

# Automatic Analysis of Brain Tissue and Structural Connectivity in MRI

Renske de Boer

# **Automatic Analysis of Brain Tissue and Structural Connectivity in MRI**

**Renske de Boer**

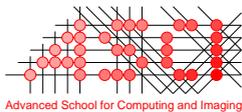
ISBN: 978-90-6464-470-2

Cover: Brain surface rendering based on an MRI scan of Renske de Boer.

Cover rendering and design by Michiel Schaap and Renske de Boer.

Printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen.

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



Advanced School for Computing and Imaging

This work was carried out in the ASCI graduate school.

ASCI dissertation series number 230.

The Rotterdam Study is supported by the Erasmus MC and Erasmus University Rotterdam, the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw), The Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry of Health, Welfare and Sports, the European Commission (DG XII) and the Municipality of Rotterdam. This study was further financially supported by the Netherlands Organization for Scientific Research (NWO) grants 948-00-010 and 918-46-615 and the Care4Me project.

Financial support for the publication of this thesis was provided by the Erasmus University Rotterdam, the Department of Radiology of the Erasmus MC, the ASCI graduate school, Alzheimer Nederland (Bunnik), and the Van Leersum Fund (Royal Netherlands Academy of Arts and Sciences, the Netherlands).

Copyright © 2011 by Renske de Boer

No part of this thesis may be reproduced, distributed, stored in a retrieval system or transmitted in any form or by any means, without the written permission of the author, or, when appropriate, of the publishers of the publications.

# AUTOMATIC ANALYSIS OF BRAIN TISSUE AND STRUCTURAL CONNECTIVITY IN MRI

AUTOMATISCHE ANALYSE VAN HERSENWEEFSELS EN STRUCTURELE  
HERSENCONNECTIVITEIT IN MRI

Proefschrift

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

Prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op  
vrijdag, 17 juni 2011 om 11:30 uur

door

Renske de Boer  
geboren te Enschede.



## **Promotiecommissie**

Promotoren: Prof.dr. W.J. Niessen  
Prof.dr. M.M.B. Breteler

Overige leden: Prof.dr. A. van der Lugt  
Prof.dr. M.A. van Buchem  
Dr. S. Ourselin

Copromotor: Dr.ir. H.A. Vrooman

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Quantitative image analysis in brain MRI studies . . . . .	2
1.1.1	Accuracy assessment in MR image analysis . . . . .	2
1.1.2	Reproducibility assessment in MR image analysis . . . . .	3
1.2	Thesis overview . . . . .	3
1.2.1	Brain tissue and white matter lesion segmentation . . . . .	4
1.2.2	Analysis of structural brain connectivity . . . . .	4
<b>I</b>	<b>Brain Tissue and White Matter Lesion Segmentation</b>	<b>7</b>
<b>2</b>	<b>White Matter Lesion Extension to Automatic Brain Tissue Segmentation</b>	<b>9</b>
2.1	Introduction . . . . .	10
2.2	Materials and methods . . . . .	11
2.2.1	Atlas data . . . . .	11
2.2.2	Test data . . . . .	12
2.2.3	Pre-processing of test data . . . . .	13
2.2.4	Brain tissue segmentation . . . . .	13
2.2.5	White matter lesion segmentation . . . . .	14
2.2.6	Post-processing . . . . .	16
2.2.7	Evaluation measures . . . . .	16
2.3	Experiments . . . . .	17
2.3.1	Optimization of $\alpha$ and $\beta$ . . . . .	17
2.3.2	Comparison of different types of atlas registration . . . . .	17
2.3.3	Quantitative analysis of brain tissue and WML segmentation . . . . .	18
2.3.4	Qualitative analysis in large dataset . . . . .	19
2.3.5	Association between age and WML volume . . . . .	19
2.4	Results . . . . .	19
2.4.1	Comparison of different types of atlas registration . . . . .	19
2.4.2	Quantitative analysis of brain tissue and WML segmentation . . . . .	21
2.4.3	Qualitative analysis in large dataset . . . . .	23
2.4.4	Association between age and WML volume . . . . .	26
2.5	Discussion and conclusion . . . . .	28

<b>3</b>	<b>Accuracy and Reproducibility of Automatic Brain Tissue Segmentation</b>	<b>31</b>
3.1	Introduction . . . . .	32
3.2	Materials and methods . . . . .	33
3.2.1	Imaging data . . . . .	33
3.2.2	Segmentation . . . . .	33
3.2.3	Experiments . . . . .	35
3.3	Results . . . . .	37
3.3.1	Segmentation . . . . .	37
3.3.2	Accuracy . . . . .	39
3.3.3	Reproducibility . . . . .	41
3.3.4	Sample size calculations for longitudinal studies . . . . .	42
3.4	Discussion and conclusion . . . . .	43
	Appendix - Results using original preprocessing . . . . .	46
<b>4</b>	<b>Comparison of Brain Tissue Segmentation of Longitudinal MR Images</b>	<b>51</b>
4.1	Introduction . . . . .	52
4.2	Materials and methods . . . . .	53
4.2.1	Data . . . . .	53
4.2.2	Brain tissue segmentation methods . . . . .	54
4.2.3	Comparison . . . . .	54
4.3	Results . . . . .	56
4.3.1	Visual inspection . . . . .	56
4.3.2	Volume change rates . . . . .	56
4.3.3	Atrophy maps . . . . .	57
4.4	Conclusion and discussion . . . . .	62
<b>II</b>	<b>Analysis of Structural Brain Connectivity</b>	<b>65</b>
<b>5</b>	<b>Statistical Analysis of Minimum Cost Path Based Structural Connectivity</b>	<b>67</b>
5.1	Introduction . . . . .	68
5.2	The SAMSCo framework . . . . .	69
5.2.1	Connectivity . . . . .	70
5.2.2	Construction of the mcp-network . . . . .	73
5.2.3	Statistical analysis . . . . .	73
5.3	Experiments . . . . .	75
5.3.1	Data . . . . .	75
5.3.2	Start and target regions . . . . .	76
5.3.3	Age prediction . . . . .	76
5.3.4	Classification of WML load and WM atrophy . . . . .	77
5.3.5	Results . . . . .	78
5.4	Conclusion and discussion . . . . .	81

---

<b>6</b>	<b>Reproducibility of the SAMSCo Framework</b>	<b>87</b>
6.1	Introduction . . . . .	88
6.2	Materials and methods . . . . .	88
6.2.1	Data . . . . .	88
6.2.2	Experiments . . . . .	89
6.3	Results . . . . .	90
6.4	Conclusion and discussion . . . . .	90
<b>7</b>	<b>Summary and Discussion</b>	<b>95</b>
7.1	Summary . . . . .	96
7.1.1	Brain tissue and white matter lesion segmentation . . . . .	96
7.1.2	Analysis of structural brain connectivity . . . . .	97
7.2	Discussion . . . . .	97
7.2.1	Brain tissue and white matter lesion segmentation . . . . .	97
7.2.2	Analysis of structural brain connectivity . . . . .	99
	<b>References</b>	<b>101</b>
	<b>Samenvatting</b>	<b>109</b>
	<b>Dankwoord</b>	<b>117</b>
	<b>Publications</b>	<b>121</b>
	<b>PhD portfolio</b>	<b>123</b>
	<b>About the author</b>	<b>125</b>



# **Chapter 1**

## **Introduction**

## 1.1 Quantitative image analysis in brain MRI studies

Magnetic resonance imaging (MRI) has become an indispensable tool in clinical practice and biomedical research, as it facilitates the non-invasive investigation of brain anatomy and function. In clinical practice, MRI scans of individual patients are used for diagnosis and therapy planning. In clinical or population imaging studies, groups of subjects are studied to provide insights into normal brain development, brain aging, and morphological and functional brain changes associated with different disorders. Several designs of such studies exist. In a case-control study a group of patients is compared to a group of healthy controls. In a cross-sectional study a population, or representative subset, is studied at a defined time. In a longitudinal study changes over time are studied using repeated measurements.

The Rotterdam Scan Study, the neuroimaging component of the Rotterdam Study (Hofman et al., 2009), includes both cross-sectional and longitudinal brain MRI studies. In this population-based study the aging brain is studied in healthy middle-aged and elderly subjects. Thousands of brain MRI scans are obtained from which a range of variables are derived like tissue volumes or measurements of white matter microstructural integrity. These variables are then related to, for example, age or cognitive function (e.g., Ikram et al., 2010, 2008; Vernooij et al., 2008, 2009).

Owing to the sheer magnitude and complexity of the MRI data, manual extraction of these variables is a laborious procedure. Volume measurements are time consuming and require trained observers. Complex analyses, such as the analysis of white matter integrity on a voxel level, are even impossible without a computer. More importantly, analysis by human observers can hamper the reproducibility by both intra- and inter-observer variability. Brain MRI studies would therefore benefit from automatic image analysis techniques that yield accurate and reproducible quantitative results.

In this thesis, we develop two neuroimage analysis techniques, namely segmentation of brain tissues and white matter lesions and analysis of structural brain connectivity. We evaluate these techniques on both accuracy and reproducibility using MRI scans from the Rotterdam Scan Study.

### 1.1.1 Accuracy assessment in MR image analysis

In medical image analysis, accuracy is defined as the ability to produce results identical to the reference standard, referred to as ground truth. These results can, for example, be the segmentation of a tissue or the quantification of a brain structure property. In neuroimaging it is difficult to obtain a ground truth as it is often unknown what the true underlying anatomy is.

In practice, medical image analysis results are often compared to results obtained by human experts. Human annotations are subject to intra- and inter-observer

variability and may be biased due to the method used by the observers, e.g. a thresholding brush. Evaluation of an automated method is therefore often carried out by comparing differences between automatic and manual analyses to the differences between two or more manual analyses. In this thesis, we utilize this approach for evaluating the accuracy of automatic segmentation methods.

For some complex analyses, e.g. the analysis of diffusion MRI, no manual in vivo reference standard can be obtained. In these cases, a direct evaluation can be carried out by relating the quantitative output of the image analysis method to different variables (e.g. age, sex, disease type) and look for known associations. An example is when automatically measured diffusion measures, like fractional anisotropy, are related to age and the found associations are compared to literature (e.g., O'Sullivan et al., 2001). This approach is applied in this thesis as evaluation of a newly developed image analysis technique and as additional evaluation of segmentation methods.

### 1.1.2 Reproducibility assessment in MR image analysis

For reproducibility in quantitative medical image analysis not only results should be similar for the same input scan, but also for a second scan of the same subject, made within a short time interval. The reproducibility of the analysis does therefore largely depend on the way the analysis method deals with variation in the input scan due to the reproducibility of the acquisition (Clark et al., 2006; Jovicich et al., 2009; Shuter et al., 2008). In addition, repositioning of the subject will result in different partial volume effects due to discretization and thereby affect the reproducibility. Also, if the second scan is not made immediately after the first scan, there can be natural variations in, for example, fluid balance which will affect the brain and thereby the results of the analysis. These influences are also present in a longitudinal study and should therefore be taken into consideration.

The reproducibility of a method is the variation of its results given the same or similar input. It influences the minimum number of samples required in a study aimed at finding a certain effect with sufficient power (Fox et al., 2000; Schott et al., 2006). The better the reproducibility, the fewer samples are required.

In this thesis, we use subjects from the Rotterdam Scan Study that were scanned twice within a short time interval to evaluate the reproducibility of previously published algorithms and algorithms that have been developed as part of this thesis.

## 1.2 Thesis overview

This thesis consists of two parts. In the first part (Chapter 2, 3 and 4) automatic methods for brain tissue and white matter lesion segmentation are developed and evaluated. In the second part (Chapter 5 and 6) a framework for statistical analysis of structural brain connectivity is presented and evaluated.

### 1.2.1 Brain tissue and white matter lesion segmentation

The brain consists of two brain tissue types, namely gray matter (GM) and white matter (WM), and is surrounded by cerebrospinal fluid (CSF). Segmenting these brain tissues and the CSF enables volume measurements. These volumes can be compared across subjects or over time in order to study, for example, brain atrophy, which is a marker for neurodegeneration (Frisoni et al., 2010). Brain tissue segmentation has also been used as a preprocessing step for other methods, e.g. voxel based morphometry (Ashburner and Friston, 2000) or cortical thickness measurements (e.g., Eskildsen and Ostergaard, 2006; Fischl and Dale, 2000).

White matter lesions (WML) are regions in the white matter that appear more hyperintense than the surrounding tissue on T2-weighted MRI scans (Malloy et al., 2007). WML occur often in the elderly and are associated with cognitive decline (de Groot et al., 2002) and increased risk of stroke (Vermeer et al., 2003) and dementia (Prins et al., 2004). Segmentation of WML enables studying the cause and effect of these lesions.

In Chapter 2 we propose a method for automatic white matter lesion segmentation, that can be used as an extension to a brain tissue segmentation method. In addition, we optimize a previously proposed brain tissue segmentation method in combination with the WML segmentation extension. We compare the accuracy and reproducibility of this brain tissue segmentation method and several previously proposed methods in Chapter 3. In Chapter 4 we compare the performance of two brain tissue segmentation methods, extended with white matter lesion segmentation, on longitudinal data.

### 1.2.2 Analysis of structural brain connectivity

The gray matter in the brain is connected through white matter bundles. Studying this connectivity increases the understanding of brain anatomy, maturation and aging. Additionally, investigating the differences in connectivity between a group of patients and a group of healthy controls can provide valuable information about pathology.

Brain connectivity can be studied in several ways. Functional MRI (fMRI) can be used to investigate which gray matter regions show the same functional activity and are therefore likely to be connected. This is referred to as functional connectivity. Alternatively, for estimating structural connectivity the white matter bundles should be studied. Diffusion MRI enables the study of white matter bundles by making use of the preferential diffusion of water molecules along these bundles (Beaulieu, 2002).

In Chapter 5 we propose a new framework for the analysis of structural brain connectivity in groups of subjects. Using diffusion MRI, we construct structural connectivity networks and analyze them using statistical methods. We perform several proof of principle experiments to demonstrate the information contained in the

networks. The reproducibility of the framework is reported in Chapter 6.



## **Part I**

# **Brain Tissue and White Matter Lesion Segmentation**



# Chapter 2

## White Matter Lesion Extension to Automatic Brain Tissue Segmentation

Based on:

Renske de Boer, Henri A. Vrooman, Fedde van der Lijn, Meike W. Vernooij, M. Arfan Ikram, Aad van der Lugt, Monique M.B. Breteler, Wiro J. Niessen, White matter lesion extension to automatic brain tissue segmentation on MRI, *NeuroImage*, 2009.

### Abstract

A fully automated brain tissue segmentation method is optimized and extended with white matter lesion segmentation. Cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM) are segmented by an atlas-based  $k$ -nearest neighbor classifier on multi-modal magnetic resonance imaging data. This classifier is trained by registering brain atlases to the subject. The resulting GM segmentation is used to automatically find a white matter lesion (WML) threshold in a fluid-attenuated inversion recovery scan. False positive lesions are removed by ensuring that the lesions are within the white matter. The method was visually validated on a set of 209 subjects. No segmentation errors were found in 98% of the brain tissue segmentations and 97% of the WML segmentations. A quantitative evaluation using manual segmentations was performed on a subset of 6 subjects for CSF, GM and WM segmentation and an additional 14 for the WML segmentations. The results indicated that the automatic segmentation accuracy is close to the interobserver variability of manual segmentations.

## 2.1 Introduction

Brain tissue segmentation on structural magnetic resonance imaging (MRI) has received considerable attention. Quantitative analysis of MR images of the brain is of interest in order to study the aging brain in epidemiological studies, to better understand how diseases affect the brain and to support diagnosis in clinical practice. Manual quantitative analysis of brain imaging data is a tedious and time-consuming procedure, prone to observer variability. Therefore, there is a large interest in automatic analysis of MR brain imaging data, especially segmentation of cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM). In the last decade several automatic brain tissue segmentation methods have been proposed, often based on T1-, T2- or proton density-weighted MR images. Some use a fixed set of labeled samples, that were derived from manual segmentations, to train the classifier (Amato et al., 2003; Anbeek et al., 2005). This has, however, some disadvantages as it is dependent on the MRI sequence, requires a laborious training stage and is limited to the MRI intensity variations captured in the training set. Therefore, some studies developed methods to obtain subject-specific training samples labeled by, for example, clustering (Barra and Boire, 2000; Harris et al., 1999), Gaussian mixture models (Lemieux et al., 2003; Ruan et al., 2000) or atlas registration (Cocosco et al., 2003; Song et al., 2006). These methods are independent of intersubject intensity variations and MRI sequence. Another option, often used nowadays, is updating both the classification and the model parameters in an iterative process (Ashburner and Friston, 2005; Awate et al., 2006; Kovacevic et al., 2002; van Leemput et al., 1999; Ruf et al., 2005; Zhang et al., 2001). This type of method is more complicated but it is also independent of intersubject intensity variations and MRI sequence. Several of these studies have evaluated their method on a reasonably large group varying between 34 to 71 subjects (Cocosco et al., 2003; Harris et al., 1999; Kovacevic et al., 2002; Lemieux et al., 2003; Song et al., 2006; Vrooman et al., 2007). The method developed by Zhang et al. (2001) is incorporated as FSL's brain tissue segmentation method, FAST, and is used in multiple studies, similar to the SPM brain tissue segmentation method by Ashburner and Friston (2005).

Besides automatic brain tissue segmentation, automatic WML segmentation has also received considerable interest. White matter lesions (WML) are commonly found in elderly subjects and are associated with cognitive decline (de Groot et al., 2002) and increased risk of stroke (Vermeer et al., 2003) and dementia (Prins et al., 2004). Recent studies often use T2-weighted or fluid-attenuated inversion recovery (FLAIR) scans in which white matter lesions are hyperintense. Several automatic segmentation methods have been developed based on intensity alone (Admiraal-Behloul et al., 2005; DeCarli et al., 2005; Jack et al., 2001) or including also spatial- (Anbeek et al., 2004), texture- (Krugger et al., 2008) or shape-information (Alfano et al., 2000). It is difficult to compare the reported accuracies of these WML segmentation methods. Often different evaluation measures

are used and some of these measures depend on the WML load of the subject (Admiraal-Behloul et al., 2005). Furthermore, automatic WML segmentations are often evaluated by comparison to manual segmentations and the evaluation is therefore influenced by the manual segmentation protocol. The robustness of an automatic segmentation method can be demonstrated by applying the method to a large dataset. Only some studies evaluated their WML segmentation method on datasets of 100 or more subjects (Admiraal-Behloul et al., 2005; Kruggel et al., 2008; Maillard et al., 2008).

In this chapter a fully automated method for CSF, GM and WM segmentation based on multimodal MRI data is optimized and extended with WML segmentation. The contribution to the existing literature is threefold. Firstly, we evaluate different atlas registration methods for a brain tissue segmentation method presented by Cocosco et al. (2003) and Vrooman et al. (2007) where atlas registration is used to automatically train a  $k$ -nearest neighbor classifier. Different types of registration are compared: single- versus multiple-atlas registration; affine versus B-spline based non-rigid registration at different control point spacings; and registration of a varying number of atlases. Secondly, the method is extended with an automatic WML segmentation. This segmentation method uses the GM classification to determine a white matter lesion intensity threshold value in the FLAIR scan. Thirdly, the method is qualitatively validated on a large dataset of 209 elderly subjects. A quantitative evaluation is performed on a small subset using manual segmentations.

## 2.2 Materials and methods

### 2.2.1 Atlas data

Twelve atlases have been obtained by manual segmentation of scans from the Rotterdam Scan Study (de Leeuw et al., 2001), that were acquired in 1995-1996. This population-based imaging study is aimed at investigating determinants of age-related neurologic diseases among elderly persons. The twelve subjects were female and had a mean age (sd) of 64 (1.8) years. MR brain imaging was performed on a 1.5 T Siemens scanner using a quadrature head coil. An inversion recovery double contrast, 3D half-Fourier acquisition single-shot turbo spin echo (HASTE) sequence was performed (TR=2800 ms, TI=440 ms, matrix=192×256, FOV=256×256 mm<sup>2</sup>, 128 contiguous sagittal slices of 1.25 mm). The voxel dimensions were 1×1×1.25 mm<sup>3</sup>. Two HASTE modules were sequentially acquired after the inversion pulse (effective TE of 29 ms and 440 ms). Each HASTE module combined non-selective radio frequency excitations to provide a short interecho spacing of 3.9 ms. We used the first HASTE module (HASTE-Odd), with contrast similar to an inverted T1-weighted image, for subsequent processing. The datasets were manually segmented by two trained physicians, using a paintbrush method in the tool 'Display' from the Montreal Neurological Institute (MNI). The labeled output contained four

**Table 2.1:** *Subjects per sex and age group*

	Total subjects included			60 - 70 year			70 - 80 year			80+ year		
	Total <sup>a</sup>	TMS <sup>b</sup>	LMS <sup>c</sup>	Total <sup>a</sup>	TMS <sup>b</sup>	LMS <sup>c</sup>	Total <sup>a</sup>	TMS <sup>b</sup>	LMS <sup>c</sup>	Total <sup>a</sup>	TMS <sup>b</sup>	LMS <sup>c</sup>
Men	110	5	11	52	4	9	39	1	2	19	0	0
Women	105	1	9	47	1	7	40	0	1	18	0	1
Total	215	6	20	99	5	16	79	1	3	37	0	1

<sup>a</sup>'Total' is the number of subjects in the corresponding sex/age group.

<sup>b</sup>'TMS' is the number of subjects with total manual segmentation by 2 observers.

<sup>c</sup>'LMS' states the number of subjects with manual WML segmentation by 1 observer.

labels, background (BG), CSF, GM and WM.

## 2.2.2 Test data

Imaging data from the Rotterdam Scan Study (Hofman et al., 2007), acquired in 2005-2006, were used for the evaluation of the method. Scans were obtained on a 1.5 T GE scanner using an 8-channel head coil. The protocol included three high-resolution axial MRI sequences, i.e. a T1-weighted (T1w) 3D fast RF spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse sequence (TR=13.8 ms, TE=2.8 ms, TI=400 ms, FOV=25×25 cm<sup>2</sup>, matrix=416×256 (interpolated to 512×512 resulting in voxel sizes of 0.49×0.49 mm<sup>2</sup>), flip angle=20°, NEX=1, bandwidth (BW)=12.50 kHz, 96 slices with a slice thickness of 1.6 mm zero-padded in the frequency domain to 0.8 mm), a proton density-weighted (PDw) sequence (TR=12,300 ms, TE=17.3 ms, FOV=25×25 cm<sup>2</sup>, matrix=416×256, NEX=1, BW=17.86 kHz, 90 slices with a slice thickness of 1.6 mm), and a fluid-attenuated inversion recovery (FLAIR) sequence (TR=8000 ms, TE=120 ms, TI=2000 ms, FOV=25×25 cm<sup>2</sup>, matrix=320×224, NEX=1, BW=31.25 kHz, 64 slices with a slice thickness of 2.5 mm).

For this study, 215 subjects were used, that were randomly chosen from age- and sex-specific strata from the total Rotterdam Scan Study population. Table 2.1 shows the subjects per age group and sex. All subjects were non-demented and none had multiple sclerosis, even though the latter was no exclusion criteria. Two independent physicians performed manual segmentations of brain tissues in six datasets, using a paintbrush method in the MNI-tool 'Display'. Scans were manually segmented into CSF, GM and WM on the T1w volumes and WMLs were segmented on the FLAIR volumes. Another 20 subjects had manual WML segmentations by one expert. Six of these 20 subjects were randomly selected for parameter optimization of the WML segmentation and were therefore not included in the final set of 209 test subjects.

### 2.2.3 Pre-processing of test data

We use T1w and PDw MR images for the segmentation of CSF, GM and WM by the previously proposed method. White matter lesions are segmented using a FLAIR scan. All scans are registered to the T1w image by rigid registration and resampled by trilinear interpolation to the resolution of the T1w image. A brain mask, to exclude for example the cerebellum, eyes and skull, is obtained by non-rigid registration of a manual segmented brain mask to the T1w image using Elastix<sup>1</sup> (Klein et al., 2010). Subsequently, the scans are corrected for intensity non-uniformity using the N3 method described by Sled et al. (1998) within the brain mask. A range matching procedure on the PDw and T1w images ensures unbiased feature weights in the subsequent  $k$ -nearest neighbor (kNN) classification. This procedure excludes 4% of the voxels with lowest intensities and 4% of the highest intensity voxels while rescaling the remaining intensities between zero and one.

### 2.2.4 Brain tissue segmentation

In the first stage of the algorithm, CSF, GM and WM are segmented using an automatically trained kNN classifier which is an extension of the work by Cocosco et al. (2003). Training samples for the kNN classifier are obtained from the subject itself by atlas-based registration. This is accomplished by registration of single- or multiple-atlases to the subject, using the Image Registration Toolkit (IRTK)<sup>2</sup> introduced by Rueckert et al. (1999). The centers of mass are aligned first, followed by an affine registration and, in some experiments, a spline-based non-rigid registration using decreasing control point spacings (20 mm, 10 mm, 5 mm, 2.5 mm). All registrations are driven by normalized mutual information. The transformations are obtained by registration of the grayscale images and are applied to the labeled images.

The registered labeled images are equally weighted in the averaging to create tissue probability maps (TPMs) of the three tissue classes and a fourth background class. These TPMs give the probability, represented by a value between zero and one, of a certain voxel belonging to a certain (tissue) class. The TPMs are thresholded in order to get candidate training samples with a predefined probability to belong to a specific label. A threshold of 0.7 is chosen, which is shown by Vrooman et al. (2007) to be a good threshold value for these tissue types, especially CSF. When the CSF TPM is thresholded at this value, it includes not only ventricular CSF but also sulcal CSF, contrary to higher thresholds. Inclusion of sulcal CSF will result in more variations in the CSF samples.

For all four classes, 7500 candidate training samples are randomly taken from the spatial locations masked by the thresholded TPMs. The features of the samples consist of the intensity values of the PDw and T1w images at the sample locations.

---

<sup>1</sup>Elastix is available at <http://elastix.isi.uu.nl/>

<sup>2</sup>IRTK is available at <http://www.doc.ic.ac.uk/~dr/software/>

A pruning step is applied to the initial set of samples to remove samples with incorrect labels (Cocosco et al., 2003). First, a minimal spanning tree of the samples in feature space is created. In an iterative process, the pruning algorithm removes connections whose length exceeds a threshold value equal to a constant multiplied with the average length of the other connections of a sample. At every iteration the threshold value decreases. This process is continued until every tissue class has a unique main cluster in feature space. A main cluster is defined as the cluster containing more samples of a certain class than the other clusters. A cluster is an unique main cluster when it is the main cluster for a single class. The final step removes all samples that are not connected or that are not in their main cluster. A  $k$ -nearest neighbor classifier performs the final classification based on the pruned sample set. A value of 45 is used for  $k$ , similar to Vrooman et al. (2007) and Cocosco et al. (2003). The kNN implementation uses a fast nearest neighbor lookup library called ANN<sup>3</sup>.

## 2.2.5 White matter lesion segmentation

Upon completion of the first step of the algorithm in which CSF, GM and WM are segmented, WMLs that are present in the brain are misclassified as GM with a ‘halo’ of WM. Figure 2.1 shows a slice of a FLAIR image and an automatic brain tissue segmentation with an outline of a corresponding manual WML segmentation. In the FLAIR image the WMLs are clearly visible as hyperintensities and thus do not resemble GM intensity. Therefore a histogram is created of all voxels in the FLAIR image that are classified as GM. An example of such a FLAIR histogram is shown in Figure 2.2. The highest peak in the histogram corresponds to the true gray matter voxels. The FLAIR intensities corresponding to white matter lesion voxels are located to the right of this peak. The histogram is smoothed by a convolution with a Gaussian kernel ( $\sigma_G = 4$  FLAIR intensity units). This makes it possible to estimate the FLAIR intensity corresponding to the center of the GM peak by the histogram bin containing most true positive gray matter voxels. The peak is approximated by a Gaussian function with the mean ( $\mu$ ) defined as the peak center location and the standard deviation ( $\sigma$ ) calculated using the full width at half maximum. The threshold  $T$  for the WML is subsequently defined as:

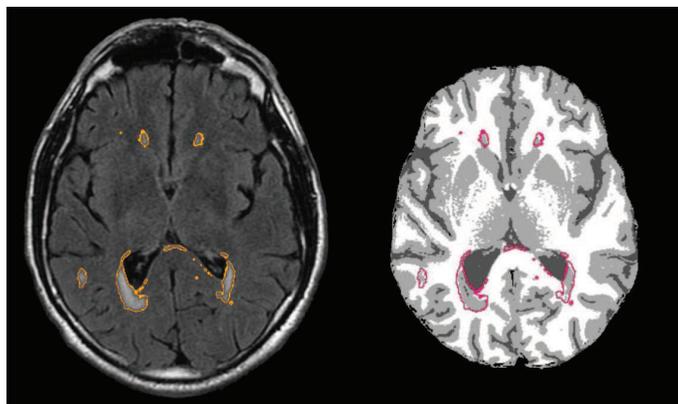
$$T = \mu + \alpha\sigma,$$

with  $\alpha$  a threshold parameter to be optimized. The threshold for the example in Figure 2.2 is shown as a dashed line. The WML segmentation is obtained by thresholding the FLAIR image.

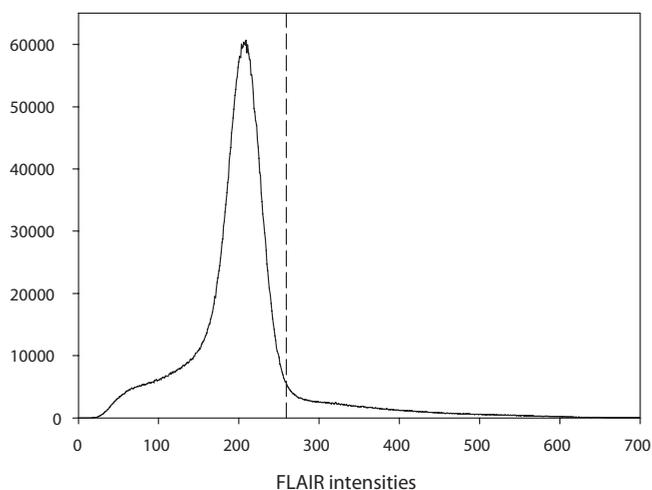
Upon thresholding, a number of regions is wrongly classified as WML. In several segmentation methods, information from neighboring voxels is utilized in order to improve segmentation performance, for example using Markov random

---

<sup>3</sup>The ANN library is available at <http://www.cs.umd.edu/~mount/ANN/>



**Figure 2.1:** Axial slice of a FLAIR image (left) and a masked automatic brain tissue segmentation. In the automatic segmentation dark gray corresponds to CSF, light gray to GM and white to WM. The area within the colored lines is the WML segmentation as performed by an observer. It is clear that hyperintensities on the FLAIR image are segmented as GM with a halo of WM.



**Figure 2.2:** Example of a histogram of FLAIR intensities of the voxels classified as GM by the brain tissue segmentation method. The final threshold is shown as a dashed line.

fields (Khayati et al., 2008; van Leemput et al., 2001, 1999; Ruan et al., 2000). In our case, most of the false positive WML are clearly located outside the white matter and a relatively simple measure suffices. For every lesion, the following fraction is calculated,

$$\text{WM fraction} = \frac{\text{number of neighboring WM voxels}}{\text{number of neighboring CSF+GM voxels}'}$$

within the one voxel wide surroundings, obtained using a 3D 18-neighborhood relation (spherical kernel with a radius of 3). A lesion is defined as a group of connected voxels, using a 3D 18-connectivity, with a WML label. If the fraction mentioned above is smaller than the optimized parameter  $\beta$  the lesion is reclassified as GM.

### 2.2.6 Post-processing

The non-rigidly transformed brain mask sometimes includes parts outside the brain. These parts, mainly dura and skull, are classified by the automatic segmentation method as mixtures of GM, WM and WML depending on their intensities. Especially false positive WML outside the brain can have a large influence on the relatively small total volume of WML. A simple post-processing step is applied to remove brain tissue and WML located outside the brain. This step uses the brain tissue and WML classification to find components of connected voxels, defined by 3D 18-connectivity, with the same label. For every component, the number of neighboring background and non-background voxels are counted. Just like the false positive WML reclassification step, a 3D 18-neighborhood relation is used for defining neighboring voxels. If the ratio of background voxels to non-background voxels is larger than a certain value the component is relabeled as background. Pilot experiments showed that the improvement of the brain mask is not very sensitive to this value. A team of experts chose the value to be 0.4 by visual inspection of several subjects.

### 2.2.7 Evaluation measures

The automatic segmentations are evaluated quantitatively by calculation of the similarity index (SI) or Dice coefficient between the automatic segmentation and the manual segmentations.

$$SI = \frac{2(S_1 \cap S_2)}{S_1 + S_2}$$

Where  $S_j$  with  $j = 1, 2$  is a segmented volume and  $(S_1 \cap S_2)$  is the overlap of  $S_1$  and  $S_2$ . The SI is also used as a measure for the interobserver variability.

The true positive fraction (TPF), or sensitivity, and the extra fraction (EF) are also used for evaluation. The extra fraction is a measure for oversegmentation. EF

and TPF are defined as follows:

$$EF = \frac{FP}{TP + FN}$$

$$TPF = \frac{TP}{TP + FN}$$

by false positives (FP), true positives (TP) and false negatives (FN). For the six subjects with total manual segmentation (CSF, GM, WM and WML) the quantitative evaluation is only performed for the voxels that are present in the manual segmentation. This means that the subjects are masked by the total area of the manual CSF, GM, WM and WML segmentations.

