Development and Validation of Automatic Tools for Segmentation of White Matter Hyperintensities

Mahsa Dadar

Faculty of Medicine
Biomedical Engineering Department
McGill University, Montreal

Match 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

ABSTRACT

Automatic methods for segmentation of various tissues and pathologies are critical for systematic studies of the brain to investigate changes that occur in different components of the phenomenon under study. White matter hyperintensities (WMHs) are one of the major components of small-vessel disease in aging and Alzheimer's disease (AD) populations that need to be assessed and monitored to estimate the vascular disease burden. In this thesis, a new fully automatic technique is proposed for segmenting WMHs from multiple contrasts of magnetic resonance (MR) brain images.

The proposed segmentation technique uses a machine learning classification scheme by combining a set of intensity and location features obtained from multi-contrast MR sequences, namely T1w, T2w, proton density (PD) and fluid attenuated inversion recovery (FLAIR) images and a linear or nonlinear classifier to detect WMHs. The segmentations are performed in the native space of the optimal contrast (e.g. FLAIR or T2w) to avoid the blurring caused by resampling the images, especially since these images generally have relatively thick slices (3-5 mm). All other contrasts are linearly registered to the optimal contrast using a 6-parameter rigid registration. The feature set includes the voxel intensity value from each available contrast, spatial probability of a voxel at a specific location being a WMH, average intensity of healthy tissue at the voxel location, histogram distribution of the healthy tissue and WMH intensities, and the ratio of the two histograms. The spatial probability, average intensity and the histogram features are calculated from a training set with manually segmented WMHs. To obtain these features, the training images are nonlinearly registered to an average template, each feature is calculated in the template space, and the results are transformed back into the native space of the optimal image. The classifiers are then trained on the training dataset with manually segmented labels. The

performance of the classifiers is assessed using Dice Kappa values as the primary outcome measure and through a 10-fold cross validation scheme.

Using the developed tool, the WMHs were segmented using different combinations of input image contrasts (i.e. T1w+T2w+PD, T1w+FLAIR, T1w) to assess the performance of the classifiers and the contribution of each of the contrasts in detecting WMHs. The question of interest was whether the WMHs loads obtained from segmentations based only on T1w images can be used as accurate estimates of the actual WMH loads. To assess this, the volumetric correlation of WMH loads in different brain lobes as well as correlation with age and cognitive measures were compared to investigate the effectiveness of each contrast in providing WMH load estimates that are highly correlated with aging and cognitive scores. The assessments revealed that the best Dice Kappa values are obtained while using the optimal FLAIR and T2w/PD contrasts. Classifications based solely on T1w images tend to undersegment the WMHs, only detecting the brightest of these lesions on FLAIR and T2w/PD images. However, the WMH loads obtained from T1w segmentations were still able to provide high correlations with age and cognitive scores.

Finally, using the developed tool, baseline WMHs were segmented in an early stage Parkinson's disease (PD) database as well as age matched healthy controls. Using longitudinal clinical assessments and cortical thickness measures, we studied the relationship between baseline WMHs and future cognitive decline and cortical thinning. PD subjects with high WMH loads were found to present with more future cognitive decline and cortical thinning in comparison with (i) PD subjects with low WMH loads and (ii) age matched control subjects with high WMH loads. These findings show that the existence of WMHs affects PD patients differently from controls.

RESUME

Les méthodes de segmentation automatique de différents tissus et pathologies sont primordiales pour l'étude systématique du cerveau afin d'examiner les changements qui se produisent dans différentes composantes du phénomène étudié. Les hyperintensités de la matière blanche (HMB) sont un des éléments caractéristiques de la maladie des petits vaisseaux qui doit être évalué chez les populations vieillissantes ou souffrant de la maladie d'Alzheimer (MA), afin d'estimer les impacts de la maladie vasculaire.

La technique de segmentation proposée utilise un procédé de classification par apprentissage automatique, en associant un ensemble de caractéristiques d'intensité et d'emplacement, obtenues à partir de séquences de RM multi-contrastes, à savoir des images T1w, T2w, de densité de protons (DP) et d'inversion-récupération atténuée par un fluide (FLAIR) ainsi qu'un modèle de classification linéaire ou non-linéaire pour détecter les HMB. Les segmentations sont effectuées dans l'espace natif du contraste optimal (FLAIR ou T2w) afin d'éviter le floutage dû au rééchantillonnage des images, en particulier parce que ces images sont en général composées de tranches épaisses (3-5 mm). Tous les autres contrastes sont recalés linéairement avec le contraste optimal en utilisant un recalage rigide à 6 paramètres. Les caractéristiques étudiées incluent la valeur d'intensité des voxels pour chaque contraste disponible, la probabilité qu'a un voxel à un emplacement spécifique de faire partie d'une zone d'HMB, l'intensité moyenne d'un tissu sain à l'emplacement du voxel, les histogrammes de distribution des intensités du tissu sain et des HMB et le ratio des deux histogrammes. La probabilité spatiale, l'intensité moyenne et les caractéristiques des histogrammes sont calculées à partir d'un jeu de données d'entrainement pour lequel les HMB ont été segmentées manuellement. Afin d'obtenir ces caractéristiques, les images d'entrainement sont recalées de façon non-linéaire sur un modèle moyen, chaque

caractéristique est calculée dans l'espace du modèle, puis les résultats sont retransformés dans l'espace natif de l'image optimale. Les modèles de classification sont ensuite entrainés sur le jeu de données dont les étiquettes ont été segmentées manuellement. La performance des modèles de classification est évaluée en utilisant des valeurs de coefficient de Dice comme mesure primaire de sortie et à travers un procédé de validation croisée en 10 étapes.

Grâce à l'outil développé, les HMB ont été segmentées en utilisant différentes combinaisons de contraste comme images d'entrée (par exemple T1w+T2w+PD, T1w+FLAIR, T1w) afin d'évaluer la performance des modèles de classification et la contribution de chacun des contrastes pour la détection des HMB. La question soulevée était de déterminer si la charge de HMB obtenue par une segmentation basée seulement sur des images T1w pouvait être utilisées comme estimé précis de la charge réelle d'HMB. Afin de déterminer la réponse à cette question, la corrélation volumétrique de charge d'HMB dans différents lobes du cerveau ainsi que la corrélation avec l'âge et les mesures cognitives ont été comparées pour examiner l'efficacité de chaque contraste à fournir un estimé de la charge d'HMB fortement corrélé avec l'âge et la performance cognitive. Cette évaluation a révélé que les meilleures valeurs de coefficient de Dice sont obtenues lors de l'utilisation des contrastes FLAIR et T2w/PD optimaux. Les classifications basées uniquement sur les images T1w ont tendance à sous-segmenter les HMB, détectant ainsi uniquement les lésions les plus intenses sur les images FLAIR et T2w/PD. Cependant, les charges de HMB obtenues à partir de segmentations sur les images T1w présentaient tout de même une forte corrélation avec l'âge et la performance cognitive.

Enfin, en utilisant l'outil développé, les HMB de base ont été segmentées pour un jeu de données collecté pour des patients à un stade peu avancé de la maladie de Parkinson (MP) ainsi que chez des sujets en bonne santé et d'âge équivalent. En utilisant des évaluations

cliniques longitudinales et des mesures de l'épaisseur du cortex, nous avons étudié le lien entre les HMB de base et le déclin cognitif à venir et l'amincissement cortical. Les patients atteints de la MP avec des charges de HMB élevées ont présenté des déclins cognitifs et un amincissement cortical plus importants en comparaison (i) des patients atteints de la MP avec des charges d'HMB basses et (ii) des sujets d'âge comparable possédant des charges élevées d'HMB. Ces résultats montrent que l'existence d'HMB affecte les patients atteints de la MP différemment qu'elle ne le fait chez les individus normaux.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my supervisor, Dr. D. Louis Collins, for his constant guidance and encouragement, without which this work would never have been possible. I am truly grateful for your unwavering support, dedication, and enthusiasm.

I am grateful to my committee members, Dr. Pedro Rosa Neto and Dr. Tal Arbel for their invaluable time, consultation, and assistance throughout my research.

I would like to express my thanks to my co-authors and collaborators, especially Dr. Josefina Maranzano for providing manual segmentations and clinical and anatomical knowledge that substantially helped my research, Dr. Vladimir S. Fonov for his technical guidance, and Drs. Owen Carmichael and Charles Decarli for providing invaluable data that helped start my project.

I would also like to thank my friends and colleagues, especially Marie Franquin, Yvonne Yau, Simon Drouin, and Azar Zandifar.

Last but not least, I would like to thank my family, especially Yashar Zeighami, for his constructive advice and continuous support.

TABLE OF CONTENTS

ABSTRACT 3
RESUME 5
ACKNOWLEDGEMENTS 8
LIST OF FIGURES 14
LIST OF TABLES 18
CHAPTER 1. Introduction 20
1.1. Motivation 20
1.2. Overview of thesis 20
1.3. Author contributions 21
1.4. Scientific contributions 23
CHAPTER 2. Review of White Matter Hyperintensities, Their Effect on Cognition, and Automated WMH Segmentation from MRI 25
Introduction 25
2.1. White Matter Hyperintensities and Their Effect on Cognitive Function 26
2.1.1. Cerebral Small Vessel Disease 26
2.1.2. White Matter Hyperintensities 27
2.1.3. WMH Locations 28
2.1.4. WMH Rating Scales 30
2.1.5. WMH Risk Factors 33
2.1.6. Impact of WMHs on Cognition 34
WMHs in cognitively normal populations 35
WMHs in individuals with mild cognitive impairment 37
WMHs in Alzheimer's disease patients 41
2.1.7. Prevention and Treatment 41
2.1.8. Conclusion 42
2.2. White Matter Hyperintensity Segmentation 42
2.2.1. Challenges in detecting WMHs 46
2.2.2. Image Pre-processing 51
2.2.3. Unsupervised WMH Segmentation Techniques 53
2.2.4. Supervised WMH Segmentation Techniques 56
2.2.5. Post-processing 58
2.2.6. Validation 59
2.2.7. Conclusion 60
2.3. WMHs as Biomarkers in Alzheimer's Disease 61
2.3.1. Biomarkers 61
2.3.2. Models of Disease Progression 62
2.3.3. WMHs as Biomarkers in Diagnosis and Prognosis 66

2.3.4. Conclusion 69
CHAPTER 3. Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease 71
Preface 71
Abstract 72
3.1. Introduction 73
3.2. Materials and methods 78
3.2.1. Subjects 78
3.2.2. MR imaging 80
3.2.3. Pre-processing 83
3.2.4. Features 83
3.2.5. Tissue Classification 86
3.2.6. Evaluation metrics 88
3.3. Experiments and Results 88
3.3.1. Qualitative results 88
3.3.2. Quantitative results 91
3.3.3. Contribution of the features 94
3.3.4. Comparison between classifiers 94
3.3.5. Impact of Size of the Training Set 94
3.4. Discussion 95
Chapter 4. Performance Comparison of 10 Different Classification Techniques in Segmenting White Matter Hyperintensities in Aging 103
Preface 103
Abstract 104
4.1. Introduction 106
4.2. Materials and methods 108
4.2.1. Subjects 108
4.2.2. MR imaging 110
4.2.3. Manual segmentation 111
4.2.4. Pre-processing 113
4.2.5. Features 113
4.2.6. Classification Methods 116
Naive Bayes 116
Discriminant Analysis 116
Logistic Regression 117
Decision Trees 118
Random forests 118
K-nearest neighbors 118
Support Vector Machines 119
Bagging 119

AdaBoost 119
4.2.7. Evaluation metrics 120
4.3. Results 122
4.3.1. Segmentation using T1w, T2w, PD, and FLAIR 122
4.3.2. Segmentation using T1w and FLAIR data 124
4.3.3. Segmentation using T1w, T2w, and PD data 127
4.3.4. Segmentation using T1w, and T2w data 130
4.3.5. Segmentation using only T1w data 132
4.3.6. Over-segmentation/Under-segmentation 135
4.3.7. Computational burden 136
4.4. Discussion 137
Chapter 5. Validation of T1w-based Segmentations of White Matter Hyperintensity Volumes in Large Scale Datasets of Aging 145
Preface 145
5.1. Introduction 147
5.2. Materials and Methods 150
5.2.1. Subjects 150
5.2.2. Clinical evaluations 152
5.2.3. MR imaging 153
5.2.4. Pre-processing 155
5.2.5. Manual segmentation 156
5.2.6. Automatic segmentation 156
5.3. Results 159
5.3.1. Comparison of Tissue Histograms 159
5.3.2. Comparisons with Manual Segmentations 160
5.3.3. Large Scale Correlations 161
5.3.4. WMHs and clinical measures 164
5.3.5. WMHs and other measures 168
5.4. Discussion 169
Chapter 6. White Matter Hyperintensities Predict Cognitive Decline in de Novo Parkinson's Disease Patients 174
Preface 174
6.1. Introduction 177
6.2. Methods 179
6.2.1. Patients 179
6.2.2. Procedures 179
6.2.3. Statistical Analysis 181
6.3. Results 183
6.3.1. Baseline WMH Load as a Predictor of Longitudinal Cognition 183
6.4. Discussion 189

Chapter 7. Conclusions and Future Work 195
7.1. Discussions 195
7.1.1. Feature Selection and Linear Classifier 195
7.1.2. Nonlinear Classifiers and Automated Pipeline 196
7.1.3. Detecting WMHs from T1w Images 198
7.1.4. WMHs in Parkinson's Disease 199
7.2. Future work 200
7.2.1. WMH Segmentation 200
7.2.2. Studying WMHs in Aging and Diseases 202
7.3. Conclusions 203
Appendix 205
Validation of the segmentations 205
Cognitive Testing 206
Baseline Information 207
Mixed-effects modeling 207
Medication Information 208
REFERENCES 210

LIST OF FIGURES

Fig. 2.1. Examples of periventricular (caps and rims) and deep (punctate and confluent) WMHs. 29
Fig 2.2. The relationship between WMHs and risk of MCI and rate of decline in global cognition. 37
Fig. 2.3. Results summary of studies investigating the relationship between MRI biomarkers and the risk of progression from MCI to AD. 39
Fig. 2.4. Association between cortical thickness or WMH probability and episodic memory in subjects with MCI. 40
Fig. 2.5. WMHs on T1w, T2w, PD, and FLAIR MR images. 45
Fig. 2.6. Tissue intensity histograms. 46
Fig. 2.7. Varying intensities and tissue contrasts in different sequences. 46
Fig. 2.8 MR image intensity noise. 47
Fig. 2.9 MR image inhomogeneity. 48
Fig. 2.10. Ringing artifact in MR images caused by subject movement. 48
Fig. 2.11. Partial volume effect. 48
Fig. 2.12. Atrophy and change in GM/WM tissue contrast in MR image. 49
Fig. 2.13. Hyperintense appearance of non-WMH tissues such as choroid plexus. 49
Fig. 2.14. Fuzzy borders of lesions. 50
Fig. 2.15. Inconsistency in lesion borders between different contrast. 51
Fig. 2.16. FLAIR image histogram. 54
Fig. 2.17. Biomarkers of AD pathological cascade 63
Fig. 2.18. Regional anatomical imaging biomarkers of AD pathological cascade 64
Fig. 2.19. Revised model of biomarkers of AD pathological cascade. 65
Fig. 2.20. Data-driven model of LOAD progression 66
Fig. 2.21. Longitudinal changes in differences of enduring WMSA MD. 69
Fig. 3.1. Histograms of WMH load ranges for the 3 datasets 82
Fig. 3.2. Flow-chart of the proposed classifier and the preprocessing steps. 86
Fig. 3.3. Comparison of the automated vs. manually segmented WMHs for a subject from ADC dataset. 89
Fig. 3.4. Comparison of the automated vs. manually segmented WMHs for a subject from PREVENT-AD dataset. 90
Fig. 3.5. Comparison of the automated vs. manually segmented WMHs for a subject from ADNI2/GO dataset. 91
Fig. 3.6. SI (Dice Kappa) vs threshold for A) ADC B) PREVENT-AD C) ADNI2/GO datasets. 92
Fig. 3.7. Boxplot diagrams of SI (Dice Kappa) (<5 CCs, 5-20 CCs, and >20 CCs) for A) ADC B) PREVENT-AD C) ADNI2/GO datasets. 93

Fig. 3.8. SI (Dice Kappa) vs manually segmented WMH loads (CCs) for A) ADC B) PREVENT-AD C) ADNI2/GO datasets. 93
Fig. 3.9. Impact of the number of training subjects on SI (Dice Kappa) and intra-class correlation (ICC). 95
Fig. 4.1. Axial slices comparing manual segmentations and T1w, T2w, PD, and FLAIR information for subjects from ADC, NACC, ADNI1, and ADNI2 datasets. 112
Fig. 4.2. Flow-chart of the preprocessing, registration, and feature selections steps. 115
Fig. 4.3. Dice Kappa (SI) for different classification methods for (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w, T2w, PD, and FLAIR information for the ADC dataset. 115
Fig. 4.4. Axial slices comparing manual and automatic segmentations using T1w, T2w, PD, and FLAIR information for a subject from ADC dataset. 123
Fig. 4.5. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. 124
Fig. 4.6. Dice Kappa (SI) for different classification methods for low (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w and FLAIR information for ADC (red), NACC (black), and ADNI2/GO (magenta) datasets. 126
Fig. 4.7. Axial slices comparing manual and automatic segmentations using T1w and FLAIR information in one subject from each of ADC, NACC, and ADNI2/GO datasets. 126
Fig. 4.8. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. 127
Fig. 4.9. Dice Kappa (SI) for different classification methods for low (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w, T2w, and PD information for ADC (red) and ADNI1 (blue) datasets. 129
Fig. 4.10. Axial slice comparing manual and automatic segmentations using T1w, T2w, and PD information for a subject from each of ADC and ADNI1 datasets. 129
Fig. 4.11. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. 130
Fig. 4.12. Dice Kappa (SI) for different classification methods for low (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w and T2w information for ADC (red) and ADNI1 (blue) datasets. 132
Fig. 4.13. Axial slice comparing manual and automatic segmentations using T1w and T2w information for ADC and ADNI1 datasets. 132
Fig. 4.14. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. 132

Fig. 4.15. Dice Kappa (SI) for different classification methods for low (<5 cm^3), medium (5-20 cm^3), and high (>20 cm^3) WMH load using only T1w information for ADC (red), NACC (black), ADNI1 (blue), and ADNI2/GO (magenta) datasets. 134
Fig. 4.16. Axial slice comparing manual and automatic segmentations using T1w information for ADC, NACC, ADNI1, and ADNI2/GO datasets 135
Fig. 4.17. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. 135
Fig. 5.1. Axial slices showing T1w, T2w, PD, and FLAIR images as well as manual (yellow), All-Contrasts (cyan), and T1w-Only (red) segmentations of subjects from ADC (top), ADNI1 (middle), and ADNI2/GO (bottom) datasets. 158
Fig. 5.2. Intensity histograms of white matter (WM), grey matter (GM), cerebrospinal fluid (CSF), and white matter hyperintensities (WMHs) for a. ADC and b. ADNI datasets. 159
Fig. 5.3. Total brain and per lobe correlations for automatic versus manually segmented WMH volumes (CCs), using all contrasts available (red) and using only T1w contrast (blue) for a. ADC, b. ADNI1, c. ADNI2/GO datasets. 161
Fig. 5.4. Total brain and per lobe correlations for automatically segmented WMHs volumes (CCs) based on all available contrast images versus only T1w contrast image for ADNI1 (left) and ADNI2/GO (right) datasets. 162
Fig. 5.5. Total brain and per lobe correlations for z-scored log transformed automatically segmented WMH volumes correlated with age (z-scored) based on all available contrasts (red) versus only T1w images (blue) for ADNI1 (left) and ADNI2/GO (right) datasets. 163
Fig 5.6. Dice Kappa and volumetric correlation between T1-Only and All-Contrasts segmentations and thresholded manual labels based on T1w image intensity. 164
Fig. 5.7. Total brain and per lobe correlations for log transformed automatically segmented WMH volumes (CCs) versus ADAS13 while controlling for age, based on all available contrasts (red) and based only on T1w images (blue) for ADNI1 (left) and ADNI2/GO (right) datasets. 167
Fig. 5.8. Boxplots of log transformed WMHLs for NC, MCI, and Dementia Cohorts. 168
Fig. 5.9. Comparing T1w hypointensities and FLAIR hyperintensities. 171
Fig. 6.1. Kaplan-Meier graph of survival showing the survival curves of control and PD patients with low versus high WMH loads demonstrating the compounded affect of PD and WMH load. 17186
Fig. 6.2. Density plots of longitudinal cognitive changes versus age and log transformed baseline WMH load.. 17187
Fig. 6.3. Differences in cortical thickness changes between high and low WMHL cohorts in PD subjects. 17189
Fig. 6.4. Correlation between WMH location and slope of MoCA (top) and Benton (bottom) score in the PD cohort, controlled for age and modality. . 17190
Fig. A.1. Axial slices of a subject with T1w, T2w, and FLAIR scans along with the automatic WMH segmentations based on T1w + FLAIR and T1w + T2w images. 205

