io
xnatpy: Pythonic XNAT client https://xnat.readthedocs.io

Other

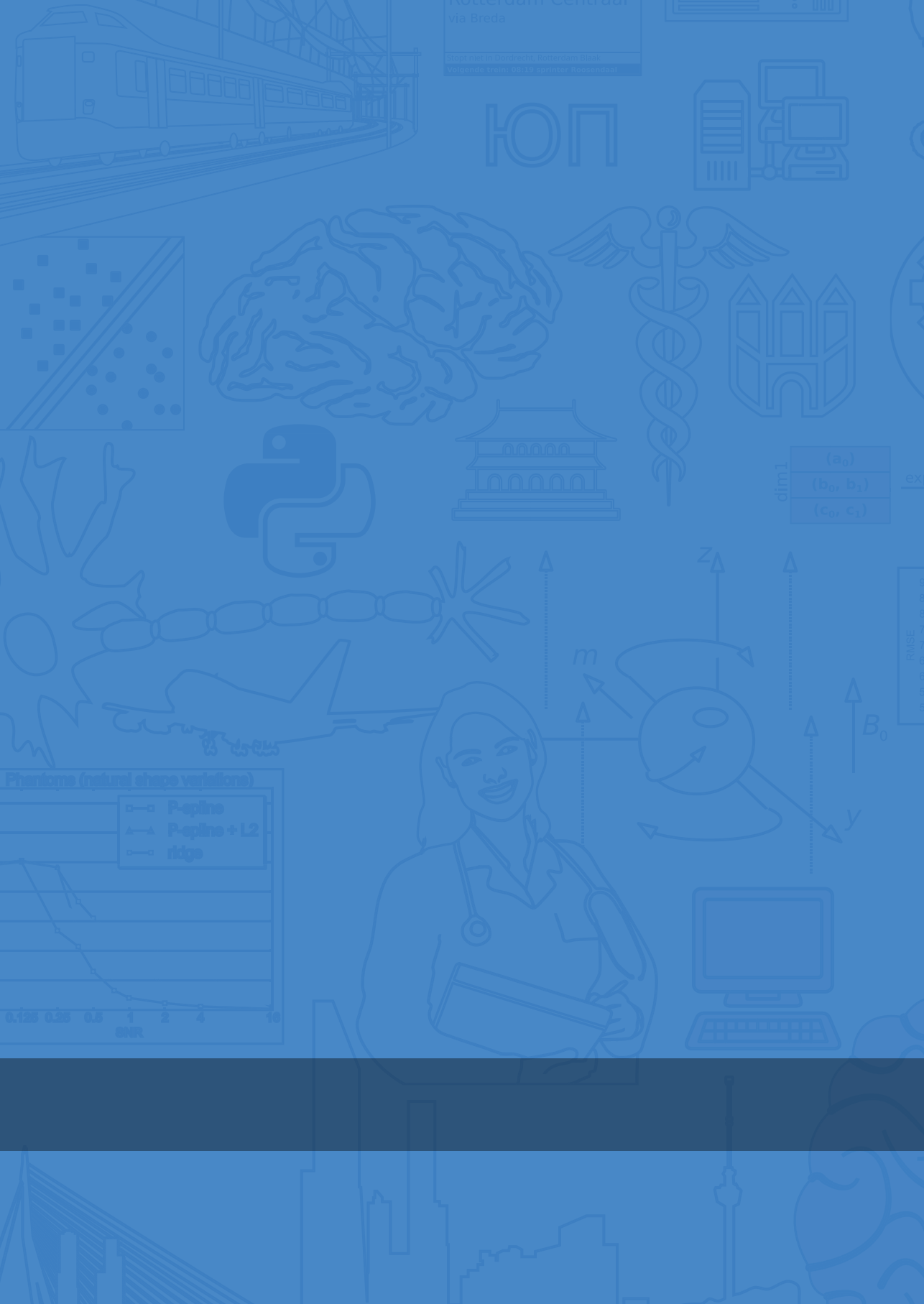
De Jonge Akademie on Wheels 2011, 2012
MICCAI student board, co-founder and president 2011-2013
MICCAI student board, advisory member 2014

About the author

Hakim Achterberg was born on 1984 in Tilburg, The Netherlands. He finished his secondary education at the Theresia Lyceum in 2003. Following high school, Hakim enrolled in the Bachelor's program in Biomedical Technology at the Eindhoven University of Technology. During this program Hakim first came in contact with Image Analysis. Having completed his Bachelor's degree, Hakim moved on to the Master's program in Biomedical Engineering and started specializing in Biomedical Image Analysis. During his Master's education, Hakim spent an exchange year at Umeå University in Sweden, where he strengthened his knowledge of computing science, image analysis and visualization further. Once back in the Netherlands Hakim performed an internships at the x-ray predevelopment group of Philips Healthcare in Best and the Image Processing and Analysis Group at Yale University led by James Duncan. In 2009 he finished his thesis "Optimal Acquisition Schemes in High Angular Diffusion Imaging" and received his M.Sc. degree from the Eindhoven University of Technology

In December 2009 Hakim started his PhD project in the BIGR group at the Erasmus Medical Centre Rotterdam. His work focused on shape analysis of brain structures in relation to neuro-degenerative diseases. He was supervised by W. Niessen (promotor) and M. de Bruijne (co-promotor). His work focused on two main points: first the analysis of the imaging data from the Rotterdam Scan Study, a large cohort study on aging, in which imaging biomarkers are related to the develop of dementia, and second there is the evaluation and improvement of methods for shape analysis. During his PhD Hakim also was involved of the creation of the MICCAI student board, a student body that organizes activities around the MICCAI conference, of which he was the president in 2012 and 2013.

As of December 2013, Hakim is working as a scientific programmer in the BIGR group, where he focusses on the IT infrastructure for large imaging studies. This includes creating software for the management, storage, transfer, processing and annotation of large imaging dataset. For the processing both automatic pipelines as well as integrated annotation tools are used. Hakim is the main developer for a number of tools, including the open-source workflow engine Fastr and python XNAT communication package xnatpy.


Rotterdam Centraal
via Breda
Stopt niet in Dordrecht, Rotterdam Blaak
Volgende trein: 08:15 sprinter Roosendaal
ЮП
Phenome (natural shape variations)
P-spline
P-spline + L2
ridge
0.125 0.25 0.5 1 2 4 16
SNR
dim1
(a_0)
(b_0, b_1)
(c_0, c_1)
z
m
B_0
y