In case of small white matter lesions, a slight oversegmentation or undersegmentation will easily lead to low similarity indices, whereas the influence on the total white matter lesion load is small. Therefore, for evaluation of WML segmentations, we also consider an evaluation measure which describes the maximum and average boundary localization error. This error is computed by calculating a distance map from the border of the lesion in one segmentation to the border of the corresponding lesion in the other segmentation, and subsequently computing the mean and maximum distance per subject. Because the distance measure is not symmetric, each WML segmentation is compared with the other and vice versa and the result is averaged. Of course the distance can only be measured for the lesions that have been identified in both segmentations and therefore we also report the total volume of the FP and FN lesions.

## 2.3 Experiments

### 2.3.1 Optimization of $\alpha$ and $\beta$

The two WML segmentation parameters  $\alpha$  and  $\beta$  were optimized on six of the 20 subjects who had only a manual WML segmentation. These subjects were not used in any of the other experiments. For these six subjects WML segmentations were obtained by the automatic method with non-rigid registration of 11 atlases with 2.5 mm control point spacing and a range of  $\alpha$  from 2.0 to 3.1 at intervals of 0.1 and of  $\beta$  from 0.02 to 0.30 at intervals of 0.02. For every subject the SI between the automatic and the manual segmentations was calculated and ranked within the parameter ranges. The parameter values with the minimal summed rank were chosen as the optimal parameter settings. This resulted in  $\alpha = 2.3$  and  $\beta = 0.26$ . These settings for  $\alpha$  and  $\beta$  were used in all other experiments.

### 2.3.2 Comparison of different types of atlas registration

To evaluate the influence of different types of atlas registration on the resulting segmentations, a comparison was made using the six subjects with manual segmenta-

tions of the brain tissues and WML performed by two observers. Registration can be time consuming, so reduction of registration time might be desirable. This can, for example, be accomplished by 1) reducing the number of registrations, 2) using affine instead of non-rigid registration or 3) increasing the control point spacing. All these alternatives were tested. First, a single average atlas registration was considered. The average atlas was created by non-rigidly registering 11 atlases to a twelfth. This twelfth atlas was identified as the best target for alignment similar to the method described in Smith et al. (2006). The non-rigid registration of the 11 atlases to the target atlas was performed with a control point spacing of 2.5 mm. The resulting probability atlas was registered to the subject by applying the transformation obtained from registering the alignment target HASTE-Odd image to the subject T1w image. Secondly, multiple-atlas registration was performed by registering all 12 atlases to the subject and averaging the result to create a probability map. Both the single-atlas and the multiple-atlas methods were tested with affine registration and non-rigid registration at control point spacings of 20 mm, 10 mm, 5 mm and 2.5 mm. In another experiment, the number of atlases was varied. All 12 atlases were used for each phase, consisting of either twelve tests with one atlas, four tests with three randomly picked atlases, or two tests with six randomly picked atlases. All tests were performed using non-rigid registration at a control point spacings of 2.5 mm. The resulting segmentations were evaluated using the similarity index for CSF, GM, WM, WML and brain (GM + WM + WML) averaged over the six subjects. The type of atlas registration yielding the optimal segmentation is determined by summing the average CSF, GM, WM and WML similarity indices and is used for the other experiments.

### 2.3.3 Quantitative analysis of brain tissue and WML segmentation

The same six subjects were used for quantitative evaluation of the brain tissue and WML segmentation method. TPF, EF and SI were used as evaluation measures. The resulting SIs between automatic and manual segmentations were compared to the interobserver SI of the manual segmentations. For WML segmentations in these six subjects the boundary localization error was assessed by calculating the distance measures. The remaining 14 subjects with white matter lesion segmentation by only one observer were used to provide insight into the amount of over- or under-segmentation depending on total lesion volume. This was accomplished by relating SI and volume differences to the total lesion volume in the segmentations. The differences between automatic and manual segmentation were further investigated by calculating the volume difference per subject. This volume difference per subject was divided into false positive volume and false negative volume and was calculated before and after the false positive WML reclassification step to investigate the influence of this reclassification step. Additionally, we investigated

whether the performance of the FP WML removal step depended on lesion volume, by plotting the contribution to the FP volume as function of lesion size before and after the reclassification. The volume differences and FP WML numbers were calculated for the 6 subjects with manual segmentations by two observers and the 14 subjects with manual WML segmentations by one observer. In the case of two manual segmentations, the per subject volume differences were averaged over the two observers and for the FP lesion volume influence a manual segmentation was randomly picked per subject.

### 2.3.4 Qualitative analysis in large dataset

All 209 subjects were segmented using the optimal method from the atlas registration study. The results were visually inspected by a team of two experts. Of every subject, three representative axial slices at 12 mm distance, of both the FLAIR image and the corresponding automatic segmentation were shown and the experts were asked to evaluate the segmentations. The CSF, GM and WM segmentations were rated as 'good', 'reasonable' or 'poor'. For the WML segmentations it was specified if there were voxels that were clearly false positive or false negative lesions. This WML over- or under-segmentation specification could be based on either the size of the white matter lesions or on extra or missing lesions. Brain mask errors were not taken into account in the segmentation evaluation because the main focus is on the tissue and WML segmentation.

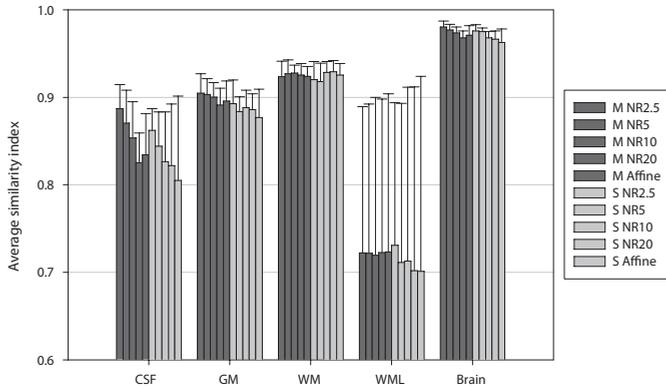
### 2.3.5 Association between age and WML volume

Finally, we assessed the relation between age and total white matter lesion volume as determined with the automatic WML segmentations. The strength of these associations is compared with previously reported values from population samples in the discussion of this work. Because of small numbers of people over the age of 80 years, we restricted our analyses on the effect of age on white matter lesion volume to persons aged 60 to 80 years. Moreover, we excluded the subjects with suboptimal WML segmentations according to the visual inspection in the previous experiment.

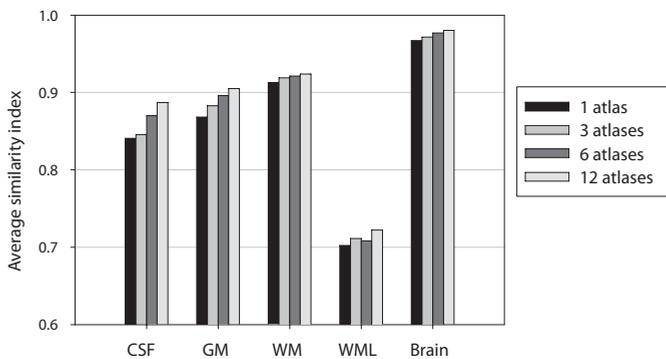
## 2.4 Results

### 2.4.1 Comparison of different types of atlas registration

The results for the accuracy study, as a function of the atlas registration method used in training, are shown in Figure 2.3. The given SIs are averages of twelve SIs obtained by comparing the automatic segmentations of the six subjects with the manual segmentations by two observers. Figure 2.4 shows the average SIs of the experiment using different numbers of atlases. The SIs obtained by registering 12 atlases are given as comparison. Non-rigid registration using 12 atlases and a



**Figure 2.3:** Average similarity index (and standard deviation) of different tissue type for several atlas-registration methods. 'NR' stands for non-rigid registration and following number refers to the control point spacing (in mm) used. 'M' refers to the multiple-atlas registration and 'S' to the single-atlas registration. SIs are averaged over 6 subjects compared to the manual segmentations by two observers.



**Figure 2.4:** Average similarity index of different tissue type for different numbers of atlases used by the segmentation method. All atlases are non-rigidly registered with a control point spacing of 2.5 mm. SIs are averaged over the tests and the 6 subjects compared to the manual segmentations by two observers.

**Table 2.2:** Average true positive fraction (TPF), extra fraction (EF) and similarity index (SI) (6 subjects, 2 observers) for different tissues

Tissue	<i>Method vs. manual</i>			<i>Manual</i>
	TPF	EF	SI	Interobserver SI
CSF	0.90	0.13	0.89	0.89
GM	0.87	0.057	0.90	0.93
WM	0.97	0.13	0.92	0.95
WML	0.79	0.50	0.72	0.75
Brain	0.98	0.019	0.98	0.98

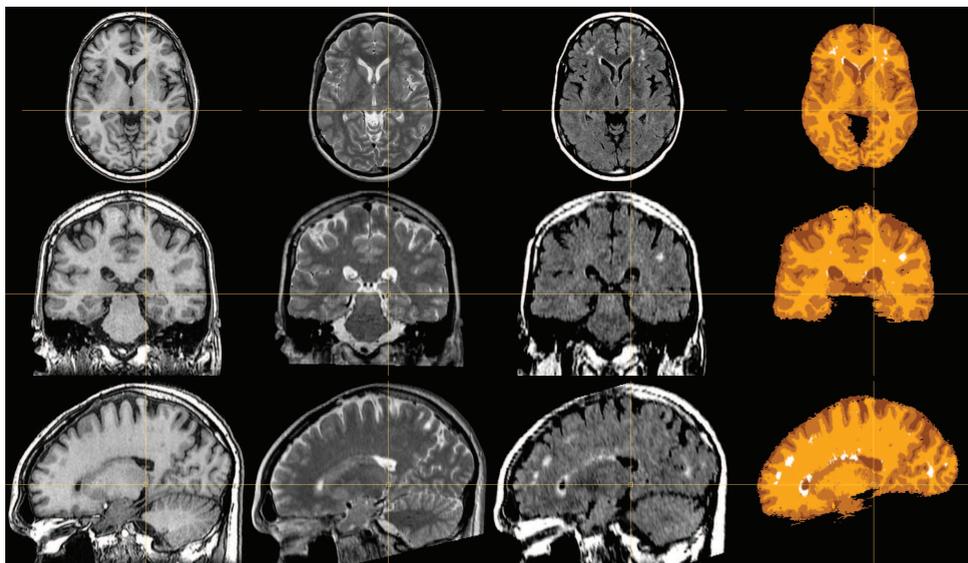
control point spacing of 2.5 mm gave the highest summed SI and is therefore used for the other experiments. Especially the CSF segmentation is influenced by the type of registration and benefits from registering at a small control point spacing and using 12 atlases.

## 2.4.2 Quantitative analysis of brain tissue and WML segmentation

Table 2.2 gives the average TPF, EF, SI, along with the interobserver SI, of the six subjects using the best performing registration method. The SI of the different tissue types is close to the interobserver SI. TPF is overall high and EF is low for all tissue types except WML. The latter is mainly due to the two subjects with lowest WML load. Without these two subjects the EF is 0.16. Example slices of a segmentation result is shown in Figure 2.5. An example segmentation image of a subject with low lesion load is given in Figure 2.6.

The maximum and mean absolute distance between two WML segmentations, obtained using the distance measure as defined in Section 2.2.7, is shown in Figure 2.7. Subjects are ordered by increasing mean manual WML segmentation volume. Whereas sometimes a large maximum error occurs (Figure 2.7(a)), Figure 2.7(b) shows that these large disagreements do not occur frequently. The mean absolute distance between segmentations is smaller than 1 mm in five out of six cases, and on average 0.4 mm. Figure 2.7(c) shows the average total volume percentage that was excluded from the distance evaluation because the corresponding lesions are not present in both segmentations.

Figure 2.8(a) shows a plot of the similarity index of the WML segmentations versus their volume in the manual segmentation for the 20 subjects with manual WML segmentations by one or more observers. The results for the six subjects with manual segmentations by two observers were averaged. As expected, SI of subjects with small WML volume is lower than the SI of subjects with large WML volume. Similarly the volume differences between the automatic and the manual segmentations



**Figure 2.5:** *T1w, PDw, FLAIR images and automatic segmentation result of subject 4. From dark to light (white) the labels represent CSF, GM, WM and WML.*

are shown in the normalized Bland-Altman plot in Figure 2.8(b). The volume differences were normalized by the average segmentation volume to emphasize that sometimes the disagreement is almost as large as the average total WML volume. This is the case for subjects with small WML volume and even the disagreement between the observers is large in these subjects.

An example of white matter lesions reclassified by the FP WML reclassification step can be seen in Figure 2.9. This example subject has, compared to observer 1, 4.92 ml FP and 0.272 ml FN WML before and 1.59 ml FP and 0.275 ml FN WML after the reclassification step. The total WML volume according to observer 1 is 9.93 ml. The reclassification step removes the FP WML in the cortical gray matter while preserving the periventricular white matter lesions. Figure 2.10(a) shows the total false positive and false negative WML volume before and after the FP WML reclassification step for 20 subjects. For the six subjects, the two manual volumes and their FP and FN volumes are averaged. It is obvious that the reclassification step decreases the FP volume while keeping the FN volume increase to a minimum. After reclassification, false positive WML volume is less than 2 ml for 18 out of 20 subjects. The graph indicates an FN volume dependence on WML load for subjects with WML load up to 6 ml. Figure 2.10(b) shows the total false positive WML volume per bin for 20 subjects with a bin size of 10 voxels (except for the last three

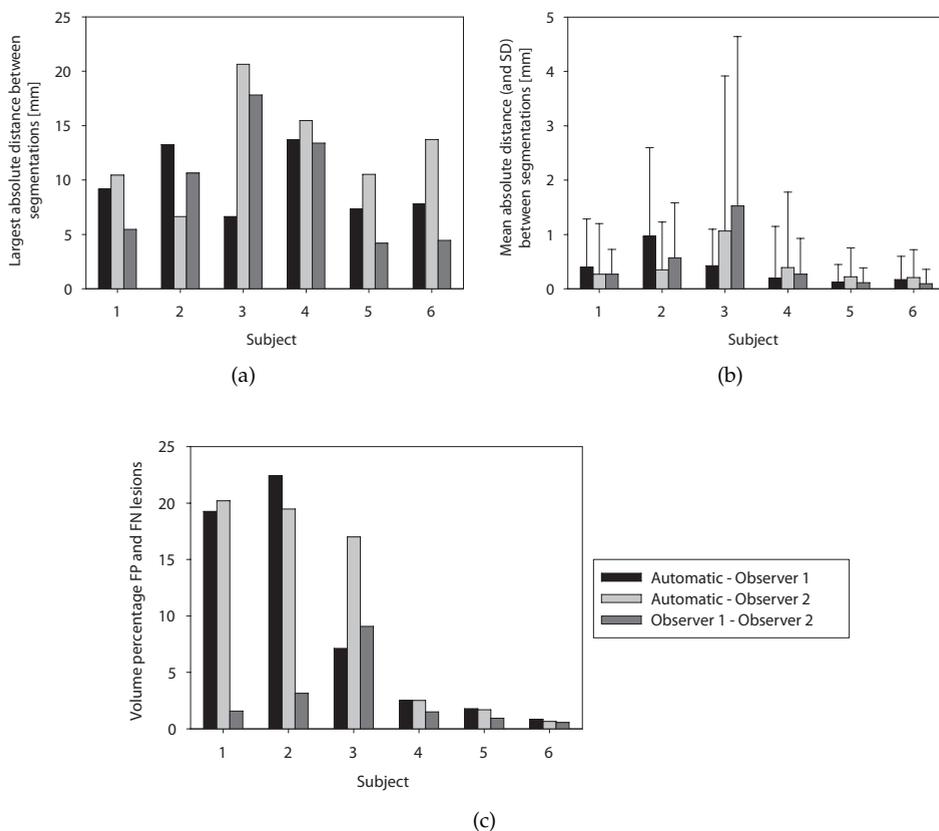


**Figure 2.6:** *T1w, PDw, FLAIR images and automatic segmentation result of subject 2. From dark to light (white) the labels represent CSF, GM, WM and WML.*

bins). For the six subjects with manual WML segmentations by two observers, a manual segmentation is picked randomly. Before the FP WML reclassification step the FP WML size has a range up to 9.3 ml and after this step the maximum FP lesion size is 0.075 ml. There is no strong indication for lesion size dependence of the FP WML reclassification step.

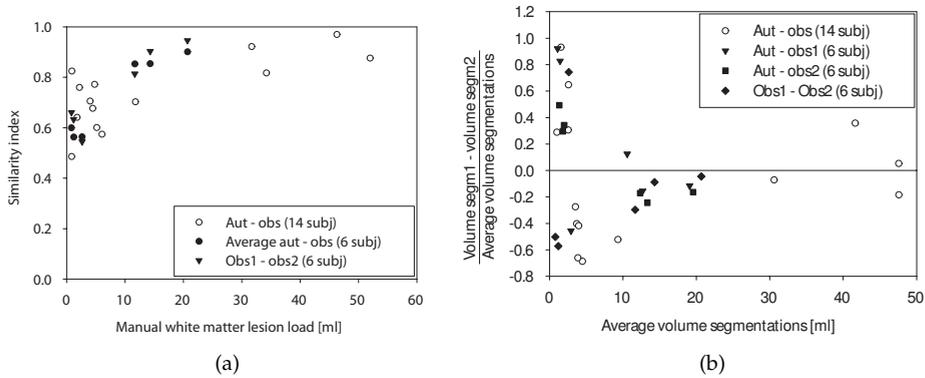
### 2.4.3 Qualitative analysis in large dataset

Table 2.3 shows the results for the visual inspection of the 209 subjects. According to the team of experts, 98% of the subjects had ‘good’ brain tissue segmentations and 97% had no obvious oversegmented or undersegmented WML. The six subjects with total manual segmentation and the 14 subjects with manual WML segmentations also had ‘good’ CSF, GM and WM segmentations and no obvious WML over- or undersegmentation. Ten subjects had ‘reasonable’ or ‘poor’ brain tissue segmentations and/or oversegmented or undersegmented WML. The MRI scans of these ten subjects were inspected fully by two experts. Seven of these subjects show motion artifacts in the T1w, PDw and/or FLAIR image. If these subjects are excluded, only one ‘reasonable’ brain tissue segmentation, one ‘oversegmented’ WML segmentation and one ‘undersegmented’ WML segmentation remain. An ex-

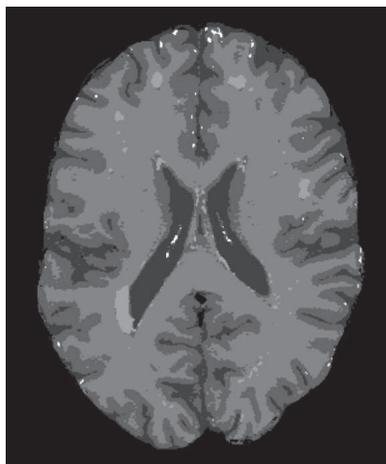


**Figure 2.7:** Maximum (a) and mean (b) distance between segmentation borders. Percentage of averaged false positive and false negative lesion volumes (c). Subjects are ordered by increasing mean manual WML segmentation volume.

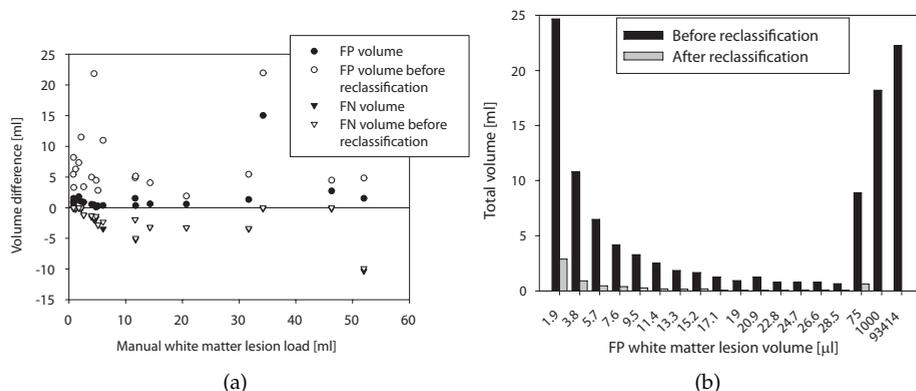
ample slice of the remaining ‘reasonable’ brain tissue segmentation is shown in Figure 2.11(a). Figure 2.11(b) shows a slice of the segmentation labeled as ‘poor’ brain tissue segmentation and ‘undersegmented’ WML. This segmentation was based on MR images with motion artifacts which effected mainly the gray matter and WML segmentations. The masking of the brain contained small errors in 11 subjects. Errors varied from inclusion of parts of the dura and skull (9 subjects) to inclusion of one or both eyes (2 subjects).



**Figure 2.8:** Similarity index (a) and normalized Bland-Altman plot (b) of WML for all subjects with manual segmentation. ‘Aut’ refers to the automatic segmentation, ‘obs’ to observer and ‘segm’ to segmentation. The 6 subjects with manual WML segmentations by two observers are shown in black. Their SI values and volumes are averaged over the two observers for the SI plot (a).



**Figure 2.9:** Example of WML reclassified as GM by the false positive WML reclassification step. From dark to light gray the labels represent CSF, GM, WM and WML. The white labels represent reclassified WML.



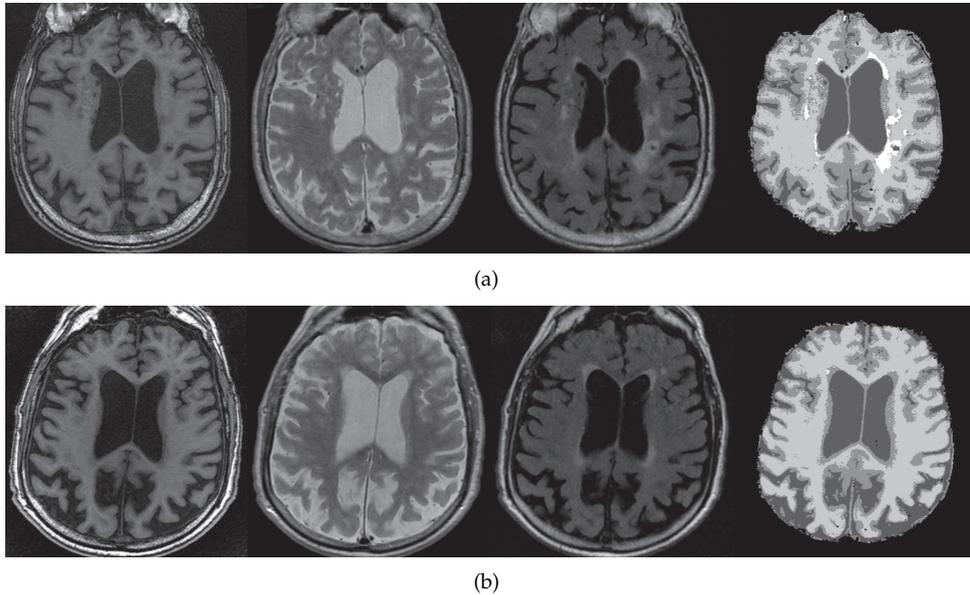
**Figure 2.10:** False positive and false negative WML volume per subject as a function of WML load (a), and contribution to FP WML volume as a function of individual lesion volume (b), before and after the FP WML reclassification step. The lesion volumes mentioned in (b) are the upper bounds of the bins.

**Table 2.3:** Result of visual inspection

	Evaluation	Number of subjects
CSF, GM, WM segmentations	Good	204
	Reasonable	4
	Poor	1
WML segmentation	No FP or FN	203
	Oversegmentation	4
	Undersegmentation	2

## 2.4.4 Association between age and WML volume

This experiment was performed on all subjects between 60 and 80 years old, excluding five subjects in this age category with suboptimal WML segmentations. Study population characteristics, mean WML volume and difference in WML volume per year increase in age are given in Table 2.4. In order to correct for individual head size, WML volume is expressed as percentage of intra-cranial volume (ICV). The WML volume is natural log transformed for the analysis because of leftward skewness of the untransformed measure. The last two rows in Table 2.4 list the mean WML volume transformed back to percentages and the exponential WML growth. A scatterplot of age versus WML volume is shown in Figure 2.12.



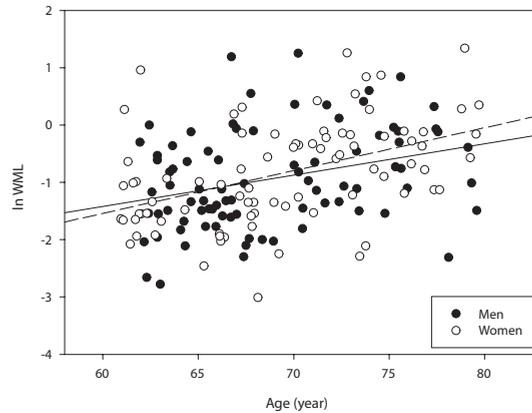
**Figure 2.11:** Examples of segmentation slices viewed by the experts and labeled as having ‘reasonable’ (a) or ‘poor’ (b) brain tissue segmentation. The corresponding MRI slices are also shown. Segmentation (b) had undersegmented WML according to the experts.

**Table 2.4:** Study population characteristics, mean (sd) WML volume and difference in WML volume per year increase of age (95% confidence interval)

	Total	Men	Women
Number of subjects	167	84	83
Age (year)	69.1 (5.2)	68.8 (5.0)	69.5 (5.5)
Intra-cranial volume (ml)	1148 (123)	1230 (99.3)	1066 (81.6)
White matter lesions <sup>a</sup>	-0.89 (0.89)	-0.94 (0.87)	-0.84 (0.91)
$\Delta$ WML volume per year <sup>a,b</sup> (95%CI)	0.066 (0.042; 0.090)	0.055 (0.018; 0.091)	0.075 (0.042; 0.11)
White matter lesions (% of ICV)	0.41	0.39	0.43
Exponential WML growth (95%CI)	1.068 (1.043; 1.094)	1.056 (1.018; 1.095)	1.077 (1.043; 1.113)

<sup>a</sup>Natural log transformed percentage of intra-cranial volume.

<sup>b</sup>Difference in WML volume per year increase in age.



**Figure 2.12:** Scatterplot of natural log transformed white matter lesion volume versus age. Volumes are expressed as percentage of intra-cranial volume. Regression lines for linear fit are shown as a solid line for men and a dashed line for women.

## 2.5 Discussion and conclusion

A fully automated method for CSF, GM and WM segmentation has been optimized, extended with WML segmentation, and quantitatively and qualitatively validated. The different brain tissues, CSF, GM and WM, are segmented by an automatically trained kNN classifier using atlas registration. The quantitative evaluation comparing the automatic segmentations to manual segmentations showed similarity indices close to the interobserver similarity index. The quantitative evaluation was performed primarily in the age category 60-70 year, as for these scans manual segmentations were available. There was no noticeable difference between age groups in the qualitative evaluation. We therefore do not expect that the accuracy would be different in different age groups.

The atlases used for the registration-based automatic training were available from an earlier study and were based on data acquired with a different scanner and scanning protocol than the data on which the method was evaluated. This shows that the method can be successfully applied to data acquired with different scanning protocols than the scanning protocol of the atlas datasets. In the study we evaluated (1) single-atlas versus multiple-atlas registration, (2) affine versus non-rigid registration at different control point spacings and (3) registration of different number of atlases. A non-rigidly registered atlas is able to capture more differences between anatomies and this leads to better training samples for classification. Increasing the number of atlases also allows for more anatomical variation to be cap-

tured in the training samples. Especially the CSF segmentation benefits from non-rigid registration with a small control point spacing and the use of more atlases. This is most likely caused by the difficulty to obtain training samples of sulcal CSF. The subarachnoid space is very narrow and especially the gyri vary widely between subjects. Non-rigid registration of 12 atlases using a control point spacing of 2.5 mm resulted in the best segmentations with accuracy close to the interobserver variability. This approach, however, also involves the most computing time. Computation time may be reduced by faster registration methods, e.g. Klein et al. (2007). Further increase of the number of atlases might improve the segmentation even more, but we could not study this due to the limited number of atlases available. The segmentation might also benefit from decreasing the control point spacing beyond 2.5 mm. This will, however, increase the computational costs even further and might give the registration algorithm too many degrees of freedom leading to sub-optimal registrations. The atlas data was from elderly subjects, although slightly younger than the average age of the test subjects. Since the atlases are only used for the brain tissue segmentation and since the obtained samples are pruned before they are used, we do not expect any problems to occur when the atlas age is mismatched as long as multiple atlases are used to correct for possible age-related variations or lesions.

White matter lesions are segmented using the brain tissue segmentation and a FLAIR scan. A WML threshold for the FLAIR image is determined by using the FLAIR intensity histogram within a GM mask, obtained from the GM segmentation. It is followed by a simple post-processing step to ensure that the lesions found are within the white matter. This is accomplished by thresholding the WM fraction of neighboring voxels for every lesion. Experiments show that this reclassification step removes a large fraction of false positive white matter lesions while keeping the increase in FN WML to a minimum. The performance of this reclassification does not strongly depend on total white matter lesion load or individual lesion volume. The acquired accuracy of the WML segmentation is not as high as for the brain tissue segmentation but it is still in the range of the interobserver variability. Most false positive WML are very small and a WML size cutoff might therefore be considered. Based on the experiments discarding white matter lesions up to  $1.9 \mu\text{l}$  (10 voxels) would remove a large fraction of false positive WML volume. Distance measures show that in general the average distance between the automatic and the manual lesion segmentation boundaries is small. There is, however, no reliable *in vivo* method to find out if a suspected lesion area in a MR image is a true lesion and it is therefore impossible to obtain a true golden standard. This results in relatively large disagreements between WML segmentations, both manually and automatically obtained. Especially for subjects with low lesion load this has a large influence on similarity index- and extra fraction-values. For these subjects, differences in volume can be almost as large as the average total volume. Other studies (van Leemput et al., 2001; Zijdenbos et al., 1998) also reported high interobserver

and intra-observer variability in WML segmentation and the resulting problem with validating automated segmentation methods.

Visual inspection of 209 segmentations showed no obvious segmentation errors in 98% of the brain tissue segmentations and 97% of the WML segmentations of the subjects. The majority of subjects with segmentation errors had motion artifacts in one or more MR images. The high percentage of correct segmentations shows that the method is robust to intersubject variations. The subject-specific classifier-training by atlas registration prevents segmentation errors due to MR intensity variations between scans and even to a certain extent within an MR image.

We also assessed the association between age and WML volume in subjects aged 60 to 80 years old with no WML segmentation errors. This association was expressed as difference in the natural log transformed WML percentages of ICV per year increase in age (and 95% confidence interval). The associations that we found in our study (0.055 (95% CI 0.018; 0.091) for men and 0.075 (95% CI 0.042; 0.11) for women) were similar to both the estimates reported by Ikram et al. (2008) from their manually corrected automatic WML segmentations in a different subset of the Rotterdam Scan Study that was scanned with a different scanning protocol, and the estimates reported by DeCarli et al. (2005) from semiautomated image segmentation analysis of more than 2200 participants of the Framingham Heart Study.

Partial volume effects may have an impact on segmentation results, as they can induce errors in both the atlas registration and the kNN classification. In order to reduce this error, it is possible to perform a probabilistic brain tissue segmentation by defining the tissue class probability as the fraction of neighbors of this class in the kNN classification. The white matter lesion segmentation method, however, is not designed to give a probabilistic output. Anisotropic voxels have more partial volume artifacts in one (or more) direction(s) than in the other. This will also influence the segmentation results. Furthermore, the WML segmentation and post-processing steps use a symmetric neighborhood definition. Anisotropic voxels will make these processing steps anisotropic too. We did not find this to have any noticeable and undesirable effect given the voxel sizes used in this study. As seen in the experiment evaluated by visual inspection, motion artifacts have a great influence on the segmentation result. Motion does not only induce a possible 'ringing artifact' in the final segmentation but it also influences the training samples obtained after atlas registration. This can result in a classification of GM voxels as WM and vice versa. It is therefore important to carefully inspect the MR images used.

In conclusion we introduced and optimized an automated method for CSF, GM, WM and WML segmentation and showed that its accuracy is close to the interobserver variability. Robustness was shown in a large study on 209 subjects. The association between age and WML volume of the subjects younger than 80 years old are comparable to reported associations found by similar studies.

# Chapter 3

## Accuracy and Reproducibility of Automatic Brain Tissue Segmentation

Based on:

Renske de Boer, Henri A. Vrooman, M. Arfan Ikram, Meike W. Vernooij, Monique M.B. Breteler, Aad van der Lugt, Wiro J. Niessen, Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods, *NeuroImage*, 2010.

### Abstract

The ability to study changes in brain morphometry in longitudinal studies majorly depends on the accuracy and reproducibility of the brain tissue quantification. We evaluate the accuracy and reproducibility of four previously proposed automatic brain tissue segmentation methods: FAST, SPM5, an automatically trained k-nearest neighbor (kNN) classifier, and a conventional kNN classifier based on a prior training set. The intensity nonuniformity correction and skull-stripping mask were the same for all methods. Evaluations were performed on MRI scans of elderly subjects derived from the general population. Accuracy was evaluated by comparison to two manual segmentations of MRI scans of six subjects (mean age  $65.9 \pm 4.4$  years). Reproducibility was assessed by comparing the automatic segmentations of 30 subjects (mean age  $57.0 \pm 3.7$  years) who were scanned twice within a short time interval. All methods showed good accuracy and reproducibility, with only small differences between methods. The conventional kNN classifier was the most accurate method with similarity indices of 0.82/0.90/0.94 for cerebrospinal fluid/gray matter/white matter, but it showed the lowest reproducibility. FAST yielded the most reproducible segmentation volumes with volume difference standard deviations of 0.55/0.49/0.38 (% of intracranial volume) respectively. The results of the reproducibility experiment can be used to calculate the required number of subjects in the design of a longitudinal study with sufficient power to detect changes over time in brain (tissue) volume. Example sample size calculations demonstrate a rather large effect of the choice of segmentation method on the required number of subjects.