Fig. A.2. WMH distribution map for the PD cohort. 208

LIST OF TABLES

Table 2.1. Visual Cross-sectional WMH Rating Scales and their definitions. 31
Table 2.2. List of similarity measures and their definitions. 60
Table. 2.3. Results of a logistic regression model, examining the effects of WMHs on cognition. 68
Table 3.1. MRI acquisition parameters for ADC, PREVENT-AD, and ADNI2/GO datasets. 81
Table 3.2. Similarity measures between the manual and automatic segmentations for ADC, PREVENT-AD, and ADNI2/GO datasets. 93
Table 3.3. Percentage of drop in SI (Dice Kappa) by removing feature sets for ADC, PREVENT-AD (periventricular-deep), and ADNI2/GO. 94
Table 3.4. Performance (SI) of LDA, LogitBoost and Random Forests classifiers. LDA= Linear Discriminant Analysis 94
Table 3.5. Comparison of SI (Dice Kappa) for different lesion loads in various studies 101
Table 4.1. Demographic information for ADC, NACC, ADNI1 and ADNI2/GO datasets. 110
Table 4.2. Scanner information and MRI acquisition parameters for ADC, NACC, ADNI1, and ADNI2/GO datasets. 110
Table 4.3. Intra-rater mean Dice Kappa, range of WMHLs, and number (N) of subjects with low (<5cm^3), medium (5-20cm^3), and high (>20cm^3) WMHLs for manual segmentations of WMHs in different datasets. 112
Table 4.4. List of similarity measures and their definitions. 121
Table 4.5. Comparison between mean Dice Kappa, detection/outline error rate (DER/OER), intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs using T1w, T2w, PD and FLAIR data in the ADC dataset. 122
Table 4.6. Comparison between mean Dice Kappa and detection/outline error rate (DER/OER) values of different classifiers for segmentation of WMHs using T1w and FLAIR data in the ADC, NACC, and ADNI2/GO datasets. 125
Table 4.7. Comparison between intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs in different datasets using T1w and FLAIR data for ADC, NACC, and ADNI2/GO datasets. 125
Table 4.8. Comparison between mean Dice Kappa (SI), detection/outline error rate (DER/OER), intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs using T1w, T2w, and PD data, ADC and ADNI1 datasets. 128

Table 4.9. Comparison between mean Dice Kappa (SI), detection/outline error rate (DER/OER), intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs using T1w, and T2w data – ADC and ADNI1 datasets. 131
Table 4.10. Comparison between mean Dice Kappa and detection/outline error rate (DER/OER) values of different classifiers for segmentation of WMHs using T1 data for ADC, NACC, ADNI1, and ADNI2/GO datasets. 133
Table 4.11. Comparison between intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) of different classifiers for segmentation of WMHs in different datasets using T1w data for ADC, NACC, ADNI1, and ADNI2/GO datasets. 134
Table 4.12. Mean ± standard deviation of WMH loads in small (<5cm^3), medium (5-20cm^3), and large (>20cm^3) groups 136
Table 4.13. Comparison between training and segmentation times (s) between different classifiers 137
Table 4.14. Comparison of SI (Dice Kappa) for different lesion loads in various studies 143
Table 5.1. Descriptive statistics for the ADNI subjects enrolled in this study. 153
Table 5.2. Scanner information and MRI acquisition parameters for ADC, ADNI1, and ADNI2/GO datasets. 155
Table 5.3. Percentage of overlap between density histograms of WMHs and GM, WM, and CSF. .. 160
Table 5.4. Coefficients of correlation between WMH loads in different lobes and cognitive measures for ADNI1 subjects, controlling for age. 165
Table 5.5. Coefficients of correlation between WMH loads in different lobes and cognitive measures for ADNI2/GO subjects, controlling for age 166
Table 5.6. Coefficients of correlation between WMH loads in different lobes and measures for ADNI1 and ADNI2/GO subjects. 169
Table 6.1. Descriptive statistics for the PPMI subjects enrolled in this study. 180
Fig. 6.1. Kaplan-Meier graph of survival showing the survival curves of control and PD patients with low versus high WMH loads demonstrating the compounded affect of PD and WMH load. 185
Fig. 6.2. Density plots of longitudinal cognitive changes versus age and log transformed baseline WMH load. 186
Table 6.2. Summary of the mixed effects models of association between baseline WMH Load and change in cognition. 187
Table A.1. Correlation of baseline cognitive measures with WMH loads. 207
Table A.2. Descriptive statistics at baseline for the PPMI low and high WMHL subjects. Data are mean (SD) of key variables. 207

CHAPTER 1. Introduction

The main purpose of this thesis is to develop automatic image segmentation techniques to detect white matter hyperintensities (WMHs) using brain magnetic resonance images (MRIs). The motivation of this study, an overview of this thesis and its scientific contributions are presented below.

1.1. Motivation

WMHs are areas of abnormality that are observed in the white matter tissue of the brain in neuroimaging data. They tend to occur primarily adjacent to the cerebral ventricles, especially around the horns of the lateral ventricles. Although they can occur in individuals that are presumed free of neurological and cerebrovascular diseases, WMHs are often associated with aging, small-vessel disease, Alzheimer's disease, and ischemic diseases. WMHs impact the cognitive function of individuals without any other comorbidities as well as patients with mild cognitive impairment and dementia. The presence of WMHs may confound diagnosis for treating physicians and potentially hinder effective care for these patients (Yoshita et al., 2005). As a result, WMHs need to be taken into account as potential confounders in diagnosis and treatment of Alzheimer's disease patients. Unfortunately, WMHs are difficult to identify and quantify with high accuracy. Manually detecting WMHs is challenging, time consuming, expensive, and variable due to human subjectivity. Thus, developing automated segmentation tools that can detect WMHs robustly and with a high sensitivity and specificity are highly advantageous.

1.2. Overview of thesis

The rest of the thesis is organized as follows. Chapter 2 provides a detailed literature review on WMHs, their effect on cognition, and automated WMH segmentation techniques.

A novel automated linear regression technique for segmenting WMHs is developed in Chapter 3. Using a set of intensity and location features, the linear regression technique provides continuous outputs that can be used as subject-specific WMH maps which can also be thresholded to obtain binary labels. The generalizability of the segmentations is further assessed using an independent dataset. Chapter 4 compares the performance of 10 linear and nonlinear classifiers in detecting WMHs using different combinations of input image contrasts, validating on four different datasets. A WMH segmentation pipeline is developed and made publicly available along with pre-trained classifiers that can be used to segment the WMHs with any of the desired image contrast combinations or classifiers. In Chapter 5, the contribution of different image contrasts in detecting WMHs is assessed. The question of interest is 1) whether WMHs detected solely on T1w images can be used as estimates of the actual WMH volumes, and 2) whether these estimates hold significant correlations with measures of cognitive decline. Next, in Chapter 6, the WMH segmentation technique is used to detect WMHs in a database of early stage Parkinson's disease patients, and the effect of WMHs in later cognitive decline and cortical thinning is assessed. Finally, Chapter 7 concludes the thesis and provides a discussion on possible future work in this area.

1.3. Author contributions

I am the first author of all four manuscripts included in this thesis. I have performed the methodological developments, software implementations, experimental design, data processing, and result analysis for all the experiments. The contributions of the co-authors include supervision of the research, providing data and manual segmentations, technical discussions, and review of manuscripts. The following summarizes the contributions of each author by manuscript.

Chapter 3. Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease

- Authors: Mahsa Dadar, Tharick A. Pascoal, Sarinporn Manitsirikul, Karen Misquitta, Vladimir S. Fonov, M. Carmela Tartaglia, John Breitner, Pedro Rosa-Neto, Owen T. Carmichael, Charles Decarli, D. Louis Collins
- Contributions: Concepts study and design: M. D., and D. L. C.; Manual Segmentation: T. A. P., S. M., K. M.; Method analysis and implementations: M. D., Providing Data: J. B., O. T. C., C. D.; Data preparation and preprocessing: M. D.; Manuscript preparation: M. D.; Manuscript revision: all authors; Editing and final version: M. D. and D. L. C.

Chapter 4. Performance Comparison of 10 Different Classification Techniques in Segmenting White Matter Hyperintensities in Aging

- Authors: Mahsa Dadar, Josefina Maranzano, Karen Misquitta, Cassandra J. Anor, Vladimir S. Fonov, M. Carmela Tartaglia, Owen T. Carmichael, Charles Decarli, D. Louis Collins
- Contributions: Concepts study and design: M. D., and D. L. C.; Manual Segmentation: J. M., K. M., C. J. A.; Method analysis and implementations: M. D., Providing Data: O. T. C., C. D.; Data preparation and preprocessing: M. D.; Manuscript preparation: M. D.; Manuscript revision: all authors; Editing and final version: M. D. and D. L. C.

Chapter 5. Validation of T1w-based Segmentations of White Matter Hyperintensity Volumes in Large Scale Datasets of Aging

- Authors: Mahsa Dadar, Josefina Maranzano, Simon Ducharme, Owen T. Carmichael, Charles Decarli, D. Louis Collins
- Contributions: Concepts study and design: M. D., and D. L. C.; Manual Segmentation: J. M., Method analysis and implementations: M. D., Providing Data: O. T. C., C. D.; Data preparation and preprocessing: M. D.; Manuscript preparation: M. D.; Manuscript revision: all authors; Editing and final version: M. D. and D. L. C.

Chapter 6. White Matter Hyperintensities and Cognitive Decline in de Novo Parkinson's Disease Patients

- Authors: Mahsa Dadar, Yashar Zeighami, Yvonne Yau, Seyed Mohammad Fereshtehnejad, Josefina Maranzano, Ronald Postuma, Alain Dagher, D. Louis Collins
- Contributions: Concepts study and design: M. D., and D. L. C.; Method analysis and implementations: M. D., Y. Z., Data preparation and preprocessing: M. D., Y. Y.; Manuscript preparation: M. D., Y. Y., S. M. F.; Manuscript revision: all authors; Editing and final version: M. D. and D. L. C.

1.4. Scientific contributions

The main original contributions of this thesis are listed below.

(i) Developing and validating fully automatic segmentation tools that can detect WMHs.

   a. Using different combinations of imaging contrasts (e.g. T1w, T1w+T2w, T1w+FLAIR, T1w+T2w+PD+FLAIR).

b. Using ten different linear and nonlinear classifier options.

c. In multi-site, multi-scanner datasets.

The developed tool as well as the pre-trained classifiers have been made publicly available.

(ii) Assessing whether WMH segmentations that are based only on T1w images can be used as accurate estimates in studies where FLAIR or T2w/PD sequences are not available.

(iii) Studying the effect of WMHs on later cognitive decline and cortical thinning in early Parkinson's Disease.

CHAPTER 2. Review of White Matter Hyperintensities, Their Effect on Cognition, and Automated WMH Segmentation from MRI

Introduction

Magnetic resonance imaging (MRI) is a non-invasive widely used medical imaging technique to create in-vivo images of the anatomy in health and disease. Fundamental particles such as protons have a quantized form of angular momentum referred to as spin, causing them to align in a static magnetic field (van der Kouwe et al. 2015). To acquire MRI images, the subject is positioned within a powerful electromagnetic field that aligns the polarity of the proton molecules in the tissues. Then, an oscillating radio frequency (RF) pulse with a perpendicular magnetic component is applied to, and absorbed by, the aligned molecules, resulting in their reorientation. Returning to their original state, the spins emit energy that is received and recorded by the scanner to produce an MRI image. Having different water molecule and therefore proton densities, different tissue types return to their original states (referred to as relaxation) with different rates. Consequently, they present with different intensities in the image (Edelman et al., 1996). By manipulating the relaxation behavior of the spins using a sequence of RF pulses, additional contrast between different tissues can be obtained (van der Kouwe et al. 2015). Since pathological processes such as demyelination, axonal degeneration, and atrophy result in changes in tissue densities, they can be detected in MR images by their differences from normal tissue intensities of healthy individuals.

MRIs are generally three dimensional, composed of a series of two dimensional slices. There exist various types of contrasts in structural MR imaging, specifically designed to emphasize the contrast between different types of tissues or pathologies, by adjusting the

parameters of the pulse sequence that excites the tissue molecules. In structural brain MRI, three primary tissue types can be distinguished: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Brain atrophy is detectable on MRI as a decrease in GM and WM tissue volumes due to the overall brain shrinkage, along with enlargement of the ventricles and widening of the subarachnoid and cisternal spaces, replacing the residual volume with additional CSF. In addition to the atrophy observed in GM and WM tissues in MRI, severe levels of axonal degeneration, neuronal loss, and increased vessel wall permeability can become visible commonly in the WM tissue, referred to as Leuko-araiosis or white matter hyperintensities (WMHs) (Debette and Markus, 2010; Hachinski et al., 1987).

2.1. White Matter Hyperintensities and Their Effect on Cognitive Function

2.1.1. Cerebral Small Vessel Disease

The term cerebral small-vessel disease (CSVD) refers to a combination of clinical and MRI findings that result from a series of pathological processes that affect the small vessels on the surface of and within the brain, i.e. the arteries, arterioles, capillaries, venules, and veins (Joutel and Faraci, 2014; Pantoni, 2010). CSVD is the most common vascular cause of dementia and a major contributor to mixed dementia (i.e. CSVD co-occurring with other pathologies) (Dubois et al., 2014; Masdeu and Pascual, 2008). In particular, CSVD frequently coexists with neurodegenerative diseases such as Alzheimer's disease (AD) in the elderly population and may increase the cognitive and physical impairments caused by neurodegeneration (Dubois et al., 2014; Honjo et al., 2012; Ogama et al., 2014; Wardlaw et al., 2013). AD and CSVD share many risk factors and both lead to cognitive decline and dementia (De la Torre, 2000; Sachdev et al., 2014; Wardlaw et al., 2013). The effect of vascular and degenerative changes in individuals with AD are reported to be cumulative, i.e.

the co-occurrence of CSVD can increase the likelihood of clinical presentation of dementia, for the same level of AD-related pathology (Masdeu and Pascual, 2008).

Furthermore, recent studies indicate that CSVD might be the starting point in a chain of events that lead to neurodegeneration and AD. The Critically Attained Threshold of Cerebral Hypoperfusion (CATCH) hypothesis of AD proposes that the blood supply and metabolism disturbances due to the hypo-perfusion and the ischemic changes associated with aging and vascular risk factors would cause neuronal energy failure (De la Torre, 2000). These changes would in turn lead to neuronal injury as well as acceleration in over-production and reduction in clearance of Aβ (a protein whose abnormal extracellular deposition is crucially involved in AD), eventually leading to AD pathology. Therefore, this process would create a chain of events that lead to the progressive cognitive deficits and neurodegeneration that characterize AD (De la Torre, 2000). Similarly, Honjo *et al.* hypothesize that hypo-perfusion and/or inflammation caused by early on vascular risk factors (which precede the Aβ increase and deposition) would lead to Aβ drainage dysfunction and start the neurodegenerative process leading to AD. Therefore, controlling vascular risk factors earlier in life and before the onset of these pathological processes could be crucial for prevention of AD (Honjo et al., 2012).

2.1.2. White Matter Hyperintensities

WMHs are one of the major MRI signs of CSVD. Other signs include infarctions or microhemorrhages in the WM or deep GM and enlargement of perivascular spaces (Joutel and Faraci, 2014). WMHs are non-specific areas of abnormality, caused by several different mechanisms that are generally observed in the WM tissue of the brain in MR images. They mainly result from chronic diffuse subclinical ischemia (i.e. a restriction in blood supply to the brain tissue, creating oxygen and glucose shortage and consequently disturbing cellular metabolism) that primarily impacts subcortical regions (Scott et al., 2014). Aβ deposition

could also increase the WMH burden by accelerating processes that are not necessarily vascular, including neuroinflammation, reactive oxygen species production, and oxidative stress (Scott et al., 2014; Snyder et al., 2015; Yamada, 2012). In such cases, an initial rise in Aβ would damage the WM, which in turn would elevate Aβ levels, leading to more WM damage in a cyclical process. Damage to the vascular system such as thickening and sclerosis of the arteries may also disrupt the clearance of molecules such as Aβ (Huang et al., 2010). Alternatively, WMHs from another process such as a head injury may initiate or accelerate the pathological effects of Aβ deposition on the WM (Wardlaw et al., 2015). Grimmer et al. (2012) showed a significant association between the baseline WMH load and the progression of Aβ between baseline and follow-up visits. This association was mostly observed in parieto-occipital regions, especially in regions closer to the brain surface. This further supports the hypothesis that impaired Aβ clearance due to dysfunction in the drainage process (as measured by WMH volume) may contribute to Aβ deposition in AD, which in turn makes WMH load a relevant factor that needs to be taken into consideration in drug trials that target Aβ (Grimmer et al., 2012).

2.1.3. WMH Locations

WMHs occur primarily adjacent to the cerebral ventricles, especially around the horns of the lateral ventricles. The WMH distribution extends outward to other brain regions as the disease progresses. Similar hyperintensities can also occur in subcortical GM structures, such as the basal ganglia, thalamus, as well as brainstem and have sometimes been analyzed alongside the hyperintensities in the WM (Wardlaw et al., 2013). WMH studies sometimes differentiate between periventricular and deep WMHs since they appear to have different histopathologies, risk factors, and clinical consequences.

Periventricular hyperintensities can also be seen in elderly individuals without clinically significant cognitive symptoms and can be categorized into *rims* and *caps*. The term rim is referred to the laterally positioned periventricular hyperintensities surrounding the lateral ventricles and caps are the hyperintensities surrounding the anterior poles of the lateral ventricles (Kertesz et al., 1988). In general, rims and caps are both subtle and follow the shape of the surface of the ventricle, with caps being thicker than rims. The pathologic findings associated with periventricular WMHs include myelin pallor, dilation of perivascular spaces, and increased extracellular spaces (Gouw et al., 2010).

Deep hyperintensities are generally categorized into *punctate* and *confluent* lesions. These punctate and/or confluent hyperintense areas are frequently seen in the deep and subcortical WM on T2w and FLAIR images, and are commonly observed in the elderly individuals, particularly in those with vascular risk factors (Fazekas et al., 1993). They are considered to be produced by chronic ischemia or brief and moderately severe repeated ischemia occurring in the subcortical WM (Jellinger et al., 2013; Matsusue et al., 2006).

Figure 2.1. shows examples of periventricular (caps and rims) and deep (punctate and confluent) WMHs in axial FLAIR image slices of individuals with CSVD.

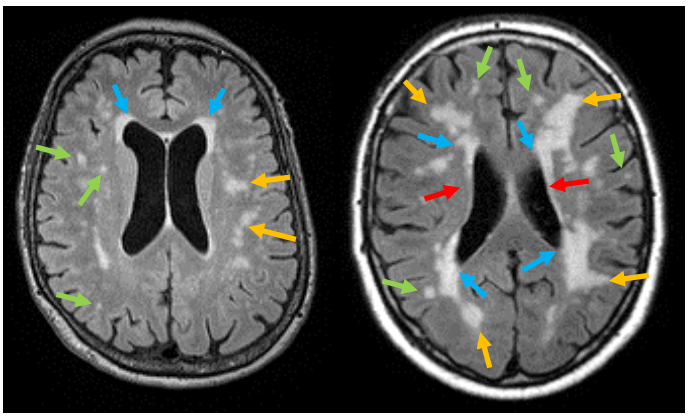


Fig. 2.1. Examples of periventricular (caps and rims) and deep (punctate and confluent) WMHs. Arrows indicate different types of WMHs. Red= rims. Cyan= caps. Green= punctate. Orange= confluent. WMHs= White Matter Hyperintensities.

2.1.4. WMH Rating Scales

WMHs can be visually scored by expert neurologists according to their level of severity. Manual rating of WMH severity is generally done using FLAIR and T2w images as the primary sequence, with the help of other available modalities to exclude artifacts and other pathologies. The Fazekas scale, the Age-Related WM Changes scale (ARWMC), and Scheltens scale are three cross-sectional visual scales for assessing the WMH burden. Table 2.1 shows scoring details for each scale.