## 3.1 Introduction

Longitudinal MR brain imaging studies provide unique insight into morphometric changes in the brain owing to the aging brain in general, neurodegenerative diseases such as Alzheimer's disease, or the effect of possible treatments. The assessment of relevant changes in brain morphometry is affected by the accuracy and reproducibility of the image acquisition and the subsequent image analysis with brain tissue segmentation tools. When designing a longitudinal study with sufficient power to detect changes in brain (tissue) volume, this reproducibility should be taken into account, as it affects the required sample size for a specific interscan interval. In a limited number of previous studies, e.g. Fox et al. (2000); Schott et al. (2006), this type of sample size calculations have been performed for clinical trials in Alzheimer's disease.

In literature, most brain tissue segmentation methods have been evaluated on their accuracy using real data (e.g., Anbeek et al., 2005; de Boer et al., 2009; Vrooman et al., 2007), simulated data (e.g., Ashburner and Friston, 2000, 2005; Kovacevic et al., 2002) or both (e.g., Amato et al., 2003; Awate et al., 2006; Cocosco et al., 2003; van Leemput et al., 1999; Song et al., 2006). Ten studies have reported the reproducibility of brain tissue segmentation methods (Cardenas et al., 2001; Chard et al., 2002; Clark et al., 2006; Fotenos et al., 2005; Harris et al., 1999; Kovacevic et al., 2002; Lemieux et al., 1999, 2003; Shuter et al., 2008; Wang et al., 1998). Most of these studies did not compare the reproducibility of several tissue segmentation methods applied to the same data sets. This makes it difficult to compare the results, as reproducibility is also influenced by imaging hardware and acquisition parameter settings. A few studies have investigated the impact of factors related to image acquisition, including scan sessions, acquisition sequences, data analyses, scanner upgrades, scanner vendors, field strengths, signal-to-noise ratio and scanner software (Jovicich et al., 2009; Shuter et al., 2008) on the segmented volumes. The study by Clark et al. (2006) is the only one to compare reproducibility of several tissue segmentation methods on the same data sets. By evaluating these on 20 MR images of the same subject they did, however, not take into account the robustness of these methods to anatomical variation.

In this study, we compare both the accuracy and reproducibility in segmenting cerebrospinal fluid (CSF), gray matter (GM), white matter (WM) and total brain (GM + WM) of four well-known segmentation methods. One of the methods also incorporates white matter lesion (WML) segmentation and is evaluated on this aspect too. Accuracy is measured by comparing each automatic segmentation result to manual segmentations performed by two observers on six datasets. The reproducibility of all methods is assessed based on the automatic brain tissue segmentations of MRI scans of 30 subjects who were scanned twice within a short time interval. All subjects were scanned on the same scanner with the same acquisition protocol and no scanner (software) updates were performed during the course of

this study. The resulting reproducibility measures can be used to estimate the required minimal number of subjects to find a certain effect in a longitudinal study with sufficient power.

## 3.2 Materials and methods

### 3.2.1 Imaging data

Imaging data from the population-based Rotterdam Scan Study (Hofman et al., 2009) acquired in 2005-2008 were used for the evaluation of the accuracy and reproducibility. Scans were obtained on a 1.5 T GE scanner using an 8-channel head coil. The protocol included three high-resolution axial MRI sequences, i.e. a T1-weighted (T1w) 3D fast RF spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse sequence (TR=13.8 ms, TE=2.8 ms, TI=400 ms, FOV=25×25 cm<sup>2</sup>, matrix=416×256 (interpolated to 512×512 resulting in voxel sizes of 0.49×0.49 mm<sup>2</sup>), flip angle=20, NEX=1, bandwidth (BW)=12.50 kHz, 96 slices with a slice thickness of 1.6 mm zero-padded in the frequency domain to 0.8 mm), a proton density-weighted (PDw) sequence (TR=12,300 ms, TE=17.3 ms, FOV=25×25 cm<sup>2</sup>, matrix=416×256, NEX=1, BW=17.86 kHz, 90 slices with a slice thickness of 1.6 mm), and a fluid-attenuated inversion recovery (FLAIR) sequence (TR=8000 ms, TE=120 ms, TI=2000 ms, FOV=25×25 cm<sup>2</sup>, matrix=320×224, NEX=1, BW=31.25 kHz, 64 slices with a slice thickness of 2.5 mm).

Brain MRI scans of six subjects were used to assess the accuracy of the different methods (one woman, five men, mean (sd) age 65.9 (4.4) years when scanned). Two physicians independently performed manual segmentations of the brain tissues in the cerebrum on all slices of these six datasets, using a paintbrush method with a locally adapted threshold in the MNI-tool 'Display'. Scans were manually segmented into CSF, GM and WM on the T1w volumes and WMLs were manually segmented on the FLAIR volumes. To evaluate the reproducibility, 30 different subjects were scanned twice within an average interval of 18.5 days (median 14.5, interquartile range 10 - 23 days). This group of subjects consisted of 16 women and 14 men and the mean (sd) age was 57.0 (3.7) years at time of the first scan.

### 3.2.2 Segmentation

#### 3.2.2.1 Preprocessing

As the methods incorporate different non-uniformity correction and skull-stripping/masking procedures it is difficult to compare the effects of the tissue segmentation. We therefore performed our experiments using the same preprocessing for all methods. In addition, we reported the results obtained with the default preprocessing in the Appendix.

The scans are corrected for intensity non-uniformity using the N3 method (Sled et al., 1998) within a mask. The skull-stripping mask was obtained by non-rigid registration of a manual segmented brain mask to the T1w image using Elastix<sup>1</sup> (Klein et al., 2010). This mask excludes the cerebellum since this structure was not always completely included in the PDw image.

### 3.2.2.2 Segmentation methods

The accuracy and reproducibility of four fully automated brain tissue segmentation methods were compared. These methods are listed below with specific details on the parameter settings that were used in the experiments.

1. FAST (Zhang et al., 2001) is a brain tissue segmentation method which is part of FSL<sup>2</sup> (Smith et al., 2004). This method is based on a hidden Markov random field model and an associated expectation-maximization algorithm. Both probabilistic and deterministic segmentations are given as output. We ran FAST version 4.1 without bias field correction and with otherwise default parameter settings, using the non-uniformity corrected and masked T1w image as input. Tissue volumes were calculated from the probabilistic images.
2. SPM5<sup>3</sup> contains a probabilistic brain tissue segmentation method (Ashburner and Friston, 2005). A model, based on a mixture of Gaussians and tissue probability maps as deformable spatial priors, is fitted in an iterative procedure. This model combines image registration, tissue classification and (if applicable) bias correction. The output images are probabilistic images per tissue class. Default parameter settings were used, except for switching off the bias field correction. The non-uniformity corrected T1w image and the mask were given as input. Subsequently, the mask is applied to all output images, since the GM and WM images are masked by default but the CSF image is not skull-stripped and includes non-CSF components. Tissue volumes were calculated from the probabilistic images. If an experiment required a deterministic segmentation result, majority voting was used by classifying a voxel as the tissue type with the highest probability.
3. A k-nearest neighbor (kNN) brain tissue segmentation method, automatically trained on the subject itself using atlas registration, and extended with white matter lesion segmentation (Cocosco et al., 2003, and Chapter 2) is the third method considered. This method uses 12 non-rigidly registered atlases to obtain locations where training samples for the kNN classifier are extracted using the T1w and PDw intensities as features. The features are normalized by

---

<sup>1</sup>Elastix is available at <http://elastix.isi.uu.nl/>

<sup>2</sup>FSL is available at <http://www.fmrib.ox.ac.uk/fsl/>

<sup>3</sup>SPM5 is available at <http://www.fil.ion.ucl.ac.uk/spm/>

a simple range matching procedure. The resulting brain tissue segmentation is used to automatically derive a subject-specific intensity threshold for white matter lesions in the FLAIR image. Before segmentation, the PDw and FLAIR images are rigidly registered to the T1w image. This method was also evaluated without the PDw input image.

4. The conventional kNN brain tissue classifier is constructed from a prior training set of atlases using the T1w and PDw intensities as features (Vrooman et al., 2007). The features are normalized by a simple range matching procedure within a brain mask. Contrary to the previous method, this kNN classifier is not trained on the subject itself. The training set is obtained from the six subjects with manual segmentations by two observers. Since the accuracy experiment uses these same six subjects, the accuracy of this method was assessed in a leave-one-out experiment. The same co-registered and non-uniformity corrected T1w and PDw images were used as for the automatically trained kNN method. Similar to automatically trained kNN classifier, this method was also evaluated using only the T1w intensities.

### 3.2.3 Experiments

#### 3.2.3.1 Accuracy

Segmentation accuracy was assessed by comparing the automatically obtained results to manual segmentations. The automatic segmentation methods that do not classify WML were only compared to the manual segmentations made on the T1w images (these segmentations also did not include WML). Accuracy is reported using four measures. The true positive fraction (TPF) and extra fraction (EF) are reported to express sensitivity and oversegmentation respectively:

$$\text{TPF} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.1)$$

$$\text{EF} = \frac{\text{FP}}{\text{TP} + \text{FN}} \quad (3.2)$$

based on true positives (TP), false negatives (FN) and false positives (FP). Third, the similarity index (SI), or Dice coefficient (Dice, 1945; Zijdenbos et al., 1994), is used to express overlap between segmentations:

$$\text{SI} = \frac{2(S_1 \cap S_2)}{S_1 + S_2} \quad (3.3)$$

where  $S_1$  and  $S_2$  denote the segmented volumes and  $(S_1 \cap S_2)$  is the overlap of  $S_1$  and  $S_2$ . In addition, the overlap measure conformity (C) (Chang et al., 2009) is used:

$$C = 1 - \frac{FP + FN}{TP} \quad (3.4)$$

### 3.2.3.2 Reproducibility

The results of the reproducibility experiments are presented in two ways. Firstly, based on the segmentations of both MRI scans, the resulting volumes for CSF, GM and WM (and WML) were compared. These volumes were expressed as percentages of intra-cranial volume (ICV = CSF + GM + WM + WML) in order to correct for differences in head size. The differences between the two sequential scans were calculated from these fractional volumes by subtracting the volume of the first scan from the volume of the second scan. The fractional volumes of the sequential scans were also used to compute the coefficient of variation (CoV). The CoV is defined as the ratio of the standard deviation to the mean, and is expressed in percentages:

$$\text{CoV} = \frac{1}{N} \sum_i \frac{\sigma_i}{\mu_i} \times 100\% \quad (3.5)$$

where  $N$  is the number of subjects,  $i$  indexes subjects,  $\sigma_i$  is the standard deviation of subject  $i$ , and  $\mu_i$  is the mean of subject  $i$ .

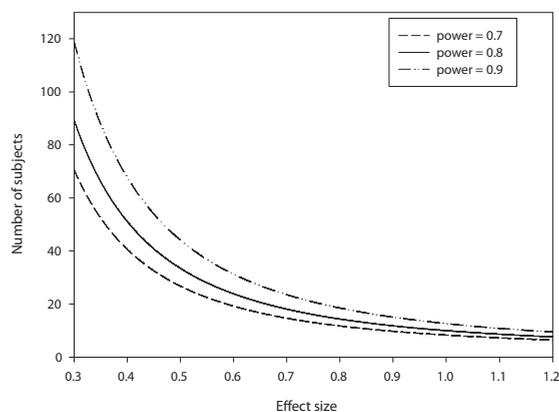
Secondly, the segmentation obtained from the second scan was transformed to the first scan by rigid registration of the T1w images using the Image Registration Toolkit (IRTK)<sup>4</sup> (Rueckert et al., 1999). The overlap of the transformed segmentation and the segmentation of the first scan is represented by the similarity index and the conformity measure. Similarly, the overlap of the skull-stripping masks was calculated and their SI and  $C$  are given as indicators of the error caused by the registration. A two-tailed paired  $t$ -test compared the methods based on the SI values.

### 3.2.3.3 Sample size calculations for longitudinal studies

Reproducibility influences the number of subjects needed in a longitudinal study. In a study, two types of statistical errors can be made. The probability of a type I error, or level of significance ( $\alpha$ ), is the probability of rejecting the null hypothesis when the null hypothesis is true. The probability of a type II error ( $\beta$ ) is the probability of accepting the null hypothesis when the null hypothesis is false. The power of a study is the probability of rejecting the null hypothesis when the alternative hypothesis is true ( $1 - \beta$ ). The number of subjects required to find an effect depends on the power, level of significance and the effect size. The effect size is defined as follows.

$$\text{effect size} = \frac{\text{mean difference over time period}}{\text{standard deviation}} \quad (3.6)$$

<sup>4</sup>IRTK is available at <http://www.doc.ic.ac.uk/~dr/software/>



**Figure 3.1:** Effect size versus required number of subjects for the design of a longitudinal study (two-tailed paired  $t$ -test,  $\alpha = 0.05$ ).

The standard deviation depends on the segmentation method and results from the reproducibility experiment as the standard deviation of the volume differences. Given the effect size (e.g. rates of tissue atrophy), power and significance level, the required number of subjects can be calculated using the program G\*Power<sup>5</sup> (Faul et al., 2007). Figure 3.1 shows a graph relating the required number of subjects to effect size. This graph is obtained with G\*Power 3, using varying powers and a level of significance of  $\alpha = 0.05$  in a paired two-tailed  $t$ -test. We give several example calculations using the results from the reproducibility experiment.

The sample size calculations presented in this paper assume a longitudinal study of a single group of subjects. If a longitudinal study is designed to compare the atrophy rates of two groups, e.g. patients with Alzheimer's disease and a control group, the required number of subjects also depends on the standard deviations and the difference in sizes of these groups.

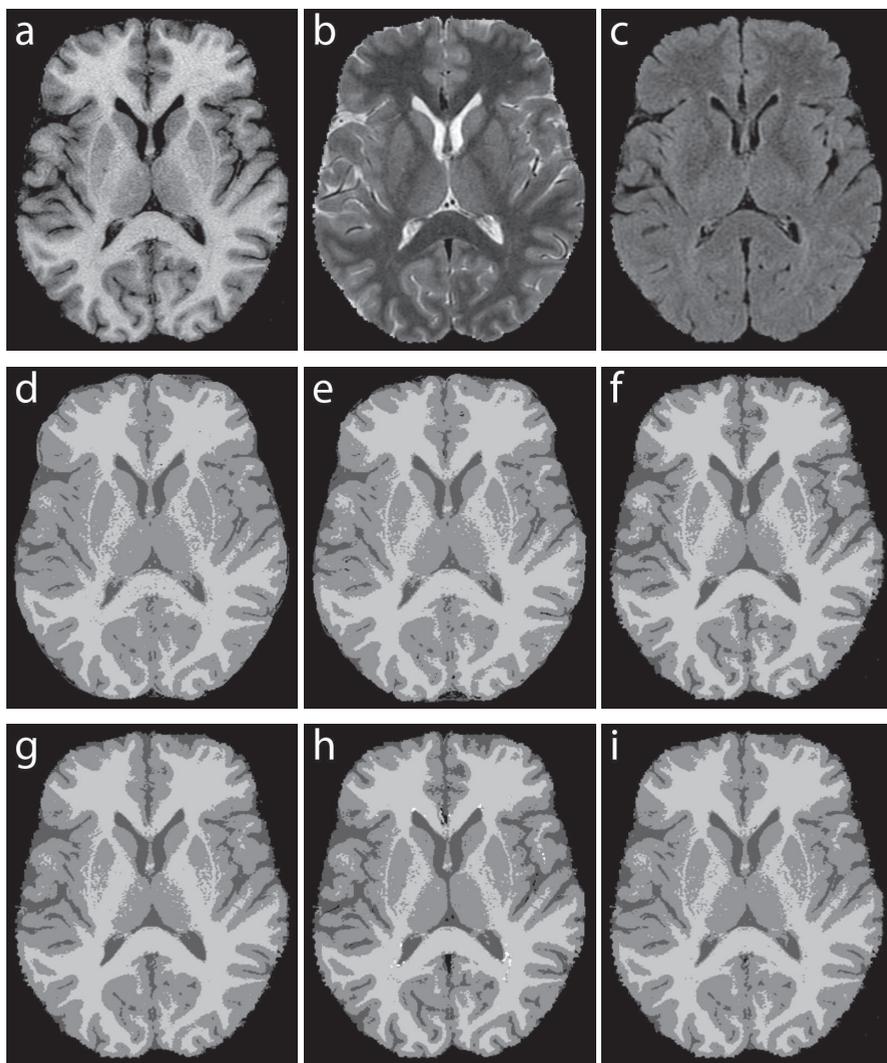
## 3.3 Results

### 3.3.1 Segmentation

Figure 3.2 shows a representative result of an axial slice of the MR images and their corresponding manual and automatic segmentations of a subject with low WML load.

One scan of one of the 30 subjects used in the reproducibility experiments was

<sup>5</sup>G\*Power 3 is available at <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>



**Figure 3.2:** Axial slice of a subject with low WML load included in the accuracy study: (a) T1w image, (b) PDw image, (c) FLAIR image, (d) manual segmentation based on T1w by observer 1, (e) manual segmentation based on T1w by observer 2, (f) FAST segmentation based on T1w, (g) SPM5 segmentation based on T1w, (h) automatically trained kNN classifier segmentation (WML in white) based on T1w/PDw/FLAIR, (i) conventional kNN classifier segmentation based on T1w/PDw.

not masked properly by the non-rigidly registered skull-stripping mask as it included part of both eyes. This resulted in an erroneous segmentation by the conventional kNN classifier, as it performs its feature normalization within the brain mask. A new brain mask was created by an alternative registration strategy and a new conventional kNN segmentation was obtained. The new mask was also used for all other segmentations of this subject.

Processing time of the FAST method was approximately 26 min on a 64-bit Linux cluster node. SPM5 processing took approximately 8 min on a 64-bit Linux system. The non-rigid atlas registration for the automatically trained kNN classifier took approximately 6 - 7 hours per atlas using IRTK and a control point spacing of 2.5 mm on a 64-bit Linux system. Comparable non-rigid registration of the same atlas using Elastix took approximately 15 min on a 64-bit Linux system yielding a final kNN segmentation with similar accuracy. After registering the atlases and creating tissue probability maps, the remaining processing time of the automatically trained kNN method (including WML segmentation) was 22 min on a 32-bit Windows desktop machine. The conventional kNN classifier took approximately 23 min on a 64-bit Linux system.

### 3.3.2 Accuracy

The results for the accuracy measures are shown in Table 3.1. The interobserver measures did not differ for the manual segmentation including or excluding WML. Overall, the different segmentation methods showed only small differences in accuracy. Most methods showed high accuracy for all tissue classes and the SIs were close to the interobserver SI of the manual segmentations. As conformity and SI are closely related, they show the same trends. However, for tissue types with less overlap, the conformity measure shows a better distinction between the segmentation methods. The conventional kNN classifier showed the highest overlap with the manual segmentation for all tissues. SPM5 and the automatically trained classifier using PDw and T1w input images showed the lowest accuracy. The accuracy of the automatically trained classifier improved if the PDw image was left out. For the conventional kNN classifier there was no clear improvement or decline in accuracy if only the T1w image was used.

Table 3.2 shows the accuracy of the white matter lesion segmentation by the automatically trained classifier. White matter lesion segmentation is a difficult task and interobserver variability is caused by both differences in detection and in segmented volumes. Due to the small volumes of the lesions, small differences in segmentations have a relatively large effect on the evaluation measures. The SI of the automatically trained kNN method was, however, close to the interobserver SI of the manual segmentations. Other accuracy evaluation measures of the WML segmentation of this method are available in Chapter 2.

**Table 3.1: Accuracy of the segmentation methods**

	CSF				Gray matter				White matter				Brain <sup>a</sup>			
	TPF	EF	SI	C	TPF	EF	SI	C	TPF	EF	SI	C	TPF	EF	SI	C
FAST	0.88 (0.03)	0.47 (0.20)	0.75 (0.07)	0.33 (0.23)	0.82 (0.02)	0.05 (0.02)	0.88 (0.01)	0.72 (0.03)	0.97 (0.02)	0.10 (0.05)	0.94 (0.02)	0.87 (0.04)	0.93 (0.02)	0.02 (0.01)	0.96 (0.01)	0.91 (0.02)
SPM5	0.87 (0.03)	0.49 (0.22)	0.75 (0.07)	0.30 (0.24)	0.83 (0.03)	0.07 (0.03)	0.87 (0.02)	0.70 (0.05)	0.93 (0.03)	0.08 (0.04)	0.93 (0.01)	0.84 (0.03)	0.93 (0.02)	0.01 (0.01)	0.96 (0.01)	0.91 (0.02)
Auto. trained kNN PDw + T1w + FLAIR	0.82 (0.07)	0.22 (0.15)	0.81 (0.03)	0.52 (0.09)	0.85 (0.02)	0.10 (0.06)	0.87 (0.03)	0.70 (0.07)	0.97 (0.02)	0.13 (0.06)	0.92 (0.02)	0.83 (0.04)	0.96 (0.01)	0.04 (0.03)	0.96 (0.01)	0.91 (0.02)
Auto. trained kNN T1w + FLAIR	0.85 (0.04)	0.26 (0.15)	0.81 (0.04)	0.51 (0.13)	0.85 (0.03)	0.07 (0.04)	0.89 (0.02)	0.74 (0.05)	0.98 (0.01)	0.14 (0.06)	0.93 (0.02)	0.84 (0.05)	0.96 (0.01)	0.03 (0.02)	0.96 (0.01)	0.93 (0.01)
Conventional kNN <sup>b</sup> PDw + T1w	0.86 (0.03)	0.24 (0.09)	0.82 (0.04)	0.55 (0.12)	0.87 (0.03)	0.07 (0.03)	0.90 (0.01)	0.78 (0.03)	0.96 (0.02)	0.09 (0.04)	0.94 (0.01)	0.87 (0.02)	0.96 (0.02)	0.03 (0.01)	0.97 (0.01)	0.93 (0.01)
Conventional kNN <sup>b</sup> T1w	0.82 (0.03)	0.22 (0.08)	0.81 (0.03)	0.51 (0.10)	0.88 (0.03)	0.09 (0.04)	0.90 (0.01)	0.77 (0.03)	0.95 (0.02)	0.08 (0.04)	0.94 (0.01)	0.87 (0.02)	0.96 (0.01)	0.03 (0.02)	0.96 (0.00)	0.93 (0.01)
Interobserver			0.89 (0.05)	0.74 (0.14)			0.93 (0.02)	0.86 (0.04)			0.95 (0.02)	0.90 (0.04)			0.98 (0.01)	0.96 (0.01)

Reported values are mean (sd).

TPF, true positive fraction; EF, extra fraction; SI, similarity index; C, conformity.

<sup>a</sup>Brain segmentation is defined as the combined segmentations of gray matter and white matter (and WML).

<sup>b</sup>Leave-one-out evaluation.

**Table 3.2:** Results for the white matter lesion segmentations

	Accuracy experiment				Reproducibility experiment 1		Reproducibility experiment 2	
	TPF	EF	SI	C	$\Delta$ (%)	CoV (%)	SI	C
Auto. trained kNN PDw + T1w + FLAIR	0.79 (0.14)	0.48 (0.60)	0.73 (0.16)	0.09 (0.74)	0.05 (0.12)	7.45	0.569 (0.137)	-0.762 (1.106)
Auto. trained kNN T1w + FLAIR Interobserver	0.80 (0.15)	0.50 (0.59)	0.72 (0.17)	0.06 (0.77)	0.01 (0.05)	5.87	0.567 (0.140)	-0.791 (1.166)
			0.75 (0.15)	0.22 (0.58)				

TPF, true positive fraction (mean (sd)); EF, extra fraction (mean (sd)); SI, similarity index (mean (sd)); C, conformity (mean (sd)),  $\Delta$ , volume difference (mean (sd)) as percentage of intracranial volume; CoV, coefficient of variation.

**Table 3.3:** Reproducibility of the segmentation methods

	CSF		Gray matter		White matter		Brain <sup>d</sup>	
	$\Delta$ (%)	CoV (%)	$\Delta$ (%)	CoV (%)	$\Delta$ (%)	CoV (%)	$\Delta$ (%)	CoV (%)
FAST	0.09 (0.55)	1.48	-0.22 (0.49)	0.69	0.13 (0.38)	0.53	-0.09 (0.55)	0.39
SPM5	-0.20 (1.05)	2.98	-0.05 (1.16)	1.50	0.25 (0.86)	1.29	0.20 (1.05)	0.72
Auto. trained kNN PDw + T1w + FLAIR	-0.14 (0.69)	2.26	-0.03 (1.02)	1.30	0.14 (0.86)	1.42	0.14 (0.69)	0.52
Auto. trained kNN T1w + FLAIR	0.29 (0.64)	2.27	-0.57 (1.17)	1.50	0.27 (1.02)	1.53	-0.29 (0.64)	0.50
Conventional kNN PDw + T1w	0.22 (0.98)	3.53	-0.33 (1.61)	2.13	0.11 (1.53)	2.29	-0.22 (0.98)	0.62
Conventional kNN T1w	0.07 (0.55)	2.16	-0.03 (0.67)	0.74	-0.04 (0.67)	0.92	-0.07 (0.55)	0.37

$\Delta$ , volume difference (mean (sd)) as percentage of intracranial volume; CoV, coefficient of variation.

<sup>d</sup>Brain volume is defined as the combined volumes of gray matter and white matter (and WML).

### 3.3.3 Reproducibility

Table 3.3 shows the results for the first reproducibility experiment. In general, the reproducibility results of the different segmentation methods only showed small differences. The segmentation methods showed small volume differences and standard deviations and low CoV for all tissue classes indicating high reproducibility. Since the volume differences are expressed as percentage of ICV, the CSF volume differences are equal to the brain volume differences. SPM5 had the lowest reproducibility for brain segmentation, as its standard deviations and CoV were highest. The conventional kNN classifier with PDw and T1w input images had the lowest reproducibility for the other classes. FAST showed the best reproducibility for all tissues. Using only the T1w image as input improved the reproducibility of the conventional kNN classifier segmentation for all tissue types. The reproducibilities of the GM and WM segmentations from the automatically trained kNN classifier were slightly worsened by leaving out the PDw image. Table 2 shows the results for both reproducibility experiments for WML.

Table 3.4 shows the SI and C values of the second reproducibility experiment

**Table 3.4:** Measured overlap of the rigidly registered segmentation of scan 2 and the segmentation of scan 1 of the reproducibility data set

	CSF		Gray matter		White matter		Brain <sup>d</sup>	
	SI	C	SI	C	SI	C	SI	C
FAST	0.796 (0.024)	0.486 (0.077)	0.845 (0.015)	0.631 (0.044)	0.904 (0.011)	0.788 (0.027)	0.960 (0.005)	0.916 (0.012)
SPM5	0.802 (0.024)	0.504 (0.076)	0.868 (0.013)	0.695 (0.036)	0.909 (0.009)	0.800 (0.022)	0.964 (0.005)	0.925 (0.011)
Auto. trained kNN	0.820 (0.022)	0.559 (0.067)	0.865 (0.012)	0.687 (0.032)	0.909 (0.010)	0.800 (0.025)	0.961 (0.005)	0.918 (0.011)
PDw + T1w + FLAIR								
Auto. trained kNN	0.773 (0.026)	0.409 (0.088)	0.844 (0.016)	0.628 (0.047)	0.894 (0.013)	0.762 (0.033)	0.958 (0.006)	0.912 (0.012)
T1w + FLAIR								
Conventional kNN	0.783 (0.029)	0.443 (0.097)	0.853 (0.013)	0.654 (0.036)	0.896 (0.013)	0.768 (0.032)	0.961 (0.005)	0.919 (0.011)
PDw + T1w								
Conventional kNN	0.752 (0.029)	0.337 (0.106)	0.846 (0.013)	0.636 (0.035)	0.890 (0.013)	0.752 (0.034)	0.959 (0.005)	0.915 (0.011)
T1w								

Reported values are mean (sd).

SI, similarity index; C, conformity.

<sup>d</sup>Brain segmentation is defined as the combined segmentations of gray matter and white matter (and WML).

where the segmentation of the second scan is rigidly transformed to the first scan. The mean (sd) SI and C of the rigidly transformed skull-stripping mask of the second scan and the mask of the first scan equaled 0.973 (0.005) and 0.945 (0.011) respectively and can be used as a reference. Although the differences between the segmentation methods were small, the  $p$ -values in Table 3.5 show that there is a significant difference between most methods at a  $p < 0.05$  level. For the automatically trained kNN classifier and conventional kNN classifier results, the paired  $t$ -tests were performed only with the T1w and PDw input images as these resulted in higher SI than with only the T1w input image. Most methods differed significantly for (almost) all tissue types. Only SPM5 and the automatically trained kNN classifier showed no significant difference in the overlap of their WM segmentations. In addition, the overlap between the brain segmentations of the automatically trained kNN and SPM5 and the automatically trained kNN and the conventional kNN had no significant differences. The automatically trained kNN classifier and SPM5 had the highest overlap between the registered segmentations and FAST and the conventional kNN classifier the lowest.

### 3.3.4 Sample size calculations for longitudinal studies

Finally, we performed an analysis to relate the estimated reproducibilities to sample size calculations for longitudinal studies into brain atrophy. The number of subjects required to find a given effect size (rate of atrophy of brain or tissue type), given a certain power and a level of significance of  $\alpha = 0.05$ , can be deduced for each method from Figure 3.1. The effect size can be calculated using the standard deviation of the volume differences mentioned in Table 3.3. An example is the design of a longitudinal study with the aim to find a brain atrophy rate of -0.45% per year.

**Table 3.5:** Results of two-tailed paired *t*-tests (*df* = 29) comparing the SI values of the registered segmentation of scan 2 and the segmentation of scan 1 of the reproducibility data set

	CSF		Gray matter		White matter		Brain <sup>d</sup>	
	<i>t</i> -value	<i>p</i> -value	<i>t</i> -value	<i>p</i> -value	<i>t</i> -value	<i>p</i> -value	<i>t</i> -value	<i>p</i> -value
FAST vs. SPM5	-12.6	< 0.001	-28.6	< 0.001	-8.4	< 0.001	-22.6	< 0.001
FAST vs. Auto. trained kNN	-14.2	< 0.001	-13.3	< 0.001	-5.5	< 0.001	-1.8	0.090
FAST vs. Conventional kNN	10.3	< 0.001	-7.7	< 0.001	14.7	< 0.001	-5.3	< 0.001
SPM5 vs. Auto. trained kNN	-10.4	< 0.001	2.2	0.038	0.1	0.913	5.6	< 0.001
SPM5 vs. Conventional kNN	13.6	< 0.001	13.9	< 0.001	13.7	< 0.001	10.0	< 0.001
Auto. trained kNN vs. Conventional kNN	17.5	< 0.001	8.0	< 0.001	11.0	< 0.001	-0.6	0.539

Bold indicates significance at a  $p < 0.05$  level.

<sup>d</sup>Brain segmentation is defined as the combined segmentations of gray matter and white matter (and WML).

If FAST is used for the brain segmentation, the reproducibility experiment shows that the standard deviation of the brain volume difference is 0.55% (Table 3.3). With an interscan interval of one year, the effect size will be  $0.45/0.55 = 0.82$ . Figure 3.1 indicates that the required number of subjects included in the analysis will be 18 for a power of 0.9 or 14 for a power of 0.8. If the conventional kNN classifier method is used the effect size will be  $0.45/0.98 = 0.46$ . In this case, the required number of subjects will be 52 for a power of 0.9 or 40 for a power of 0.8, as indicated by the graph.

Similar calculations can be made for GM, WM or WML. Table 3.6 shows the required number of subjects for several example volume differences of different tissues for a longitudinal study with a one-year interscan interval, a power of 0.9 and a level of significance of 0.05. Although the segmentation methods only showed small differences in reproducibility, the effect on the required number of subjects can be large. If the aimed volume difference is relatively small compared to the reproducibility standard deviation, the required number of subjects increases dramatically, as is the case for the example gray matter atrophy rates. Since FAST has the lowest reproducibility standard deviation, it requires the least subjects for longitudinal studies on volume differences of all tissue types.

### 3.4 Discussion and conclusion

Comparing accuracy and reproducibility of segmentation methods based on literature can be difficult due to the use of different evaluation measures, different manual segmentation protocols and most importantly different imaging data. We compared both the accuracy and the reproducibility of several previously proposed brain tissue segmentation methods on the same datasets. All scans were made with the same acquisition protocol on the same scanner, without any scanner updates during the course of the study. The rescans, for the reproducibility exper-

**Table 3.6:** Example estimations of the required number of subjects of a longitudinal study (one-year interscan interval, power of 0.9, level of significance 0.05) aimed to find the selected volume differences

Volume difference <sup>b</sup> (% of ICV per year)	Gray matter			White matter			Brain <sup>a</sup>			White matter lesions		
	-0.05	-0.10	-0.15	-0.3	-0.4	-0.5	-0.3	-0.4	-0.5	0.05 <sup>c</sup>	0.06 <sup>c</sup>	0.07 <sup>c</sup>
FAST	1012	255	115	19	12	9	38	22	15			
SPM5	5658	1416	631	89	51	34	131	75	49			
Auto. trained kNN PDw + T1w (+ FLAIR)	4375	1096	488	89	51	34	58	34	23	63	44	33
Auto. trained kNN T1w (+ FLAIR)	5756	1441	642	124	71	46	50	29	20	13	10	8
Conventional kNN PDw + T1w	10,897	2725	1213	276	156	101	115	66	43			
Conventional kNN T1w	1889	474	212	55	32	21	38	22	15			

<sup>a</sup>Brain volume is defined as the combined volumes of gray matter and white matter (and WML).