(i) The Fazekas scale was designed to quantify the WMHs caused by chronic small-vessel ischemia, dividing the WMHs into periventricular and deep WM (PVWM and DWM, respectively) and assigning a grade between 0 (none) and 3 (significant amount) to each category depending on the size and confluence of the WMHs (Fazekas et al., 1987).

(ii) The ARWMC scale divides the WMHs into WM lesions and basal ganglia lesions (striatum, globus pallidus, thalamus, internal/external capsule, and insula), giving a 0 (no lesions) to 3 (confluent lesions) score to each category (Wahlund et al., 2001).

(iii) The Scheltens scale rates WMHs in the periventricular region on a 0 (absent) to 6 scale (significant amount), and in the subcortical regions on a 0 (no abnormalities) to 24 (large confluent lesions) scale, on the basis of the size and number of the lesions (Scheltens et al., 1993).

Table 2.1. Visual Cross-sectional WMH Rating Scales and their definitions. WMH= White Matter Hyperintensities. ARWMC= Age-Related WM Changes scale. WM= White Matter.

Scale	Details			
Fazekas	**Periventricular hyperintensities**			Score
	Absence			0
	Caps or pencil-thin lining			1
	Smooth halo			2
	Irregular PVH extending into the deep WM			3
	Deep WMHs			
	Absence			0
	Punctate foci			1
	Beginning confluence of foci			2
	Large confluent areas			3
ARWMC	**WMHs**			Score
	No lesions (including symmetrical, well-defined caps or bands)			0
	Focal lesions			1
	Beginning confluence of lesions			2
	Diffuse involvement of the entire region/with or without involvement of U fibers			3
	Basal Ganglia Lesions			
	No lesions			0
	One focal lesion (≥ 5 mm)			1
	More than one focal lesion			2
	Confluent lesions			3
Scheltens		Number	Definition	Score
	Periventricular hyperintensities		0: absent	**0-6**
	Occipital caps		1: ≤ 5mm	0-2
	Frontal caps		2: > 5mm & < 10 mm	0-2
	Lateral ventricle bands			0-2
	WMHs		0: absent	**0-24**
	Frontal	n ≤ 5	1: < 3 mm	0-6
	Parietal	n > 6	2: < 3 mm	0-6
	Occipital	n ≤ 5	3: 4-10 mm	0-6
	Temporal	n > 6	4: 4 mm	0-6
	Basal Ganglia hyperintensities	n > 1	5: > 11 mm	**0-30**
	Caudate Nucleus		6: confluent	0-6
	Putamen			0-6
	Globus Pallidus			0-6
	Thalamus			0-6
	Internal Capsule			0-6
	Infra-tentorial foci of hyperintensity			**0-24**
	Cerebellum			0-6
	Mesencephalon			0-6
	Pons			0-6
	Medulla			0-6

Rotterdam Progression Scale and Schmidt Progression Scale are two longitudinal scales designed for assessing the progression of WMHs.

(i) The Rotterdam scale rates decrease, absence or presence of progression (-1, 0 or +1) in three periventricular regions (frontal caps, occipital caps, and bands) resulting in a periventricular score of -3 to +3, and four subcortical WM regions (frontal, parietal, occipital, temporal) resulting in a subcortical score of -4 to +4. Decrease is defined as disappearance or shrinkage of a previously visible lesion, absence of progression is defined as no significant change in the appearance of the lesions, and increase is defined as enlargement of a previously visible lesion or occurrence of a new lesion. The time between the two scans is not factored into this progression scale (Leeuw et al., 2001; Prins et al., 2004a).

(ii) The Schmidt progression scale rates the change in WMHs by counting and categorizing the number of the lesions into (i) zero, (ii) one to four, (iii) five to nine, and (IV) more than nine lesions. Change of WMHs in grade and number from the baseline is then determined by direct scan comparison. Progression of WMHs is graded as absent (no WMHs in either scan), minor (a change from baseline by one to four punctate lesions), or marked (a difference of more than four lesions or a transition to early-confluent or confluent WMHs) (Schmidt et al., 1999).

The most important disadvantage of these visual WMH severity scales is their non-linearity, i.e. two subjects with significantly different WMH loads (generally high loads) can receive the same scores, i.e. ceiling effect (Fazekas et al., 2002). This limitation makes it difficult to accurately assess the progression of WMHs or population differences in WMH

burdens, since all subjects with high WMH burden would receive the same scores, while they may be still significantly different in the actual WMH loads.

Additionally, the same scans might be rated quite differently by different scales and raters, consequently showing different associations with clinical risk factors and cognitive measures (Mäntylä et al., 1997). Comparing the five rating scales, Gouw et al. reported mean intra-rater Kappa values of 0.79, 0.91, and 0.84 for the baseline and 0.82, 0.92, and 0.88 for the follow-up visits for Fazekas, ARWMC, and Scheltens scores, respectively for 2 subjects (Gouw et al., 2008). They also verified the inter-rater reliability for 4 subjects, showing Kappa values of 0.70, 0.88, and 0.80 for the baseline and 0.72, 0.86, and 0.82 for the follow-up visits for Fazekas, ARWMC, and Scheltens scores, respectively. They further showed that the Rotterdam progression scale had a better sensitivity to WMH change over time compared with the other four scales (Gouw et al., 2008). Another study reported inter-observer Kappas for periventricular and subcortical WMHs of 0.37 and 0.84 on the Fazekas scale and 0.64, 0.90 on the Rotterdam scale, and 0.56 and 0.84 on the Scheltens scale (Prins et al., 2004a).

Different scaling systems also have different time requirements. For example, Fazekas scale is relatively simple and fast, while Scheltens scale is more time-consuming, requiring the rater to count the number/types of lesions (Mäntylä et al., 1997).

In conclusion, although useful, visual ratings of WMH severity are generally time consuming, susceptible to inter-rater and intra-rater variabilities, and less sensitive to subtle difference/changes in the WMHs.

2.1.5. WMH Risk Factors

Although WMHs can occur in individuals without clinically significant neurodegenerative and cerebrovascular diseases, they are often associated with aging, AD,

and ischemic diseases. The number and extent of WMHs increase from normal aging to mild cognitive impairment and to dementia. Age, a history of ischemia, CSVD, and hypertension (i.e. elevated blood pressure level) have been reported to be strong predictors of WMH load (Au et al., 2006; Gunning-Dixon and Raz, 2000). Doi et al. report subjects with mild cognitive impairment (MCI) that have lower mobility to suffer from higher WMH volumes (Doi et al., 2015). Higher baseline WMH load is associated with increased longitudinal progression of WMHs (Silbert et al., 2008). Furthermore, WMHs are associated with reduced cerebral and frontal lobe metabolism, higher systolic blood pressure (even in subjects whose blood pressures are in a normal range), brain atrophy, and smaller brain volumes (DeCarli et al., 1995a). Benedictus et al. report a significant association between WMH load and cerebral blood flow in AD patients, but not in controls (Benedictus et al., 2014).

Scott et al. have shown that in subjects with hypertension, higher cerebral Aβ was associated with a greater WMH load than subjects with normal blood pressure (Scott et al., 2014). Those with both elevated Aβ and hypertension had greater WMH loads than could be accounted for by either of these factors alone. This finding suggests that both vascular dysfunction and Aβ deposition may independently increase the WMH burden or alternatively, vascular dysfunction might accelerate both the Aβ deposition and accumulation of WMHs, creating a downstream association between WMH loads and Aβ levels (Scott et al., 2014). They thus conclude that both Aβ deposition and vascular pathology are (possibly independent) contributors to WM damage in cognitively normal individuals.

2.1.6. Impact of WMHs on Cognition

WMHs impact the cognitive function of individuals without any clinical symptoms as well as patients with MCI and dementia (Yoshita et al., 2005). Many studies include total WMH loads in their analyses when studying the cognitive performance and decline in aging

and diseased populations. Other measures of interest include location and number of these lesions (Wardlaw et al., 2013). Here we present a summary of studies that have used WMH related measures as correlates of impairment in different cognitive and clinical domains in cognitively normal, MCI, and AD populations.

WMHs in cognitively normal populations

In a quantitative review of 23 studies that investigate the effects of WMHs in cognitively normal individuals, Gunning-Dixon and Raz found WMH burden to be associated with poorer performance on various cognitive tasks. More specifically, WMHs were reported to be associated with lower memory, processing speed, executive function, and gait. They also found that periventricular WMHs were less likely to be associated with cognitive decline than deep WMHs, particularly in tasks that concern processing speed and widely distributed neural networks in the brain (Gunning-Dixon and Raz, 2000).

In a cohort of 100 cognitively normal subjects followed up longitudinally for an average of 9.1±4.0 years, Silbert et al. reported that periventricular WMH load was associated with gait difficulties, while the subcortical WMH load was associated with decline in cognition and increased rate of memory decline, even after adjusting for the rates of cerebral or hippocampal atrophy (Silbert et al., 2008).

Studying 1077 participants without dementia at baseline followed up for an average of 5.2 years (Rotterdam Scan Study), Prins et al. observed that a higher severity of WMHs (specially periventricular WMHs) at baseline increased the risk of developing dementia later on, adjusted for age and sex and independent of risk factors (including hypertension, diabetes mellitus, smoking, APOE genotype, and history of stroke) and other structural brain changes (Prins et al., 2004b).

Using 1820 dementia and stroke free participants from the Framingham offspring study, Au et al. showed that individuals with high WMH loads perform worse on cognitive tests that are associated with frontal lobe systems and to a lesser extent, the medial temporal areas. The battery of assessments included measures of attention, planning and initiation of complex activity, cognitive factors of visuospatial memory and organization and visual scanning and motor speed as well as new learning. All findings were adjusted for the effect of age, sex, education, height, and Framingham stroke risk profile (Au et al., 2006).

In a cohort of 354 older individuals that were initially cognitively normal with an average of 4.1 years of follow-up, Boyle et al. showed that an individual with a high WMH load was approximately 2.7 times more likely to develop MCI compared with a person with a low load. Also, every 1 standard deviation increase in WMH load was associated with a 1.43 increase in risk of MCI (Fig. 2.2). Additionally, WMH load did not correlate with cognition at baseline, but it was associated with increased rate of cognitive decline as well as decline in perceptual speed, working memory, episodic memory, and semantic memory. WMH load was significantly correlated with other cognitive domains at older ages. The correlations remained statistically significant after adjusting for total GM volume, vascular risk factors, and vascular diseases (Boyle et al., 2016).

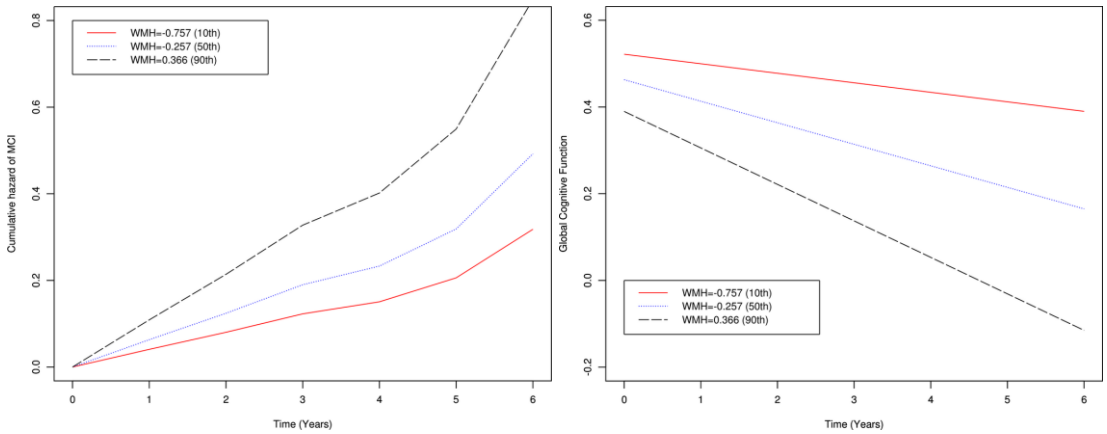


Fig 2.2. The relationship between WMHs and risk of MCI and rate of decline in global cognition. The figure shows the cumulative hazard of developing MCI (left) and rate of decline in global cognition (right) for subjects with low (red), medium (blue), and high (black) WMH loads (Boyle et al., 2016). WMHs= White Matter Hyperintensities. MCI= Mild Cognitive Impairment. *Original figure from Boyle et al. 2016 (replicated under CC BY-NC-ND 4.0 license).*

Using Cox proportional hazard modeling in a cohort of 67 cognitively normal and 156 MCI subjects followed up for 6.0±4.1 years, Smith et al. reported that adjusting for age, sex, education, smoking, and APOE status, high WMH load (defined as a log-transformed WMH volume more than 1 standard deviation above the mean value) predicted progression from normal cognition to MCI, but not conversion from MCI to dementia (Smith et al., 2008).

WMHs in individuals with mild cognitive impairment

Studying 514 pairs of male twin participants in the National Heart, Lung, and Blood Institute (NHLBI) Twin Study, Decarli et al. observed that subjects with MCI had higher WMH volumes. Additionally, WMHs were associated with significantly increased risk of MCI (DeCarli et al., 2001). Similarly, in a cohort of 3608 participants from the Cardiovascular Health Study (CHS) Cognition Study, Lopez et al. found that WMHs to be associated with symptoms of MCI (Lopez et al., 2003). In a relatively small cohort (N=41), Nordahl et al. showed that compared with cognitively normal subjects, MCI subjects with hippocampal atrophy and MCI subjects with severe WMHs were both impaired on an

episodic memory task, but MCI subjects with WMHs were additionally impaired on verbal and spatial working memory and attention control tests (Nordahl et al., 2005).

In a recent review of 12 studies investigating the imaging and neuropsychological correlates of WMHs in different MCI subtypes, the authors report cognitive correlates of WMH loads in the amnestic MCI subtype for memory, language, psychomotor speed, attention and executive functions. They also hypothesize that cognitive reserve and WM plasticity may modulate the effect of WMHs on neurodegenerative diseases (Lam et al., 2017).

Li et al. performed a meta-analysis on 19 studies aiming to identify risk factors for progression from MCI to AD. The meta-analysis included studies that investigate MRI-derived biomarkers such as WMHs, atrophy in hippocampal, medial temporal lobe and entorhinal regions. The results showed that MCI subjects with WMHs had a significantly higher chance of conversion to AD compared with subjects without WMHs. Other significant predictors included CSF p-tau, CSF tau/Aβ, hippocampal, entorhinal and medial temporal lobe atrophy as well as depression, diabetes, hypertension, older age, and female gender (Fig 2.3) (Li et al., 2016).

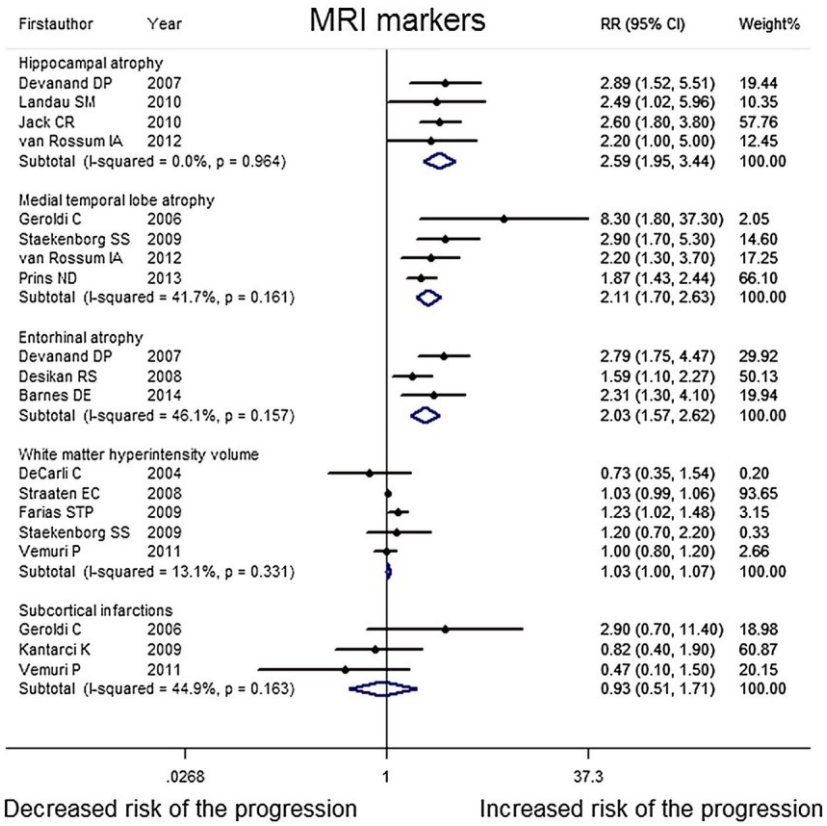


Fig 2.3. Results summary of studies investigating the relationship between MRI biomarkers and the risk of progression from MCI to AD. MCI= Mild Cognitive Impairment. AD= Alzheimer's Disease. RR= Relative Risk (Li et al., 2016). *Original figure reproduced with permission from Li et al. 2016.*

Studying the effect of WMHs on cognition in an MCI cohort of 374 individuals, Tosto et al. investigated whether baseline WMH load and lower entorhinal cortex (EC) volume increase the risk for aggressive cognitive decline (ACD), defined as either a 3-point decline in MMSE score over 6 months or a 6-point decline in MMSE over 12 months. They also investigated whether there is an interaction between the two factors, or their effect is independent (Tosto et al., 2014). The results showed that a higher baseline WMH load, APOE4 status, and smaller EC volume at baseline were associated with an increased risk for ACD. In a survival analysis, WMH volume was also found to modify the effect of EC volume on the risk of ACD, i.e. individuals with low WMH loads and high EC volumes were at particularly low risk of ACD. In addition, individuals with ACD were more likely to convert to AD during a follow-up period of 48 months (Tosto et al., 2014).

In a cohort of 136 normal controls and 186 MCI individuals, Fujishima et al. found MCI, poorer episodic memory and late-life depression to be associated with WMHs as well as cortical thinning (Fig. 2.4) (Fujishima et al., 2014).

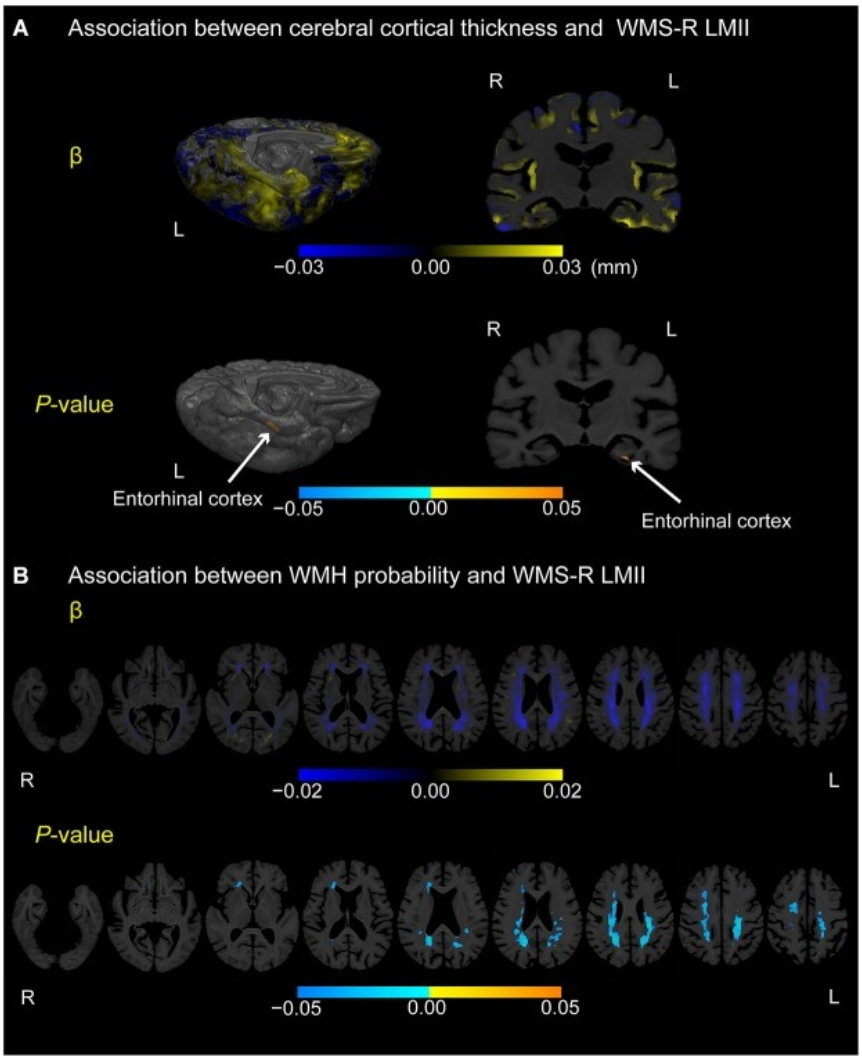


Fig 2.4. Association between cortical thickness or WMH probability and episodic memory in subjects with MCI. (A) regression coefficients maps in cortical regions and FWE-corrected *p*-values. (B) regression coefficients maps for WMHs and FWE-corrected *p*-values. These results are adjusted for effects of age, gender, and years of education. WMHs= White Matter Hyperintensities. MCI= Mild Cognitive Impairment. FWE= Family Wise Error (Fujishima et al., 2014). *Original figure from Fujishima et al. 2014 (replicated under CC BY license).*

WMHs in Alzheimer's disease patients

Using the Fazekas, ARWMC, and Scheltens scales (Fazekas et al., 1987; Scheltens et al., 1993), as well as WMH volumes in 108 AD patients, 23 individuals with MCI and 34 normal controls (NC), Gao et al. found that WMHs volumes, and ARWMC and Scheltens scores show significant negative correlations with cognitive scores controlling for age, education, global and hippocampal atrophy in AD patients. Additionally, they found that the Fazekas scale significantly separated AD, MCI and NC cohorts (Gao et al., 2011).