<sup>b</sup>Volume differences are based on cross-sectional and a limited number of longitudinal studies, e.g., DeCarli et al. (2005); Fotenos et al. (2005); Ge et al. (2002); Ikram et al. (2008).

<sup>c</sup>Natural log transformed.

iment, were made on average 18.5 days after the first scan. This period ensures that the natural variations in, for example, fluid balance in the brain are captured, while no significant brain changes take place. In general, all segmentation methods showed good accuracy and reproducibility. There were, however, small differences between the various methods. The conventional kNN classifier method performed best in the accuracy experiment and worst in the reproducibility experiment. FAST showed the best reproducibility, but its accuracy was relatively low for CSF and GM.

Despite its high accuracy, the conventional kNN classifier method has several weaknesses. Due to the fixed training set, any changes in image acquisition require a laborious training stage. Furthermore, the MRI contrast between GM and WM tissues changes with age (Cho et al., 1997) and might therefore result in an age-related bias in tissue volumes obtained with the conventional kNN classifier.

The accuracy experiment compared the automatic segmentations to manual segmentations based on the T1w scan. This might induce a bias for the methods, as they are based on the same T1w image. The automatically trained kNN classifier results support this hypothesis as the accuracy increased when the PDw image was left out of the analysis.

In our experiments, we performed segmentations on T1w and PDw data, or on T1w data only. It is also possible to segment the brain tissues based on only the PDw image. The contrast between GM and WM, however, is lower on the PDw image than on the T1w image. Furthermore, PDw sequences are less commonly used nowadays. This experiment was therefore not performed in this study.

The reproducibility experiment is influenced by the reproducibility of the image acquisition. A difference between the segmentations of the two sequential scans must therefore be expected, no matter how reproducible the segmentation method. Since all methods are tested on the same data, the image acquisition variation is the

same for every method.

In some studies the correlation coefficient between segmented volumes at two time points is used as evaluation measure for reproducibility (e.g., Cardenas et al., 2001; Harris et al., 1999; Wang et al., 1998). We decided not to use this measure, as it depends on the dispersion of the segmented volume measurements. Instead, the coefficient of variation was used. Different definitions of CoV are used in the literature. We used the same definition as Cardenas et al. (2001); Wang et al. (1998), but they used absolute volumes instead of fractional volumes for their calculations. The results can therefore not be compared.

The reproducibility experiment measuring the overlap between the registered segmentation of the second scan and the segmentation of the first scan showed less overlap for CSF and GM compared to WM. As CSF and GM are not as compact as WM, they have more boundary voxels. Since boundary voxels are more difficult to segment due to partial volume effects, a lower overlap can be expected, as is supported by the accuracy evaluation. In addition, this reproducibility evaluation suffers from registration errors. Especially for CSF and GM this is a disadvantage, since they are on the outside of the brain, close to the border of the skull-stripping mask. Errors in mask registration will therefore mainly be reflected in less CSF and GM overlap.

The subjects in this study are participants of a population-based cohort study of the elderly. As white matter lesions are commonly found in elderly subjects (de Leeuw et al., 2001), they are also present in the subjects used for the evaluation. Three of the subjects with manual segmentations had a low WML load with a mean (sd) of the manual segmentations of 1.54 (1.09) ml and a mean (sd) of the automatic segmentations of 1.96 (0.31) ml. The other three accuracy subjects had a high WML of 15.56 (4.32) ml according to the manual segmentations and 13.63 (3.34) ml according to the automatic segmentations. The 30 subjects for the reproducibility experiment were picked randomly. These subjects had no manual segmentations but according to the automatic segmentations of their first scan, their mean (sd) WML load was 5.00 (6.84) ml with a range of 0.97 - 34.81 ml.

White matter lesions can take up a considerable part of the white matter, so excluding these WMLs might improve the WM segmentation. Furthermore, WMLs may be of interest themselves as they are associated with cognitive decline and increased risk of stroke and dementia. FSL 4.1 and SPM5 have no automatic WML segmentation method included. FAST v4.1 and the conventional kNN classifier are capable of multi-modal segmentation and it is possible to add an extra segmentation class. There are, however, no processing steps that ensure that the additional class will contain only WMLs, so other components might get the same label. SPM5 is not capable of multi-modal brain tissue segmentation.

Besides the automatic WML segmentation method evaluated in this study, several automatic WML segmentation methods have been proposed. Some of these studies evaluated the reproducibility of their method (Admiraal-Behloul et al.,

2005; Jack et al., 2001). Both of these studies performed the rescan with repositioning within several minutes of the first scan. Admiraal-Behloul et al. reported only intra-class correlation coefficients and no CoVs. Jack et al. reported CoVs based on absolute volumes and subjects with larger WML volumes and are therefore not comparable to our results.

Since reproducibility depends on the image acquisition variables, a change in these variables might influence the required number of subjects in the design of a longitudinal study. The sample size estimates in this study can, in that case, be used as example calculations for a new reproducibility study. In addition, they demonstrate the influence of the choice of the segmentation method on the required number of subjects.

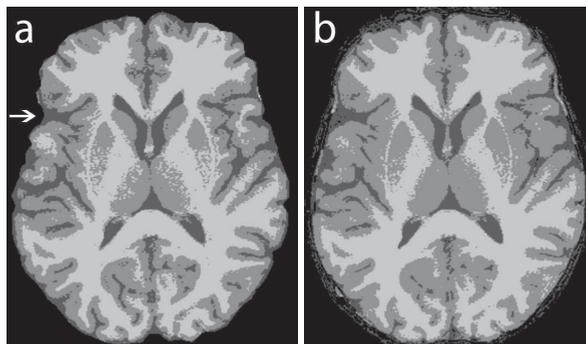
The small differences in reproducibility of the different segmentation methods have a rather large effect on the required number of subjects in a study that aims to detect a certain longitudinal change in tissue volume. Especially if the aim of the study is to find a tissue volume difference that is relatively small compared to the reproducibility standard deviation of the segmentation method, the choice of the segmentation method has a large influence on the required number of subjects. Segmentation methods with lower reproducibility standard deviations require fewer subjects to find the same tissue volume difference in a longitudinal study than methods with higher reproducibility standard deviations.

Currently, data on the association between age and gray matter volume are inconsistent. Several studies report a decline in GM volume from early adulthood onwards e.g. Fotenos et al. (2005); Ge et al. (2002), while others find no such decline e.g. Ikram et al. (2008). If there is an association between GM volume and age, its rate is likely to be small. Consequently our power calculations show that the required number of subjects to find a small GM atrophy rate in a longitudinal study would be very large.

In conclusion, we compared the accuracy and reproducibility of four known brain tissue segmentation methods. Overall, the accuracy and reproducibility were good and there were only small differences between the methods. The small differences in reproducibility do, however, have a relatively large effect on the required number of subjects in the design of a longitudinal study with sufficient power.

## Appendix - Results using original preprocessing

Since most researchers use the publicly available methods with their default preprocessing, we also report the results of our experiments with the intensity non-uniformity correction and skull-stripping as provided by the methods FAST and SPM5. FAST intrinsically corrects for spatial intensity variations, and the input image is skull-stripping using FSL's Brain Extraction Tool (BET) (Smith, 2002) version 2.1. Contrary to the skull-stripping mask used in this study, the BET mask



**Figure 3.3:** Axial slice of a subject with low WML load included in the accuracy study: (a) FAST segmentation based on T1w with default preprocessing, (b) SPM5 segmentation based on T1w with default preprocessing and probabilistic CSF threshold of 50%.

does include the cerebellum. SPM5 combines the bias field correction with the image registration and tissue classification. As mentioned before, SPM5 applies no skull-stripping to its CSF segmentation. The cerebellum is included in the GM and WM segmentations. In an attempt to obtain a SPM5 CSF segmentation that can be used for our experiments, we thresholded the CSF segmentation. Visual inspection showed that there is no optimal threshold that excludes all non-CSF components while maintaining the actual CSF. We chose a CSF threshold of 50%.

Figure 3.3 shows the resulting segmentations of the same axial slice as Figure 3.2. Since the manual segmentations do not include the cerebellum, the accuracy measures were only calculated within the manual mask. This manual mask is defined as all voxels with a CSF, GM or WM label in the manual segmentation. Contrary to the accuracy experiment mentioned before, this evaluation has a bias since the automatic segmentations can have no false positives outside the manual mask. Table 3.7 shows the results of this accuracy experiment, including the mean interobserver values determined within the manual mask. The low accuracy of the FAST CSF segmentation is partly due to differences in masking. The BET mask used by FAST excludes more sulcal CSF than the manual segmentations. An example of this is indicated by the arrow in Figure A1a.

Tables 3.8 and 3.9 show the results of both reproducibility experiments. The tissue volume difference standard deviations in Table 3.8 can be used for sample size calculations.

Keep in mind that due to differences in masking and in the accuracy evaluation, the results in this appendix cannot be compared to the values in the Results section. Also, the results in Tables 3.8 and 3.9 cannot be compared between methods due to differences in masking.

**Table 3.7:** Accuracy of the segmentation methods with default preprocessing

	CSF				Gray matter				White matter				Brain <sup>a</sup>			
	TPF	EF	SI	C	TPF	EF	SI	C	TPF	EF	SI	C	TPF	EF	SI	C
FAST	0.72 (0.08)	0.34 (0.16)	0.70 (0.07)	0.12 (0.31)	0.79 (0.02)	0.01 (0.01)	0.88 (0.01)	0.72 (0.04)	0.98 (0.01)	0.11 (0.04)	0.94 (0.02)	0.87 (0.05)	0.92 (0.02)	0.00 (0.00)	0.96 (0.01)	0.92 (0.02)
SPM5	0.93 (0.05)	0.19 (0.13)	0.88 (0.04)	0.72 (0.11)	0.88 (0.03)	0.06 (0.04)	0.91 (0.01)	0.80 (0.04)	0.94 (0.02)	0.09 (0.06)	0.93 (0.02)	0.84 (0.05)	0.97 (0.02)	0.01 (0.02)	0.98 (0.00)	0.96 (0.01)
50% CSF-threshold			0.93 (0.05)	0.83 (0.13)			0.95 (0.02)	0.89 (0.04)			0.95 (0.02)	0.90 (0.04)			0.99 (0.01)	0.98 (0.01)
Interobserver																

Evaluation is performed within the manual mask.

Reported values are mean (sd).

TPF, true positive fraction; EF, extra fraction; SI, similarity index; C, conformity.

<sup>a</sup>Brain segmentation is defined as the combined segmentations of gray matter and white matter.

**Table 3.8:** *Reproducibility of the segmentation methods with default preprocessing (including cerebellum)*

	CSF		Gray matter		White matter		Brain <sup>a</sup>	
	$\Delta$ (%)	CoV (%)	$\Delta$ (%)	CoV (%)	$\Delta$ (%)	CoV (%)	$\Delta$ (%)	CoV (%)
FAST	0.22 (0.52)	1.35	-0.36 (0.57)	0.87	0.14 (0.41)	0.64	-0.22 (0.52)	0.32
SPM5 50% CSF-threshold	0.18 (1.50)	3.09	-0.53 (1.34)	1.22	0.35 (0.99)	1.54	-0.18 (1.50)	0.76

$\Delta$ , volume difference (mean (sd)) as percentage of intracranial volume; CoV, coefficient of variation.

<sup>a</sup>Brain volume is defined as the combined volumes of gray matter and white matter.

**Table 3.9:** *Measured overlap of the rigidly registered segmentation of scan 2 and the segmentation of scan 1*

	CSF		Gray matter		White matter		Brain <sup>a</sup>	
	SI	C	SI	C	SI	C	SI	C
FAST	0.765 (0.021)	0.385 (0.073)	0.838 (0.017)	0.614 (0.048)	0.893 (0.011)	0.759 (0.027)	0.961 (0.005)	0.919 (0.010)
SPM5 50% CSF-threshold	0.770 (0.029)	0.399 (0.099)	0.880 (0.017)	0.726 (0.047)	0.895 (0.011)	0.764 (0.028)	0.966 (0.007)	0.929 (0.016)

Segmentations obtained with default preprocessing (including cerebellum).

Reported values are mean (sd).

SI, similarity index; C, conformity.

<sup>a</sup>Brain segmentation is defined as the combined segmentations of gray matter and white matter.



# Chapter 4

## Comparison of Brain Tissue Segmentation of Longitudinal MR Images

Renske de Boer, Benjamin F.J. Verhaaren, Stefan Klein, Henri A. Vrooman, M. Arfan Ikram, Meike W. Vernooij, Monique M.B. Breteler, Aad van der Lugt, Wiro J. Niessen

### Abstract

We compared two automatic brain tissue segmentation methods for assessing brain tissue and white matter lesion volume changes in longitudinal MRI data of middle-aged and elderly subjects. The first method is a conventional  $k$ -nearest neighbor (kNN) classifier, which is trained on a prior training set and the second method is a kNN classifier automatically trained on the scan to be segmented. The methods were compared in three experiments: visual inspection of the segmentations of 894 baseline and 749 follow-up scans, comparison of the derived volume change rates based on 567 scan pairs, and visual comparison of averaged atrophy maps of 567 scan pairs. The visual inspection ratings showed a slight preference for the automatically trained kNN method. The gray and white matter volume change rates differed between the methods. Literature is unclear about the true volume change rates for gray and white matter and results are difficult to compare due to study differences. The rates obtained with the automatically trained kNN were, however, very similar to the rates obtained in a cross-sectional study of a comparable population. Furthermore, the atrophy maps obtained with the automatically trained kNN classifier showed more resemblance to atrophy maps estimated with non-rigid registration of the longitudinal image data. In conclusion, the effect of the brain tissue segmentation method used on the derived volume change rates and atrophy locations can be substantial.

## 4.1 Introduction

Longitudinal brain MRI studies are important for studying maturation, aging, disease progression or other brain changes over time. In these studies, every participant is scanned multiple times during the study period. Imaging biomarkers like brain atrophy are extracted from the longitudinal MRI data and are studied in relation to either determinant or outcome variables such as age or test scores (e.g., Scahill et al., 2003; Schott et al., 2008). The rate of brain atrophy can be deduced by manual segmentation, semi-automated or automatic. Manual interaction is subject to inter- and intra-expert variability and can be very time-consuming. Especially in large study populations automatic algorithms are therefore preferred. Several approaches have been proposed for automated deduction of brain atrophy.

Some dedicated shift algorithms for measuring brain atrophy have been proposed. These methods compute the magnitude of the brain boundary shift between two MRI scans. Two well-known examples are the boundary shift integral (BSI) (Freeborough and Fox, 1997) and SIENA (Smith et al., 2001, 2002). Another method uses longitudinal MRI data to segment the brain tissues (Xue et al., 2006). They use multiple scans (over time) of a single subject to perform brain tissue segmentation at each timepoint. The difference between brain volumes deduced from the segmentations represents brain atrophy. As an alternative scans made at different time points may be segmented independently of each other. Segmenting a single scan is relatively fast and many automatic brain tissue segmentation methods have been proposed (e.g., Anbeek et al., 2005; Ashburner and Friston, 2005; Cocosco et al., 2003; Zhang et al., 2001), some of which are publicly available. Brain tissue segmentation, based on longitudinal or on a single scan, has the advantage that atrophy of both gray and white matter can be studied in addition to whole brain atrophy. The outcome is influenced by the accuracy and reproducibility of the method. Additionally, the method should be capable of handling possible changes in the scanner or its software during the time of the study, e.g. scanner drift.

A number of studies have compared the accuracy and reproducibility of brain tissue segmentation methods or other atrophy measuring methods on the same datasets. Clark et al. (2006) investigated the influence of several steps during the segmentation process on the final segmentation result. They used 20 scans of the same subject with negligible interscan interval for evaluating reproducibility. The 20 scans were combined into a 'golden standard' scan which was used for accuracy evaluation. In de Boer et al. (2010b) we compared the accuracy and reproducibility of several brain tissue segmentation methods. We used six scans with manual segmentations for the accuracy experiment and 30 subjects scanned twice within a short time interval for reproducibility assessment. The study reported in de Bresser et al. (2011) is the only one to compare volume change measurements like atrophy. They used the baseline and follow-up scans with an interscan interval of four years of ten subjects. Two follow-up scans were made and these were used

for the reproducibility evaluation. Accuracy in brain atrophy was assessed using manual segmentations of the baseline scan and the first follow-up scans.

In this study, we investigate the effects of different brain tissue segmentation methods on the assessment of magnitude and location of atrophy and changes in white matter lesion volume. We compare two brain tissue segmentation methods on longitudinal brain MRI data of 749 subjects. The evaluation comprises of three parts: a comparison of the results of visual inspection of the brain tissue segmentations, a comparison of the atrophy rates resulting from the brain tissue volumes of the longitudinal scans, and a comparison of white matter and gray matter atrophy maps averaged over all subjects.

## 4.2 Materials and methods

### 4.2.1 Data

Imaging data from the Rotterdam Scan Study (Vernooij et al., 2009) is used for the comparison. A group of 894 subjects was scanned in 2005-2006. This group consisted of 442 men and had a median age of 66.3 years (interquartile range 63.5 - 69.2 years). A subgroup of 749 subjects underwent a second scan in 2008-2010. Mean (sd) interscan interval was 3.47 (0.16) years. This subgroup consisted of 371 men and had a median age at first scan of 66.1 years (interquartile range 63.4 - 68.7 years). All subjects were non-demented at time of scanning.

Both scans were obtained on the same 1.5 Tesla GE scanner using an 8-channel head coil. No software or hardware upgrades were performed between scans. The scanning protocol included three axial MRI sequences, i.e. a T1-weighted (T1w) 3D fast RF spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse sequence (TR=13.8 ms, TE=2.8 ms, TI=400 ms, FOV=25×25 cm<sup>2</sup>, matrix=416×256 (interpolated to 512×512 resulting in voxel sizes of 0.49×0.49 mm<sup>2</sup>), flip angle=20°, NEX=1, bandwidth (BW)=12.50 kHz, 96 slices with a slice thickness of 1.6 mm zero-padded in the frequency domain to 0.8 mm), a proton density-weighted (PDw) sequence (TR=12,300 ms, TE=17.3 ms, FOV=25×25 cm<sup>2</sup>, matrix=416×256, NEX=1, BW=17.86 kHz, 90 slices with a slice thickness of 1.6 mm), and a fluid-attenuated inversion recovery (FLAIR) sequence (TR=8000 ms, TE=120 ms, TI=2000 ms, FOV=25×25 cm<sup>2</sup>, matrix=320×224, NEX=1, BW=31.25 kHz, 64 slices with a slice thickness of 2.5 mm).

Brain MRI scans of six subjects, acquired in 2005-2006 on the same scanner with the same scanning protocol, were used as atlases. Two physicians independently performed manual brain tissue segmentation using a paintbrush method with locally adapted threshold in the MNI-tool 'Display'. Scans were manually segmented into cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM) on the T1w volumes, and white matter lesions (WML) were manually segmented on the FLAIR volumes.

## 4.2.2 Brain tissue segmentation methods

For our comparison we segmented all scans with two brain tissue segmentation methods. Both methods are based on  $k$ -nearest neighbor (kNN) classification with four classes: background, CSF, gray matter and white matter. The conventional kNN classifier (Vrooman et al., 2007) was trained on the six atlases. The automatically trained kNN classifier was trained by registering the six atlases to the scan to be segmented and subsequently choosing training samples from locations with high tissue probability according to the atlases (Cocosco et al., 2003; Vrooman et al., 2007, and Chapter 2). Both classifiers used the T1w and PDw intensities as features. The methods were extended with the WML segmentation method from Chapter 2 that uses the FLAIR scan. We applied a brain mask that includes only the cerebrum.

These methods were previously compared on accuracy and reproducibility in Chapter 3. In that study the conventional kNN classifier was not extended with white matter lesion segmentation. The differences in accuracy and reproducibility between the two methods were small and both showed good accuracy and high reproducibility. The conventional kNN method showed a slightly better overlap with manual segmentations for all tissue types, but the automatically trained kNN classifier outperformed it on reproducibility.

## 4.2.3 Comparison

### 4.2.3.1 Visual inspection

All segmentations were visually inspected by an expert by scrolling through all slices. The FLAIR scan was shown as reference for the WML segmentation. The brain tissue segmentation (CSF, GM, and WM) was rated as ‘poor’, ‘reasonable’, or ‘good’. The WML segmentation was separately rated as ‘FP’ (containing false positives), ‘FN’ (containing false negatives), or ‘none’ (no false positives or false negatives). If false positive WML, false negative WML or the brain mask required corrections, this was indicated. It was also indicated if a segmentation should be excluded due to (motion) artifacts.

The conventional kNN segmentations of the baseline scans were inspected in February 2009. The automatically trained kNN segmentations of the baseline scans and the segmentations by both methods of the follow-up scans were inspected in August 2010. The two segmentations of the follow-up scans were shown simultaneously. The rater was unaware of the method used to create the segmentation as they were shown in random order. Additionally to the standard ratings, the rater could also indicate which of the two segmentations, if any, was preferred over the other.

#### 4.2.3.2 Volume change rates

We selected all subjects with two scans and segmentations that were not excluded and did not require corrections for either method. This resulted in a selection of 567 subjects (median age at first scan 65.9 years, interquartile range 63.4 - 68.0 years; 273 men; mean (sd) interscan interval 3.47 (0.16) years).

Volumes per tissue type were deduced from the segmentations and were expressed as percentage of the intracranial volume ( $ICV = CSF + GM + WM + WML$ ) in order to correct for head size. The change in volume between the two scans was computed by subtracting the volume of the second scan from the volume of the first scan. Volume change rate was then calculated by dividing the volume change by the interscan interval in years.

We performed one-sample *t*-tests to test if the mean volume change rate for each method equals zero. Additionally we used paired *t*-tests to test if the mean volume change rates of the two segmentation methods were equal.

#### 4.2.3.3 Atrophy maps

To study localization of possible atrophy, we compared the average white matter and gray matter atrophy maps of both methods in a common space, using the same selection of subjects as for the volume change rates comparison. This common space was created in several steps. In the first step, the baseline T1w scans of all subjects were registered to the baseline scan of a single subject by using affine registration followed by b-spline registration. This subject was selected because its age (65.8 years) was close to the median age and no large amounts of atrophy or WML appeared on its scan. In the second step, the deformation fields of the transformations of all subjects were averaged. Finally, the averaged deformation field was inverted and the corresponding transformation was applied to the scan of the selected subject. This creates a common space, to which the scans in this study can be transformed to with, on average, small transformations.

We generated atrophy maps for GM, WM (+ WML), and brain (= GM + WM + WML). White matter lesions were included in the white matter segmentation to exclude WML effects. To correct for pose, the follow-up scans were first rigidly registered to the baseline scans. Subtraction of the rigidly aligned follow-up and baseline segmentations yielded GM, WM, and brain atrophy images. The atrophy images of all subjects were transformed to common space by b-spline registration of the baseline T1w image. Subsequently, we averaged the atrophy images over all subjects, yielding maps where a voxel value indicates the fraction of the subjects with atrophy or growth at the voxel position. If some subjects show atrophy while others show growth at a certain voxel position, this averages out and only the remaining fraction is shown. Atrophy or growth is indicated by the sign of the voxel value: a positive value indicates atrophy and a negative value indicates growth. The averaged atrophy maps of both methods were visually compared.

Additionally, we estimated local atrophy using non-rigid registration. We registered the follow-up scan to the baseline scan using b-spline non-rigid registration and transformed the GM, WM, and brain segmentations according to the corresponding b-spline transformation. We subtracted the rigidly transformed segmentations from the non-rigidly transformed segmentations to yield atrophy images. These atrophy images were generated for both segmentation methods and averaged in the common space.

We also show the amount of compression or expansion in the b-spline transformation between follow-up and baseline scan using the determinant of the Jacobian matrix. For every subject an image of the determinant of the Jacobian matrix was created and transformed to common space based on the non-rigid registration of the baseline T1w image. The median value over all subjects was taken for every voxel in the transformed Jacobian determinant images. The result is an image showing the median local compression, expansion or volume preservation in all subjects.

All registrations were performed with Elastix<sup>1</sup> (Klein et al., 2010).

## 4.3 Results

### 4.3.1 Visual inspection

Table 4.1 and 4.2 show the visual rating results of the segmentations of respectively the baseline scans and the follow-up scans. In both baseline and follow-up segmentation, the automatically trained kNN segmentations received better ratings and required fewer corrections or exclusions. Also, the automatically trained kNN segmentations of the follow-up scans were more frequently rated as better than the conventional kNN segmentations.

The difference between the ratings of the automatically trained kNN segmentations and the ratings of the conventional kNN segmentations is larger for the baseline scans than for the follow-up scans. The percentages show that the rating of the automatically trained kNN segmentations is similar for the baseline scans and the follow-up scans. The conventional kNN segmentations of the follow-up scans got better ratings than the baseline segmentations by the same method.

### 4.3.2 Volume change rates

Table 4.3 shows the mean volume change per year for the different brain tissues and the white matter lesions. All mean volume change rates differ significantly from zero. The two segmentation methods agree only on the brain volume change rates. Especially the gray and white matter volume changes per year differ between the two segmentation methods. The conventional kNN segmentations show a decrease

---

<sup>1</sup>Elastix is available at <http://elastix.isi.uu.nl/>

**Table 4.1:** Rating results of the visual inspection of the segmentations of the baseline scans ( $n = 894$ ).

	Rating	Conventional kNN	Auto. trained kNN	Agreement
Brain tissues	Good	808 (90.4%)	876 (98.0%)	805
	Reasonable	63 (7.0%)	10 (1.1%)	4
	Poor	23 (2.6%)	8 (0.9%)	6
White matter lesions	None	814 (91.1%)	860 (96.2%)	796
	FP	80 (8.9%)	34 (3.8%)	16
	FN	0	0	-
Require correction		104 (11.6%)	53 (5.9%)	25
Exclude		22 (2.5%)	10 (1.1%)	10

**Table 4.2:** Rating results of the visual inspection of the segmentations of the follow-up scans ( $n = 749$ ).

	Rating	Conventional kNN	Auto. trained kNN	Agreement
Brain tissues	Good	720 (96.1%)	727 (97.1%)	718
	Reasonable	19 (2.5%)	13 (1.7%)	12
	Poor	10 (1.3%)	9 (1.2%)	7
White matter lesions	None	706 (94.3%)	718 (95.9%)	696
	FP	42 (5.6%)	27 (3.6%)	21
	FN	1 (0.1%)	4 (0.5%)	0
Require correction		67 (8.9%)	50 (6.7%)	37
Exclude		18 (2.4%)	15 (2.0%)	11
Better segmentation		15 (2.0%)	66 (8.8%)	

in GM volume with an increase of age, while the automatically trained kNN segmentations show a slight increase in GM volume with age. White matter volume decreased with age according to both segmentation methods, but the automatically trained kNN shows a larger WM volume change per year. The automatically trained kNN method also shows a greater increase in WML volume with increasing age.

### 4.3.3 Atrophy maps

Figure 4.1 shows the gray matter atrophy maps of both methods and the gray matter atrophy maps estimated by non-rigid registration. If white matter atrophy occurs, it is expected that the cortex shifts inwards and the gray matter around the ventricles shifts outwards. A shift is shown in the atrophy maps as loss on one side and growth on the other side. This effect can be seen in the gray matter atrophy

**Table 4.3:** Mean annual volume change rates per (tissue) type ( $n = 567$ ).

(Tissue) type	Conventional kNN		Automatically trained kNN		<i>p</i> -value <sup>c</sup>
	Mean <sup>a</sup>	95% CI <sup>b</sup>	Mean <sup>a</sup>	95% CI <sup>b</sup>	
Gray matter	-0.21	(-0.26;-0.16)	0.09	(0.05;0.14)	< <b>0.001</b>
White matter	-0.15	(-0.20;-0.11)	-0.41	(-0.45;-0.38)	< <b>0.001</b>
White matter lesions <sup>d</sup>	0.048	(0.043;0.053)	0.057	(0.052;0.062)	< <b>0.001</b>
Brain	-0.33	(-0.36;-0.29)	-0.29	(-0.33;-0.25)	0.097

Volume change rates are expressed as percentage of intra-cranial volume per year increase of age.

<sup>a</sup>Mean volume change rate.

<sup>b</sup>95% confidence interval of one-sample *t*-test comparing the mean value to zero.

<sup>c</sup>*p*-value of two-tailed paired *t*-test comparing the mean values of the two segmentation methods. Bold indicates significance at a  $p < 0.05$  level.

<sup>d</sup>Natural log transformed

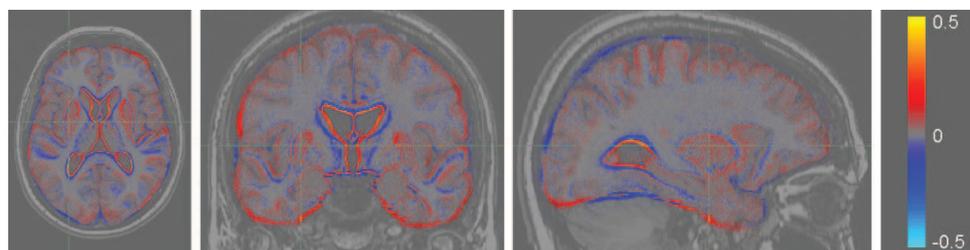
maps, especially around the ventricles but also at the cortex. The gyri and sulci differ between subjects and these differences are difficult to capture using non-rigid registration. This results in less agreement and therefore lower voxel values in the cortex regions of the atrophy maps.

The automatically trained kNN GM atrophy map and the conventional kNN GM atrophy map are quite different. The conventional kNN map shows less shifting and more loss at the cortex. It also shows less consistency at the outer gray matter border, as growth and loss alternate in the axial and sagittal views. The automatically trained kNN map bears more resemblance to the atrophy maps estimated by non-rigid registration than the conventional kNN map does. The two registration estimated atrophy maps appear very similar. This is to be expected because the atrophy is based on the non-rigid registration and the only difference is in the segmentation used for masking.

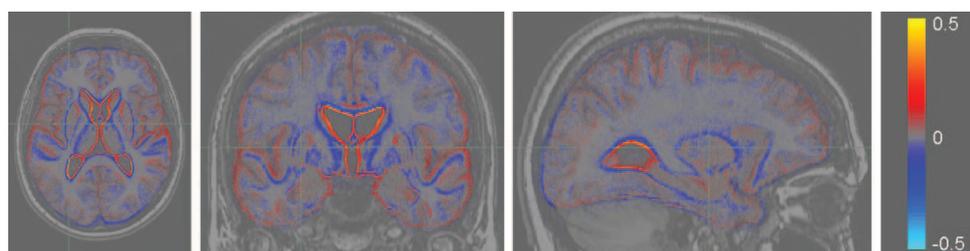
Figure 4.2 shows the white matter atrophy maps of both methods and the white matter atrophy maps estimated by non-rigid registration. Again there are differences between the two segmentation methods. The conventional kNN map shows more growth compared to the other WM atrophy maps. The WM atrophy map generated with the automatically trained kNN segmentation appears similar to the registration estimated WM atrophy maps. These three maps all show primarily WM loss, especially near the cortex and around the ventricles.

The brain atrophy maps for both methods and the maps estimated by registration are shown in Figure 4.3. These atrophy maps show less differences. Brain loss and growth occur at similar locations. The conventional kNN atrophy map shows more brain loss in the inferior regions of the brain.

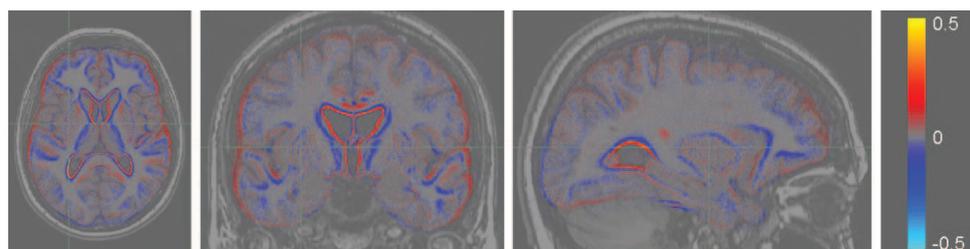
The median Jacobian image with and without gray matter outline is shown in Figure 4.4. As expected, expansion occurs in the ventricles and compression in the



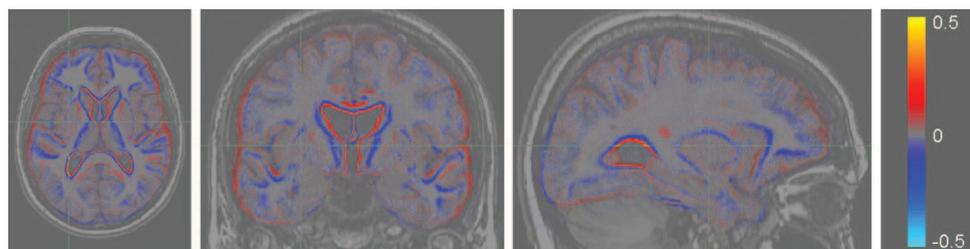
(a) Conventional kNN.



(b) Automatically trained kNN.

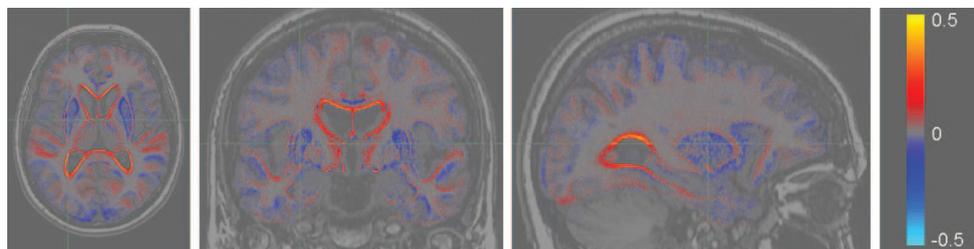


(c) Registration estimated atrophy (conventional kNN).

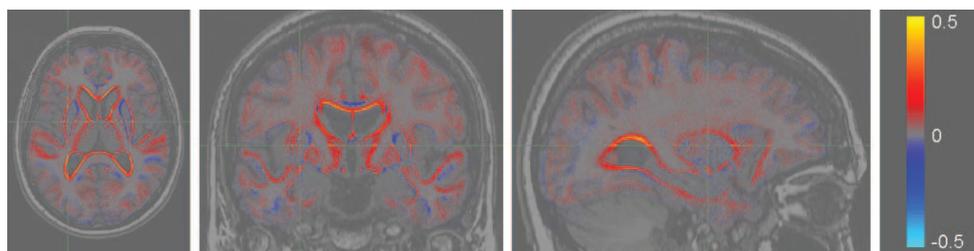


(d) Registration estimated atrophy (automatically trained kNN).

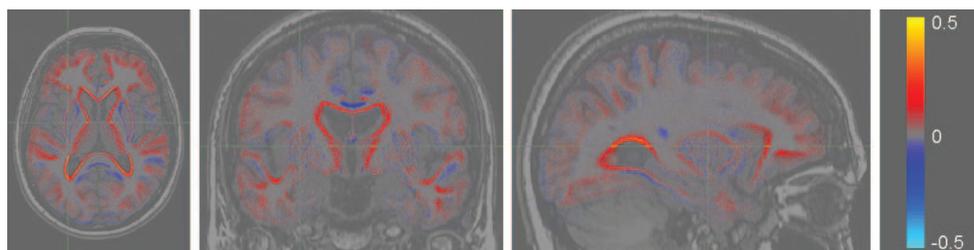
**Figure 4.1:** Gray matter atrophy maps based on 567 subjects. Red indicates tissue loss, blue indicates growth.



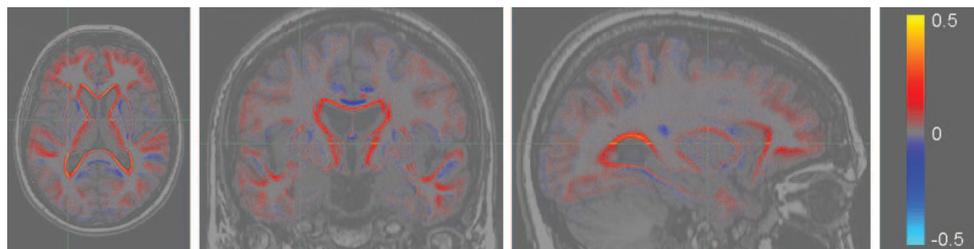
(a) Conventional kNN.



(b) Automatically trained kNN.

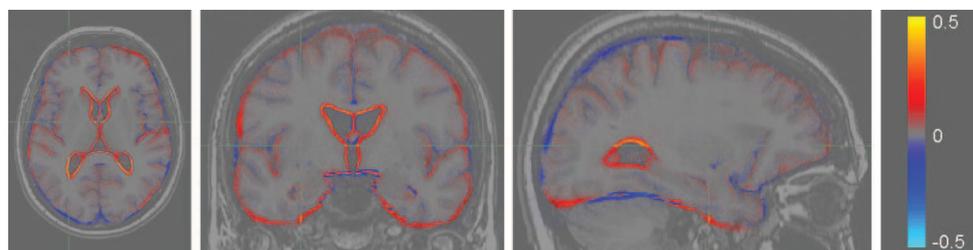


(c) Registration estimated atrophy (conventional kNN).

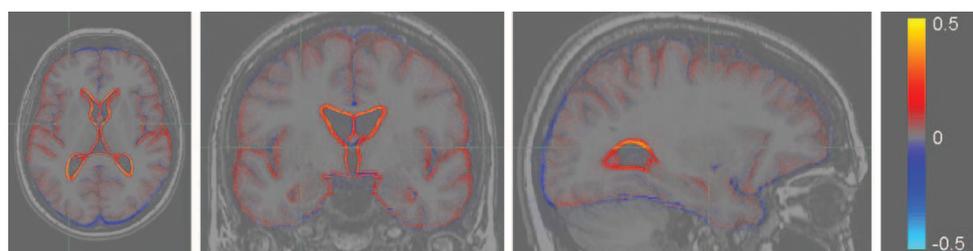


(d) Registration estimated atrophy (automatically trained kNN).

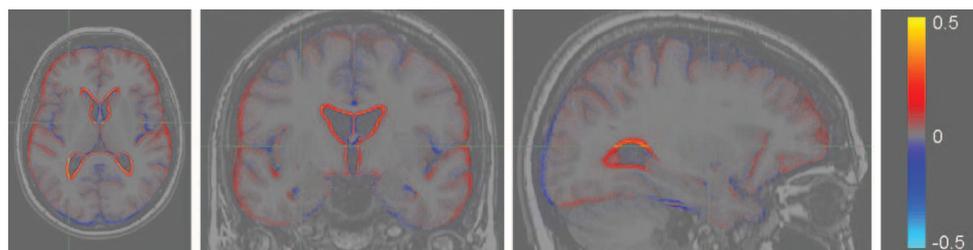
**Figure 4.2:** White matter atrophy maps based on 567 subjects. Red indicates tissue loss, blue indicates growth.



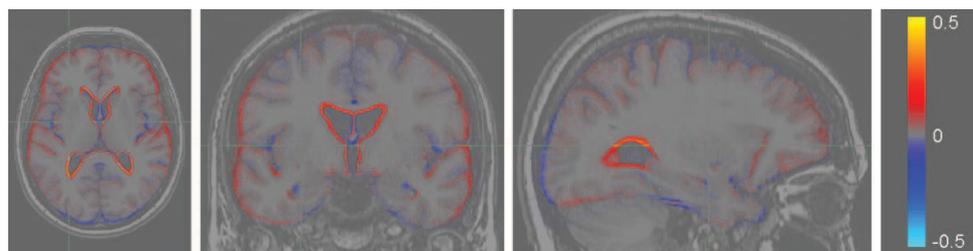
(a) Conventional kNN.



(b) Automatically trained kNN.

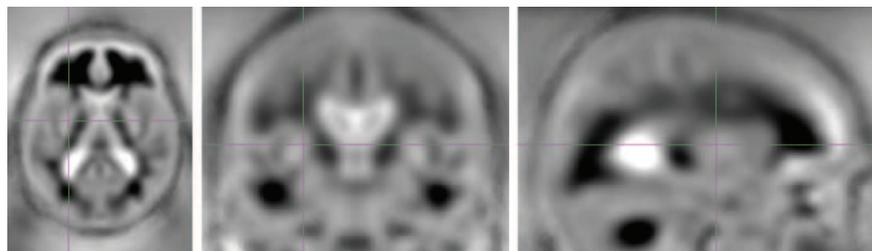


(c) Registration estimated atrophy (conventional kNN).

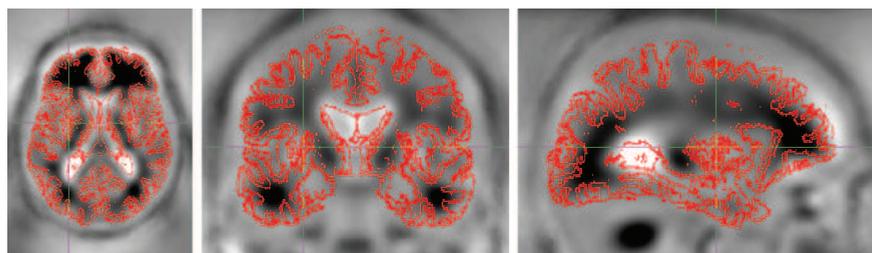


(d) Registration estimated atrophy (automatically trained kNN).

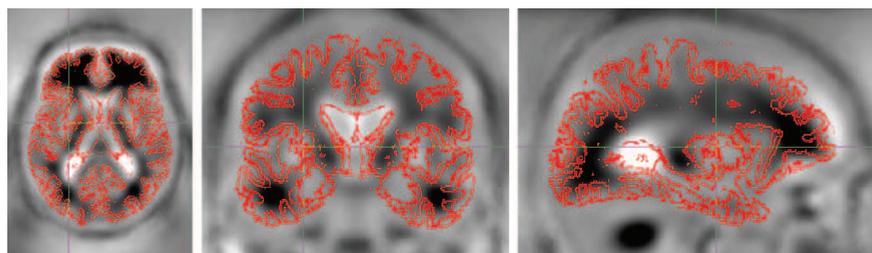
**Figure 4.3:** Brain atrophy maps based on 567 subjects. Red indicates tissue loss, blue indicates growth.



(a) Median Jacobian.



(b) Median Jacobian with conventional kNN gray matter outline.



(c) Median Jacobian with automatically trained kNN gray matter outline.

**Figure 4.4:** The median over 567 subjects of the determinant of the Jacobian matrix. The Jacobian matrix is computed based on the  $b$ -spline transformation between follow-up and baseline scans. Black indicates compression and white indicates expansion.

white matter.

## 4.4 Conclusion and discussion

We compared two brain tissue segmentation methods on longitudinal brain MRI data. Previous research showed that both methods have good accuracy and reproducibility with small differences between methods. This study showed that these differences between the methods can have a large influence on the segmentation of

longitudinal data. Both the resulting volume change rates of gray and white matter and the location of the gray and white matter changes differed. The two methods did not disagree on the magnitude and location of total brain atrophy, indicating that the difference is in the segmentation of the boundary between gray and white matter.

Visual inspection of the segmentations resulted in small differences in the ratings of the two methods, with the automatically trained kNN method slightly favored. The large difference between ratings of the methods in the baseline segmentation is most likely due to the time difference of more than a year between the inspections. Part of the difference may also be explained by a learning effect of the subjects, causing the follow-up scans to have less motion artifacts than the baseline scans. The rater observed that the automatically trained kNN classifier is less sensitive to motion artifacts than the conventional kNN classifier. This can possibly cause a larger difference in ratings between the baseline and follow-up scans of the conventional kNN classifier compared to the automatically trained kNN classifier.

The volume change rates obtained in our study might be biased due to the exclusion of segmentations that require corrections and scans with (motion) artifacts. The reported volume change rates are, however, only reported for the purpose of comparing the two methods. An analysis of actual volume changes requires an unbiased selection of scans and subjects. The GM, WM and WML volume change rates obtained in this study show large differences between the two methods. Even though the exact rates might be biased, the trends found can be compared to literature.

Comparing atrophy rates to figures reported in literature can, however, be difficult due to differences in study population, method or reporting. The age range of a study population and the selection of the subjects, for example the inclusion of demented subjects, can influence the resulting volume change rates. Other examples that can complicate the comparison are the use of absolute or relative (to ICV) volumes or differences in brain masking. Most atrophy rates reported in literature are based on volumes including the cerebellum. In this study, the brain mask used included only the cerebrum. Furthermore, the difference between cross-sectional studies and longitudinal studies should be taken into account. Cross-sectional studies estimate the volume change per year based on volume data of many subjects measured at a defined time point. Longitudinal studies measure volume changes over time for multiple subjects. As such, this can affect the volume change rates and especially their variance.

The conventional kNN method yielded GM volumes that decreased with increasing age, while the GM volumes obtained with the automatically trained kNN method showed a slight increase with increasing age. There is some discussion in literature about gray matter atrophy. Several studies report gray matter atrophy with age, while others find no such atrophy. A similar discussion exists for white matter atrophy. See Greenberg et al. (2008); Ikram et al. (2008) for discussions and

references on the topics of GM and WM atrophy.

Ikram et al. (2008) performed a cross-sectional study on volume change rates using a comparable study population (with more older subjects), an identical brain mask, and volumes relative to ICV. The segmentation method used was similar to the conventional kNN segmentation method used in this study, but it was trained on different scans because the MR images were acquired with a different scanning protocol. White matter lesions were classified as an additional class by the kNN classifier and were manually corrected where necessary. For men/women, Ikram et al. (2008) reported a non-significant increase in GM volume of 0.06/0.02% per year, a decrease in WM volume of -0.42/-0.43% per year, an increase in (log transformed) WML volume of 0.066/0.061% per year, and a decrease in brain volume of -0.31/-0.32 % per year. These results are very similar, and in some cases identical, to the volume change rates obtained with the automatically trained kNN method in this study.

The comparison of the two segmentation methods using atrophy maps also showed a preference for the automatically trained kNN method, as the GM and WM atrophy maps obtained with this method showed more similarities to the registration estimated atrophy maps. Additionally, the conventional kNN segmentations showed more inconsistency at the outer GM border. As also indicated by the volume change rates, the conventional kNN GM atrophy maps showed more loss and the automatically trained kNN GM maps more growth and shifting. The WM atrophy maps show primarily loss based on the automatically trained kNN and more growth based on the conventional kNN.

The sulci and gyri are difficult to capture with registration. This affects the atrophy maps as they are the average of the atrophy images of all subjects. The atrophy map voxel values near the gyri and sulci are therefore lower than the voxel values near the ventricles. Moreover, it has an effect on the registration estimated atrophy maps. Atrophy estimated by subtracting brain tissue segmentations of subsequent scans is more accurate because a brain tissue segmentation is better capable of modeling the sulci and gyri.

Automated training of the classifier on the scan(s) to be segmented has the advantage that the classification can adapt to differences in the scans. In longitudinal MRI data intensity differences can occur due to scanner drift or contrast changes with age. Cho et al. (1997) reported changes in MRI contrast between GM and WM with age. Such an effect will influence the brain tissue segmentation if the same training set is used for all scans.

In conclusion, we compare two automatic brain tissue segmentation methods on segmenting longitudinal data. The differences in the volume change rates of different tissues and the location of the atrophy is quite large. Overall, the automatically trained kNN segmentation method seems to be preferred over the conventional kNN segmentation method.

## **Part II**

# **Analysis of Structural Brain Connectivity**



# Chapter 5

## Statistical Analysis of Minimum Cost Path Based Structural Connectivity

Based on:

Renske de Boer, Michiel Schaap, Fedde van der Lijn, Henri A. Vrooman, Marius de Groot, Aad van der Lugt, M. Arfan Ikram, Meike W. Vernooij, Monique M.B. Breteler, Wiro J. Niessen, Statistical analysis of minimum cost path based structural brain connectivity, *NeuroImage*, 2011.

### Abstract

We present a framework for the construction of weighted structural brain networks, containing information about connectivity, which can be effectively analyzed using statistical methods. Connectivity is established using a minimum cost path (mcp) method with an anisotropic local cost function based directly on diffusion weighted images. We refer to this framework as Statistical Analysis of Minimum cost path based Structural Connectivity (SAMSCo) and the weighted structural connectivity networks as mcp-networks. In a proof of principle study we investigated the information contained in mcp-networks by predicting age based on the mcp-networks of a group of 974 middle-aged and elderly subjects. Using SAMSCo, age was predicted with a mean error of 3.7 years. This was significantly better than predictions based on fractional anisotropy or mean diffusivity averaged over the whole white matter or over the corpus callosum, which showed prediction errors of at least 4.8 years. Additionally, we classified subjects into groups of low and high white matter lesion (WML) load, correcting for age, sex and atrophy. SAMSCo outperformed the diffusion measures with a classification accuracy of 76.0% versus 63.2%. We also performed a classification in groups of mild and severe atrophy, correcting for age, sex and WML load. In this case, mcp-networks and diffusion measures yielded similar classification accuracies of 68.3% and 67.8% respectively. The SAMSCo prediction and classification experiments indicate that the mcp-networks contain information regarding age, WML load and white matter atrophy, and that in case of age and WML load the mcp-network based models outperformed the predictions based on diffusion measures.

## 5.1 Introduction

Structural or anatomical connectivity is a concept that aims to describe the white matter connection between brain regions. Structural connectivity can be studied with diffusion MRI and is of great interest when studying brain physiology or pathology. The first connectivity related studies using diffusion MRI looked for a relation between region of interest averaged diffusion MRI based measures such as fractional anisotropy (FA) and mean diffusivity (MD) with age (O'Sullivan et al., 2001; Salat et al., 2005; Vernooij et al., 2008).

More recently, structural brain connectivity has been modeled by a network whose nodes represent brain regions and connections are obtained with tractography. Streamline tractography establishes tracts along the principal direction of a second order diffusion tensor model that is fitted to the diffusion MRI data. This type of tractography is often incapable of finding a connection in regions of direction ambiguity due to, for example, mixed fiber populations or noise. Probabilistic tractography has been proposed to overcome this problem. In probabilistic tractography, multiple flow vectors are chosen from a distribution of directions to account for the uncertainty in the principal diffusion direction (e.g., Behrens et al., 2003; Parker et al., 2003). In High Angular Resolution Diffusion Imaging (HARDI) higher order models of diffusion, such as the orientation diffusion function, are used to overcome problems due to direction ambiguity. Both streamline and probabilistic tractography methods have been introduced for these diffusion models (Descoteaux et al., 2009).

Tractography based connectivity networks have been analyzed using graph theoretical approaches (Bullmore and Sporns, 2009). Network measures like clustering coefficient, path length and efficiency are defined to study the topology of the network. Hagmann et al. (2007) were the first to propose a method for the construction of structural networks based on diffusion MRI and analyze their topology with graph theoretical approaches. Since then, structural brain networks and their topology measures have been used in multiple studies (Iturria-Medina et al., 2010; Li et al., 2009; Shu et al., 2009; Yan et al., 2011). Usually the structural brain networks analyzed with graph theoretical analysis are binary, i.e. there either does or does not exist a connection between two nodes. Sometimes connection weights are assigned based on the number of tracts between two nodes (e.g., Iturria-Medina et al., 2010; Li et al., 2009; Yan et al., 2011). Recently, Robinson et al. (2010) proposed machine learning as an alternative to graph theoretical analysis of connectivity networks. They express the 'strength' of a connection by the mean diffusion anisotropy over the corresponding tract and classify subjects into two age groups based on the resulting weighted networks.

We present an alternative framework to study connectivity based on a minimum cost path method with a local anisotropic cost function based directly on diffusion weighted imaging (DWI) data. Such a minimum cost path method can

establish connections without being hindered by uncertainty in the directionality (Fletcher et al., 2007; Jackowski et al., 2005; Melonakos et al., 2008; Pichon et al., 2005). Minimum cost path methods are designed to find globally optimal paths given a cost function. Probabilistic tractography is more likely to end at a local optimum because of limited flow vector sampling in a modeled distribution. Also, minimum cost path methods contain no random factor and will therefore give reproducible results.

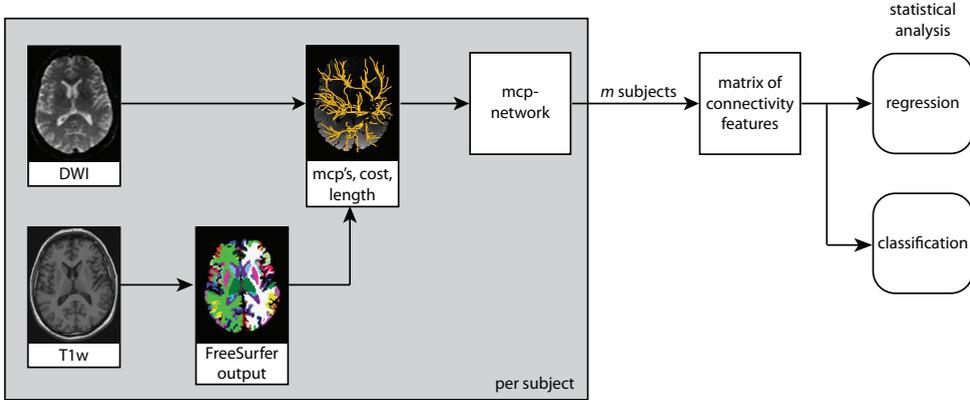
Minimum cost path methods have been used before to obtain a brain connectivity matrix in a single subject (Jbabdi et al., 2008). Tuch et al. also constructed a brain connectivity matrix, using a simulated annealing algorithm (Tuch, 2002; Tuch et al., 2001), but this algorithm is not guaranteed to find the global optimum of the objective function. The main difference of our work with the work by Jbabdi et al. and Tuch et al. is that we establish correspondence between subjects, thereby enabling group analysis.

The proposed framework extracts connectivity information, derived from the minimum cost paths, and stores it in networks referred to as minimum cost path (mcp)-networks. These mcp-networks may encode valuable information about brain physiology or pathology such as differences due to normal aging, neurodegenerative disease or psychiatric disorders. As minimum cost path methods will always find a connection between two regions, the mcp-networks cannot be analyzed with conventional graph theory. They are, however, very suitable for analysis with statistical methods such as multivariate regression or classification. The proposed framework employs these statistical methods to examine the information encoded in the mcp-networks. We refer to the framework as Statistical Analysis of Minimum cost path based Structural Connectivity (SAMSCo).

In this proof of principle study, we investigate the added value of the mcp-networks over other diffusion measures regarding age, white matter atrophy and white matter lesion load information. We do so by predicting age and classifying subjects into groups of low or high white matter lesion load and groups of mild or severe white matter atrophy, based on mcp-networks. We compare the results to predictions and classifications based on regional averaged FA or MD.

## 5.2 The SAMSCo framework

Figure 5.1 shows a schematic overview of the SAMSCo framework discussed in the following sections.



**Figure 5.1:** Schematic overview of the SAMSCo framework for statistical group analysis of structural brain connectivity. Connectivity is established through minimum cost paths (mcp's) that are constructed using diffusion weighted images. The mcp's run from start to target regions defined by FreeSurfer segmentation and cortical parcellation. The image shows the mcp's starting at the left putamen and a slice of the corresponding cumulative cost image. Per subject an mcp-network is constructed based on the mcp's and the cumulative cost over, and path length of, these mcp's. The mcp-networks of all  $m$  subjects are combined into a matrix of connectivity features for statistical analysis.

## 5.2.1 Connectivity

### 5.2.1.1 Minimum cost paths

Connectivity between two brain regions is established using minimum cost paths with a local anisotropic cost function.  $\Gamma : [0, L] \rightarrow \mathbb{R}^3$  is a path with Euclidean length  $L$  that runs between start region  $R$  and point  $\mathbf{p}$ . The connection between  $R$  and  $\mathbf{p}$  is defined by the path  $\hat{\Gamma}$  having minimum cumulative cost  $u(R, \mathbf{p})$  given by

$$u(R, \mathbf{p}) = \min_{\Gamma} \int_0^L \psi(\Gamma(s), \Gamma'(s)) ds, \quad (5.1)$$

where  $s$  is the arc length along  $\Gamma$ ;  $\Gamma'(s)$  is the unit local direction of the path; and  $\psi(\mathbf{x}, \mathbf{v})$  is the local anisotropic cost function, defining the cost at position  $\mathbf{x}$  in direction  $\mathbf{v}$ .

In case of a local isotropic cost function  $\psi(\mathbf{x})$ , independent of local direction, Equation 5.1 satisfies the Eikonal equation. In this case, Equation 5.1 can be solved using the single-pass Fast Marching method (Sethian, 1999). Diffusion MRI data does, however, contain directionality information that can be exploited, in which case a local anisotropic cost function is required and conventional Fast Marching can not be used.

Jbabdi et al. (2008) proposed an alternative Fast Marching algorithm that is capable of handling a local anisotropic cost function. We use, however, the Fast Sweeping algorithm, originally proposed in Kao et al. (2004), to numerically solve Equation 5.1 for local anisotropic cost functions based on DWI (Melonakos et al., 2008). We restrict the minimum cost paths to the brain by constraining the algorithm within a gray and white matter mask. The Fast Sweeping algorithm iteratively updates for each voxel in the mask the cumulative cost  $u(R, \mathbf{p})$  by choosing the optimal direction  $\mathbf{v}$  from a discrete set of directions. In this work, we use 100 (interpolated) directions distributed uniformly over a sphere.

### 5.2.1.2 Local cost function

For the minimum cost paths to run through white matter bundles, the local cost function should be low both on and in the direction of these white matter bundles. To this end, different cost functions have been proposed, of which some are based on the second order diffusion tensor model (Fletcher et al., 2007; Jackowski et al., 2005). These second order diffusion tensor models are, however, not suitable for modeling regions of multiple fiber populations, e.g. in the event of crossing fibers. Pichon et al. (2005) and Melonakos et al. (2008) proposed local cost functions based on DWI data rather than the diffusion tensor. We use the local cost function proposed in Melonakos et al. (2008):

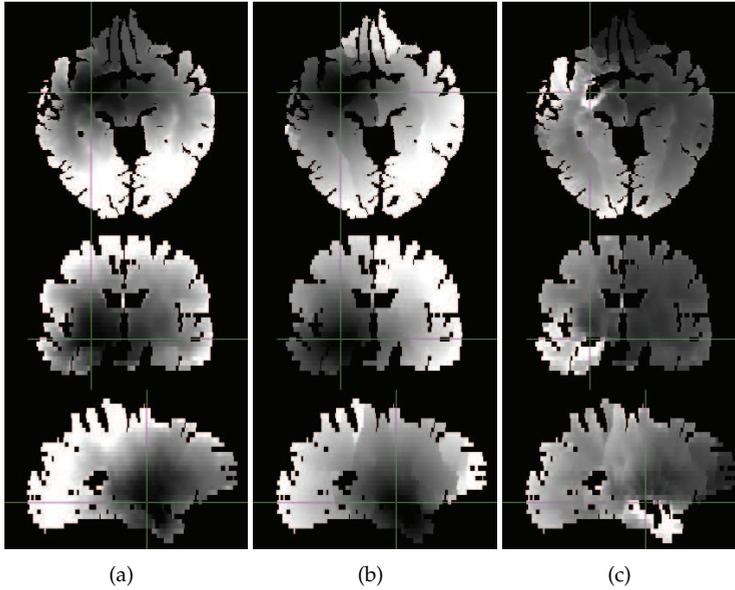
$$\psi(\mathbf{x}, \mathbf{v}) = \left( \frac{S(\mathbf{x}, \mathbf{v})}{\int_{\mathbf{w} \perp \mathbf{v}} \frac{S(\mathbf{x}, \mathbf{w})}{S(\mathbf{x}, \mathbf{0})} d\mathbf{w}} \right)^3, \quad (5.2)$$

where  $S(\mathbf{x}, \mathbf{v})$  is the signal intensity of the diffusion weighted image at position  $\mathbf{x}$  and interpolated (diffusion gradient) direction  $\mathbf{v}$ ; and  $S(\mathbf{x}, \mathbf{0})$  is the signal intensity of the image without diffusion weighting, i.e. the  $b_0$  intensity, at position  $\mathbf{x}$ . Due to diffusion-related signal loss,  $S(\mathbf{x}, \mathbf{v})$  is low if the diffusion at position  $\mathbf{x}$  in direction  $\mathbf{v}$  is high. The costs are therefore small if the diffusion in direction  $\mathbf{v}$  is high compared to directions perpendicular to  $\mathbf{v}$ .

### 5.2.1.3 Quantifying connectivity

A weight, representing the strength of the connection, is assigned to every connection in the mcp-network. Connectivity can be represented by different measures. While constructing the minimum cost path  $\hat{\Gamma}$ , defined by  $\psi(\mathbf{x}, \mathbf{v})$ , it is possible to integrate a local measure  $g(\mathbf{x}, \mathbf{v})$  from the start region  $R$  to point  $\mathbf{p}$ . Dividing the resulting cumulative measure by the Euclidean path length  $L$  yields a mean measure over the minimum cost path,  $\bar{g}_{\hat{\Gamma}}$ :

$$\bar{g}_{\hat{\Gamma}} \triangleq \frac{1}{L} \int_0^L g(\hat{\Gamma}(s), \hat{\Gamma}'(s)) ds. \quad (5.3)$$



**Figure 5.2:** Example slices with right amygdala (cross hair) start region: Dividing (a) cumulative costs  $u(R, \mathbf{p})$  by (b) path length  $L$  yields (c) average costs over the minimum cost paths  $\bar{u}_{\hat{\Gamma}}$ .

The local measure can either be dependent or independent of direction. It is therefore possible to use Equation 5.3 to average the (direction-independent) measures FA and MD over the minimum cost path. Both these measures are, however, based on the tensor model, which has shortcomings as discussed before. We prefer a measure that is not based on a tensor model and use in this proof of principle study  $g(\mathbf{x}, \mathbf{v}) = \psi(\mathbf{x}, \mathbf{v})$ . Our connectivity measure is the mean cost  $\bar{u}_{\hat{\Gamma}}$  which equals:

$$\bar{u}_{\hat{\Gamma}} = \frac{1}{L} \int_0^L \psi(\hat{\Gamma}(s), \hat{\Gamma}'(s)) ds = \frac{u(R, \mathbf{p})}{L}. \quad (5.4)$$

The local cost function depends on both local anisotropy and diffusivity and its average over the minimum cost path is therefore a suitable connectivity measure. We chose to normalize by the length of the minimum cost path in order to correct for differences in head size and/or brain atrophy. Figure 5.2 shows, for three example slices, the cumulative costs, path length, and average costs over the minimum cost paths starting in the right amygdala.

## 5.2.2 Construction of the mcp-network

To enable statistical analysis of mcp-networks, corresponding brain regions should be defined in all subjects. We use the FreeSurfer software package<sup>1</sup>, which is capable of segmenting subcortical structures (Fischl et al., 2004b) and parcellating the cortex (Fischl et al., 2004a) based on T1-weighted (T1w) images. The T1w scan is rigidly registered to the b0 diffusion image using Elastix<sup>2</sup> (Klein et al., 2010). The FreeSurfer segmentation and parcellation are transformed to DWI space according to the resulting transformation. The gray and white matter mask, used to restrict the Fast Sweeping algorithm to the brain, is defined by the FreeSurfer segmentation.

Additionally, the FreeSurfer segmentation of the subcortical structures and cortical parcellation define the start and target regions. In this chapter, we use the term connection to refer to an mcp-network connection between two nodes, or (the trajectory of) the corresponding minimum cost path. This does not necessarily correspond to a direct anatomical connection between the two brain regions. Moreover, in an anatomical connection the minimum cost path runs from start region to target region, which is not necessarily the same direction as signals are transported along the white matter path.

An mcp-network consists of  $n$  connections that are weighted by  $\bar{u}_{\hat{f}}$ . As every voxel in the target region has a different  $\bar{u}_{\hat{f}}$ , one value needs to be defined to represent the connection between the start and target region. We use the value of  $\bar{u}_{\hat{f}}$  of the voxel with minimum cumulative cost  $u(R, \mathbf{p})$ . All the minimum cost paths running to the target region most probably run through the same white matter bundle. Of these paths, the minimum cost path running to the voxel with minimum cumulative cost is the most optimal.

Using the  $\bar{u}_{\hat{f}}$  at the representing voxels, mcp-networks are obtained for the  $m$  subjects. All of these mcp-networks are combined into an  $m \times n$  matrix of connectivity features for statistical analysis.

## 5.2.3 Statistical analysis

SAMSCo uses statistical analysis to investigate whether the matrix of connectivity features contains information regarding connectivity changes e.g. due to normal aging or neurodegenerative disease. Based on this matrix, we investigated the prediction of variables such as subject age using regression, and the classification of subjects into groups defined by markers of brain tissue degeneration.

---

<sup>1</sup>FreeSurfer is available at <http://surfer.nmr.mgh.harvard.edu/>

<sup>2</sup>Elastix is available at <http://elastix.isi.uu.nl/>

### 5.2.3.1 Regression

Multivariate regression can be used to predict a particular variable  $y$ , e.g. a disease severity index or subject age, based on the matrix of connectivity features. In linear regression, the predicted value  $\hat{y}$  depends on the vector of input variables  $\mathbf{f} = (f_1, f_2, \dots, f_N)^T$ :

$$\hat{y} = \sum_{j=1}^N f_j \beta_j + \beta_0, \quad (5.5)$$

with  $\beta_j$  the regression coefficients, and  $\beta_0$  the intercept. When the matrix of connectivity features is used as regression input, the length of the input vector,  $N$ , is equal to the number of connections  $n$ . In this case, it is often necessary to perform some sort of regularization to prevent overfitting, as there generally will be more features than subjects ( $n > m$ ).

We perform regression based on two types of shrinkage methods, namely ridge regression and lasso (Hastie et al., 2009). Furthermore two regression methods that use derived input, namely principal component regression (PCR) and partial least squares regression (PLS) (Hastie et al., 2009), will be considered. Shrinkage methods impose a penalty on the magnitude of the regression coefficients. Ridge regression uses an  $L_2$ -norm penalty, while lasso uses an  $L_1$ -norm penalty. In PCR, regression is performed using the principal components of the input data. PLS also constructs a new space for regression, but does so based on both the input variables and the dependent variable(s). Ridge regression and lasso have a shrinkage parameter that needs to be optimized. In PCR and PLS the number of components requires optimization.

SAMSCo regression can for example be used to study the possibility of predicting dementia. It is also possible to use SAMSCo regression to relate a certain variable, e.g. age, to the mcp-network and corresponding minimum cost paths. Even though the minimum cost paths do not necessarily correspond to direct anatomical connections, they contain information about the brain regions they pass through. It can therefore be of interest to study the minimum cost paths related to this variable in more detail.

In a regression model, a positive  $\beta_j$  expresses an increase in the mean cost  $\bar{u}_{\hat{f}}$  of the corresponding connection with an increase of the outcome variable. An increase in  $\bar{u}_{\hat{f}}$  indicates a decrease in diffusivity in the path direction compared to the diffusivity in the perpendicular directions, which could be due to the deterioration of white matter at these locations. The connections with high absolute  $\beta_j$  have the largest influence on the prediction. Permutation tests allow for testing of significance of the contributions of these connections.