Investigating the potential effects of APOE status on WMHs and their effect on cognition, Morgen et al. reported that among 201 AD patients, *APOE* ε4 non-carriers showed significantly higher WMH loads as well as an association between WMH load and cognition (Morgen et al., 2015).

In two mixed populations of vascular dementia and AD patients (N_1= 87, N_2= 66), Lange et al. used a lesion shape irregularity score to reflect the level of WMH severity and showed that among subjects with total WMH loads higher than 13.5 CCs, cognitive performance in mental speed and fluid ability domains was more strongly associated with the shape irregularity score than the total volume or number of lesions (Lange et al., 2016).

2.1.7. Prevention and Treatment

Prevention and treatment of vascular risk factors (and WMHs downstream) is a promising avenue to slow down or prevent the subsequent cognitive decline, especially in early stages when the individuals are still mostly cognitively asymptomatic. One potential treatment strategy is to control elevated blood pressure. Hypertension is a leading cause of WMHs and effective usage of anti-hypertensive medication has been shown to limit the severity of

WMHs and subsequently reduce the risk of cognitive decline (Hutton et al., 2009; de Leeuw et al., 2002; Schrag et al., 2017). In a randomized trial administering either anti-hypertensive medication or placebo to CSVD patients, active lowering of blood pressure in the former cohort was shown to stop or lower the progression of WMHs over 3 years of follow-up (Dufouil et al., 2005).

2.1.8. Conclusion

In summary, WMHs are clinically significant markers of CSVD in aging as well as neurodegenerative diseases as they reflect the level of vascular pathology and are associated with increase in Aβ levels and decline in various cognitive domains (Carmichael et al., 2010; DeCarli et al., 1995a; Pantoni et al., 2006; van Straaten et al., 2008). WMHs are used as outcome measures in clinical trials that investigate CSVD in the context of stroke and dementia (Debette and Markus, 2010) and can also be a potential biomarker of preclinical risk for developing AD (Brickman et al., 2012; Deoni et al., 2013; Provenzano et al., 2013).

2.2. White Matter Hyperintensity Segmentation

Image segmentation is a means of differentiating the voxels of interest in an image from the background and plays an essential role in medical image analysis. Many algorithms have been developed to segment different structures as well as healthy or pathologic tissue from brain MR images. Overviews on WMH segmentation techniques can be found in several review papers (Caligiuri et al., 2015; García-Lorenzo et al., 2013). In this section, we review automated techniques for segmenting WMHs from MRIs.

The MRI segmentation task can generally be defined as partitioning an image into non-overlapping regions where the pixels/voxels inside each region have similar characteristics (e.g. intensity, shape, texture). In the case of WMHs, the segmentation task is defined as assigning either a WMH or a background label to each voxel in the MRI image.

Segmentation of WMHs is generally performed on Fluid-attenuated inversion recovery (FLAIR) or T2w MR images by expert raters. A FLAIR image is essentially a T2w image, with the CSF signal nulled, resulting in an image in which the WMHs appear as the brightest intra-cranial tissue type, making FLAIR an ideal contrast for detecting WMHs (Rydberg et al., 1994, 1995).

Manual WMH Segmentation

Manual segmentation is used as the gold standard for MRI WMH segmentation. It requires a detailed segmentation protocol for what should be considered as WMHs as well as expert knowledge of brain anatomy on MR images. Accurate and consistent segmentation of WMHs is a complicated task due to the inherent heterogeneity in their texture and patterns as well as the fact that these lesions often have fuzzy borders. These complications make manual detection of WMHs challenging, time consuming, expensive and inconsistent due to high inter-rater and intra-rater variability. In a recent review, García-Lorenzo et al. report that inter-rater and intra-rater agreements in lesion segmentation tasks are generally modest at best, even when compared within the same protocol (García-Lorenzo et al., 2013). Challenges in manually segmenting WMHs are mainly caused by the fact that the boundary between healthy tissue and WMHs is indistinct and difficult to precisely determine. As a result, different raters draw different arbitrary distinctions between the two tissues around the edges. Even the same rater never achieves perfect intra-rater agreement. Therefore, these inconsistencies introduce an inevitable level of noise into the manually segmented WMHs.

Automated WMH Segmentation

Conversely to manual identifications, automated methods always apply the same policy to the edge distinctions, providing consistent segmentations. In addition, the extensive number of images being collected makes the human cost of manual identification prohibitive. These

make automated segmentation tools that can detect WMHs accurately highly advantageous since with their objectivity and reproducibility, they would essentially eliminate the intra-rater variability and make it possible to reliably follow individual subjects over time, or segment WMHs in large scale studies and clinical trials.

Since different MRI modalities have different contrasts across tissues, integrating information from multiple modalities can reduce uncertainty and consequently increase segmentation accuracy, both in manual and automatic segmentations. In automatic segmentations, MR image contrasts can be weighted differently to reflect specific properties of different anatomical structures, regions, or pathologies. The MRI contrasts that are commonly used in automatically detecting WMHs include:

(i) T1w or spin lattice relaxation images for which the magnetization is allowed to recover before measuring the MR signal by changing the repetition time (TR). The higher T_1 relaxation time in the CSF followed by GM (compared with WM) causes them to appear darker in T1w images (van der Kouwe et al. 2015). Similarly, WMHs appear hypointense on T1w images. Due to their excellent contrast between GM, WM, and CSF as well as their high spatial resolution, T1w images are mostly used for co-registration purposes in the literature.

(ii) T2w or spin-spin relaxation images, for which magnetization is allowed to decay (characterized by a time constant, T_2, also called spin-spin relaxation time) before measuring the MR signal by changing the echo time (TE) (van der Kouwe et al. 2015). The CSF followed by GM have higher T_2 relaxation times than WM, causing them to appear brighter on T2w images. WMHs also appear bright on T2w images, making them easy to differentiate from WM, but more difficult to distinguish from CSF.

(iii) Proton density (PD), an intermediate sequence, which reflects the actual density of protons, with a long TR and a short TE that shares some characteristics of both T1w and T2w images. CSF has the highest PD, followed by GM and WM appearing the brightest on PD images.

(iv) Fluid attenuated inversion recovery (FLAIR) images for which the inversion time (TI) is manipulated to suppress the effect of CSF on the image to outline the WMHs more clearly.

As mentioned before, WMHs appear hyperintense on T2w, PD, and FLAIR MR images and hypointense on T1w images. Figure 2.5 shows co-registered axial slices of T1w, T2w, PD and FLAIR images as well as manually segmented WMH labels for one subject.

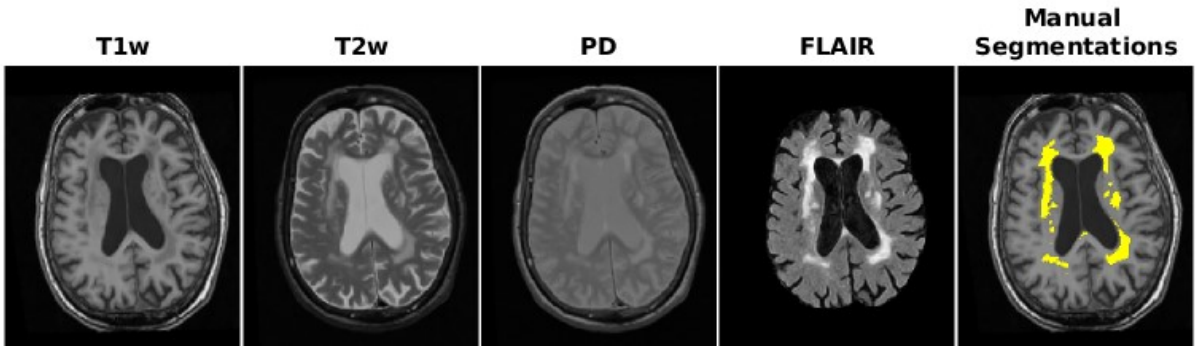


Fig. 2.5. WMHs on T1w, T2w, PD, and FLAIR MR images. WMHs= White Matter Hyperintensities. PD= Proton Density. FLAIR= Fluid Attenuated Inversion Recovery. The WMHs manually detected by an expert rater using these four sequences are indicated by yellow color.

The typical intensity histograms of GM, WM, CSF, and WMHs in each of the four contrasts (Fig. 2.6) show that WMH histograms have significant intensity overlap with GM, WM and CSF. However, WMH intensities have the least amount of overlap with other tissue histograms in FLAIR and T2w images, making them the optimal sequences for detecting WMHs both manually and automatically.

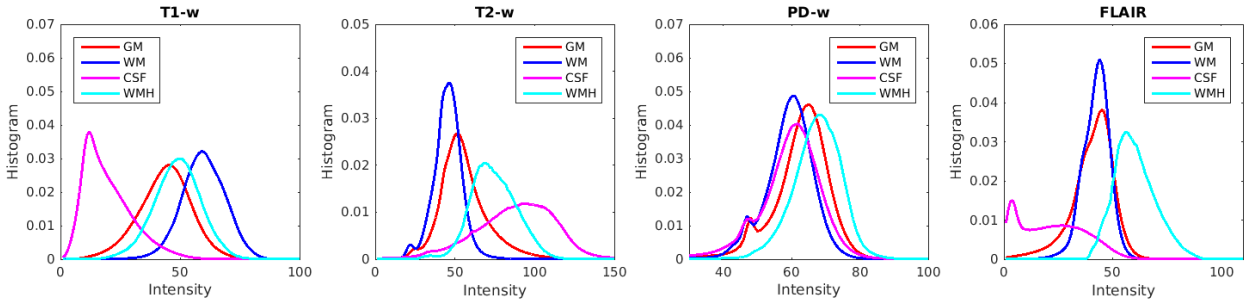


Fig. 2.6. Tissue intensity histograms. GM= Gray matter. WM= White Matter. CSF= cerebrospinal fluid. WMH= White Matter Hyperintensity. PD= proton density. FLAIR= Fluid Attenuated Inversion Recovery.

2.2.1. Challenges in detecting WMHs

Automated segmentation (and to some extent manual segmentation) of WMHs present a number of challenges, including:

(i) Varying intensities and GM, WM, and CSF tissue contrasts between MR images that are obtained across different scanners, sequences, field strengths, and timepoints (Fig. 2.7).

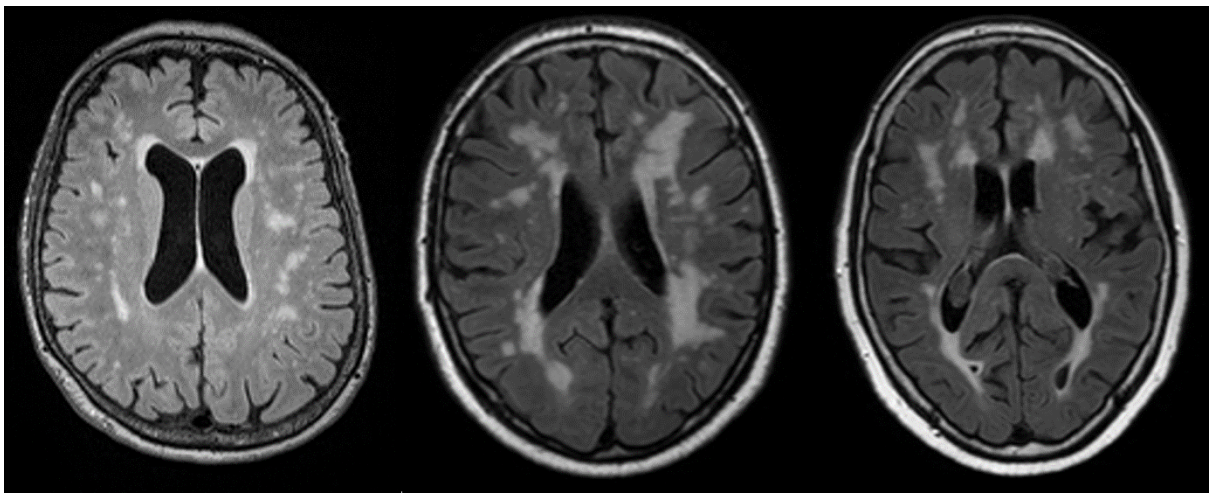


Fig. 2.7. Varying intensities and tissue contrasts in different sequences acquired on different scanners. Note the difference in contrast between the WMHs, GM and WM in different sequences. The left image shows an axial slice of a 3D sagittal T2 space FLAIR image. The middle and right images show axial slices of 2D axial turbo spin echo T2-FLAIR images. WMH= White Matter Hyperintensity. GM= Gray Matter. WM= White Matter.

(ii) Image intensity noise caused by the electrical current running through the electromagnet coil and artifacts caused by sensor noise (Fig. 2.8).

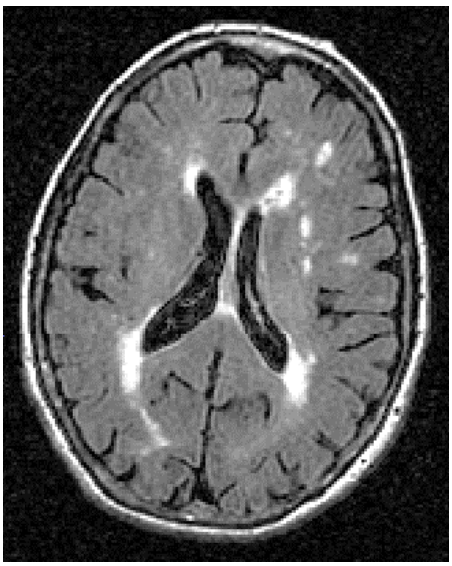


Fig. 2.8. MR image intensity noise in an axial slice of a FLAIR image.

(iii) Bias field (i.e. image intensity inhomogeneity) caused by spatially varying coil sensitivity, multi-coil reconstruction errors, induced currents and standing waves, magnetic settings, subject positioning inside the scanner, and other factors (Fig. 2.9).

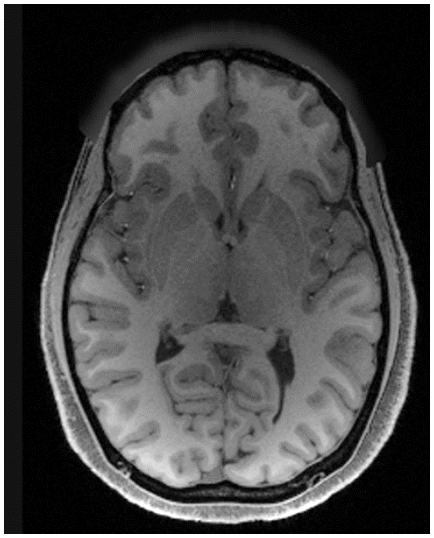


Fig. 2.9. MR image inhomogeneity in an axial slice of a T1w image. Note the dark to light trend from top to bottom.

(iv) Motion artifacts caused by breathing as well as subject movement. These artifacts are generally increased in the elderly and diseased populations due to disease related tremors, subject discomfort as well as forgetfulness (Fig. 2.10).

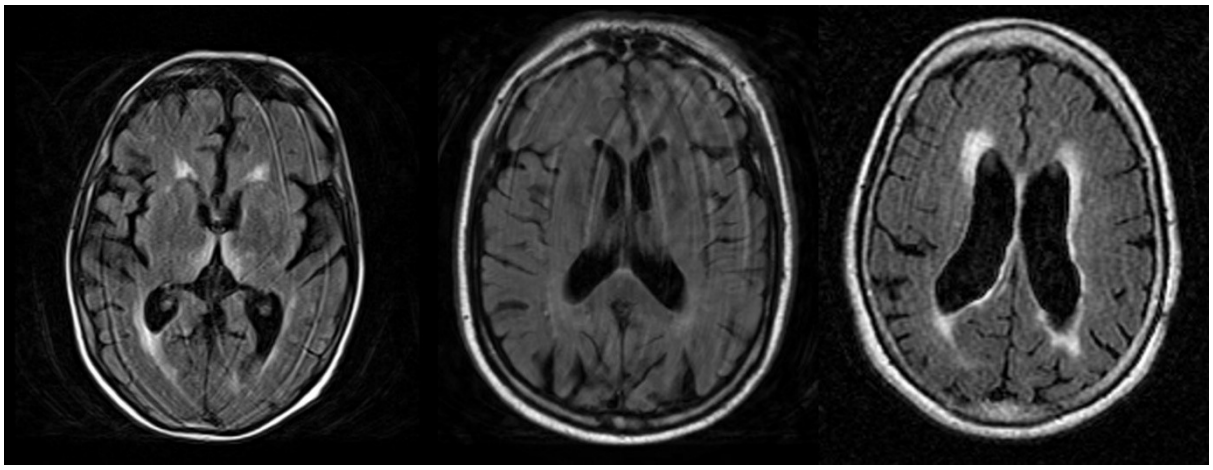


Fig. 2.10. Ringing artifact in MR images caused by subject movement in axial slices of FLAIR images.

(v) Partial volume effect caused by the discretization of the continuous signal in the sampling (voxel) resolution blurring the tissue border intensities (Fig. 2.11).

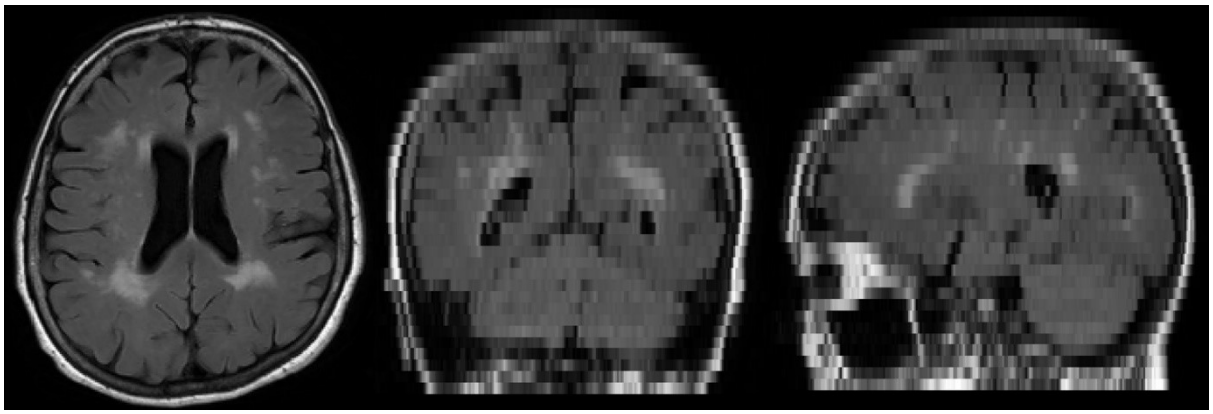


Fig. 2.11. Partial volume effect. Axial (left), coronal (middle), and sagittal (right) slices of an individual's FLAIR scan.

(vi) Atrophy and changes in GM/WM tissue contrasts in the aging population, i.e. lower GM/WM contrast as age increases (Fig. 2.12).

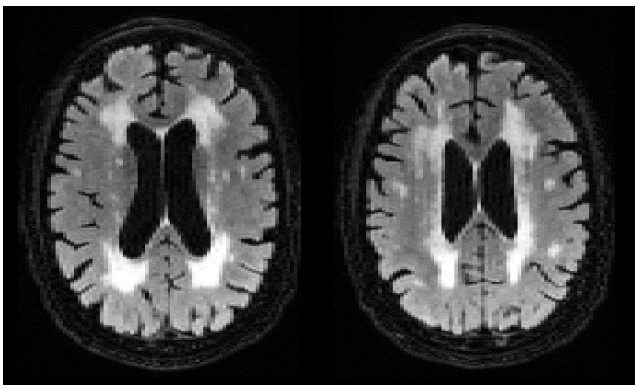


Fig. 2.12. Atrophy (large symmetric ventricles and subarachnoid spaces) and change in GM/WM tissue contrast in FLAIR MR images.

(vii) Hyperintense appearance of healthy tissues such as the choroid plexus, blood vessels, venous sinuses, the tail of the caudate nucleus, the tangential plane of transition to bone at the base of the skull, and the fat signal at the diploe level of the skull bones, depending on the image modality and the amount of partial volume effects (García-Lorenzo et al., 2013) (Fig. 2.13).

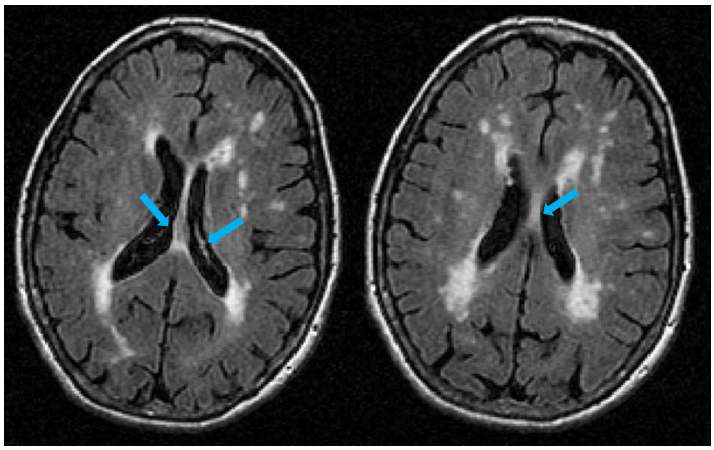


Fig. 2.13. Hyperintense appearance of non-WMH tissues such as choroid plexus in FLAIR images.

(viii) Fuzzy borders of the lesions exhibiting a degree of hyperintense signal that decreases gradually towards the surrounding healthy tissue (Fig. 2.14). Some groups identify this a *dirty white matter* (Moore et al., 2008).

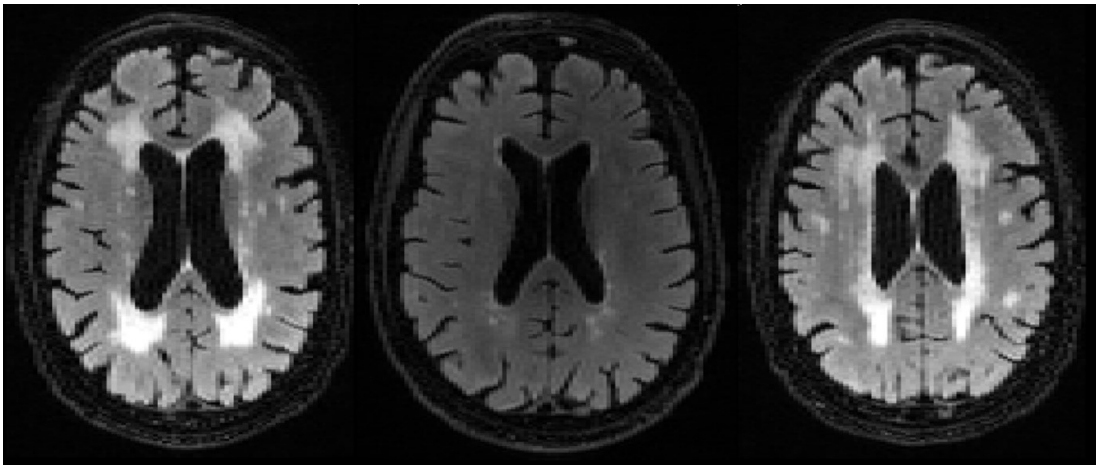


Fig. 2.14. Fuzzy borders of lesions in FLAIR images. Note the difference between the sharpness of the borders in different lesions.

(ix) Inconsistency in lesion borders on different MR contrasts, generally leading to more generous segmentations on FLAIR images in comparison with T2w based labels (Ciccarelli et al., 2002) (Fig 2.15).

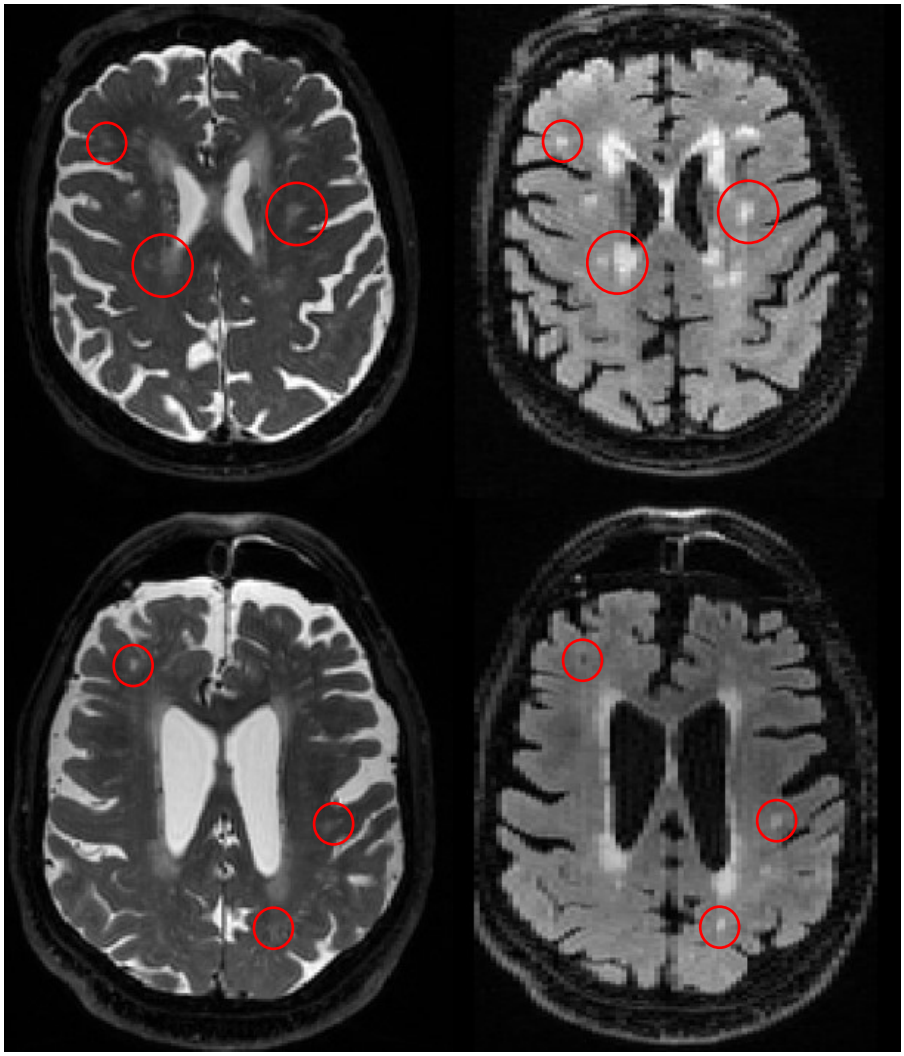


Fig. 2.15. Inconsistency in lesion borders between different contrast. First row: subject 1. Second row: subject 2. Left: T2w MR image. Right: FLAIR MR image.

2.2.2. Image Pre-processing

Regardless of the type of automated algorithm that is adopted to perform the segmentation task, the raw MRI images initially need to go through a pre-processing pipeline that generally includes,

(i) Brain extraction to remove skull and non-brain tissues and obtain a brain mask (Eskildsen et al., 2012; Smith, 2002).

(ii) Image noise reduction to diminish the effects of sensor noise (Bao and Zhang, 2003; Coupe et al., 2008; Gerig et al., 1992).

(iii) Image intensity non-uniformity correction to correct the inhomogeneity of the static or applied magnetic fields within the scanner (Brinkmann et al., 1998; Sled et al., 1998; Van Leemput et al., 1999; Vovk et al., 2007).

(iv) Intensity range normalization to a predefined intensity range (generally 0-100) (Nyúl et al., 2000; Caligiuri et al., 2015).

For techniques that use multi-modality images, different contrasts also need to be co-registered, using a rigid body registration (Collins et al., 1994; Maes et al., 1997), and for techniques that use population characteristics, the images need to be (nonlinearly) registered to a standard coordinate system or standard brain template. Depending on the method, segmentation can be performed either in the native image space to avoid the additional blurring that is caused by resampling the images into the template space, or based on the non-linearly registered images since the original data generally has thick slices (3-5 mm).

Most automated lesion segmentation methods in the literature have been developed for detection of lesions in Multiple Sclerosis (MS) patients (García-Lorenzo et al., 2013; Mortazavi et al., 2012). These methods generally use a set of features such as multi-modality image intensities as well as normal tissue statistics and spatial priors as inputs to various classifiers to segment the lesions automatically. Such classifiers can be divided into two main categories: unsupervised and supervised. This section is divided into two sub-sections to describe unsupervised and supervised WMH segmentation methods.

2.2.3. Unsupervised WMH Segmentation Techniques

Unsupervised classifiers do not require pre-labeled data or user inputs to draw inferences. Unsupervised algorithms usually perform some form of clustering analysis to find patterns in the data, so that the data points within a cluster are closer to each other according to a similarity metric, and as distinct as possible from data points in other clusters. As a result, the clusters are obtained from an optimization driven by the data rather than a set of labels, as is the case for supervised classification.

Thresholding Techniques

Thresholding techniques are generally used in an unsupervised manner for segmenting WMHs. To achieve this, an intensity histogram is created from the optimal modality (generally FLAIR or T2w images). The central peak in the histogram of the FLAIR image is assumed to correspond to the normal tissue intensity. WMHs and CSF form the right and left tails of the histogram, respectively (Fig. 2.16).

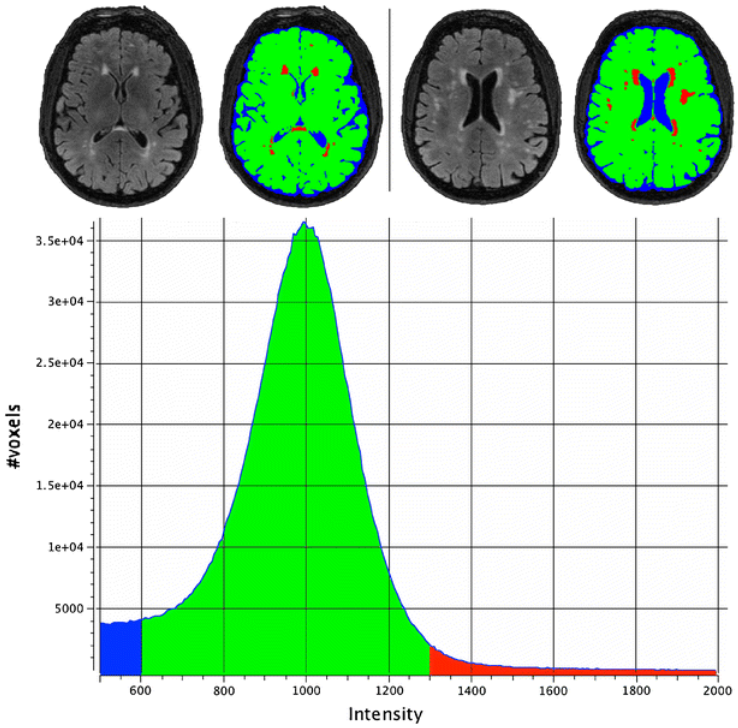


Fig. 1.16. FLAIR image histogram. Top: two axial slices of a FLAIR image and segmentation into CSF, normal tissue, and WMHs. Bottom: histogram of the image intensities for the three tissue types (Caligiuri et al., 2015). *Original figure from Caligiuri et al. 2015 (replicated under CC BY license).*

After calculating the histogram, the segmentation task involves determining a threshold value to separate WMHs from normal tissue. Jack et al. used 1/3 of the mode value of the histogram and a stepwise regression model estimated based on phantom image data to calculate the optimal cut-off threshold for differentiating WMHs from healthy tissue (Jack et al., 2001). Smart et al. use 1.45 times the modal pixel intensity in the GM+WM tissue as a threshold to detect WMHs (Smart et al., 2011). Similarly, de Boer et al. used GM segmentation results to create the histogram, which was then approximated with a Gaussian function. The threshold for WMHs was estimated as a linear function of the mean and standard deviation of this Gaussian function (de Boer et al., 2009).

Fuzzy Segmentation

Fuzzy segmentation is performed based on a fuzzy inference methodology (Mamdani and Assilian, 1975; Takagi and Sugeno, 1985), which involves,

(i) Determining a set of fuzzy rules.

(ii) Fuzzifying the inputs using membership functions.

(iii) Combining the fuzzified inputs to establish a fuzzy rule.

(iv) Combining the rule and membership functions to find the output distribution.

(v) Defuzzifying the output distribution using a threshold.

Admiraal-Behloul et al. used T2w and FLAIR image intensities (dark, medium-bright, and bright) and voxel position (intracranial and WM) as fuzzy rules and developed a fuzzy inference technique to segment WMHs (Admiraal-Behloul et al., 2005). Wu et al. initially identify lesion seeds using the image intensity histogram, and a fuzzy algorithm using intensity and adjacency information to segment lesions (Wu et al., 2006a).

Bayesian Clustering

Bayesian classification can be performed through probabilistic clustering in an unsupervised manner (i.e. without using training data to determine the priors). The conditional probability density function is generally modeled as a weighted sum of N Gaussian density functions (Gaussian Mixture Models or GMMs) and an expectation-maximization (EM) algorithm is then used to iteratively estimate the parameters of the distribution (Dempster et al., 1977). Freifeld et al. segmented MS lesions as outlier components of a GMM and used EM to perform parameter estimation (Freifeld et al., 2009). Similarly, Leemput et al. defined MS lesions as outliers to a Markov Random Field (MRF) tissue classification technique and used EM to iteratively perform lesion classification and parameter estimation (Leemput et al., 2001).

2.2.4. Supervised WMH Segmentation Techniques

Supervised techniques use manually labeled training data to generate models to segment new examples. In the following, the supervised techniques that are generally used for lesion segmentation applications are reviewed, including Bayesian classifiers, decision trees, random forests, k-nearest neighbors (k-NN), support vector machines (SVM), and AdaBoost.

Bayesian Classification

Supervised Bayesian learning uses a combination of the prior knowledge and observed data. Let x be a data point with an unknown class label. Supposing H is the hypothesis that x belongs to class y, the goal is to estimate the probability that hypothesis H is true given the observed data sample x (i.e. $p(y|x)$). MRFs can be considered as a generalization of Markov processes by replacing the time axis with a spatial axis (Kindermann and Snell, 1980). The Markov property indicates that a variable in a MRF is conditionally independent of all other variables, given its neighbors. The MRF theory provides a convenient tool to capture local spatial or contextual dependencies for highly correlated features such as image voxel intensities (Geman and Geman, 1984). MRF methods have also been popular in detecting MS lesions (Khayati et al., 2008), and WMHs in the elderly population (Schwarz et al., 2009). Sajja et al. used a Parzen window classification method for lesion segmentation in MS and minimized the false negative lesion classifications using a MRF with EM optimization (Sajja et al., 2006). One variant of MRFs is the conditional random field (CRF), in which each random variable may also be conditioned upon a set of global observations. Karimaghaloo et al. used CRFs combined with a variety of potential functions to detect MS lesions (Karimaghaloo et al., 2012).

Decision Trees

Decision trees were first proposed for performing induction by Hunt et al. (Hunt et al., 1966) and later extended by Quinlan for performing classification tasks (Quinlan, 1986). Decision tree classifiers map the feature vector to decisions about the target using a tree structure in which the leaves indicate class labels and the nodes indicate the corresponding partitionings of the feature space. The decision tree is generally constructed in 2 phases; a recursive, top-down procedure that "grows" a tree to fit the training data, and a "pruning" phase to prevent overfitting. Decision tree classifiers have since been used for tissue classification (Chao et al., 2009) and lesion segmentation in MS (Kamber et al., 1992, 1995).

Random forests

A single decision tree may have a large number of nodes and leaves and tend to overfit the training data, resulting in poor generalizability to features that are not observed by the decision tree (Breiman et al., 1984). To overcome these limitations, the concept of random forests was proposed to introduce some degree of randomization and consequently increase the generalizability of the predictions. Initially introduced by Breiman (Breiman, 2001), Random forests perform classification and regression by constructing a series of independent decision trees and using the mode or mean of their predictions as the output for classification or regression tasks, respectively. They have since been widely used for lesion segmentation in MS (Geremia et al., 2011; Maier et al., 2015; Mitra et al., 2014; Akselrod-Ballin et al., 2009) as well as for WMH segmentation in aging and AD populations (Ithapu et al., 2014).

K-nearest neighbors

The K-nearest neighbours (KNN) is a non-parametric instance based algorithm developed by Altman for classification and regression (Altman, 1992; Cover and Hart, 1967). The KNN classifier uses majority voting between the labels for the K closest data points in the feature space from the training data to assign a label to the new unseen test data. The distance metric used for determining

the closest data points is generally the Euclidian distance for continuous variables or Hamming distance for discrete variables. Due to its simplicity, KNN has been popular for various applications including segmentation of MS lesions (Wu et al., 2006b) and WMHs (Anbeek et al., 2004).

Support Vector Machines

Support vector machines (SVMs) were proposed by Boser et al. (Boser et al., 1992) to perform classification by finding a maximum-margin hyperplane that separates the two classes while maximizing the distance between the nearest points from either class. SVMs have been widely used for lesion segmentation tasks in MS populations (Abdullah et al., 2011; Ferrari et al., 2003) as well as for WMH segmentation in aging and AD populations (Ithapu et al., 2014; Lao et al., 2008; Quddus et al., 2005).

AdaBoost

Adaptive Boosting or AdaBoost was developed by Freund and Schapire (Freund et al., 1999). AdaBoost performs classification by aggregating the outputs of other classifiers (also called weak learners) into a weighted sum that determines the final output. The weak learners are iteratively updated to improve the performance on the instances that were misclassified by previous classifiers to improve classification accuracy. AdaBoost has been used for MS lesion segmentation (Wels et al., 2008), interactive lesion segmentation (Li et al., 2007), as well as segmentation of WMHs (Beare et al., 2009; Ghafoorian et al., 2016a; Quddus et al., 2005).

2.2.5. Post-processing

Many segmentation methods tend to over-segment image noise and imaging artifacts that appear hyperintense as WMHs, especially in cortical areas and around the 4th ventricle, where a large percentage of false positives are detected (Caligiuri et al., 2015). Segmentation techniques that are based on single MR contrasts are more susceptible to over-segmenting artifacts, since they will not be able to use information from other contrasts to increase the certainty of the WMH detection and

reduce the effect of noise and artifacts (Caligiuri et al., 2015). Therefore, such techniques generally need post-processing steps to remove the false positives. This post-processing generally includes,

(i) Using WM masks or other masks of regions of interest (de Boer et al., 2009; Griffanti et al., 2016; Ong et al., 2012; Simões et al., 2013; Yoo et al., 2014).

(ii) Removing lesions smaller than a certain size (Admiraal-Behloul et al., 2005; de Boer et al., 2009; Khayati et al., 2008; Sajja et al., 2006; Steenwijk et al., 2013; Yoo et al., 2014).

However, applying these post-processing techniques has the additional drawback of losing small WMHs as well as the lesions outside of the imposed WM mask. The latter can significantly affect the performance of the method since WM masks are generally obtained based solely on the T1w images, on which the WMH appear hypointense and are often mis-classified by tissue segmentation techniques as GM.

2.2.6. Validation

The common measures used for evaluating WMH segmentation methods include Dice Kappa similarity metric (SI) (Dice, 1945), intra-class correlation coefficient reflecting the volumetric correspondence between the two segmentations (ICC), sensitivity, Jaccard index (JI), detection error rate measuring agreement in detecting the same regions (DER), and outline error rate measuring agreement of the raters in outlining of the same lesion (OER) (Wack et al., 2012). Table 2.2 shows a list of these metrics as well as their definitions.