### 5.2.3.2 Classification

If it is possible to classify subjects into two or more groups based on their mcp-network, the mcp-networks contain information regarding the variable used to separate the groups. Groups can, for example, consist of patients versus control subjects. Similar to regression, overfitting should be prevented in classification. We perform classifications using support vector machines (SVMs) (Joachims, 1999) with a linear kernel. SVMs find the best separating hyperplane with the largest margin between two classes, and are not very sensitive to overfitting.

It is straightforward to compute the separating hyperplane of an SVM with a linear kernel. For a vector of input variables  $\mathbf{f} = (f_1, f_2, \dots, f_N)^T$ , a two-class linear SVM classification satisfies:

$$c = \sum_{j=1}^N f_j w_j - b, \quad (5.6)$$

with  $\mathbf{w} = (w_1, w_2, \dots, w_N)^T$  the normal of the separating hyperplane;  $\frac{b}{\|\mathbf{w}\|}$  the offset of the separating hyperplane from the origin along  $\mathbf{w}$ ; and  $N$  the length of the input vector, in our case  $N = n$ . For  $c$  equal to zero, the equation describes the separating hyperplane. Input is classified according to a negative or positive  $c$  value.

Similar to regression, it can be of interest to study the connections with large influence on the classification. Connections with a positive weight  $w_j$  have higher  $\bar{u}_{\hat{f}}$  in the positive class, and connections with a negative weight  $w_j$  have higher  $\bar{u}_{\hat{f}}$  in the negative class. Again, higher  $\bar{u}_{\hat{f}}$ , compared to other subjects, can indicate a region of deteriorated white matter along the minimum cost path.

## 5.3 Experiments

### 5.3.1 Data

Imaging data from the population-based Rotterdam Scan Study (Vernooij et al., 2009), acquired in 2005-2006, was used for the proof of principle experiments. Of the 1,094 subjects with completed MRI examinations, 120 subjects had to be excluded because of (motion) artifacts, cortical infarcts or FreeSurfer errors. The remaining 974 subjects consisted of 468 men and 506 women. Median age was 65.9 years with an interquartile range of 62.9 - 72.9 years. All subjects were non-demented.

Scans were obtained on a 1.5 Tesla GE scanner using an 8-channel head coil. All subjects were scanned on the same scanner and no software or hardware upgrades were performed during the study period. DWI images were acquired with a single shot, diffusion-weighted spin echo echo-planar imaging sequence (TR = 8000 ms, TE = 68.7 ms, FOV =  $21 \times 21$  cm<sup>2</sup>, matrix =  $96 \times 64$ , zero-padded in the frequency domain to  $256 \times 256$ ), 36 contiguous slices with slice thickness 3.5 mm, applying parallel imaging with acceleration factor = 2. Maximum b-value was 1000 s/mm<sup>2</sup>.

in 25 non-collinear directions (NEX = 1), and one volume was acquired without diffusion weighting ( $b_0$ ). The final voxel size used for analysis was  $0.8 \times 0.8 \times 3.5$  mm.

Additionally we performed a T1-weighted (T1w) 3D fast RF spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse sequence (TR = 13.8 ms, TE = 2.8 ms, TI = 400 ms, FOV =  $25 \times 25$  cm<sup>2</sup>, matrix =  $416 \times 256$  (zero-padded to  $512 \times 512$ ), flip angle =  $20^\circ$ , NEX = 1, bandwidth = 12.50 kHz, 96 contiguous slices with a slice thickness of 1.6 mm zero-padded to 0.8 mm). The final voxel size was  $0.49 \times 0.49 \times 0.8$  mm.

### 5.3.2 Start and target regions

Every additional start region requires a separate run of the computationally intensive Fast Sweeping algorithm, therefore we used a limited number of start regions. From the FreeSurfer segmentation, 17 subcortical regions were selected as start region: the brain stem, L/R-thalamus, head of the L/R caudate nucleus, L/R-putamen, L/R-pallidum, L/R-hippocampus, L/R-amygdala, L/R-accumbens area and L/R-ventral diencephalon (where L/R denotes left/right). The 16 subcortical regions not in use as start region were used as target regions. Additionally 75 cortical regions per hemisphere, as defined by the FreeSurfer parcellation based on the Destrieux 2009 atlas (Destrieux et al., 2010), were used as target regions. The resulting mcp-network consisted of  $2 \times 75 + 17 = 167$  nodes and  $n = \text{number of start regions} \times \text{number of target regions} = 17 \times (16 + 2 \times 75) = 2822$  connections.

### 5.3.3 Age prediction

We used the regression techniques explained in section 5.2.3.1 to predict subject age. We compared the results of the different methods to those of age prediction based on linear regressions with mean FA or mean MD as input. FA and MD were averaged over the entire white matter (WM) and additionally over the corpus callosum (CC), which should be less affected by possible registration errors than the WM.

Subjects were randomly divided over a test set, containing 30% of the subjects, and a training set. Regression parameters were optimized in a 5-fold cross validation on the training set. The optimal regression parameters were used for training of the models on the entire training set. The resulting models were tested on the hold-out test set. The same subdivision of subjects into test set, training set and folds, was used for every regression method. Prediction accuracy was evaluated by mean absolute difference ( $|\Delta|$ ) between predicted and actual subject age. A two-tailed Wilcoxon signed rank test was used to test for significant differences between the absolute differences of the prediction models. All models were also compared to a model assuming the median age of the training set for each subject.

In population studies of the elderly, gender can be distributed unevenly over all ages. As gender is related to the matrix of connectivity features (de Boer et al., 2010a), it should be corrected for when predicting age. Gender correction was performed by multiple linear regression of the training set with gender as input variable and the mcp-network connection weights as outcome variables. The residuals of the connection weights were used as input for the age regression. A similar correction was performed for the FA and MD variables.

### 5.3.4 Classification of WML load and WM atrophy

Vernooij et al. (2008) previously showed that age related FA changes on the major white matter tracts can be in large part explained by existence of white matter lesions (WMLs) and white matter atrophy. With the classification technique explained before, we classified our subjects into two classes based on either white matter lesion load or white matter atrophy. As FreeSurfer does not provide an accurate white matter lesion segmentation, we used a conventional  $k$ -nearest neighbor based brain tissue segmentation method (Chapters 3 and 4) extended with the white matter lesion segmentation method from Chapter 2. All brain tissue and WML segmentations were inspected and corrected when necessary. Twelve segmentations were not corrected because the required corrections were too extensive and the corresponding subjects were excluded. The remaining subjects had a median WM volume of 35.0% of the intracranial volume (ICV) (the sum of the cerebrospinal fluid, gray matter, white matter and white matter lesion volumes) and an interquartile range of 32.6 - 37.1 % of the ICV. Their median WML volume was 0.35% of the ICV and the interquartile range was 0.21 - 0.68 % of the ICV.

White matter lesion load was expressed as the natural log transformed percentage of ICV. White matter atrophy was defined by normal appearing white matter volume as percentage of ICV. Both variables were corrected for age and sex. Additionally white matter lesion load was corrected for white matter atrophy and vice versa. The correction was performed by linear regression with the confounding variables as input and the classification variable as output. The residuals of this regression were used to obtain two classes defined as the upper quartile and the lower quartile. The resulting classes consisted of 241 subjects for both white matter lesion load and white matter atrophy.

The matrix of connectivity features was normalized to zero mean and unit variance. Half of the subjects were used for optimizing the SVM parameter that determines the trade-off between training error and margin, in a 5-fold cross validation. Subsequently this set of samples was used to determine the separating hyperplane, which was then tested on the remaining 50% of the samples. Classification accuracy was defined as the percentage of correctly classified subjects. Classification was also performed with an SVM given the following four diffusion measures as input: WM averaged FA and MD, and corpus callosum averaged FA and MD. The diffusion

measures based classification used the same subdivision of subjects as the classification based on the matrix of connectivity features. The classification experiment was repeated 50 times with new training and test sets. The resulting classification accuracies of the connectivity based model and the diffusion measures based model were compared using a two-tailed paired t-test.

## 5.3.5 Results

### 5.3.5.1 Age prediction

Mean absolute difference between predicted age and actual age of the holdout test set for all models is reported in Table 5.1. The prediction errors of all models were compared to a model assuming the median age of the training set (66.1 years) for each subject, resulting in a mean absolute difference of 5.3 years. Both FA based models performed slightly worse than the median age model. Bonferroni correction adapts the threshold for significance for multiple comparisons by dividing the significance level by the number of tests performed. In this case, this results in a significance threshold of 0.006. The difference between the WM averaged FA model and the median age model was therefore significant after Bonferroni correction ( $p$ -value  $< 0.001$ ). The WM averaged MD based model resulted in a mean absolute difference of 4.9 years, which is better than the median age model but the difference was not significant ( $p$ -value = 0.09). The models based on the matrix of connectivity features resulted in the best age predictions, which differed significantly from the median age model ( $p$ -values  $\leq 0.002$ ). The ridge regression model yielded the lowest mean absolute difference of 3.7 years. Table 5.2 shows the  $p$ -values of the Wilcoxon signed rank tests comparing the results of the eight prediction models. The four diffusion measure based models differed significantly from the four connectivity based models, with the exception of the lasso and WM averaged MD models. The four connectivity based models constructed from different multivariate regression methods did not differ significantly when compared to each other, except for the lasso model differing from the ridge regression model.

Figure 5.3 shows the predicted age versus the actual age of all subjects in the holdout test set. Predictions are obtained using the ridge regression model. Scatterplots of the predictions from the other models based on the matrix of connectivity features versus the actual age were comparable. For all the four models, the correlation between the predicted and actual age was significantly different from zero ( $p$ -values  $< 10^{-46}$ ). The correlation coefficients were 0.70 (lasso and PCR), 0.71 (PLS), and 0.74 (ridge regression).

The regression models are still related to the original minimum cost paths. As an example, we describe the mcp-connections with maximum positive and minimum negative regression coefficients. These connections were not tested for significance of their contribution to the age prediction. Table 5.3 lists these connections and their corresponding regression coefficients per regression model. The positive regression

**Table 5.1:** Mean (sd) absolute difference in predicted age and *p*-value of Wilcoxon signed rank test per model

	Model	$ \Delta $ (years)	<i>p</i> -value <sup>a</sup>
	Median age assumption <sup>b</sup>	5.3(0.30)	
Diffusion measures	WM-FA	<b>5.8(0.23)</b>	< <b>0.001</b>
	WM-MD	4.9(0.23)	0.09
	CC-FA	5.8(0.23)	0.007
	CC-MD	5.2(0.23)	0.7
Connectivity features	Ridge regression	<b>3.7(0.18)</b>	< <b>0.001</b>
	Lasso	<b>4.1(0.18)</b>	<b>0.002</b>
	PCR	<b>3.8(0.19)</b>	< <b>0.001</b>
	PLS	<b>3.8(0.19)</b>	< <b>0.001</b>

Bold indicates significance at the 0.05 level, Bonferroni corrected for multiple comparisons (*p*-value < 0.006).

<sup>a</sup>*p*-value of Wilcoxon signed rank test comparing prediction results to median age assumption.

<sup>b</sup>Model assumes median age of training set.

coefficients were in all cases higher than the absolute value of the negative regression coefficients, but the differences were small, especially for the lasso model. The models showed similarities in that the connection with the maximum coefficient was the same in the PCR and PLS models, and the connection with the minimum coefficient was the same in the ridge regression and lasso models. Additionally, the connections with the minimum coefficients in the PCR and PLS models have the same start and target regions but differ in hemisphere. As in SAMSCo the strength of the connection between regions is of primary interest, it is not required to extract the actual trajectory of the path running between the start and target region. However, it is of course possible to obtain this trajectory, by following the local minimum cost path directions, for visualization purposes. Figure 5.4(a) shows the minimum cost paths associated with the maximum and minimum regression coefficients in the PCR model in an example subject.

### 5.3.5.2 Classification of WML load and WM atrophy

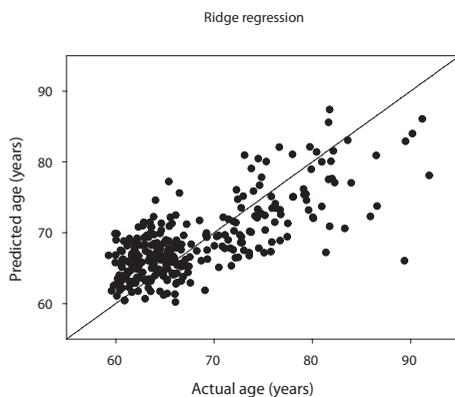
Over the 50 white matter lesion load classification experiments, the connectivity based classifier had a mean (sd) of 76.0 (2.2) % correctly classified subjects. The diffusion measures based classifier classified 63.2 (3.7) % of the subjects correctly. Figure 5.5(a) shows the percentages of correctly classified subjects of both classifiers for the 50 experiments. The classifier based on connectivity features outperformed the classifier based on the four diffusion measures in every experiment. The differ-

**Table 5.2:** *p*-values of paired, two-sided Wilcoxon signed rank test comparing the results of the age prediction models

	WM-FA	WM-MD	CC-FA	CC-MD	Ridge	Lasso	PCR
WM-MD	< 0.0001						
CC-FA	0.97	< 0.0001					
CC-MD	0.0001	0.038	< 0.0001				
Ridge regression	< 0.0001	0.0001	< 0.0001	< 0.0001			
Lasso	< 0.0001	0.0027	< 0.0001	< 0.0001	< 0.0001		
PCR	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.31	0.024	
PLS	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.44	0.015	0.55

The ridge regression, lasso, PCR and PLS models are based on the matrix of connectivity features.

Bold indicates significance at the 0.05 level, Bonferroni corrected for multiple comparisons (*p*-value < 0.0018).

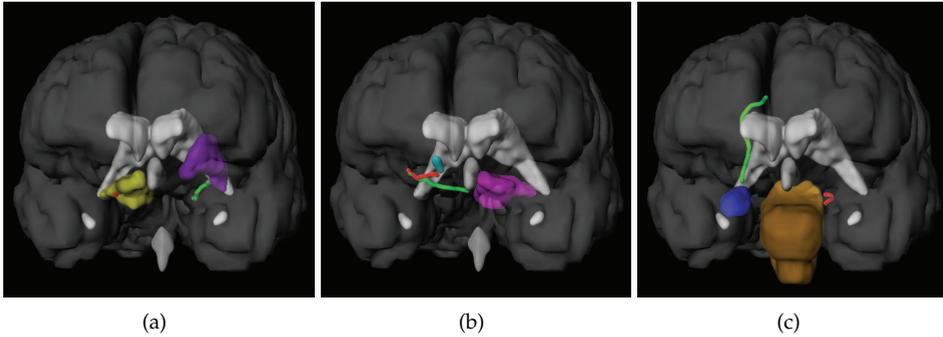


**Figure 5.3:** Predicted age, based on the ridge regression model, versus actual age for all subjects in the test set.

**Table 5.3:** Per regression model, connections with maximum positive or minimum negative regression coefficients in age prediction

Model	Start region	End region	Regression coefficient <sup>a</sup>
Ridge regression	head of left caudate nucleus	left pallidum	0.011
	left amygdala	brain stem	-0.0070
Lasso	head of left caudate nucleus	left superior temporal sulcus	0.033
	left amygdala	brain stem	-0.032
PCR	right ventral diencephalon	right amygdala	0.0071
	left putamen	left opercular part of the inferior frontal gyrus	-0.0064
PLS	right ventral diencephalon	right amygdala	0.0075
	right putamen	right opercular part of the inferior frontal gyrus	-0.0058

<sup>a</sup>Regression coefficients are listed to enable comparison of the maximum and minimum values.



**Figure 5.4:** Frontal oblique view of minimum cost paths with maximum or minimum coefficients or weights in (a) age prediction, according to the PCR model; (b) WML load classification; and (c) WM atrophy classification in an example subject. Minimum cost paths (mcp's) with maximum positive regression coefficient or SVM weight are shown in red, and mcp's with minimum negative regression coefficients or SVM weight in green. Only the start regions are visualized. Age prediction mcp's shown are: right ventral diencephalon (yellow) to right amygdala, and left putamen (purple) to left opercular part of the inferior frontal gyrus. WML load classification mcp's shown are: right accumbens area (cyan) and left ventral diencephalon (magenta) to right medial wall. WM atrophy classification mcp's shown are: brain stem (orange) to left parahippocampal gyrus, and right amygdala (blue) to right middle-posterior part of the cingulate gyrus and sulcus.

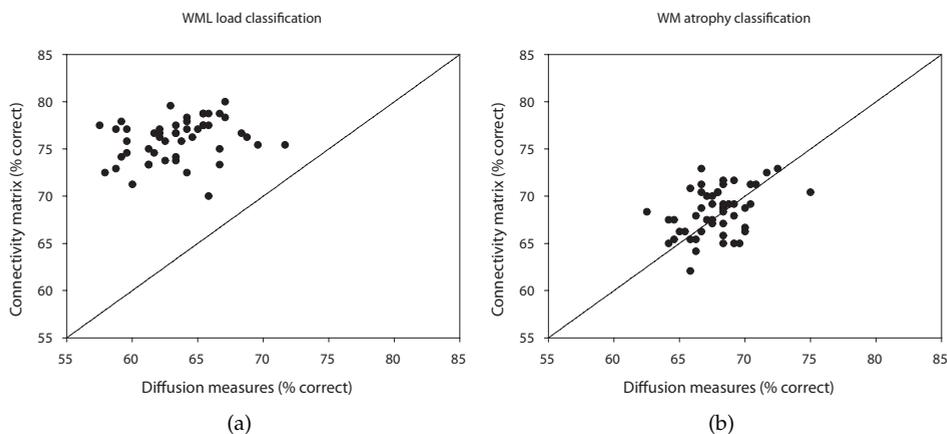
ence between the two classifiers was significant ( $p$ -value  $\ll 0.0001$ ).

The white matter atrophy classification based on connectivity had a mean (sd) of 68.3 (2.5) % correctly classified subjects over the 50 experiments. The classifier based on the four diffusion measures had a mean (sd) correctly classified subjects of 67.8 (2.3) %. Figure 5.5(b) shows the results for the 50 WM atrophy classifications. The results of both classifiers were comparable and did not differ significantly according to the paired t-test ( $p$ -value = 0.17).

As an example, Table 5.4 lists the connections that were most frequently associated with the maximum positive or minimum negative SVM weights in the 50 experiments. As white matter atrophy was expressed as white matter volume, a negative SVM weight indicates an increase in costs with an increase in white matter atrophy. Figures 5.4(b) and 5.4(c) show the connections in an example subject. The minimum negative connection in WM atrophy classification between the right hippocampus and the right temporal pole is not shown.

## 5.4 Conclusion and discussion

We proposed a framework, referred to as SAMSCo, for the construction and analysis of weighted structural connectivity networks. Connectivity is established using



**Figure 5.5:** Classification results for (a) white matter lesion load and (b) white matter atrophy of the classifier based on connectivity features and the classifier based on the four diffusion measures (global white matter averaged FA and MD, and corpus callosum averaged FA and MD).

**Table 5.4:** Connections most frequently associated with maximum positive or minimum negative SVM weights in 50 runs of white matter lesion load classification and white matter atrophy classification

Classified variable	Start region	End region	SVM weight
WML load	right accumbens area	right medial wall region	maximum positive
	left ventral diencephalon	right medial wall region	minimum negative
WM volume	brain stem	left parahippocampal gyrus	maximum positive
	right hippocampus	right temporal pole	minimum negative
	right amygdala	right middle-posterior part of the cingulate gyrus and sulcus	minimum negative

a minimum cost path method with a local anisotropic cost function based on diffusion weighted data. Network nodes are defined by FreeSurfer segmentation and parcellation based on T1w images. The resulting mcp-networks can be analyzed by statistical methods such as regression or classification.

Contrary to binary tractography based networks, mcp-networks contain connection weights representing the structural integrity of the white matter along the minimum cost paths. In this proof of principle study, we used the mean cost over the minimum cost path. It is, however, also possible to use other parameters in the SAMSCo framework. Any local DWI parameter (dependent or independent of direction) can be integrated over the minimum cost path. Alternatively, the information along the path can be represented by, for example, the path length below a certain threshold or the maximum of a parameter over the path.

The use of connection weights in mcp-networks allows for studying more than topology. Statistical methods can be applied to the connectivity features, similar

to the experiments in this study, or to network measures derived from the mcp-networks. Topological analysis is possible if a binary mcp-network is created, for example, by thresholding the maximum cost on the minimum cost path.

Both streamline and probabilistic tractography may yield connections in some subjects that are not found in others. Minimum cost path methods will always find a connection between two regions. This allows for analysis of all connections in the mcp-network instead of only a subset. The disadvantage is, however, that it might include biologically implausible connections. Connections obtained with the proposed framework may also include indirect connections, connecting two regions through a third.

We tested SAMSCo in several proof of principle experiments. Firstly, we predicted subject age based on the mcp-networks of 974 middle-aged and elderly subjects with an average error of 3.7 years. In a narrow age distribution, median age can be quite an accurate estimate. Our age prediction based on mcp-networks was, however, better than assuming the median age of the training set for each subject. Additionally, predicted age versus actual age showed high correlation coefficients. Our predictions on the mcp-networks outperformed predictions based on global WM or corpus callosum averaged FA or MD. We tested several regression techniques on the mcp-networks, which resulted in similar age prediction accuracies. Only the model obtained with lasso performed slightly worse than the ridge regression model.

Secondly, we classified subjects based on their mcp-network into two groups defined by white matter lesion load, while correcting for age, sex and white matter atrophy. And lastly, we used SAMSCo to perform classification of groups defined by white matter atrophy, while correcting for age, sex and white matter lesion load. Both classification results were better than random classification (50% correct), indicating that the mcp-networks contain information regarding white matter lesion load and white matter atrophy.

The SAMSCo classification of white matter lesion load clearly outperformed classification based on WM and corpus callosum averaged FA and MD, demonstrating the additional information encoded in the mcp-networks. We hypothesize that the regional information encoded in the mcp-networks lead to better classification results. White matter lesions can occur everywhere in the white matter. There are however some locations that are more prone to exhibit white matter lesions, for example the areas adjacent to the ventricles. Although areas affected by WMLs show differences in FA and MD, it is likely that these differences disappear when FA and MD are averaged over the entire white matter. Even though the corpus callosum is a small region compared to the global white matter, it is not likely to be affected by WMLs and is therefore not beneficial in the classification.

The regional information encoded in the mcp-networks did not improve the white matter atrophy classification. White matter atrophy is a more global effect than WMLs, affecting a larger portion of the white matter (Vernooij et al., 2008). It

is therefore not surprising that global WM and corpus callosum averaged FA and MD yielded similar WM atrophy classification results as mcp-networks.

The regression and classification models can be translated back to minimum cost paths. As an example, we reported the minimum cost paths with maximum positive or minimum negative regression coefficients or SVM weights. The contributions of these minimum cost paths to the regression or classification were, however, not tested for significance. If one wants to study which minimum cost paths have a relevant contribution to the regression or classification, permutation testing is necessary to test the significance of these contributions.

In the experiments, we use rigid registration to transform T1w images to diffusion space. This kind of co-registration may not be sufficient due to Eddy current effects and susceptibility artifacts. It has, however, successfully been used before on this data, at which time it was visually checked (Vernooij et al., 2008). Due to the limited Eddy currents and the relatively low field strength of the scanner, we experienced little problems with this registration. If there was misregistration of the T1w image to the b0 image, this would affect the global white matter averaged FA and MD. Therefore, all analyses were also performed with the corpus callosum averaged FA and MD, as these measures should be less affected by possible misregistrations.

When interpolating, the Fast Sweeping algorithm takes the size of the voxel in all three dimensions into account. Anisotropic voxel sizes are therefore not an issue for the computation of the minimum cost paths. Tests show that the cumulative cost and Euclidean path length of the minimum cost path running from region A to region B are very similar, but not identical, to those of the minimum cost path running from region B to region A. The difference is due to spatial and directional discretization. Especially on short minimum cost paths, this can have a significant effect. Removing short connections might therefore be considered.

The time required for the minimum cost path computation differs between start regions and subjects. This is due to differences in local anatomy, because the number of iterations of the Fast Sweeping algorithm depends on the number of turns in the minimum cost path trajectory (Melonakos et al., 2008). In this study, the majority of the minimum cost path computations, for a single start region and subject, took approximately half an hour. About 10% took more than an hour.

Because these computation times can add up to large amounts for a large number of start regions, we included only the subcortical structures as start regions for this proof of principle study. As the cortico-cortical connections might also hold valuable information, it might be beneficial to include these connections in the mcp-networks by adding the cortical regions as start regions. For efficient calculation this would require computation optimization. As Fast Sweeping can be efficiently implemented in a parallel fashion, we believe a significant speed-up can be obtained by implementing it on a parallel processing architecture such as a modern graphics processing unit (GPU) (see, for example, Jeong and Whitaker (2007) for

related work).

Including cortico-cortical connections yields a tenfold increase of the number of connections. Because, in most studies, the number of subjects will most likely be considerably smaller, care should be taken to prevent overfitting. Alternatively, only a subset of all connections can be used. Tractography or fMRI can, for example, indicate connections or brain regions that are more likely to be associated with the variable at interest.

In conclusion, we presented SAMSCo, a framework for construction and analysis of mcp-networks. The mcp-network connections represent connectivity based on DWI data. These brain networks can be used to study connectivity changes related to, for example, aging or neurodegenerative disease. In a proof of principle study, we demonstrated that the mcp-networks contain information regarding age, white matter lesion load and white matter atrophy.



# Chapter 6

## Reproducibility of the SAMSCo Framework

Renske de Boer, Michiel Schaap, Henri A. Vrooman, Aad van der Lugt, M. Arfan Ikram, Meike W. Vernooij, Monique M.B. Breteler, Wiro J. Niessen

### **Abstract**

We assessed the reproducibility of the Statistical Analysis of Minimum cost path based Structural Connectivity (SAMSCo) framework. We used scans of 26 subjects that were scanned twice within a short time interval (median 14.5 days). The Mahalanobis distance between the mcp-networks of the scan-rescan pairs was compared to the Mahalanobis distance between the mcp-networks of different subjects. The intra-subject Mahalanobis distance was significantly smaller than the inter-subject Mahalanobis distance. Additionally, age was predicted based on the mcp-networks using several trained prediction models. The median difference in predicted age of the scan-rescan pairs did not differ significantly from zero. The median absolute difference was in the range of 1.6 - 2.3 years, depending on the model used.

## 6.1 Introduction

Structural brain connectivity can be studied using brain networks. In Chapter 5 we presented a framework for the statistical analysis of brain networks based on minimum cost paths (mcp's), referred to as SAMSCo. In proof of principle experiments, we demonstrated that these mcp-networks contain information regarding age, white matter lesion load, and white matter atrophy.

Minimum cost path methods are designed to find the global optimum paths given a cost function. They contain no random factor and will therefore yield reproducible results given identical input. Magnetic resonance images do, however, contain noise, discretization causes partial volume effects, and motion artifacts may occur. In case of diffusion MRI, this will result in mixed fiber populations in a single voxel. A rescan, with repositioning and slight intra-individual variation over time, will therefore differ from the original scan and (slightly) different minimum cost paths will be found.

In this chapter, we perform several experiments to assess the reproducibility of the SAMSCo framework. We focus on scan-rescan reproducibility and do not test the influence of acquisition changes. We use 26 subjects that were scanned twice within a short time interval. We evaluated the reproducibility of the mcp-networks and the age prediction models.

## 6.2 Materials and methods

### 6.2.1 Data

Imaging data from the population-based Rotterdam Scan Study (Vernooij et al., 2009) was used. Twenty-six subjects were scanned twice within a short time interval in 2010. This group had a median age at first scan of 76.8 years with an interquartile range of 74.9 - 80.0 years and a total age range of 70.3 - 84.1 years. The group consisted of 13 men and 13 women. Median inter-scan interval was 14.5 days with an interquartile range of 14.0 - 18.0 days. Additionally, we used the scans of 974 subjects, which were also used for the proof of principle experiments in Chapter 5. These scans were acquired in 2005-2006. Median age was 65.9 years with an interquartile range of 62.9 - 72.9 years. The total age range was 59.0 - 96.7 years. The two groups of subjects did not overlap and all subjects were non-demented at time of scanning.

Scans were obtained on a 1.5 Tesla GE scanner using an 8-channel head coil. All subjects were scanned on the same scanner and no software or hardware upgrades were performed during the study period. DWI images were acquired with a single shot, diffusion-weighted spin echo echo-planar imaging sequence (TR = 8000 ms, TE = 68.7 ms, FOV =  $21 \times 21$  cm<sup>2</sup>, matrix =  $96 \times 64$ , zero-padded in the frequency domain to  $256 \times 256$ ), 36 contiguous slices with slice thickness 3.5 mm, applying

parallel imaging with acceleration factor equal to 2. Maximum b-value was  $1000 \text{ s/mm}^2$  in 25 non-collinear directions (NEX = 1), and one volume was acquired without diffusion weighting (b0). The final voxel size used for analysis was  $0.8 \times 0.8 \times 3.5 \text{ mm}$ .

Additionally we performed a T1-weighted 3D fast RF spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse sequence (TR = 13.8 ms, TE = 2.8 ms, TI = 400 ms, FOV =  $25 \times 25 \text{ cm}^2$ , matrix =  $416 \times 256$  (zero-padded to  $512 \times 512$ ), flip angle =  $20^\circ$ , NEX = 1, bandwidth = 12.50 kHz, 96 contiguous slices with a slice thickness of 1.6 mm zero-padded to 0.8 mm. The final voxel size was  $0.49 \times 0.49 \times 0.8 \text{ mm}$ ).

## 6.2.2 Experiments

To investigate the reproducibility of the SAMSCo framework, we computed the mcp-networks for both scans of the 26 subjects. We used the same definition of the start and target regions as described in section 5.3.2. The start regions consisted of 17 subcortical structures defined by the FreeSurfer segmentation (Fischl et al., 2004b). These regions were also used as target regions in addition to 75 cortical regions per hemisphere as defined by the FreeSurfer parcellation of the cortex (Fischl et al., 2004a).

Similarity of the two mcp-networks was assessed by calculating the distance between the networks. Because of the high-dimensionality of the connectivity feature space, we computed this distance in the subspace spanned by the  $x$  eigenvectors with the largest eigenvalues. As the eigenvectors have different variances it is necessary to normalize them. The Mahalanobis distance takes these variances into account and is therefore a suitable distance measure.

We calculated the Mahalanobis distance using the generalized inverse matrix,  $\hat{\Sigma}^{-1}$ , of the covariance matrix,  $\Sigma$ , of all connectivity features,

$$\hat{\Sigma}^{-1} = VD^{-1}V^T, \quad (6.1)$$

where  $D$  is a diagonal matrix with the  $x$  largest eigenvalues of  $\Sigma$  on the diagonal and  $V$  is a matrix with the corresponding  $x$  eigenvectors as columns. These eigenvectors were obtained using the connectivity features of the 974 subjects studied in Chapter 5 and both scans of the 26 subjects. The Mahalanobis distance  $M_{i,j}$  between sample  $i$  and  $j$  is then calculated as follows:

$$M_{i,j} = \sqrt{(X_i - X_j)^T \hat{\Sigma}^{-1} (X_i - X_j)} \quad (6.2)$$

with  $X_i$  the connectivity features for sample  $i$ . We computed the Mahalanobis distances between all possible scan pairs of the 26 subjects. A two-sample t-test, assuming unequal variances, was performed to compare the mean of the intra-subject Mahalanobis distance to the mean of the inter-subject Mahalanobis distance. To

investigate the influence of the number of included eigenvectors, we performed this t-test for every number of included eigenvectors. The maximum number of included eigenvectors with non-zero eigenvalues is equal to the number of input scans ( $974 + 2 \times 26 = 1026$ ).

Additionally we assessed the reproducibility by predicting the age of the 26 subjects using the age prediction models from Chapter 5. These models were based on four regression methods, namely ridge regression, lasso, principal component regression (PCR) and partial least squares regression (PLS). The models were trained on 70% of the 974 subjects. Age was predicted for all scan pairs. The difference in age prediction ( $\Delta$ ) was defined as the difference between the age based on the second scan minus the age based on the first scan. We tested if the prediction differences had zero median using a Wilcoxon signed rank test. Additionally we tested if the differences in prediction differences between the models had zero median ( $\text{median}(\Delta_{\text{model1}} - \Delta_{\text{model2}}) = 0$ ) using a paired Wilcoxon signed rank test.

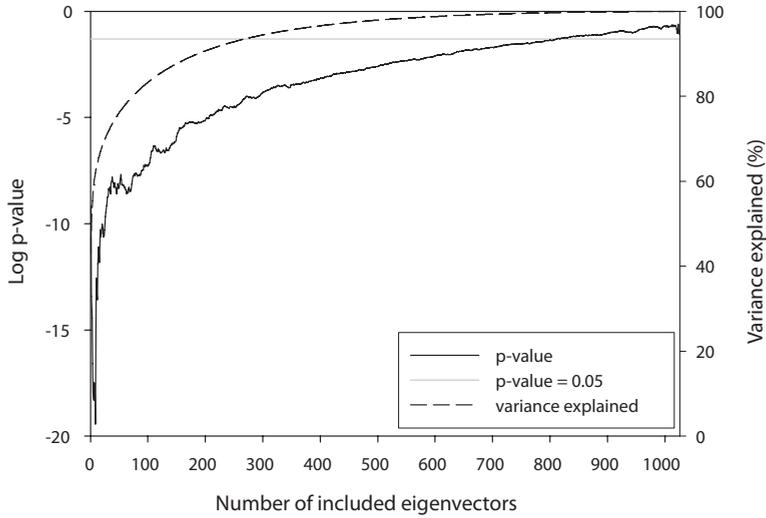
### 6.3 Results

Figure 6.1 shows the p-values for testing the difference between intra- and inter-subject Mahalanobis distances as function of the number of eigenvectors included in the calculation of the covariance matrix. Ninety percent of the variation in the data was explained by the first 188 eigenvectors. Figure 6.2 shows the Mahalanobis distance, based on these 188 eigenvectors, between the two scans of the 26 subjects. The intra-subject Mahalanobis distance appears smaller than the inter-subject distance, which is confirmed by the t-test which indicates that the means of the intra-subject and inter-subject distances are significantly different ( $p\text{-value} < 10^{-5}$ ).