Table 2.2. List of similarity measures and their definitions. The metrics are listed in the table below using the following abbreviations: true positive (TP), true negative (TN), false positive (FP), false negative (FN), true positive rate (TPR), Mean Square Within samples based upon the ANOVA (MSW), Mean Square F Statistic Regression Slope (MSR). CR_1, CR_2, and C_{12} represent region from only rater 1, region from only rater 2, and the combination of both raters, respectively. $|cr|$ represents area of the connected region, $cr \in CR_1 \; or \; CR_2$ represents the set of connected regions that can be labeled either as CR_1or CR_2. $|R_1(cr)|$, $|R_2(cr)|$ represent the areas of rater 1 and rater 2 regions within cr, respectively (Wack et al., 2012).

Measure Name	Abbreviation	Equation
Dice Kappa	SI	$\frac{2 \times TP}{FP + FN + 2 \times TP}$
Intra-class correlation	ICC	$\frac{MSR - MSW}{MSR + MSW}$
Sensitivity	TPR	$\frac{TP}{TP + FN}$
Outline Error Rate	OER	$\sum_{cr \in C_{12}} \lvert cr \rvert - \lvert R_1(cr) \cap R_2(cr) \rvert$
Detection Error Rate	DER	$\sum_{cr \in CR_1 \; or \; CR_2} \lvert cr \rvert$
Jaccard Index	JI	$\frac{TP}{TP + FP + FN}$

2.2.7. Conclusion

Given the promising performance of the supervised classification techniques in segmenting WMHs from multi-contrast MR images, in this thesis, we further explore the use of different machine learning classifiers in accurate segmentation of WMHs from a feature set of multi-contrast MR image intensities as well as several other intensity distribution and location based features.

2.3. WMHs as Biomarkers in Alzheimer's Disease

2.3.1. Biomarkers

The term biomarker is referred to a physiological, biochemical, or anatomical variable that can be objectively, accurately, and reproducibly measured as an indicator of normal/pathological biological processes, or response to a therapeutic intervention (Jack and Holtzman, 2013). In the case of pathologies/diseases, biomarkers that can indicate abnormality before significant clinical symptoms appear can be very useful, since interventions and earlier treatments of the patients before too much irreversible damage occurs is generally more likely to be effective (Gauthier et al., 2016; Sperling et al., 2011a). In the specific case of AD, studies have shown that abnormal changes can be present in multiple biomarkers in an individual, sometimes 10-20 years before the onset of clinical symptoms (Iturria-Medina et al., 2016; Klunk et al., 2004; Sperling et al., 2011b; Storandt et al., 2009). To facilitate the development of therapies to prevent or slow down the progression of AD, in vivo biomarkers that can enable accurate diagnosis of the patients in earlier asymptomatic stages of the disease as well as monitoring of the progression and effectiveness of the therapies are highly advantageous.

The clinical stages of AD have been divided into three phases (Jack Jr et al., 2010):

(i) Pre-symptomatic phase: individuals in this phase are cognitively normal, but have some AD related pathological changes. The hypothetical assumption is that many of these individuals will eventually develop AD later in life, if they live long enough.

(ii) Prodromal AD or mild cognitive impairment (MCI): individuals in this phase present with early cognitive impairment symptoms, typically episodic memory deficits. However, these deficits are not severe enough to meet the criteria for dementia.

(iii) AD dementia: individuals in this phase show severe impairment in multiple cognitive domains.

Imaging and cerebrospinal fluid (CSF) biomarkers of AD can be used to predict the disease stage and progression to AD. These biomarkers can be divided into two categories: biomarkers of Aβ deposition including CSF $A\beta_{1-42}$, positron emission tomography (PET) Aβ imaging (both CSF $A\beta_{1-42}$ and PET Aβ are correlates of Aβ plaque deposition), and biomarkers of pathological changes and neurodegeneration including CSF tau (a correlate of presence of neurofibrillary tangles), FDG PET (measuring brain glucose metabolism, reflecting synaptic dysfunction), and structural MRI (measuring cerebral atrophy). Using these biomarkers, individuals that will eventually progress to AD can be identified in earlier stages of the disease (Jack Jr et al., 2010).

An important difference between Aβ biomarkers and biomarkers of neurodegeneration is their specificity to AD. CSF $A\beta_{1-42}$ and amyloid PET are specific for Aβ deposition, while neurodegenerative biomarkers are not always specific to AD (Jack and Holtzman, 2013). Another important element of biomarker-based modeling of AD is the temporal sensitivity of different biomarkers.

2.3.2. Models of Disease Progression

Biomarkers can be used to propose and study hypothetical models of disease progression. In such models, the extent of abnormality in biomarker levels compared to their expected values in normal populations reflects disease progression.

Jack et al. used evidence of biomarker abnormality to order different imaging and clinical biomarkers in AD progression (Jack Jr et al., 2010). They proposed that the earliest AD biomarkers are Aβ, and Tau-mediated neuronal injury and dysfunction, followed by brain

structural changes, memory deficits, and clinical function impairment (Fig. 2.17). In this model, it was simplistically assumed that all the biomarkers have a sigmoidal shape as a function of time with the same slope, and approach a plateau as the individual progresses to AD.

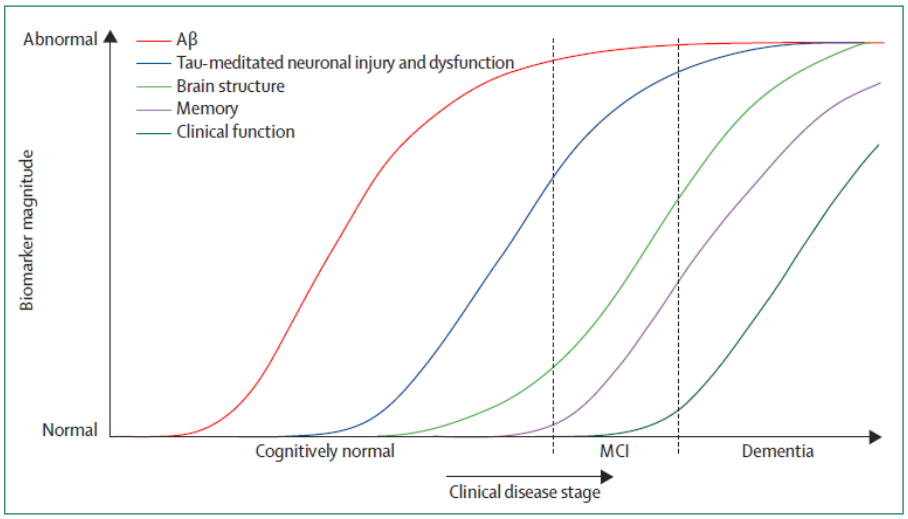


Fig. 2.17. Biomarkers of AD pathological cascade (Jack Jr et al., 2010). Aβ indicates CSF $A\beta_{42}$ or amyloid PET. Tau-mediated neuronal injury and dysfunction indicates CSF tau or FDG-PET. Brain structure indicates atrophy measured from structural MRI. AD= Alzheimer's Disease. Aβ=amyloid β. MCI=mild cognitive impairment. FDG= fluorodeoxyglucose. PET= Positron Emission Tomography. CSF= Cerebrospinal Fluid. *Original figure reproduced with permission from Jack Jr et al. 2010.*

Different brain regions are known to be affected at different stages of the disease in AD. Since FDG PET and structural MRI measure regional information, they can provide regional biomarkers, reflecting different stages of AD progression with potentially higher sensitivity. As seen in Fig. 2.18, Jack et al. suggest that posterior cingulate, lateral temporal cortex, and frontal lobe are affected first as seen on FDG-PET and that later, structural MRI shows AD-related changes in medial temporal, lateral temporal, and frontal lobes.

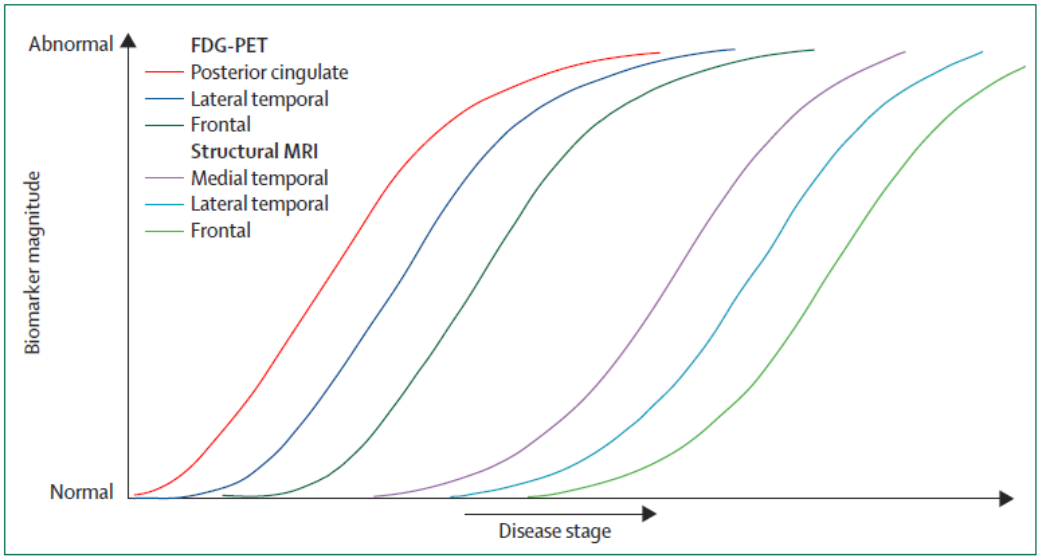


Fig. 2.18. Regional anatomical imaging biomarkers of AD pathological cascade (Jack Jr et al., 2010). AD= Alzheimer's Disease. FDG=fluorodeoxyglucose. PET= Positron Emission Tomography. *Original figure reproduced with permission from Jack Jr et al. 2010.*

In a later study, Jack et al. refined their initial model by allowing different slopes for different biomarkers, since later evidence showed that MRI and FDG-PET curves continue to significantly change in AD patients (Jack Jr et al., 2013). The horizontal axis was also changed to time in years (instead of disease stage) to enable traversing the disease pathway at a specific time. Age was not used as the horizontal axis since the specific age when a person starts this process can vary considerably. In addition, a range of possible cognitive trajectories was considered for the MCI stage, since different individuals with the same level of AD-related pathology have a range of possible cognitive performances, depending on genetic risk factors, cognitive reserve, lifestyle, or other comorbid pathological changes (Fig. 2.19). At any given time (T), CSF $A\beta_{42}$ is the most abnormal biomarker, followed by CSF Tau, Amyloid PET, MRI+FDG PET, and cognitive impairment. The cognitive impairment curve is shifted to the left (right) for subjects that are at high (low) risks of AD-related pathologies.

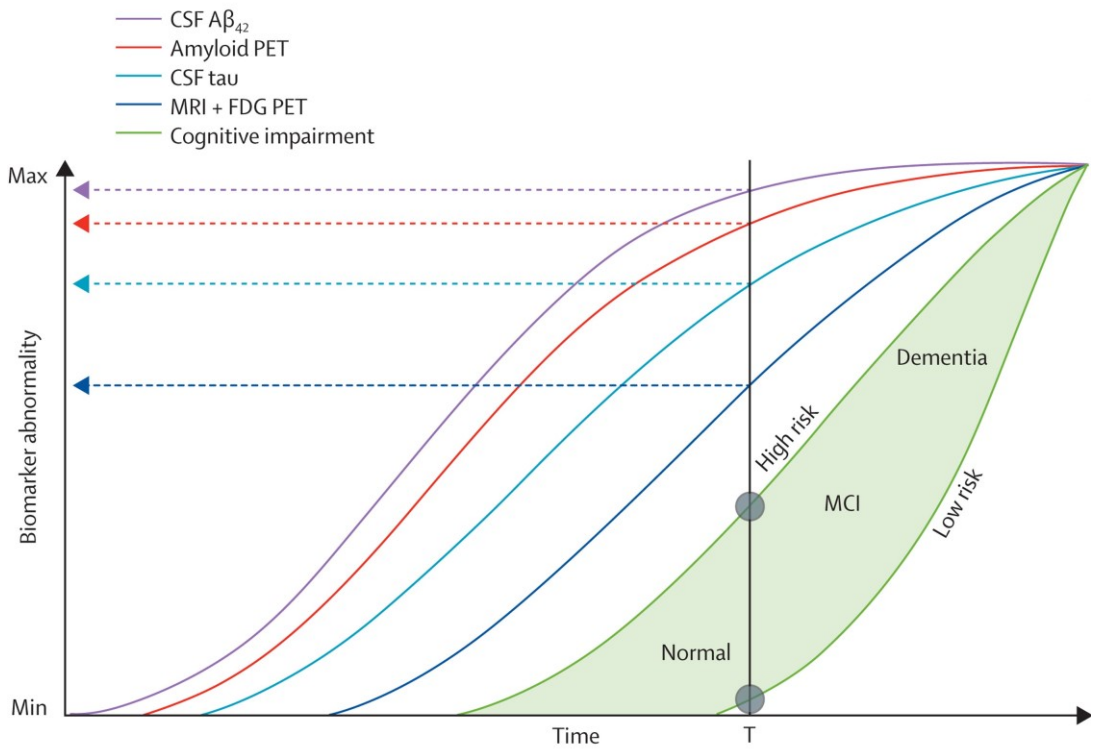


Fig. 2.19. Revised model of biomarkers of AD pathological cascade (Jack Jr et al., 2013). MRI+FDG PET indicates neurodegeneration. AD= Alzheimer's Disease. Aβ=amyloid β. FDG=fluorodeoxyglucose. PET= Positron Emission Tomography. MCI=mild cognitive impairment. *Original figure reproduced with permission from Jack Jr et al. 2013.*

More recently, Iturria-Medina et al. have used a multifactorial data-driven analysis approach to further validate and improve the model initially proposed by Jack et al., where changes in Aβ, metabolism, vascular regulation, resting state functional activity, structural MRI and levels of various protein biomarkers during the progression of AD are ordered according to their temporal characteristics (Iturria-Medina et al., 2016). They demonstrated that under the assumption that biomarkers represent physiological processes, cerebrovascular dysregulation is the earliest pathologic biomarker associated with progression to AD, followed by Aβ deposition, glucose metabolism dysregulation, functional impairment, and GM atrophy (Fig. 2.20). In a more recent study, using six different neuroimaging modalities, they further confirm that vascular dysregulation may be the most-likely initial pathologic event leading to AD, followed by functional dysregulation, glucose metabolism impairment, Aβ deposition/propagation and structural atrophy (Iturria-Medina et al., 2017).

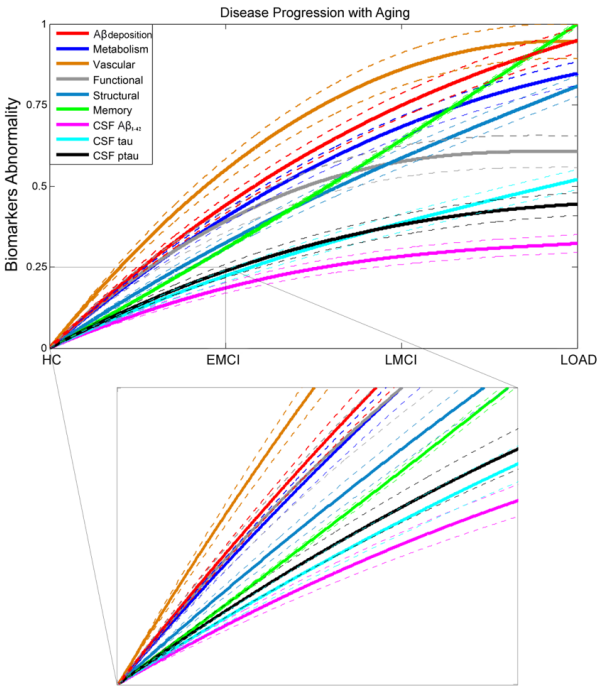


Fig. 2.20. Data-driven model of LOAD progression (Iturria-Medina et al., 2016). LOAD=Late onset Alzheimer's disease. CSF= Cerebrospinal Fluid. Aβ=amyloid β. *Original figure from Iturria-Medina et al. 2016 (replicated under CC BY license).*

2.3.3. WMHs as Biomarkers in Diagnosis and Prognosis

In addition to modeling disease progression, different clinical and imaging biomarkers can be used in combination with machine learning classifiers to predict the current (diagnosis) or future (prognosis) status of a subject. Similarly, they can be used in combination with regression techniques to predict the rate or amount of cognitive decline of the subjects in the following years. Here we provide a short review of the few studies that use WMH-related features to predict the current or future status of subjects in prediction tasks specifically for AD.

In an elderly cohort (age at baseline 79.55±5.20 years), Brickman *et al*. showed that WMH load in the parietal lobes (and not hippocampal atrophy) predicted AD, suggesting a primary role of CSVD in AD that is independent of hippocampal atrophy (Brickman et al., 2012). Later, they report that WMH load was associated with increased frequency of AD,

independent of APOE ε4 status. In addition, APOE ε4 carriers were more likely to have AD if they also had increased parietal WMH levels (Brickman et al., 2014).

In a later study, they used a structural equation modeling framework to show that smaller baseline hippocampal volume, change in hippocampal volume (i.e., hippocampal atrophy), higher baseline parietal lobe WMH, and increasing parietal lobe WMH volume independently predicted progression to AD. They also report that higher baseline WMH volumes were associated with rapid increases in the WMH volumes in follow-up visits (Brickman et al., 2015).

In a study to assess the impact of CSVD as well as amyloid pathology on the clinical expression of AD, Provenzano et al. showed that Pittsburgh compound B (PIB) positivity and increased total WMH volume independently differentiated AD, MCI, and NC subjects, using a logistic regression classifier. In the PIB-positive cohort, AD patients had greater WMH loads than NC subjects. Using a cut-off threshold of 1.25 cm^3 for the total WMH volume, they obtained a sensitivity and specificity of 83% and 64%, respectively for AD versus NC classification. In the MCI cohort, both WMH and PIB status at baseline increased the risk for conversion to AD (Provenzano et al., 2013). Using longitudinal data with a mean follow-up of 29.73±12.75 months, they found a significant increase in the proportion of MCI patients who converted to AD across PIB^+/PIB^- and WMH_{High}/WMH_{Low} groups. Results from a logistic regression analysis showed that PIB/WMH status and the time between baseline and diagnosis or latest visit (but not age) were associated with conversion to AD (Provenzano et al., 2013).

In a similar study, Gordon et al. have shown that deep and periventricular Fazekas WMH scores differentiate between cognitively normal and demented individuals,

independent of Aβ level (Table. 2.3), suggesting that WMH scales can be used in addition to Aβ for diagnosis of AD (Gordon et al., 2015).

Table. 2.3. Results of a logistic regression model, examining the effects of WMHs on cognition. WMHs=White Matter Hyperintensities. CBP = mean Cortical Binding Potential. PVWMH = Periventricular White Matter Hyperintensity. DWMH = Deep White Matter Hyperintensity (Gordon et al., 2015). *Table reproduced from Gordon et al. 2015* (*under CC BY-NC-ND license).*

		B	Std. error	*p*	Odds ratio
CDR 0 vs. CDR>0	Intercept	-9.16	1.67	0.00001	
	Age	0.07	0.02	0.001	1.08
	Gender	1.25	0.34	0.001	3.50
	MCBP	1.30	0.25	0.0001	3.68
	PVWMH	0.71	0.19	0.001	2.03
CDR 0 vs. CDR>0	Intercept	-10.44	1.64	0.0001	
	Age	0.10	0.02	0.001	1.10
	Gender	1.28	0.34	0.001	3.36
	MCBP	1.29	0.25	0.0001	3.63
	DWMH	0.42	0.20	0.036	1.52

Lindemer et al. observed that WMH load was considerably higher in AD subjects in comparison with MCI and NC cohorts, but was not significantly different between converter and non-converter MCI populations. Furthermore, they defined the Mahalanobis distance (MD) of WMHs from normal appearing white matter (NAWM) using T1w, T2w, and PD scans as MD quality and showed in a longitudinal analysis, that the MD quality changes faster in the converter MCI cohort in comparison to the matched non-converters, from 18 months before conversion to AD (Fig. 2.21). The strongest difference occurred in the period from 6 months before to 6 months after conversion to AD. The rate of change in MD quality was similar to the rate of change in hippocampal volume for the same period. They suggested that WMHs are a critical component for conversion to AD (Lindemer et al., 2015).

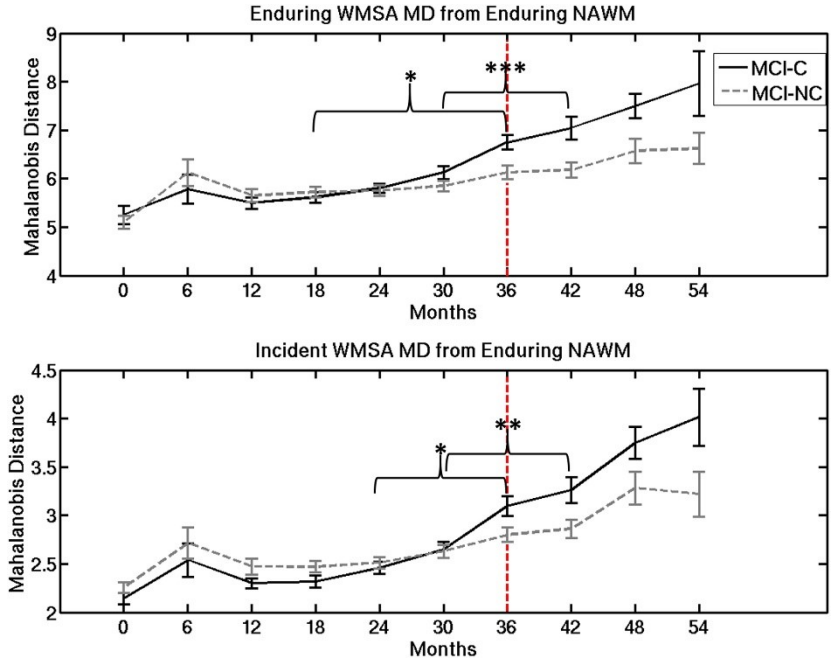


Fig. 2.21. Longitudinal changes in differences of enduring WMSA MD from enduring NAWM (top) and incident WMSA from enduring NAWM (bottom) in converter and non-converter MCI subjects (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.0001$). Red lines indicate the time of conversion to AD for converter MCI subjects. MCI-C= Mild Cognitive Impairment Converters. MCI-NC= Mild Cognitive Impairment Non-Converters. MD= Mahalanobis Distance. NAWM= Normal-Appearing White Matter. WMSA= White Matter Signal Abnormality (Lindemer et al., 2015). *Original figure reproduced with permission from Lindemer et al. 2015.*

In an attempt to predict conversion to AD from baseline to 24 months, Callahan et al. investigated whether using WMH loads in combination with other MRI and CSF biomarkers would improve prediction accuracy, using a logistic regression classifier. Although WMH loads and hippocampal volumes were significantly different between the converter and non-converters, adding either of them to the episodic memory test score features did not significantly increase the prediction accuracy (Callahan et al., 2015).

2.3.4. Conclusion

WMHs are clinically significant biomarkers of CSVD in aging as well as neurodegenerative diseases (Carmichael et al., 2010; DeCarli et al., 1995a; Pantoni et al., 2006; van Straaten et al., 2008). Using WMHs, as one of the major signs of vascular

dysregulation, can improve the predictive power of models of disease progression and increase the diagnosis and prognosis accuracy of prediction models in AD (Brickman et al., 2012; Deoni et al., 2013; Provenzano et al., 2013).

CHAPTER 3. Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease

Preface

In this chapter, we propose and validate a linear regression technique for segmenting WMHs from multi-contrast MR images. The proposed method combines a series of intensity and location features along with a spatial prior to detect WMHs from T1w, FLAIR and T2w/PD sequences if available. The different MR sequences provide complementary information that will reduce uncertainty and increase segmentation accuracy.

The performance of the proposed technique is verified on 3 different datasets, one of which consists of subjects from the multi-site and multi-scanner ADNI study. The ADNI dataset is used as an independent validation set over which the performance of different techniques is compared to assess their generalizability to new previously unseen data from other scanners. The results show that although nonlinear classification techniques such as AdaBoost and Random Forests have significantly higher performances on data from the same scanner, when validated on data from a different scanner, their performance drops drastically, compared to the proposed linear technique.

This work has been published as:

Dadar, Mahsa, Tharick A. Pascoal, Sarinporn Manitsirikul, Karen Misquitta, Carmela Tartaglia, John Brietner, Pedro Rosa-Neto, Owen Carmichael, Charles DeCarli, and D. Louis Collins. "Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease." *IEEE Transactions on Medical Imaging* (2017).

Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease

Mahsa Dadar, Tharick A. Pascoal, Sarinporn Manitsirikul, Karen Misquitta, Vladimir S. Fonov, M. Carmela Tartaglia, John Breitner, Pedro Rosa-Neto, Owen T. Carmichael, Charles Decarli, D. Louis Collins

Abstract

Segmentation and volumetric quantification of white matter hyperintensities (WMHs) is essential in assessment and monitoring of the vascular burden in aging and Alzheimer's disease (AD), especially when considering their effect on cognition. Manually segmenting WMHs in large cohorts is technically unfeasible due to time and accuracy concerns. Automated tools that can detect WMHs robustly and with high accuracy are needed. Here we present and validate a fully automatic technique for segmentation and volumetric quantification of WMHs in aging and AD. The proposed technique combines intensity and location features from multiple magnetic resonance imaging (MRI) contrasts and manually labeled training data with a linear classifier to perform fast and robust segmentations. It provides both a continuous subject specific WMH map reflecting different levels of tissue damage and binary segmentations. The method was used to detect WMHs in 80 elderly/AD brains (ADC dataset) as well as 40 healthy subjects at risk of AD (PREVENT-AD dataset). Robustness across different scanners was validated using 10 subjects from ADNI2/GO study. Voxel-wise and volumetric agreements were evaluated using Dice similarity index (SI) and intra-class correlation (ICC), yielding ICC=0.96, SI=0.62±0.16 for ADC dataset and ICC=0.78, SI=0.51±0.15 for PREVENT-AD dataset. The proposed method was robust in the independent sample yielding SI=0.64±0.17 with ICC=0.93 for ADNI2/GO subjects. The proposed method provides fast, accurate and robust segmentations on previously unseen data

from different models of scanners, making it ideal to study WMHs in large scale multi-site studies.

Index terms- White matter hyperintensities, segmentation, aging, Alzheimer's disease

3.1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia that currently affects 44 million people worldwide and is increasing in prevalence (Prince et al., 2014). AD is clinically characterized by gradual and progressive decline in memory as well as other cognitive functions. The hallmark neuropathology of AD consists of extracellular deposition of amyloid β plaques and intracellular neurofibrillary tangles made of tau (Selkoe, 2005). In addition to these major contributing factors, accumulating evidence shows that progressive loss of white matter integrity due to the loss of axons and their neurons, synapses and dendrites plays an important role in the development of AD (Lee et al., 2016). Very often and with a higher prevalence among older subjects, AD co-occurs with cerebral small vessel disease (CSVD), hypertension, hypercholesterolemia and diabetes. Such subjects typically present additional deficits in comparison with AD in subjects without these co-morbidities (Dubois et al., 2014). CSVD is represented on MRI as white matter hyperintensities (WMHs). There is accumulating evidence that the WMH load is related to ischemic damage along with microbleeds and lacunar infarcts (Conklin et al., 2014; Sam et al., 2016; Gouw et al., 2010). WMHs can also be associated with other underlying mechanisms, such as dilation of perivascular spaces in the frontal and/or parietal subcortical white matter (DeCarli et al., 1995b), increased extracellular spaces, glial cell responses, vessel wall leakage, and collagen deposition in the vessel walls. WMHs are highly prevalent in AD patients as well as the elderly population in general. They primarily occur adjacent to the cerebral ventricles, especially around the posterior horns of the lateral ventricles (DeCarli et al., 1995b).

Clinical studies commonly distinguish between periventricular WMHs and WMHs in the deep white matter tissue. The former are identified with thin hyperintense lines, smooth halos or irregular bands/caps around the ventricles while the latter are categorized as punctate, early confluent and confluent WMHs (Gouw et al., 2010). While mild periventricular WMHs are often seen in elderly individuals with no clinical symptoms, larger periventricular WMHs volumes have been reported to be associated with gait difficulties and lower motor performance (Silbert et al., 2008). Furthermore, the total volume of subcortical WMHs has been associated with decline in cognition and faster rate of memory decline, even after adjusting for rate of cerebral or hippocampal atrophy (Yoshita et al., 2005). This evidence suggests that accounting for the WMH burden in addition to the AD related pathologies can improve prediction of memory and cognitive decline.

Manual segmentation of WMHs is generally performed on Fluid-attenuated inversion recovery (FLAIR) MR images by expert raters. Accurate and consistent segmentation of WMHs is a complicated task due to the heterogeneity in their texture and pattern as well as the fact that these lesions often have fuzzy borders. Manually detecting WMHs is challenging, time consuming, expensive and inconsistent due to inter-rater and intra-rater variability. As a result, inter-rater and intra-rater agreement is generally modest at best (García-Lorenzo et al., 2013), since the boundary between WMH and non-WMH tissue is difficult to determine precisely and different raters draw different arbitrary distinctions between the two, whereas automated methods always apply the same policy to this distinction. In addition, the huge number of images being collected makes the human cost of manual identification prohibitive. These make automated segmentation tools that can detect WMHs robustly and with high sensitivity and specificity highly advantageous since with their objectivity and reproducibility they would essentially eliminate the intra-rater variability and make it possible to follow individual subjects over time, or segment WMHs in large scale

studies with 1000s of subjects, (e.g. clinical trials). The MRI contrasts that are commonly used in detecting WMHs include T1w (mostly used for co-registration purposes) on which WMHs appear hypointense, and T2w, proton density (PD), and FLAIR on which WMHs appear hyperintense. Since different MRI modalities have different contrasts across tissues, integrating information from multiple modalities can reduce uncertainty and consequently increase segmentation accuracy.

Most automated lesion segmentation methods in the literature have been developed for detection of lesions in Multiple Sclerosis (MS) patients (Mortazavi et al., 2012), (García-Lorenzo et al., 2013). These methods generally use a set of features such as multi-modality image intensities as well as normal tissue statistics and spatial priors and input this information into various classifiers to segment the WMHs automatically. Such classifiers can be divided into two main categories: unsupervised and supervised. Unsupervised classifiers do not require labeled data to draw inferences. Such algorithms usually perform some form of clustering analysis to find patterns in the data. Thresholding techniques are generally in this category. To detect WMHs, Jack et al. used a histogram segmentation of FLAIR images by finding a cut-off threshold for differentiating WMHs from normal tissue (Jack et al., 2001). Similarly, de Boer et al. used tissue segmentation results to automatically find an optimal threshold for WMHs in FLAIR images (de Boer et al., 2009). Smart et al. use 1.45 times the modal pixel intensity after skull stripping as a threshold to detect WMHs and removed isolated pixels from the segmentation afterwards (Smart et al., 2011). Admiraal-Behloul et al. combined multispectral intensity images with tissue spatial distribution probability maps and used a fuzzy inference technique to segment WMHs (Admiraal-Behloul et al., 2005). Wu et al. initially identify lesion seeds using the image intensity histogram, and a fuzzy connected algorithm to segment lesions and iteratively update seeds (Wu et al., 2006a). Leemput et al. defined and detected MS lesions as outliers to a Markov Random Field tissue classification

technique (Leemput et al., 2001). Freifeld et al. used a similar approach and segmented MS lesions as outlier components of a Gaussian mixture model (Freifeld et al., 2009).

While unsupervised techniques are favored since they do not require manual segmentations for the initial training, they are usually outperformed by supervised methods since the former often over-segment imaging artifacts as lesions (i.e. flow artifacts in the 4th ventricle) and need extensive post-processing to remove false positives (Caligiuri et al., 2015). Supervised techniques use manually labeled training data to draw inference. The supervised techniques that are generally used for lesion segmentation applications include k-nearest neighbors (k-NN), regression classifiers, graph cuts, neural networks, Bayesian classifiers, and support vector machines (SVM). Anbeek et al. used a k-NN technique to segment white matter lesions from a feature space of voxel intensities and spatial information (Anbeek et al., 2004). Similarly, Steenwijk et al. optimized intensity normalization and used spatial tissue type priors to improve k-NN classification of WM lesions (Steenwijk et al., 2013). Wu et al. combined an intensity-based statistical k-NN method with template-driven segmentation and partial volume artifact correction to segment MS lesions (Wu et al., 2006b). Garcia-Lorenzo et al. used an automated graph cuts method with expectation maximization to segment MS lesions (García-Lorenzo et al., 2009). Zijdenbos et al. used intensity information, spatial priors and neural networks to obtain a classification algorithm for MS WMHs (Zijdenbos et al., 2002). Mechrez et al. used a multichannel spatially consistent path-based technique to segment MS lesions (Mechrez et al., 2016). Beare et al. used morphological segmentation and an adaptive boosting statistical classifier, obtaining a two-phase method (Beare et al., 2009). First, they used a morphological watershed to produce overly inclusive segmentations of WMHs. In the second phase, they used statistical classifiers to distinguish between real and false WMHs by examining the properties of each region. There has also been major interest in using Bayesian classifiers with Markov random field

methods to detect WMHs in MS (Khayati et al., 2008), and the elderly population (Schwarz et al., 2009). Sajja et al. used a Parzen window classification method for lesion segmentation in MS and minimized the false negative lesion classifications using HMRF-EM (hidden Markov Random Field with expectation maximization) (Sajja et al., 2006). Karimaghaloo et al. used a conditional random field method and combined a variety of potential functions to detect lesions with various shapes (Karimaghaloo et al., 2012). Lao et al. have used support vector machines (SVMs) to create a classification algorithm for detecting WMHs (Lao et al., 2008). Ghafoorian et al. have developed a technique for detecting WMHs in CSVD across a large sample of patients by separating small and large lesions and training two size-specific AdaBoost classifiers to detect these lesions (Ghafoorian et al., 2016a). The lesion growth algorithm (LGT), a publicly available tool for segmentation of MS lesions from 3T T1w and FLAIR images by Schmidt et al. uses FLAIR intensity distribution in tissue classes to detect outliers which are then expanded toward a more liberal segmentation under certain conditions (Schmidt et al., 2012). Ithapu et al. have also developed a publicly available MATLAB toolbox for segmentation of WMHs in AD and aging by combining texture features generated by filter banks and SVM and Random Forests classifiers (Ithapu et al., 2014).

Although many different lesion segmentation techniques have been proposed, most methods have been trained and validated using data obtained from small populations, all scanned with the same MRI imaging protocol. This simplifies the problem greatly, and may lead to overfitting. As a result, these techniques cannot be widely used for other datasets due to the unreliability and high variability of results across data that is scanned with different acquisition protocols (García-Lorenzo et al., 2013; Caligiuri et al., 2015). Also, methods that have been designed for lesion detection in MS populations do not generally perform as well in segmenting WMHs in the elderly populations for two main reasons. First, the MRI contrast between gray matter and white matter tissues decreases with age. Second, the boundaries of

MS lesions are generally sharper than those of WMHs, which makes the segmentation task more challenging for the latter (Caligiuri et al., 2015). Due to these limitations, despite the number of proposed methods, an optimal algorithm has not yet been identified, leaving lesion segmentation in general and WMH segmentation in particular an open problem (García-Lorenzo et al., 2013; Caligiuri et al., 2015).

The goal of this study is to validate a robust and generalizable automatic technique for segmentation of WMHs in MRIs from elderly subjects and patients with AD to assess and monitor their vascular burden. To achieve this goal, we have investigated the performance of our technique across three different populations with different scanners and acquisition protocols. In this paper, our novel contributions are:

(i) To describe a set of discriminative features to identify WMHs

(ii) To describe a processing pipeline that implements a linear regression classifier

(iii) Evaluation on three heterogeneous multi-site datasets, including images scanned by different scanners and different scan-parameters to show robustness

(iv) To obtain results that are as good or better to previously published results

(v) To compare our classifier to publicly available FSL, SPM, and W2MHs WMH segmentation tools

3.2. Materials and methods

3.2.1. Subjects

The method was implemented and validated based on 3 datasets to ensure robustness and generalizability.

(i) The first dataset (ADC) consists of 80 elderly individuals who received a full clinical workup and structural MR scans including T1w, double-echo PD/T2w, and FLAIR scans at their times of enrollment into the University of California, Davis Alzheimer's Disease Center (ADC) (Hinton et al., 2010). Subjects were 70-90 years old with either normal cognition, mild cognitive impairment (MCI), or AD.

(ii) The second dataset (PREVENT-AD) consists of 40 cognitively normal subjects at risk of AD aged 55-75 years obtained from "Pre-symptomatic Evaluation of Novel or Experimental Treatments for Alzheimer's Disease" program data release 1.0., a longitudinal cohort study of healthy persons with a parental history of AD dementia. The PREVENT-AD subjects had T1w, T2*, and FLAIR MRIs (Tremblay-Mercier et al., 2014).

(iii) The final dataset includes T1w and FLAIR scans of 10 subjects, selected to have different loads of WMHs from ADNI2/GO study which was used to show the performance of the method on independent data from different scanners that was not previously used in the training and parameter optimization of the method. This data was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial MRI and other biomarkers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

3.2.2. MR imaging

We evaluated the proposed technique on datasets from three studies that were acquired with different MR contrasts to show the robustness of the classifier. This section describes scanner information and image acquisition parameters for the abovementioned datasets. Table 3.1. shows the summary of this information for each dataset.

(i) ADC: MRI data was acquired on two 1.5T MRI scanners: a GE MEDICAL SYSTEMS Signa scanner located at UCD Medical Center (Sacramento, CA), and a Philips Eclipse scanner located at the Veterans Administration Northern California Health Care System (Martinez, CA). Analogous sequences were installed on both scanners.

(ii) PREVENT-AD: MRI data was acquired on a 3T SIEMENS MAGNETOM TrioTim syngo MR scanner (version B17). All patients had the same MRI protocol for T1w, T2* and FLAIR scans.

(iii) ADNI2/GO: The MRI data used was acquired on two different models of GE MEDICAL SYSTEMS scanners: Signa HDxt, and DISCOVERY MR750. All patients had similar MRI protocols for T1w and FLAIR scans, acquired with gradient-recalled echo and spin echo inversion recovery sequences, respectively.

Table 3.1. MRI acquisition parameters for ADC, PREVENT-AD, and ADNI2/GO datasets.

	Parameter (unit)	ADC	PREVENT-AD	ADNI2/GO
T1w	Slice thickness (mm)	1.5	1	1.2
	No. of slices	128	176	196
	Field of view (cm^2)	250×250	256×256	256×256
	Scan Matrix (cm^2)	256×256	256×256	256×256
	TR: Repetition time (ms)	9	2300	7.2
	TE: Echo time (ms)	2.9	2.98	3.0
	Pulse Sequence	FSPGR	IR	MPRAGE
T2w/T2*	Slice thickness (mm)	3	2	
	No. of slices	42	52	
	Field of view (cm^2)	240×240	200×200	
	Scan Matrix (cm^2)	256×256	512×512	
	TR: Repetition time (ms)	2420	650	
	TE: Echo time (ms)	90	20	
	Pulse Sequence	DSE	IR	
PD	Slice thickness (mm)	3		
	No. of slices	42		
	Field of view (cm^2)	240×240		
	Scan Matrix (cm^2)	256×256		
	TR: Repetition time (ms)	2420		
	TE: Echo time (ms)	20		
	Pulse Sequence	DSE		
FLAIR	Slice thickness (mm)	3	1	5
	No. of slices	48	176	42
	Field of view (cm^2)	220×220	256×256	256×256
	Scan Matrix (cm^2)	256×192	256×256	256×256
	TR: Repetition time (ms)	11000	5000	11000
	TE: Echo time (ms)	144	388	150
	Pulse Sequence	FSE	IR	SE/IR

Label segmentation: For all datasets, the WMHs were segmented independently based solely on the FLAIR scans by raters who were blinded to clinical symptoms of the subjects. Three different manual segmentation techniques were used:

(i) ADC: a strongly validated, semi-automated method was used to detect WMHs based on the FLAIR scans and human input (Yoshita et al., 2005). In short, a threshold-based automated method identified potential WMH lesions and the expert rater eliminated false positives.

(ii) PREVENT-AD: WMHs were manually segmented by two experts using FLAIR images. Union of the two segmentations was then used as the gold standard. Periventricular and deep WMHs were identified with different labels, and thus enabled a comparison between segmenting all lesions together or segmenting these two classes of lesions separately (but only in the PREVENT-AD cohort).

(iii) ADNI2/GO: an expert rater manually painted the lesions on the native FLAIR scans. The manual segmentations were then reviewed and corrected by a second investigator.

The cohorts presented large ranges of lesion loads: ADC (0.50-40.3 CCs), PREVENT-AD (0.29-23.6 CCs), and ADNI2/GO (3.56-128.12 CCs). In the experiments below, we evaluated the performance of the classifier across 3 different white matter lesion loads (WMLL): large (WMLL > 20 CCs), medium (5-20 CCs) and small (WMLL < 5 CCs). Fig. 3.1. shows the number of subjects in the different categories for each dataset.