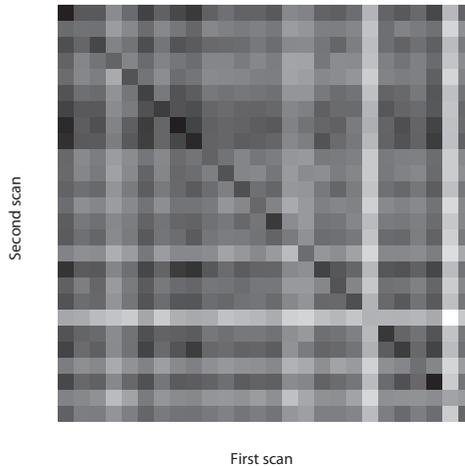
Figure 6.3 shows the histograms of the differences in age predictions for the four models. Table 6.1 lists the median absolute difference in age prediction for the 26 subjects. The PLS model had the lowest median absolute difference. The median difference in age prediction between first and second scan did not differ from zero for any of the models according to the Wilcoxon signed rank tests ( $p\text{-values} \geq 0.2$ ). The median differences in  $\Delta$ 's between the models were not significantly different from zero at a 0.05 significance level, with the exception of the differences between the  $\Delta$ 's of the PCR and PLS models ( $p\text{-value} = 0.03$ ). If the significance level was, however, corrected for multiple comparisons using Bonferroni ( $p\text{-value} < 0.008$ ), this difference between PCR and PLS model was also not significant.

### 6.4 Conclusion and discussion

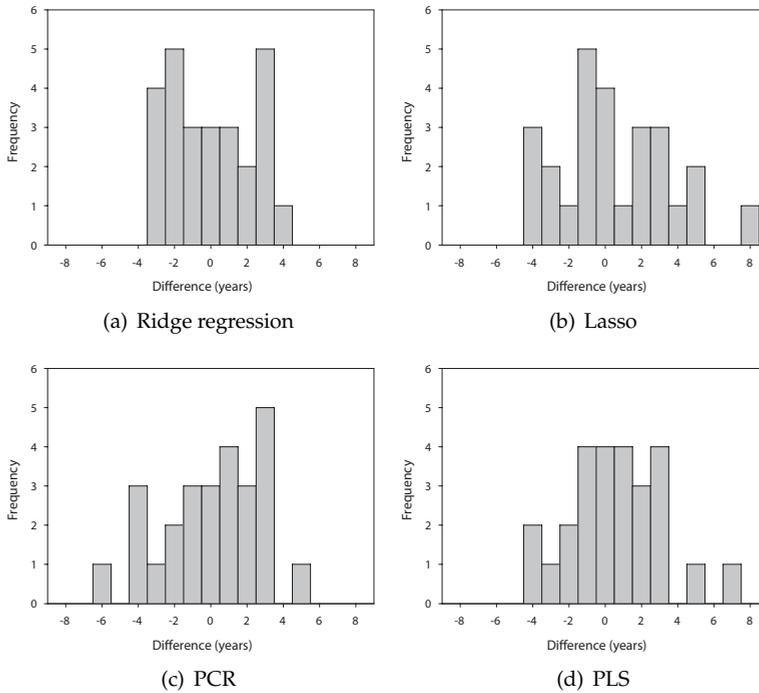
We assessed the reproducibility of the SAMSCo framework using 26 subjects that were scanned twice within a short time interval. No significant brain changes are assumed to take place within this time period. Natural variations in, for example,



**Figure 6.1:** Comparison of intra- and inter-subject Mahalanobis distances in the 26 subjects using different numbers of eigenvectors. The explained variance for these eigenvectors is also shown.



**Figure 6.2:** Mahalanobis distance between the first and second scan of the 26 subjects. The subjects are sorted according to their age. Gray-scale from black to white indicates increased distance; the dark diagonal thus illustrates smaller intra- than inter-subject distance.



**Figure 6.3:** Histograms of the differences in age prediction between scan and rescan for the four prediction models.

**Table 6.1:** Difference in age prediction between scan and rescan.

Model	Median absolute difference (years)
Ridge regression	2.1
Lasso	2.3
PCR	2.2
PLS	1.6

fluid balance do occur, but these variations are also captured in longitudinal studies and should therefore be included in estimating reproducibility. Other factors that may influence reproducibility are excluded as much as possible by using the same scanner and the same scanning protocol for all subjects.

The intra-subject Mahalanobis distance is smaller than the inter-subject Mahalanobis distance. This indicates that the connectivity features of the scan-rescan pairs are more alike than the scans of different subjects, which is a condition for good reproducibility. Additionally, the age prediction models from Chapter 5 were used to predict age for both scans. The median absolute difference between the two age predictions was 2 years. The PLS model had the lowest absolute difference of the four regression models used, but the differences between the models were small. The median difference between the models' prediction differences did not differ from zero.

The 26 subjects used for the reproducibility experiments were older than the majority of the subjects used to train the age prediction models. The predicted age is therefore not optimal, but the difference between the age predictions of the scan-rescan pairs can be used to assess the reproducibility.

In conclusion, we tested the reproducibility of the SAMSCo framework. Intra-subject differences between scan and rescan mcp-networks were smaller than the differences between the mcp-networks of different subjects. The median absolute difference in age prediction between first and second scan was 2 years.



# **Chapter 7**

## **Summary and Discussion**

In this thesis we focused on the automatic segmentation of brain tissues and white matter lesions from structural MRI data and the analysis of structural connectivity from diffusion MRI data. We presented new methods for these tasks and evaluated these together with previously proposed methods on their accuracy and reproducibility within the scope of a large population imaging study. This last chapter provides a summary and a discussion of this work and future research.

## 7.1 Summary

### 7.1.1 Brain tissue and white matter lesion segmentation

Segmentation enables volume measurements and can be a preprocessing step for further analysis and visualization. In Chapter 2 we proposed a method for automatic white matter lesion (WML) segmentation that can be used as an extension to a brain tissue segmentation method. This extension uses a brain tissue segmentation to determine a FLAIR intensity threshold for white matter lesions and to subsequently ensure that the individual lesions are surrounded by white matter. We optimized a previously proposed brain tissue segmentation method in combination with the white matter lesion segmentation extension. This automatically trained brain tissue segmentation method uses atlas registration to obtain training samples from regions with high tissue probability. The final classification is performed by a  $k$ -nearest neighbor (kNN) classifier using the intensities of the input scans as features. The quantitative evaluation showed that the accuracy of the brain tissue and white matter lesion segmentation is close to the interobserver variability.

Many brain tissue segmentation methods have been proposed, some of which are publicly available. There are, however, few studies that compare the performance of different segmentation methods on the same data. In Chapter 3 we compared several brain tissue segmentation methods on their accuracy and reproducibility. We evaluated two publicly available methods, namely FAST and SPM5, and two methods based on a kNN classifier, namely the automatically trained kNN method optimized in Chapter 2 and a conventional kNN method. All methods showed good accuracy and reproducibility with only small differences between the methods. The conventional kNN method was the most accurate but showed the lowest reproducibility. FAST yielded the most reproducible brain tissue segmentation volumes. Although the differences in reproducibility were small, they had a relatively large effect on the required number of subjects in the design of a longitudinal study with sufficient power.

In Chapter 4 we compare the performance of the two kNN segmentation methods from Chapter 3 on longitudinal MRI data. Both methods were extended with the white matter lesion segmentation method from Chapter 2. The different methods yielded significantly different volume change rates per year increase of age for gray and white matter and white matter lesions. Gray and white matter atro-

phy maps differed also between the methods, while the whole brain atrophy maps were comparable. Overall, the performance of the automatically trained segmentation method appeared to be better.

### 7.1.2 Analysis of structural brain connectivity

Most analysis of structural brain connectivity is limited to studying the topology of a binary brain network constructed using tractography. In Chapter 5 we presented a framework for the construction of weighted structural brain connectivity networks which can be effectively analyzed using statistical methods. This allows for studying more than just the topology of the obtained networks. The weighted networks are obtained using a minimum cost path (mcp) method with an anisotropic local cost function based on diffusion weighted images. We refer to the framework as Statistical Analysis of Minimum cost path based Structural Connectivity (SAMSCo) and the weighted networks as mcp-networks. In a proof of principle study, we performed age prediction based on the mcp-networks, thereby demonstrating that there is age information encoded in these mcp-networks. We additionally showed that the mcp-networks contain information regarding white matter lesion load and white matter atrophy.

We evaluated the SAMSCo framework on its reproducibility in Chapter 6. The Mahalanobis distance between the mcp-networks of the same subject were smaller than the distance between mcp-networks of different subjects. Additionally, we assessed the reproducibility of the age prediction. The median absolute difference between the age predictions based on two scans of the same subject was 2 years.

## 7.2 Discussion

### 7.2.1 Brain tissue and white matter lesion segmentation

The automatic brain tissue and white matter lesion segmentation method optimized in this thesis performed well. Accuracy evaluation showed a high overlap between the automatic and manual segmentations, which was comparable to the overlap between manual segmentations by different observers. Reproducibility of the segmentation and the resulting volumes was good. The accuracy and reproducibility of the method was comparable to other brain tissue segmentation methods. There are, however, a few considerations regarding the evaluation of accuracy and reproducibility that should be taken into account.

Evaluating brain tissue and white matter lesion segmentations is difficult as it is not known what the true anatomy is. Manual segmentations suffer from intra- and interobserver variability and are often biased due to the use of a thresholding tool. Also, they are laborious and time-consuming to generate, so evaluations will usually be limited to comparison to a small number of manual segmentations.

When compared to manual segmentations, most automatic brain tissue segmentation methods are at an accuracy level close to the interobserver variability, as indicated in Chapter 3.

White matter lesion segmentations usually consist of several small components with a small total volume and large surface area. Overlap, either in volume or in segmentation, between manual and automatic segmentations is therefore often low. This affects the accuracy evaluation measures causing WML segmentation accuracy to appear poor compared to the large volumes of gray and white matter. Interobserver variability is, however, also high and the automatic WML segmentation method performs comparable to this interobserver variability. In Chapter 2 we additionally used other accuracy evaluation measures based on the distance between the boundary of the lesions. These evaluations also indicated that the distance between the automatic and the manual WML segmentations was comparable to the distance between manual segmentations by two observers. Since the development of the white matter lesion segmentation method, it has been applied to thousands of scans. Visual inspection of the results indicate that the method is quite robust. Approximately 4% of the MRI datasets yield white matter lesion segmentations containing false positive or false negative lesions (Sections 2.4.3 and 4.3.1). It can, however, be argued that the effect of these errors in WML segmentations will be small when investigating a large cohort. A study comparing a set of uncorrected automatic WML segmentation to a set including corrections would be required to investigate the magnitude of this effect.

It is also possible to assess the accuracy of a segmentation method using simulated MRI data, in which case the ground truth is known. This can be a valuable addition, but should not be a substitute for the evaluation on real MRI data. Realistic MRI data are, if possible, extremely difficult to generate and an evaluation on simulated data alone is therefore not sufficient. Additionally, by applying a segmentation method to a large set of real MRI scans, the method is also tested on its robustness to variation in anatomy, pathology, and scanner hardware and software. The robustness can be assessed by visual inspection of the resulting segmentations. Visual inspection is, however, only a global evaluation and it is very subjective. This can, for example, be seen in the differences between the rating results reported in Chapter 4.

Segmentations can also be evaluated with respect to their ability to study associations between derived volumes and other variables such as age. We used this type of evaluation in Chapter 2 for the white matter lesions and in Chapter 4 for all tissue types. It is, however, difficult to interpret the found associations. In some cases previously reported results can be replicated. In other cases, such as in the comparison in Chapter 4, different results are obtained for different methods and it is unclear which result is more reliable.

We evaluated reproducibility of the brain tissue and white matter lesion segmentations and derived volumes with the goal of understanding its influence on a

longitudinal study. We used subjects that were scanned twice within a time interval of a few weeks. A number of previous studies scan subjects multiple times on the same day, either with or without repositioning in the scanner. A time interval of at least several days ensures that natural variations in, for example, fluid balance are included in the reproducibility assessment. Because these variations are also present in a longitudinal study, as is the repositioning of the subject, we have chosen for such a study design.

Many brain tissue segmentation methods have been proposed in literature and most give comparable results both with respect to accuracy and reproducibility. The accuracy and reproducibility of these methods are often high, as shown in this thesis, but small differences in the accuracies and reproducibilities do have a relatively large effect on the volume change rates found in a longitudinal study (Chapter 4). Different methods can even disagree on the occurrence of atrophy. These differences may be caused by differences in dealing with the age-related changes in MRI contrast between gray and white matter (Cho et al., 1997). An extensive evaluation study is required to obtain the true volume change rates. Such an evaluation study can, for example, consist of a comparison of MRI scans, and corresponding segmentations, to histological sections.

It is probably possible to improve currently existing automatic segmentation methods, especially the segmentation methods that are only based on intensity such as the kNN based methods. An improvement using additional information is, however, very likely to increase computation times significantly. Furthermore, the evaluation of an improved method requires a suitable way of assessing the accuracy as the accuracy of current automatic methods is already close to the inter-observer variability. An improvement in segmentation accuracy would therefore be difficult to measure.

We applied the brain tissue and white matter lesion segmentation methods to MRI scans from the Rotterdam Scan Study. The data from the Rotterdam Scan Study are very homogeneous, as no scanner (software) updates are performed and the same scanner and scanning protocol is used for all subjects. This homogeneity benefits the analysis and the methods do not need to be robust to large variations in input data. It is, however, expected that the automatically trained kNN method and the WML segmentation extension from Chapter 2 are capable of handling different kinds of data, because they train on the scans to be segmented. The WML segmentation method might require re-optimization of its parameters if the scanner or scanning sequence parameters change.

## 7.2.2 Analysis of structural brain connectivity

With the SAMSCo framework we presented a new way of analyzing structural brain connectivity (Chapter 5). The framework was evaluated by proof of principle experiments in which we predicted age or classified subjects into groups based

on their WML load or level of white matter atrophy. We thereby demonstrated that the obtained mcp-networks contain information regarding age, white matter lesion load and white matter atrophy. In addition, we evaluated the reproducibility of the mcp-networks and the age prediction models.

Several factors have influenced the evaluation studies that were performed. The subjects used in Chapter 5 for training and testing of the age prediction model had a skewed age distribution. The prediction model is therefore not optimally trained for all ages in the age range. The age prediction of the older subjects used in Chapter 6 had for example a larger prediction error than the test subjects from Chapter 5. It is expected that a more evenly distributed age range will improve age prediction.

The reproducibility of the framework was assessed in Chapter 6 using subjects that were scanned twice. Like the reproducibility of the brain tissue segmentation methods, the influences of image acquisition, repositioning, and variations in fluid balance are included in this study by scanning subjects with a interscan interval of a few weeks. The results give an indication of the variation that can be expected to occur in mcp-networks and age prediction.

As stated previously, the evaluation studies carried out in this thesis only provide a proof of concept: Age prediction and classification based on white matter atrophy or WML load are of course not relevant brain connectivity research topics. However, the results of these proof of principle studies are encouraging. The framework is generic and the generated weighted connectivity networks allow for statistical analysis within or between groups. In a cross-sectional or longitudinal study it can provide valuable insight into brain maturation, aging or disease progression. In these studies, the framework can indicate which minimum cost paths differ most between groups or show changes within a group, thereby generating new hypotheses. Furthermore, the framework can potentially be used to predict neurological diseases before clinical symptoms occur.

In the proof of principle experiments the start and target regions were defined by the FreeSurfer segmentation and cortical parcellation. The choices for the region boundaries in the FreeSurfer parcellation are based on the sulci and gyri in the cortex. It can not be assumed that this corresponds to boundaries of regions with similar connectivity. The use of start and target regions defined by equal connectivity can improve the analysis of structural connectivity. Functional MRI (fMRI) can provide regions with corresponding functional activity, which are assumed to be connected. The SAMSCo framework can be used to investigate the structural connectivity between these regions and hence can be used for the integrated analysis of fMRI and diffusion MRI data.

## References

- Admiraal-Behloul, F., van den Heuvel, D.M.J., Olofsen, H., van Osch, M.J.P., van der Grond, J., van Buchem, M.A., Reiber, J.H.C., 2005. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage* 28, 607–617.
- Alfano, B., Brunetti, A., Larobina, M., Quarantelli, M., Tedeschi, E., Ciarmiello, A., Covelli, E.M., Salvatore, M., 2000. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. *J Magn Reson Imaging* 12, 799–807.
- Amato, U., Larobina, M., Antoniadis, A., Alfano, B., 2003. Segmentation of magnetic resonance brain images through discriminant analysis. *J Neurosci Methods* 131, 65–74.
- Anbeek, P., Vincken, K.L., van Bochove, G.S., van Osch, M.J.P., van der Grond, J., 2005. Probabilistic segmentation of brain tissue in MR imaging. *Neuroimage* 27, 795–804.
- Anbeek, P., Vincken, K.L., van Osch, M.J., Bisschops, R.H., van der Grond, J., 2004. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Med Image Anal* 8, 205–215.
- Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821.
- Ashburner, J., Friston, K.J., 2005. Unified segmentation. *Neuroimage* 26, 839–851.
- Awate, S.P., Tasdizen, T., Foster, N., Whitaker, R.T., 2006. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. *Med Image Anal* 10, 726–739.
- Barra, V., Boire, J.Y., 2000. Tissue segmentation on MR images of the brain by possibilistic clustering on a 3D wavelet representation. *J Magn Reson Imaging* 11, 267–278.
- Beaulieu, C., 2002. The basis of anisotropic water diffusion in the nervous system - a technical review. *NMR Biomed* 15, 435–455.
- Behrens, T.E.J., Woolrich, M.W., Jenkinson, M., Johansen-Berg, H., Nunes, R.G., Clare, S., Matthews, P.M., Brady, J.M., Smith, S.M., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn Reson Med* 50, 1077–1088.
- de Boer, R., Schaap, M., van der Lijn, F., Vrooman, H.A., de Groot, M., Vernooij, M.W., Ikram, M.A., van Velsen, E.F.S., van der Lugt, A., Breteler, M.M.B., Niessen, W.J., 2010a. Statistical analysis of structural brain connectivity. *Med Image Comput Comput Assist Interv* 13, 101–108.
- de Boer, R., Vrooman, H.A., Ikram, M.A., Vernooij, M.W., Breteler, M.M.B., van der Lugt, A.,

- Niessen, W.J., 2010b. Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *Neuroimage* 51, 1047–1056.
- de Boer, R., Vrooman, H.A., van der Lijn, F., Vernooij, M.W., Ikram, M.A., van der Lugt, A., Breteler, M.M., Niessen, W.J., 2009. White matter lesion extension to automatic brain tissue segmentation on MRI. *Neuroimage* 45, 1151–1161.
- de Bresser, J., Portegies, M.P., Leemans, A., Biessels, G.J., Kappelle, L.J., Viergever, M.A., 2011. A comparison of MR based segmentation methods for measuring brain atrophy progression. *Neuroimage* 54, 760–768.
- Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* 10, 186–198.
- Cardenas, V.A., Ezekiel, F., Sclafani, V.D., Gomberg, B., Fein, G., 2001. Reliability of tissue volumes and their spatial distribution for segmented magnetic resonance images. *Psychiatry Res* 106, 193–205.
- Chang, H.H., Zhuang, A.H., Valentino, D.J., Chu, W.C., 2009. Performance measure characterization for evaluating neuroimage segmentation algorithms. *Neuroimage* 47, 122–135.
- Chard, D.T., Parker, G.J.M., Griffin, C.M.B., Thompson, A.J., Miller, D.H., 2002. The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. *J Magn Reson Imaging* 15, 259–267.
- Cho, S., Jones, D., Reddick, W.E., Ogg, R.J., Steen, R.G., 1997. Establishing norms for age-related changes in proton T1 of human brain tissue in vivo. *Magn Reson Imaging* 15, 1133–1143.
- Clark, K.A., Woods, R.P., Rottenberg, D.A., Toga, A.W., Mazziotta, J.C., 2006. Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. *Neuroimage* 29, 185–202.
- Cocosco, C.A., Zijdenbos, A.P., Evans, A.C., 2003. A fully automatic and robust brain MRI tissue classification method. *Med Image Anal* 7, 513–527.
- DeCarli, C., Massaro, J., Harvey, D., Hald, J., Tullberg, M., Au, R., Beiser, A., D’Agostino, R., Wolf, P.A., 2005. Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. *Neurobiol Aging* 26, 491–510.
- Descoteaux, M., Deriche, R., Knösche, T.R., Anwander, A., 2009. Deterministic and probabilistic tractography based on complex fibre orientation distributions. *IEEE Trans Med Imaging* 28, 269–286.
- Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15.
- Dice, L., 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 297–302.
- Eskildsen, S.F., Ostergaard, L.R., 2006. Active surface approach for extraction of the human cerebral cortex from MRI. *Med Image Comput Comput Assist Interv* 9, 823–830.
- Faul, F., Erdfelder, E., Lang, A.G., Buchner, A., 2007. G\*Power 3: a flexible statistical power

- analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 39, 175–191.
- Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci U S A* 97, 11050–11055.
- Fischl, B., van der Kouwe, A., Destrieux, C., et al., 2004a. Automatically parcellating the human cerebral cortex. *Cereb Cortex* 14, 11–22.
- Fischl, B., Salat, D.H., van der Kouwe, A.J.W., et al., 2004b. Sequence-independent segmentation of magnetic resonance images. *NeuroImage* 23 Suppl 1, S69–S84.
- Fletcher, P.T., Tao, R., Jeong, W.K., Whitaker, R.T., 2007. A volumetric approach to quantifying region-to-region white matter connectivity in diffusion tensor MRI. *Inf Process Med Imaging* 20, 346–58.
- Fotinos, A.F., Snyder, A.Z., Girton, L.E., Morris, J.C., Buckner, R.L., 2005. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology* 64, 1032–1039.
- Fox, N.C., Cousens, S., Scahill, R., Harvey, R.J., Rossor, M.N., 2000. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects. *Arch Neurol* 57, 339–344.
- Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Trans Med Imaging* 16, 623–629.
- Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* 6, 67–77.
- Ge, Y., Grossman, R.I., Babb, J.S., Rabin, M.L., Mannon, L.J., Kolson, D.L., 2002. Age-related total gray matter and white matter changes in normal adult brain. Part I: volumetric MR imaging analysis. *AJNR Am J Neuroradiol* 23, 1327–1333.
- Greenberg, D.L., Messer, D.F., Payne, M.E., Macfall, J.R., Provenzale, J.M., Steffens, D.C., Krishnan, R.R., 2008. Aging, gender, and the elderly adult brain: an examination of analytical strategies. *Neurobiol Aging* 29, 290–302.
- de Groot, J.C., de Leeuw, F.E., Oudkerk, M., van Gijn, J., Hofman, A., Jolles, J., Breteler, M.M.B., 2002. Periventricular cerebral white matter lesions predict rate of cognitive decline. *Ann Neurol* 52, 335–341.
- Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V.J., Meuli, R., Thiran, J.P., 2007. Mapping human whole-brain structural networks with diffusion MRI. *PLoS One* 2, e597.
- Harris, G., Andreasen, N.C., Cizadlo, T., Bailey, J.M., Bockholt, H.J., Magnotta, V.A., Arndt, S., 1999. Improving tissue classification in MRI: a three-dimensional multispectral discriminant analysis method with automated training class selection. *J Comput Assist Tomogr* 23, 144–154.
- Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2009. *The elements of statistical learning: data mining, inference and prediction*. Springer. second edition.

- Hofman, A., Breteler, M.M., van Duijn, C.M., Janssen, H.L., Krestin, G.P., Kuipers, E.J., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., Witteman, J.C., 2009. The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol* 24, 553–572.
- Hofman, A., Breteler, M.M.B., van Duijn, C.M., Krestin, G.P., Pols, H.A., Stricker, B.H.C., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., Witteman, J.C.M., 2007. The Rotterdam Study: objectives and design update. *Eur J Epidemiol* 22, 819–829.
- Ikram, M.A., Vrooman, H.A., Vernooij, M.W., den Heijer, T., Hofman, A., Niessen, W.J., van der Lugt, A., Koudstaal, P.J., Breteler, M.M.B., 2010. Brain tissue volumes in relation to cognitive function and risk of dementia. *Neurobiol Aging* 31, 378–386.
- 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.B., 2008. Brain tissue volumes in the general elderly population. the Rotterdam Scan Study. *Neurobiol Aging* 29, 882–890.
- Iturria-Medina, Y., Fernández, A.P., Morris, D.M., Canales-Rodríguez, E.J., Haroon, H.A., Pentón, L.G., Augath, M., García, L.G., Logothetis, N., Parker, G.J.M., Melie-García, L., 2010. Brain hemispheric structural efficiency and interconnectivity rightward asymmetry in human and nonhuman primates. *Cereb Cortex* 21, 56–67.
- Jack, C. R., J., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J., 2001. FLAIR histogram segmentation for measurement of leukoaraiosis volume. *J Magn Reson Imaging* 14, 668–676.
- Jackowski, M., Kao, C.Y., Qiu, M., Constable, R.T., Staib, L.H., 2005. White matter tractography by anisotropic wavefront evolution and diffusion tensor imaging. *Med Image Anal* 9, 427–40.
- Jbabi, S., Bellec, P., Toro, R., Daunizeau, J., Péligrini-Issac, M., Benali, H., 2008. Accurate anisotropic fast marching for diffusion-based geodesic tractography. *Int J Biomed Imaging* 2008, 320195.
- Jeong, W.K., Whitaker, R.T., 2007. A fast iterative method for a class of Hamilton-Jacobi equations on parallel systems. Technical Report. University of Utah.
- Joachims, T., 1999. Making large-scale SVM learning practical, in: Schölkopf, B., Burges, C., Smola, A. (Eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.
- Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D., Maguire, P., Rosas, D., Makris, N., Gollub, R., Dale, A., Dickerson, B.C., Fischl, B., 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage* 46, 177–192.
- Kao, C.Y., Osher, S., Qian, J., 2004. Lax-Friedrichs sweeping scheme for static Hamilton-Jacobi equations. *J Comput Phys* 196, 367–391.
- Khayati, R., Vafadust, M., Towhidkhan, F., Nabavi, M., 2008. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and markov random field model. *Comput Biol Med* 38, 379–390.

- Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging* 29, 196–205.
- Klein, S., Staring, M., Pluim, J.P.W., 2007. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Trans Image Process* 16, 2879–2890.
- Kovacevic, N., Lobaugh, N.J., Bronskill, M.J., Levine, B., Feinstein, A., Black, S.E., 2002. A robust method for extraction and automatic segmentation of brain images. *Neuroimage* 17, 1087–1100.
- Kruggel, F., Paul, J.S., Gertz, H.J., 2008. Texture-based segmentation of diffuse lesions of the brain's white matter. *Neuroimage* 39, 987–996.
- van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans Med Imaging* 20, 677–688.
- van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. *IEEE Trans Med Imaging* 18, 897–908.
- de Leeuw, F.E., de Groot, J.C., Achten, E., Oudkerk, M., Ramos, L.M., Heijboer, R., Hofman, A., Jolles, J., van Gijn, J., Breteler, M.M., 2001. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the Rotterdam Scan Study. *J Neurol Neurosurg Psychiatry* 70, 9–14.
- Lemieux, L., Hagemann, G., Krakow, K., Woermann, F.G., 1999. Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. *Magn Reson Med* 42, 127–135.
- Lemieux, L., Hammers, A., Mackinnon, T., Liu, R.S.N., 2003. Automatic segmentation of the brain and intracranial cerebrospinal fluid in T1-weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. *Magn Reson Med* 49, 872–884.
- Li, Y., Liu, Y., Li, J., Qin, W., Li, K., Yu, C., Jiang, T., 2009. Brain anatomical network and intelligence. *PLoS Comput Biol* 5, e1000395.
- Maillard, P., Delcroix, N., Crivello, F., Dufouil, C., Gicquel, S., Joliot, M., Tzourio-Mazoyer, N., Alperovitch, A., Tzourio, C., Mazoyer, B., 2008. An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology* 50, 31–42.
- Malloy, P., Correia, S., Stebbins, G., Laidlaw, D.H., 2007. Neuroimaging of white matter in aging and dementia. *Clin Neuropsychol* 21, 73–109.
- Melonakos, J., Pichon, E., Angenent, S., Tannenbaum, A., 2008. Finsler active contours. *IEEE Trans Pattern Anal Mach Intell* 30, 412–23.
- O'Sullivan, M., Jones, D.K., Summers, P.E., Morris, R.G., Williams, S.C., Markus, H.S., 2001. Evidence for cortical "disconnection" as a mechanism of age-related cognitive decline. *Neurology* 57, 632–638.

- Parker, G.J.M., Haroon, H.A., Wheeler-Kingshott, C.A.M., 2003. A framework for a streamline-based probabilistic index of connectivity (PICO) using a structural interpretation of MRI diffusion measurements. *J Magn Reson Imaging* 18, 242–254.
- Pichon, E., Westin, C.F., Tannenbaum, A.R., 2005. A Hamilton-Jacobi-Bellman approach to high angular resolution diffusion tractography. *Med Image Comput Comput Assist Interv* 8, 180–187.
- Prins, N.D., van Dijk, E.J., den Heijer, T., Vermeer, S.E., Koudstaal, P.J., Oudkerk, M., Hofman, A., Breteler, M.M.B., 2004. Cerebral white matter lesions and the risk of dementia. *Arch Neurol* 61, 1531–1534.
- Robinson, E.C., Hammers, A., Ericsson, A., Edwards, A.D., Rueckert, D., 2010. Identifying population differences in whole-brain structural networks: a machine learning approach. *Neuroimage* 50, 910–919.
- Ruan, S., Jaggi, C., Xue, J., Fadili, J., Bloyet, D., 2000. Brain tissue classification of magnetic resonance images using partial volume modeling. *IEEE Trans Med Imaging* 19, 1179–1187.
- Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Trans Med Imaging* 18, 712–721.
- Ruf, A., Greenspan, H., Goldberger, J., 2005. Tissue classification of noisy MR brain images using constrained GMM. *Med Image Comput Comput Assist Interv* 8, 790–797.
- Salat, D.H., Tuch, D.S., Greve, D.N., van der Kouwe, A.J.W., Hevelone, N.D., Zaleta, A.K., Rosen, B.R., Fischl, B., Corkin, S., Rosas, H.D., Dale, A.M., 2005. Age-related alterations in white matter microstructure measured by diffusion tensor imaging. *Neurobiol Aging* 26, 1215–1227.
- Scahill, R.I., Frost, C., Jenkins, R., Whitwell, J.L., Rossor, M.N., Fox, N.C., 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Arch Neurol* 60, 989–994.
- Schott, J.M., Crutch, S.J., Frost, C., Warrington, E.K., Rossor, M.N., Fox, N.C., 2008. Neuropsychological correlates of whole brain atrophy in Alzheimer’s disease. *Neuropsychologia* 46, 1732–1737.
- Schott, J.M., Frost, C., Whitwell, J.L., Macmanus, D.G., Boyes, R.G., Rossor, M.N., Fox, N.C., 2006. Combining short interval MRI in Alzheimer’s disease: Implications for therapeutic trials. *J Neurol* 253, 1147–1153.
- Sethian, J., 1999. Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. Cambridge Univ Pr.
- Shu, N., Liu, Y., Li, J., Li, Y., Yu, C., Jiang, T., 2009. Altered anatomical network in early blindness revealed by diffusion tensor tractography. *PLoS One* 4, e7228.
- Shuter, B., Yeh, I.B., Graham, S., Au, C., Wang, S.C., 2008. Reproducibility of brain tissue volumes in longitudinal studies: effects of changes in signal-to-noise ratio and scanner software. *Neuroimage* 41, 371–379.

- Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging* 17, 87–97.
- Smith, S.M., 2002. Fast robust automated brain extraction. *Hum Brain Mapp* 17, 143–55.
- Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., Behrens, T.E.J., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* 31, 1487–1505.
- Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23 Suppl 1, S208–S219.
- Smith, S.M., Stefano, N.D., Jenkinson, M., Matthews, P.M., 2001. Normalized accurate measurement of longitudinal brain change. *J Comput Assist Tomogr* 25, 466–475.
- Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* 17, 479–489.
- Song, Z., Tustison, N., Avants, B., Gee, J.C., 2006. Integrated graph cuts for brain MRI segmentation. *Med Image Comput Comput Assist Interv* 9, 831–8.
- Tuch, D.S., 2002. Mapping cortical connectivity with diffusion MRI, in: *IEEE International Symposium on Biomedical Imaging*, pp. 392–394.
- Tuch, D.S., Wiegell, M.R., Reese, T.G., Belliveau, J.W., Wedeen, V.J., 2001. Measuring cortico-cortical connectivity matrices with diffusion spectrum imaging, in: *Proc Int Soc Magn Reson Med*, p. 502.
- Vermeer, S.E., Hollander, M., van Dijk, E.J., Hofman, A., Koudstaal, P.J., Breteler, M.M.B., Study, R.S., 2003. Silent brain infarcts and white matter lesions increase stroke risk in the general population: the Rotterdam Scan Study. *Stroke* 34, 1126–1129.
- Vernooij, M.W., de Groot, M., van der Lugt, A., Ikram, M.A., Krestin, G.P., Hofman, A., Niessen, W.J., Breteler, M.M.B., 2008. White matter atrophy and lesion formation explain the loss of structural integrity of white matter in aging. *Neuroimage* 43, 470–477.
- Vernooij, M.W., Ikram, M.A., Vrooman, H.A., Wielopolski, P.A., Krestin, G.P., Hofman, A., Niessen, W.J., der Lugt, A.V., Breteler, M.M.B., 2009. White matter microstructural integrity and cognitive function in a general elderly population. *Arch Gen Psychiatry* 66, 545–553.
- 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., 2007. Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. *Neuroimage* 37, 71–81.
- Wang, D., Galloway, G.J., de Zubicaray, G.I., Rose, S.E., Chalk, J.B., Doddrell, D.M., Semple, J., 1998. A reproducible method for automated extraction of brain volumes from 3D human head MR images. *J Magn Reson Imaging* 8, 480–486.

- Xue, Z., Shen, D., Davatzikos, C., 2006. CLASSIC: consistent longitudinal alignment and segmentation for serial image computing. *Neuroimage* 30, 388–399.
- Yan, C., Gong, G., Wang, J., Wang, D., Liu, D., Zhu, C., Chen, Z.J., Evans, A., Zang, Y., He, Y., 2011. Sex- and brain size-related small-world structural cortical networks in young adults: a DTI tractography study. *Cereb Cortex* 21, 449–458.
- Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 20, 45–57.
- Zijdenbos, A., Forghani, R., Evans, A., 1998. Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT, in: *Med Image Comput Comput Assist Interv*, pp. 439 – 448.
- Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging* 13, 716–24.

# Samenvatting

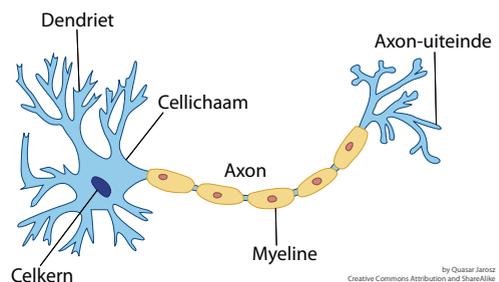
## Inleiding

In dit proefschrift worden methodes besproken om hersen-MRI-scans automatisch te analyseren. Van deze methodes moet bepaald worden hoe goed en betrouwbaar de resultaten zijn. Voor deze evaluatie worden de methodes toegepast op MRI-scans van de hersenen van oudere mensen. Deze inleiding op de samenvatting geeft wat korte achtergrondinformatie over de hersenen, veel voorkomende aandoeningen bij veroudering van de hersenen, MRI en MRI-onderzoeken.

## De hersenen

De hersenen bestaan uit twee soorten hersenweefsels, namelijk *grijze stof* en *witte stof*, en worden omringd door *hersenvocht*. Het hersenvocht bevindt zich ook in de holtes in het midden van de hersenen, die ventrikels worden genoemd. De grijze stof bestaat uit de cellichamen en dendrieten van de zenuwcellen. De witte stof bestaat uit de uitlopers, genaamd axonen, van de zenuwcellen en dankt zijn naam aan de myeline rond de axonen dat het signaaltransport versnelt. Figuur 7.1 laat een schematische weergave van een zenuwcel met dendrieten en axon zien. De grijze-stof-gebieden in de hersenen zijn met elkaar verbonden via de axonen. Een bundel van deze axonen heet een zenuwbaan. De dendrieten ontvangen de signalen van de axonen en geven ze door aan het cellichaam.

Twee veel voorkomende aandoeningen die optreden bij veroudering van de hersenen zijn *witte-stof-laesies* en *atrofie*. Witte-stof-laesies zijn beschadigingen van de witte stof. Het is aangetoond dat witte-stof-laesies verband houden met cognitieve achteruitgang en een verhoogde kans op beroerte en dementie. Atrofie is het krimpen van de hersenen en ook hier is een verband aangetoond met dementie.



**Figuur 7.1:** Zenuwcel met dendrieten en gemyeliniseerd axon.

## Hersenen-MRI

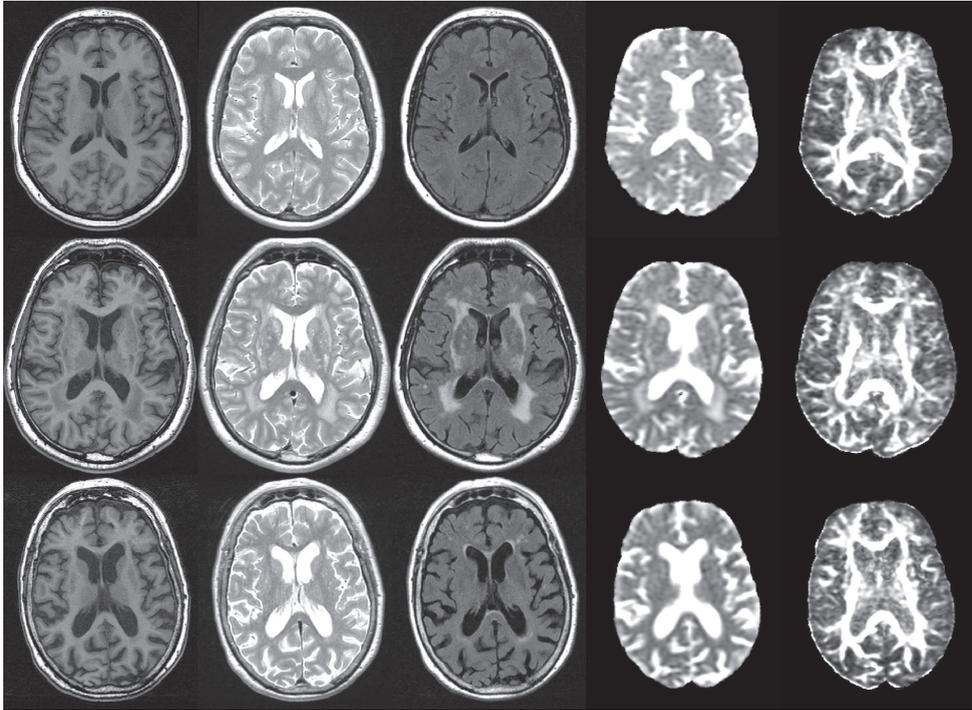
MRI staat voor *magnetic resonance imaging* en is een beeldvormende techniek die met name zachte weefsels zichtbaar maakt. De techniek is erg geschikt om de hersenen te bestuderen. Met een MRI-scanner kunnen verschillende soorten beelden van de hersenen gemaakt worden. De eerste drie kolommen in Figuur 7.2 laten drie soorten MRI-scans van de hersenen zien. Het verschil in de scans wordt veroorzaakt door de eigenschappen van de hersenweefsels en de instellingen voor het scannen. Het hersenvocht is te zien als zwart op de eerste scan en wit op de tweede scan. Witte stof is lichter dan grijze stof op de eerste scan en omgekeerd voor de tweede scan. Witte-stof-laesies zijn te zien als witte vlekken op de derde scan, een zogenaamde FLAIR-scan. Atrofie is met name op de onderste beelden duidelijk te zien als een vergroting van de ventrikels en de sulci (de 'gleuven' in de gevouwen hersenschors).

In *diffusie-MRI* wordt de diffusie van watermoleculen in de hersenen gemeten. In de witte stof wordt deze diffusie gehinderd door de zenuwbanen. De watermoleculen zullen daardoor netto meer in de richting van de zenuwbanen diffunderen dan loodrecht daar op. Als diffusie niet gelijk is in alle richtingen heet dit *anisotrope* diffusie. Door de diffusie van de watermoleculen te onderzoeken kan er iets gezegd worden over de richting van de zenuwbanen en de conditie van de witte stof. In beschadigde witte stof treedt namelijk meer diffusie op en diffunderen de watermoleculen vrijer. Diffusie-MRI-scans zijn hoog-dimensionaal en het is daarom alleen mogelijk om afgeleide beelden te laten zien. Figuur 7.2 laat bijvoorbeeld de grootte en de anisotropie van de diffusie zien. Hoe lichter het beeld plaatselijk is, hoe groter de diffusie of de anisotropie daar is. Het is duidelijk te zien dat de diffusie in witte-stof-laesies groter is en de anisotropie lager.

Hersenen-MRI-scans worden gebruikt bij het stellen van diagnose en in onderzoeksstudies. Zo kan bijvoorbeeld een groep patiënten vergeleken worden met een groep gezonde personen in een zogenaamd patiënt-controle-onderzoek of een representatief deel van de bevolking kan onderzocht worden. Dat laatste kan op twee manieren gedaan worden: In een dwarsdoorsnede-studie wordt er van ieder proefpersoon één MRI scan gemaakt en wordt er naar trends met bijvoorbeeld leeftijd of risicofactoren gekeken. In een longitudinale studie worden er meerdere scans gemaakt van de proefpersonen en wordt er gekeken naar veranderingen over de tijd.

### De Rotterdam Scan Study

De *Rotterdam Scan Study* is een grote populatiestudie waarbinnen zowel dwarsdoorsnede-studies als longitudinale studies gedaan worden. In dit onderzoek wordt er gekeken naar veroudering van de hersenen en bijbehorende degeneratieve aandoeningen. Voor dit onderzoek zijn duizenden middelbare en oudere personen uit een wijk in Rotterdam uitgenodigd om deel te nemen. Er wordt medisch-



**Figuur 7.2:** Van links naar rechts: drie verschillende soorten scans (T1-gewogen-scan, proton-density-gewogen scan, FLAIR-scan) en de gemiddelde diffusiegrootte en anisotropie van de diffusie. De twee rechterbeelden zijn gemaskeerd zodat alleen het brein zichtbaar is. Van boven naar onder: een persoon met weinig atrofie en witte-stof-laesies, een persoon met veel witte-stof-laesies en een persoon met veel atrofie.

relevante informatie over ze verzameld en testen bij ze afgenomen, waaronder hersen-MRI-scans en cognitieve testen. Vervolgens wordt er onder andere onderzocht hoe het volume aan de witte-stof-laesies verandert met leeftijd of hoe het hersenvolume samenhangt met cognitieve functie.

## Analyse van hersen-MRI-scans

Om variabelen als hersenvolume te kunnen meten, moeten de hersen-MRI-scans geanalyseerd worden. Een veelvoorkomende analyse-techniek is *segmentatie*. Hierbij wordt voor ieder beeldelement aangegeven of het bij het te segmenteren onderdeel, bijvoorbeeld een weefseltype, hersenstructuur of bloedvat, hoort. Het re-

sultaat kan weergegeven worden als een plaatje waarbij ieder beeldelement een label heeft gekregen. Segmentaties worden gebruikt voor onder andere volumemetingen, vormanalyses en visualisatie. Naast segmentatie zijn nog veel andere analyse-technieken mogelijk. Het doel hiervan is om eigenschappen van bijvoorbeeld een weefsel of hersenstructuur te kwantificeren. De resulterende getallen kunnen dan gebruikt worden in een vervolganalyse waar bijvoorbeeld een verband gezocht wordt met leeftijd of een vergelijking gedaan wordt tussen patiënten en gezonde personen.

Analyse van hersen-MRI-scans kan door experts gedaan worden, maar dit wordt bemoeilijkt door een aantal zaken. Zo is het lastig om informatie van verschillende MRI scans te combineren of om de hoog-dimensionale diffusie-MRI data te interpreteren. Ook is het natuurlijk erg arbeidsintensief om grote hoeveelheden scans te moeten analyseren. Maar belangrijker nog is dat verschillende experts dezelfde scans verschillend kunnen analyseren en dat zelfs dezelfde expert op een ander moment een ander analyseresultaat kan geven. Studies waarbij hersen-MRI-scans worden verkregen zijn daarom gebaat bij het gebruik van automatische methodes voor het analyseren van de MRI-scans. Methodes die hersen-MRI-scans automatisch analyseren moeten nauwkeurige en reproduceerbare resultaten leveren.

De nauwkeurigheid van een methode is het vermogen om resultaten te produceren die gelijk zijn aan de referentiestandaard. In medische beeldverwerking is het moeilijk om een referentiestandaard te verkrijgen omdat het vaak onduidelijk is wat de echte onderliggende anatomie is. Resultaten van de analyse door een expert worden daarom vaak als referentiestandaard gebruikt. De overeenkomst tussen de resultaten van een automatische methode en de analyse van een expert wordt dan vergeleken met de overeenkomst tussen de resultaten van de analyses van twee of meer experts.

Automatische analyse-technieken geven over het algemeen dezelfde of vergelijkbare resultaten als dezelfde MRI-scan als input wordt gegeven. De reproduceerbaarheid van een analyse-techniek is echter ook het vermogen voor het verkrijgen van vergelijkbare resultaten met een tweede scan, gemaakt binnen een kort tijdsinterval, van dezelfde persoon. Deze reproduceerbaarheid geeft niet alleen de reproduceerbaarheid van de analyse weer, maar ook die van onder andere de MRI-scanner. Omdat dit ook een rol speelt in longitudinale studies, moet hier rekening mee gehouden worden. De totale reproduceerbaarheid beïnvloedt het aantal proefpersonen dat nodig is voor de statistische analyse in een longitudinale studie. Hoe beter de reproduceerbaarheid is, hoe minder proefpersonen er nodig zijn.

In dit proefschrift wordt er naar twee automatische technieken voor analyse van hersen-MRI-scans gekeken, namelijk segmentatie van hersenweefsels en witte-stoflaesies en analyse van structurele hersenconnectiviteit. We evalueren de nauwkeurigheid en reproduceerbaarheid van deze technieken met behulp van MRI-scans van de Rotterdam Scan Study.

## Segmentatie van hersenweefsels en witte-stof-laesies

In Hoofdstuk 2 introduceren we een methode voor automatische segmentatie van witte-stof-laesies op basis van een hersenweefselsegmentatie en een FLAIR-scan (derde kolom van Figuur 7.2). Deze methode segmenteert alle lichte regio's in de FLAIR-scan, die grenzen aan witte stof, als witte-stof-laesies. Deze segmentatiemethode voor witte-stof-laesies kan gebruikt worden als extensie van een hersenweefselsegmentatiemethode. We optimaliseren een bestaande hersenweefselsegmentatiemethode in combinatie met deze extensie. Deze hersenweefselsegmentatiemethode wordt automatisch getraind op de scans die gesegmenteerd moeten worden. Figuur 7.3 laat een segmentatie zien die automatisch verkregen is met deze combinatie van methodes. We hebben de hersenweefselsegmentatiemethode en de extensie voor witte-stof-laesie-segmentatie geëvalueerd op nauwkeurigheid door te vergelijken met twee expertsegmentaties. De overeenkomst tussen de automatische segmentatie en de expertsegmentatie blijkt dichtbij de overeenkomst tussen twee expertsegmentaties te zitten.



**Figuur 7.3:** Automatische segmentatie van de scans van de persoon met veel witte-stof-laesies in Figuur 7.2. Van donker naar wit vertegenwoordigen de segmentatielabels hersenvocht, grijze stof, witte stof en witte-stof-laesies.

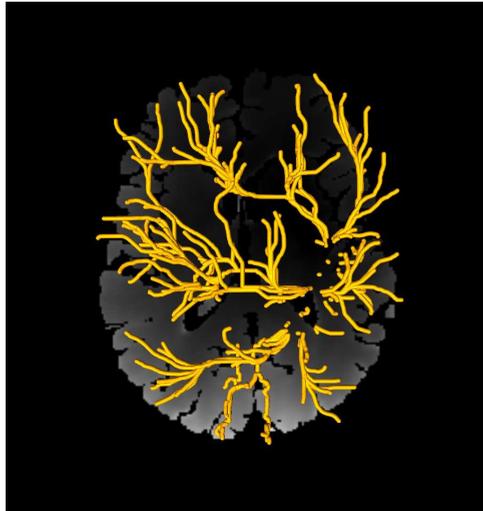
Er zijn in de literatuur meerdere hersenweefselsegmentatiemethodes geïntroduceerd en sommigen zijn publiekelijk beschikbaar. Er zijn echter maar weinig studies die de nauwkeurigheid en reproduceerbaarheid van deze methodes vergelijken op dezelfde MRI-scans. MRI-scans kunnen erg verschillen afhankelijk van de gebruikte scanner en scanmethode en zijn dus van invloed op de analyseresultaten. We vergelijken daarom de nauwkeurigheid en reproduceerbaarheid van vier verschillende hersenweefselsegmentatiemethodes in Hoofdstuk 3. Twee van deze methodes zijn publiekelijk beschikbaar, de derde is de automatisch getrainde segmentatiemethode uit Hoofdstuk 2 en de vierde methode lijkt op deze methode maar gebruikt een andere manier om te trainen. Alle methodes hebben een goede nauwkeurigheid en reproduceerbaarheid met alleen kleine verschillen tussen de methodes. De meest nauwkeurige methode had de laagste reproduceerbaarheid. Terwijl de verschillen in reproduceerbaarheid tussen de methodes klein zijn, heeft dit een relatief grote invloed op het vereiste aantal proefpersonen voor de statistische analyse van een longitudinale studie.

In Hoofdstuk 4 vergelijken we de prestaties van twee methodes uit Hoofdstuk 3 op longitudinale data. Beide methodes zijn uitgebreid met de extensie voor witte-stof-laesie-segmentatie uit Hoofdstuk 2. We gebruiken voor de vergelijking MRI-scans van proefpersonen van middelbare en oudere leeftijd die twee keer gescand zijn met een gemiddelde tijdsperiode van 3,5 jaar tussen de scans. De verwachting is dat er bij proefpersonen van deze leeftijd atrofie optreedt en het witte-stof-laesie-volume toeneemt in deze tijdsperiode. Met beide segmentatiemethodes worden de volumeveranderingen gemeten voor grijze stof, witte stof, witte-stof-laesies en de gehele hersenen (grijze- en witte stof samen). Ondanks de kleine verschillen in nauwkeurigheid en reproduceerbaarheid (Hoofdstuk 3), vinden de twee methodes statisch significant verschillende volumeveranderingen voor grijze stof, witte stof en witte-stof-laesies. Visualisaties van de locaties waar veranderingen in de weefsels optreden, laten zien dat dit ook kan verschillen tussen de twee methodes.

## Analyse van structurele hersenconnectiviteit

Met behulp van diffusie-MRI-scans kan bestudeerd worden hoe de zenuwbanen in de hersenen lopen en hoe grijze-stof-gebieden met elkaar verbonden zijn. Dit heet het bestuderen van de structurele connectiviteit van de hersenen. In Hoofdstuk 5 presenteren we een methode, genaamd *Statistical Analysis of Minimum cost path based Structural Connectivity (SAMSCo)*, voor het statistisch analyseren van deze structurele hersenconnectiviteit. Op basis van de diffusie-MRI-scans bepalen we de verbindingspaden tussen grijze-stof-gebieden. Figuur 7.4 laat als voorbeeld een aantal paden tussen grijze-stof-gebieden zien. Per proefpersoon vatten we de structurele connectiviteit samen in een zogenaamd connectiviteitsnetwerk. De punten in dit netwerk vertegenwoordigen de grijze-stof-gebieden. Een netwerkverbinding tussen twee punten krijgt een gewicht dat afhangt van de grootte en de anisotropie van de diffusie langs het bijbehorende pad. De connectiviteitsnetwerken zijn erg geschikt om te analyseren met statistische methodes. Zo kan er met de SAMSCo methode gekeken worden naar verschillen in structurele hersenconnectiviteit tussen patiënten en gezonde proefpersonen of kunnen de veranderingen in connectiviteit als gevolg van ontwikkeling of veroudering van de hersenen onderzocht worden in een dwarsdoorsnede- of longitudinale studie. In een demonstratie-experiment, bepalen we de leeftijd van de proefpersonen op basis van hun connectiviteitsnetwerken, waarmee we aantonen dat de connectiviteitsnetwerken leeftijdsinformatie bevatten. Daarnaast tonen we aan dat de connectiviteitsnetwerken ook informatie bevatten over witte-stof-atrofie en witte-stof-laesie-volume.

In Hoofdstuk 6 evalueren we de reproduceerbaarheid van de SAMSCo methode. Hiervoor gebruiken we MRI-scans van proefpersonen die twee keer kort achter elkaar gescand zijn. Voor beide scans berekenen we de connectiviteitsnetwerken. Vervolgens vergelijken we connectiviteitsnetwerken van dezelfde persoon en van verschillende personen onderling. Hieruit blijkt dat het verschil tussen con-



**Figuur 7.4:** *Bovenaanzicht van de (3-dimensionale) verbindingspaden van de hersenstructuur 'putamen' naar de andere grijze-stof-gebieden.*

nectiviteitsnetwerken van dezelfde persoon kleiner is dan het verschil tussen connectiviteitsnetwerken van verschillende personen. Daarnaast onderzoeken we de reproduceerbaarheid van de leeftijdsbepaling uit Hoofstuk 5.



## Dankwoord

The research presented in this PhD thesis is the result of many collaborations. I want to thank all of you that contributed directly or indirectly to this research. I learned a lot and it was a very valuable time for me. Some of you I would like to thank in particular.

### *Promotoren en co-promotor*

Wiro, ontzettend bedankt voor alle adviezen, kansen en vrijheid die je me gaf. Door jouw optimistische en enthousiaste instelling worden onze methodes nu steeds meer toegepast en dat is natuurlijk geweldig. Je weet uit je eigen promotietijd nog goed dat promoveren meer is dan alleen hard werken; je stimuleert de conferentie-bezoeken, borrels en andere leuke extra's en dat heb ik erg gewaardeerd.

Monique, ik heb veel geleerd van de samenwerking met jou en jouw groep. Het was fantastisch om met de data van de door jou opgezette Rotterdam Scan Study te kunnen werken en ik vind het erg leuk dat ik een bijdrage heb kunnen leveren aan de analyse van de data.

Henri, je stond altijd klaar om te helpen met het verwerken van de data. Je hebt de segmentatiemethode, die beschreven wordt in dit proefschrift, verwerkt in een pipeline die nu simpel en (bijna) zonder problemen op alle data kan worden toegepast. Bedankt voor al je hulp en de gezellige film(festival)gesprekken.

### *Promotiecommissie*

I would like to thank all members of my PhD committee: professor van der Lugt, professor van Buchem, dr. Ourselin, professor Rombouts, professor Koudstaal and professor Florack. Aad, bedankt voor de goede adviezen voor met name Hoofdstuk 3. Dr. Ourselin, thank you for traveling to Rotterdam for my defense. Luc, in Eindhoven was je mijn afstudeerbegeleider en ik vind het dan ook heel leuk dat je nu in mijn promotiecommissie plaats wilt nemen.

### *Colleagues*

I am very grateful that I have such wonderful colleagues. When I moved to Rotterdam I hardly knew anyone living here, so it was great to be able to find friends at BIGR and the departments of Epidemiology, Radiology and Medical Informatics. Thank you for all the nice 'borrels', BBQs, trips, and movie nights; I really enjoyed them.

BIGR is in z'n geheel een geweldige groep, maar een aantal BIGRs heb ik gedurende conferenties en cursussen wat beter leren kennen. We begonnen ongeveer tegelijkertijd met onze promotie-onderzoeken en zaten dus vaak in dezelfde situaties. Fedde, Coert, Michiel, Reinhard en korte tijd later ook Marius: ik vond het erg waardevol om ongeveer gelijk met jullie dit traject te doorlopen. Meike, weliswaar geen BIGR, maar wat mij betreft hoorde je er helemaal bij.

Alhoewel ik altijd op dezelfde kamer heb gezeten, heeft deze kamer veel verschillende 'bewoners' gekend. Ik ben jullie allemaal erg dankbaar voor alle keren dat ik jullie als klankbord kon gebruiken of kon bestoken met vragen. En natuurlijk voor de gezelligheid!

Het was geweldig om samen te werken met andere afdelingen en zo te zien hoe de door ons ontwikkelde methodes ook daadwerkelijk gebruikt werden. De afdeling epidemiologie, en dan met name het 'imaging meeting'-deel, was wel de belangrijkste partner. Allen ontzettend bedankt voor de leuke en leerzame samenwerking.

### *Vrienden*

Etentjes, bioscoopbezoeken, vakanties, festivals, feestjes en andere activiteiten waren welkome afleidingen tijdens het promoveren. Ik vond het altijd erg gezellig! Maar ik heb ook veel aan jullie steun gehad bij tegenvallers en moeilijke periodes. Ik wil bij deze al mijn vrienden dan ook graag bedanken.

Marieke, we hebben samen veel beleefd tijdens onze studie en onze promoties. Van een rondreis door Duitsland, als beginnende snowboarders in Frankrijk, tot de RSNA in Chicago en natuurlijk je promotie. Ik vond het erg leuk dat ik je paranimf mocht zijn en ben blij dat je straks naast me staat tijdens mijn verdediging.

Michiel, al vanaf het begin ben je een zeer goede vriend. Ik kon altijd alles met je bespreken en ik heb veel aan je steun gehad. Bovendien was het leuk en erg leerzaam om met je samen te werken aan het laatste deel van mijn proefschrift. Ik ben blij dat je mijn paranimf wilt zijn.

### *Familie*

Gaandeweg kwam ik er achter dat het helemaal niet zo vanzelfsprekend is om een familie te hebben die (enigszins) weet wat promoveren inhoudt. Ik ben jullie erg

---

dankbaar voor jullie steun, medeleven en interesse. Dat laatste zal niet altijd makkelijk zijn geweest met zo'n technisch en onbekend promotie-onderwerp. Ik vind alle familiebijeenkomsten altijd reuzegezellig en ben blij met zo'n geweldig leuke familie.

Geeke en Bas, Auke, Jaap en Yvonne, jullie waren altijd erg begripvol als ik het druk had. Ook jullie waren tijdens deze periode bezig met opleidingen en cursussen en ik heb veel bewondering voor jullie doorzettingsvermogen om te bereiken wat je wilt. Bedankt voor jullie steun, ook tijdens de moeilijke tijden.



## Publications

### Journal Papers

- **R. de Boer**, H. Vrooman, F. van der Lijn, M.W. Vernooij, M.A. Ikram, A. van der Lugt, M.M.B. Breteler and W.J. Niessen, White matter lesion extension to automatic brain tissue segmentation on MRI, *NeuroImage*, 2009.
- **R. de Boer**, H. Vrooman, M.A. Ikram, M.W. Vernooij, M.M.B. Breteler, A. van der Lugt and W.J. Niessen, Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods, *NeuroImage*, 2010.
- M. Schuur, J.C. van Swieten, S. Schol-Gelok, M.A. Ikram, M.W. Vernooij, F. Liu, A. Isaacs, **R. de Boer**, I. de Koning, W.J. Niessen, H. Vrooman, B.A. Oostra, A. van der Lugt, M.M.B. Breteler and C.M. van Duijn, Genetic risk factors for cerebral small vessel disease in hypertensive patients from a genetically isolated population, *Journal of Neurology, Neurosurgery and Psychiatry*, 2011.
- **R. de Boer**, M. Schaap, F. van der Lijn, H. Vrooman, M. de Groot, A. van der Lugt, M.A. Ikram, M.W. Vernooij, M.M.B. Breteler and W.J. Niessen, Statistical analysis of minimum cost path based structural brain connectivity, *NeuroImage*, 2011.
- M. Fornage\*, S. Debette\*, J.C. Bis\*, H. Schmidt\*, M.A. Ikram\*, C. Dufouil\*, S. Sigurdsson\*, T. Lumley, A.L. DeStefano, F. Fazekas, H.A. Vrooman, D.K. Shibata, P. Maillard, A. Zijdenbos, A.V. Smith, H. Gudnason, **R. de Boer**, M. Cushman, B. Mazoyer, G. Heiss, M.W. Vernooij, C. Enzinger, N.L. Glazer, A. Beiser, D.S. Knopman, M. Cavalieri, W.J. Niessen, T.B. Harris, K. Petrovic, O.L. Lopez, R. Au, J.C. Lambert, A. Hofman, R.F. Gottesman, M. Garcia, S.R. Heckbert, L.D. Atwood, D.J. Catellier, A.G. Uitterlinden, Q. Yang, N.L. Smith, T. Aspelund, J.R. Romero, K. Rice, K.D. Taylor, M.A. Nalls, J.I. Rotter, R. Sharret, C.M. van Duijn, P. Amouyel, P.A. Wolf, V. Gudnason, A. van der Lugt, E. Boerwinkle, B.M. Psaty, S. Seshadri, C. Tzourio\*, M.M.B. Breteler\*, T.H. Mosley\*, R. Schmidt\*, W.T. Longstreth\*, C. DeCarli\*, L.J. Launer\*, Genome-wide association studies of cerebral white matter lesion burden: the CHARGE Consortium, *Annals of Neurology*, in press.

---

\*These authors contributed equally to this work.

- B.F.J. Verhaaren, **R. de Boer**, M.W. Vernooij, F. Rivadeneira, A.G. Uitterlinden, A. Hofman, G.P. Krestin, A. van der Lugt, W.J. Niessen, M.M.B. Breteler, M.A. Ikram, Replication study of chr17q25 with cerebral white matter lesion volume, submitted.

## Conference Papers

- **R. de Boer**, F. van der Lijn, H. Vrooman, M.W. Vernooij, M.A. Ikram, M.M.B. Breteler and W.J. Niessen, Automatic segmentation of brain tissue and white matter lesions in MRI, *IEEE International Symposium on Biomedical Imaging: Macro to Nano*, 2007.
- M. de Groot, M.W. Vernooij, S. Klein, A. Leemans, **R. de Boer**, A. van der Lugt, M.M.B. Breteler and W.J. Niessen, Iterative Co-Linearity Filtering and Parameterization of Fiber Tracts in the Entire Cingulum, *Medical Image Computing and Computer-Assisted Intervention*, 2009.
- **R. de Boer**, M. Schaap, F. van der Lijn, H. Vrooman, M. de Groot, M.W. Vernooij, M.A. Ikram, E.F.S. van Velsen, A. van der Lugt, M.M.B. Breteler and W.J. Niessen, Statistical Analysis of Structural Brain Connectivity, *Medical Image Computing and Computer-Assisted Intervention*, 2010.

## Conference Abstracts

- **R. de Boer**, H. Vrooman, F. van der Lijn, M.W. Vernooij, M.A. Ikram, M.M.B. Breteler and W.J. Niessen, Automatic multiple-atlas-based segmentation of brain tissue and white matter lesions in MRI, *Scientific Assembly and Annual Meeting of the Radiological Society of North America*, 2007.
- **R. de Boer**, H. Vrooman, M.A. Ikram, M.W. Vernooij, M.M.B. Breteler, A. van der Lugt and W.J. Niessen, Reproducibility of automatic MRI brain tissue segmentation, *Annual Meeting of the Organization for Human Brain Mapping*, 2009.
- M. de Groot, M.W. Vernooij, S. Klein, A. Leemans, **R. de Boer**, A. van der Lugt, M.M.B. Breteler and W.J. Niessen, Asymmetry analysis along the entire cingulum in the general population, *Annual Meeting of the Organization for Human Brain Mapping*, 2009.

## PhD portfolio

<b>Courses</b>	<b>Year</b>
MeVisLab course (MeVis)	2006
Introduction to data-analysis (Nihes)	2006
Advanced morphological segmentaton (ASCI)	2006
Advances in clinical neuroepidemiology (Nihes)	2007
FSL & FreeSurfer course (Cardiff University)	2007
Regression analysis (Nihes)	2008
Knowledge-driven image segmentation (ASCI)	2008
Advanced pattern recognition (ASCI)	2008
R course (Erasmus MC)	2008
 <b>Attending international conferences</b>	
IEEE International Symposium on Biomedical Imaging (ISBI)	2007
Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA)	2007
European Congress of Radiology (ECR)	2008
Annual Meeting of the Organization for Human Brain Mapping (HBM)	2009
IEEE International Symposium on Biomedical Imaging (ISBI)	2010
International conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)	2010
 <b>Presentations at international conferences</b>	
Oral presentation at ISBI	2007
Poster presentation at RSNA	2007
Poster presentation at HBM	2009
Poster presentation at MICCAI	2010

**Attending international workshops**

Feature extraction and classification (ISBI)	2007
Structural and diffusion MRI (HBM)	2009
International Workshop on Machine Learning in Medical Imaging (MICCAI)	2010

**Other presentations**

Research seminars of the Biomedical Imaging Group Rotterdam and the department of Medical Informatics	2006 - 2010
Presentation at the Medical Informatics PhD days	2008
Presentation at the Dutch conference on Bio-Medical Engineering	2009

**Teaching**

Evaluation of scientific presentations by first year medical students	2006 - 2008
Introduction to image analysis for medical students	2007 - 2010
Supervisor MSc project Eugene Leung	2009 - 2010

**Other**

Demonstrations at the Medical Delta booth at ECR	2008
NWO Talent Class 'Creatief denken'	2008
De Jonge Akademie on Wheels	2008
Reviewing for NeuroImage, IEEE Transactions on Medical Imaging, Computers in Biology and Medicine, MICCAI, and the International Conference on Image Processing	2009 - 2011

## About the author

Renske de Boer was born on October 7, 1981 in Enschede, the Netherlands. She finished her VWO education in 1999 after which she traveled around Australia for almost a year. Upon her return, she studied Biomedical Engineering at the Eindhoven University of Technology. In 2004, she did an internship on three-dimensional statistical shape analysis of teeth at the 3D-Lab in Copenhagen, Denmark. Her MSc graduation project was on the closure of gaps within elongated structures in medical images. She obtained her MSc degree in 2006. That same year she started as a PhD student at the Biomedical Imaging Group Rotterdam at the Erasmus MC. The result of that work is described in this thesis. She will continue this research at the Biomedical Imaging Group Rotterdam as a postdoctoral researcher.