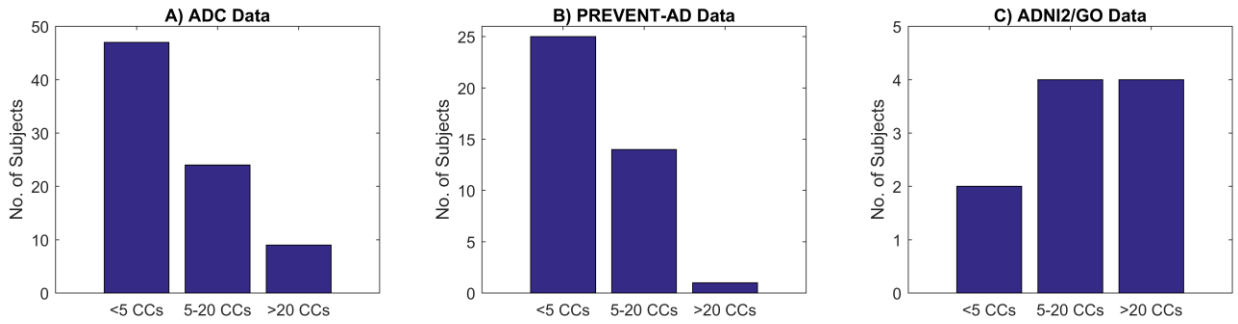


Fig. 3.2. Histograms of WMH load ranges for the 3 datasets (<5 CCs, 5-20 CCs, and >20 CCs). A) ADC B) PREVENT-AD C) ADNI2/GO.

3.2.3. Pre-processing

All MRI scans were pre-processed using our standardized pipeline. Images were denoised using an automatic and multithreaded denoising method based on non-local means filtering (Manjón et al., 2010). The bias field and intensity inhomogeneity were estimated and corrected using a nonparametric non-uniform intensity normalization (N3) tool (Sled et al., 1998). The final preprocessing step included linear intensity scaling using histogram matching to a template obtained from 150 subjects (50 normal control, 50 mild cognitively impaired and 50 dementia subjects) in the ADNI database (www.loni.ucla.edu\ADNI) (Fonov et al., 2011a). The T2w, PD, and FLAIR scans were then coregistered to the structural T1w scan of the same subject using a six-parameter rigid body registration (Collins et al., 1994). The T1w scans were nonlinearly registered to the ADNI template based on intensity correlation coefficient (Collins and Evans, 1997). Using the T1w-to-template transformations (i.e., linear + nonlinear), the other modalities (e.g., FLAIR, T2w, PD) were registered to the ADNI template as well. The manually segmented lesion maps were also registered to the ADNI template using the transformations of their corresponding FLAIR images.

3.2.4. Features

In order to reduce the feature space dimension and consequently the computational burden, each image voxel was treated as a separate data point. A feature set was defined based on a variety of intensity and probability parameters. The following features were used as inputs to the classifier:

(i) Voxel intensity for each of the available modalities, e.g. T1w, T2w, PD, and FLAIR

(ii) Spatial probability, i.e. the probability of the voxel in its specific location being a WMH

(iii) Intensity probability of the normal healthy and WMH tissues independently for each modality (P_H and P_{WMH}), i.e. the probability of the voxel with its specific intensity being normal healthy or WMH tissue, calculated for each modality independently

(iv) Average intensity of healthy tissue at voxel for each modality

(v) The probability of each voxel being a WMH divided by the probability of it being healthy tissue obtained from the intensity probabilities for the different modalities ($\frac{P_{WMH}}{P_H}$), calculated for each modality independently

All the features (except for the MRI intensities) were calculated based on training data in the cross validation step to avoid overfitting. The intensity probabilities of WMH and normal healthy tissues (P_H and P_{WMH}) were obtained by calculating histograms of intensity ranges within the manually segmented WMH masks and non-WMH brain regions, respectively.

The intensity features from the different MRI contrasts are generally used in all WMH segmentation techniques, as they provide basic intensity information for the specific voxel. The spatial probability feature can inform the classifier of how likely it is for the voxel in this specific location in the brain to be a lesion, e.g. a hyperintense voxel in the periventricular regions is more likely to be a lesion, whereas a voxel with a similar intensity in the cortical regions is less likely to be so. This feature is most informative when the training dataset is large and reflects the prevalence of WMH across different brain regions accurately. The intensity probability features reflect the likelihood of the abnormality of the intensity of the current voxel, i.e. how likely it is for a voxel with such intensity to be either WMH or normal

tissue. The division of these two features can distinguish the tails of the distributions and provide yet another measure to reflect the likelihood of being a lesion. The average intensity of the healthy tissue feature can provide a standard of what intensity is considered normal in this specific location of the brain.

Using the information from multiple contrasts can decrease uncertainty and increase classification accuracy, especially in cases where one modality has certain artifacts. For example, proximity to bones might cause an increase in the signal in the optimal FLAIR image due to susceptibility. As a result, there may be non-WMH voxels that are hyperintense on the FLAIR image, but not on the other contrasts. Integrating the information from multiple contrasts can eliminate these false positives.

Since different features have different ranges, feature normalization was performed by variance scaling, i.e. subtraction of the mean and division by standard deviation. This results in zero mean and unit variance in the normalized feature set. Fig. 3.2. shows the flowchart for the preprocessing and feature selection steps.

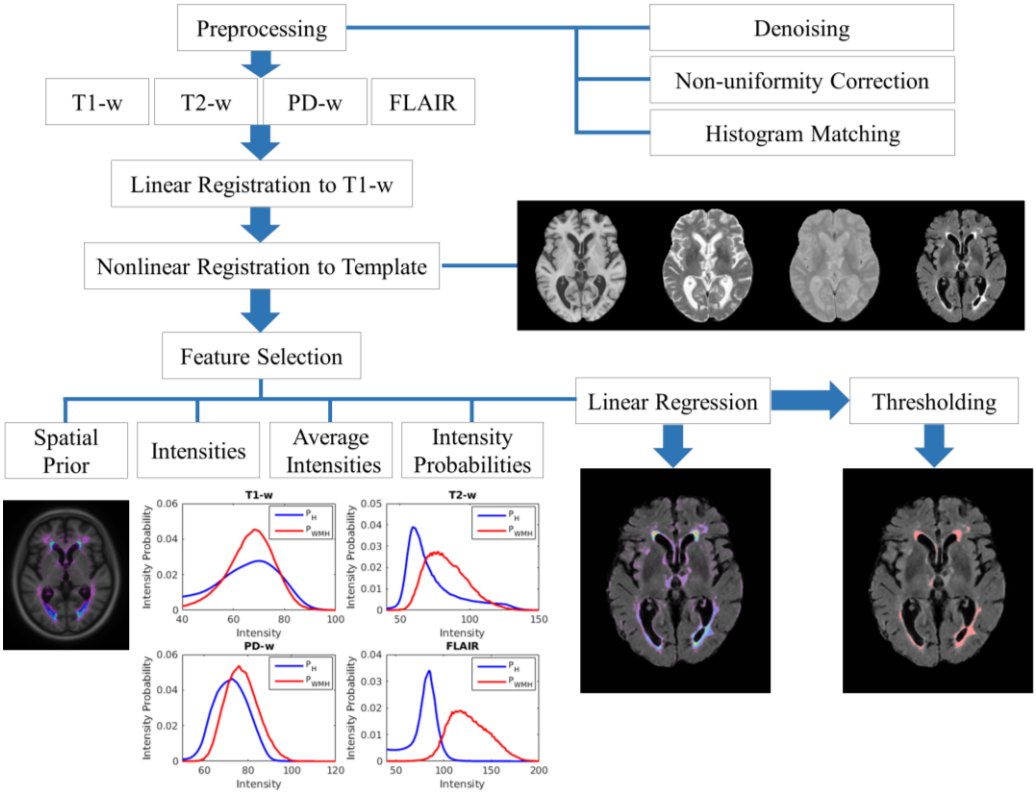


Fig. 3.3. Flow-chart of the proposed classifier and the preprocessing steps. The preprocessing includes denoising, image intensity non-uniformity correction, intensity range normalization, co-registration of T2w, PD and FLAIR to T1w scans, and stereotaxic registration of T1w. All modalities were then non-linearly warped to a template obtained from the ADNI dataset. Spatial prior, intensity and distribution features then served as inputs to the linear regression classifier.

3.2.5. Tissue Classification

The main post-processing step for lesion segmentation is assigning a label (i.e. WMH or non-WMH) to each voxel. The segmentation method was evaluated in a 10-fold cross validation manner, defining different training and testing subjects for each experiment. The training and testing subjects were selected from the same dataset in ADC and PREVENT-AD studies. For the ADNI2/GO segmentations, the training data was selected from the ADC study while testing data came from ADNI2/GO to show the robustness of the method across different scanners. The training dataset was generated from a large number of manually

labeled voxels; i.e. all voxels inside the brain mask for the subjects that were selected for training were used to create the training set - this includes all positive (WMH) and negative (non-WMH) example voxels. (Note that subjects used for testing were not used to estimate any of the features, probabilities or spatial priors, and thus serve truly as independent test data without double dipping). After training, a classifier can segment the image voxels of new subjects from the test dataset either by comparing their features with the features in its current training set or by creating a model to estimate a relationship between the output label and the input features of the training set. A variety of classification algorithms can be used for this purpose, such as neural networks (Hornik et al., 1989; Zijdenbos et al., 2002), k-NN (Denoeux, 1995; Anbeek et al., 2003; Steenwijk et al., 2013; Wu et al., 2006b), and support vector machines (Cortes and Vapnik, 1995; Lao et al., 2008). In this work, we selected a linear regression classifier with thresholding due to its low variance, high accuracy and lower computation time compared with other classifiers. The model parameters were calculated based on a least-squares estimation

$$\beta = (X^T X)^{-1}(X^T Y) = \left(\sum_{i=1}^{N} X_i X_i^T\right)^{-1}\left(\sum_{i=1}^{N} X_i Y_i\right)$$

Where β, X and Y denote the estimated weights, the feature matrix and target labels, respectively. X_i and Y_i denote the feature set and target labels for subject i and N is the number of subjects in the training set. The output of the linear regression model for a new subject j ($L_j = X_j\beta$) can be considered as a probability map that reflects the likelihood of the input voxel being a WMH. This value can later be thresholded to create a binary lesion map. The value of the threshold can determine the sensitivity and specificity of the segmentations; i.e. choosing lower threshold values can increase the sensitivity of the segmentations with the

price of decreasing the specificity, and vice versa. The optimal threshold value for creating binary segmentation maps can be obtained through cross validation as described below.

3.2.6. Evaluation metrics

To evaluate the accuracy of the automatic segmentations with respect to the gold standard manual labels, we used a variety of volumetric as well as spatial correspondence measures since no single measure is capable of reflecting all the desired information regarding the quality of segmentations (Caligiuri et al., 2015). To assess the volumetric correspondence between the automated and manual labels, we used intra-class correlation coefficient (ICC) for total lesion volume. The per-voxel spatial correspondence between two segmentations was evaluated using Dice similarity index (SI) as well as true and false positive rates (TPR and FPR), and positive prediction value (PPV) (Caligiuri et al., 2015). A high TPR (sensitivity) indicates that the automatic segmentation corresponds well to manual labels. A low FPR indicates that the procedure does not over-segment; i.e. identify non-WMH voxels as WMHs. A small PPV implies that many of the positive results are false positives. True positive (TP) and true negative (TN) indicate agreement whereas false negative (FN) and false positive (FP) indicate disagreement between the two segmentations. In cross-validations, SI was regarded as the primary outcome measure; i.e. the parameters were optimized based on SI values.

3.3. Experiments and Results

3.3.1. Qualitative results

Fig. 3.3. shows the segmentation results for a subject from the ADC dataset. In each row, 5 axial slices are shown, containing from top to bottom: the FLAIR image, the manual segmentations overlaid on the FLAIR, probability maps outputted by the linear regression

classifier, and the binary segmentations obtained by thresholding the probability maps with the optimal threshold based on SI values.

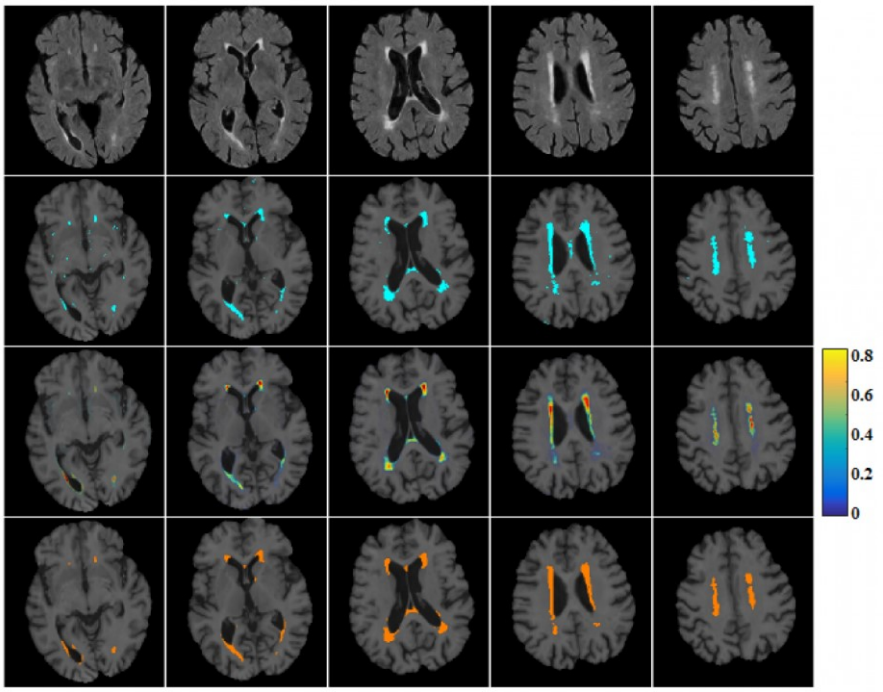


Fig. 4.3. Comparison of the automated vs. manually segmented WMHs for a subject from ADC dataset. Rows from top to bottom: A) axial FLAIR slices B) WMH labels obtained from manual segmentations C) probability maps obtained from the proposed automated method D) WMH labels obtained by thresholding the probability map. The color bar indicates the continuous output of the classifier before thresholding.

Fig. 3.4 shows similar segmentation results for a subject from the PREVENT-AD dataset. The method was trained to segment the periventricular and deep WMHs separately. Note the difference between the probability maps for the periventricular and deep WMHs and the fact that the probabilities are higher for areas closest to the ventricles for the former and lower for the latter. As a result, there is only a slight spatial overlap between the two segmentations (SI=0.05±0.04).

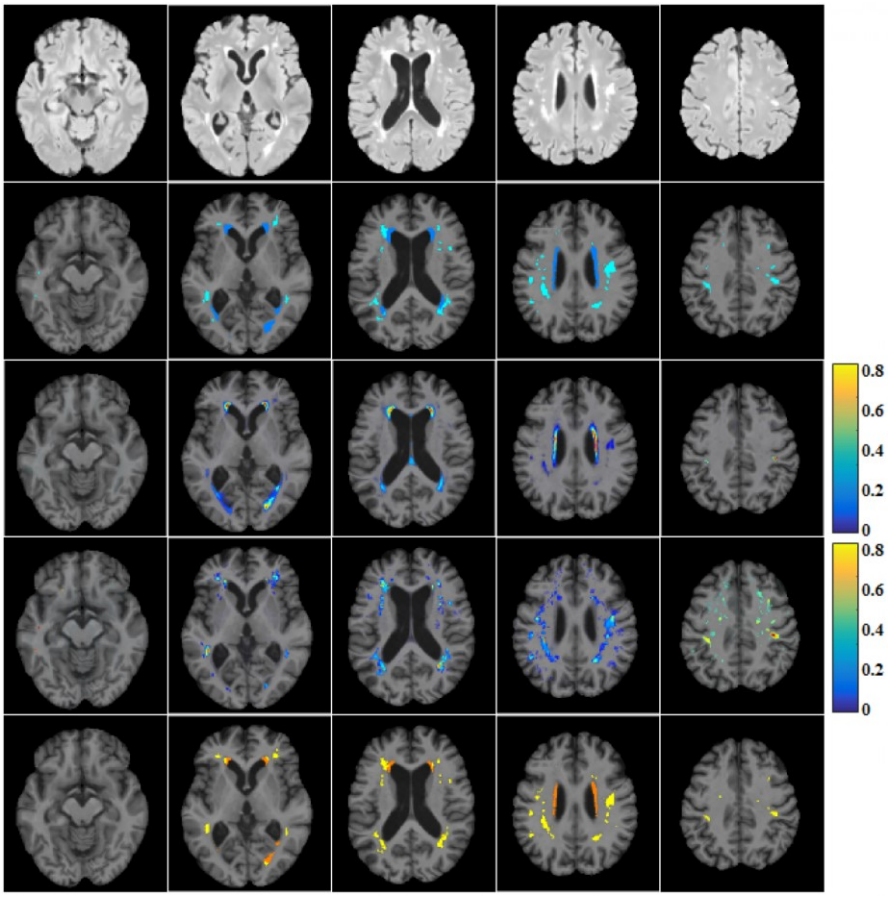


Fig. 3.4. Comparison of the automated vs. manually segmented WMHs for a subject from PREVENT-AD dataset. Rows from top to bottom: A) axial FLAIR slices B) Periventricular (dark blue) and deep (light blue) WMH labels obtained from manual segmentations C) Periventricular and D) deep probability maps obtained from the proposed automated method, respectively E) Periventricular (orange) and deep (yellow) WMH labels obtained by thresholding the probability map. The color bar indicates the continuous output of the classifier before thresholding.

Fig. 3.5 shows the segmentation results for a subject from ADNI2/GO dataset. One can see that in each case, the automatic output is very similar to the manual labels.

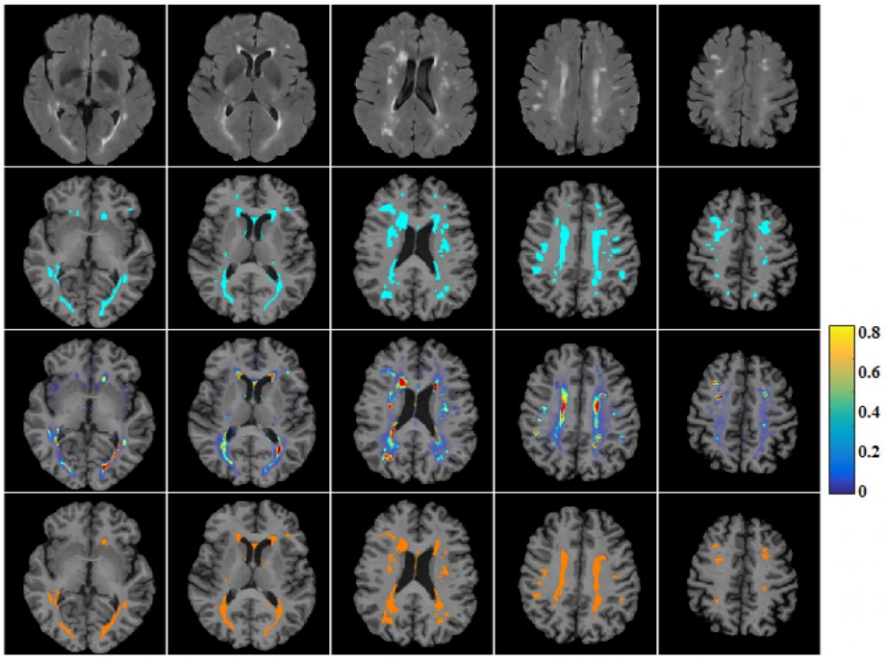


Fig. 3.5. Comparison of the automated vs. manually segmented WMHs for a subject from ADNI2/GO dataset. Rows from top to bottom: A) axial FLAIR slices B) WMH labels obtained from manual segmentations C) probability maps obtained from the proposed automated method D) WMH labels obtained by thresholding the probability map. The color bar indicates the continuous output of the classifier before thresholding.

3.3.2. Quantitative results

The performance of the method was evaluated on 3 different populations with 80, 40, and 10 subjects. We investigated 3 categories of lesion load since the different datasets had different ranges of WMH loads. In the ADC dataset, 57.5%, 31.5%, and 11.25% of the population had small, medium, and large lesion loads respectively. In the PREVENT-AD dataset, 62.5%, 35%, and 2.5% of the population had small, medium, and large lesion loads. In the ADNI2/GO dataset, 20%, 40%, and 40% of the population had small, medium, and large lesion loads, respectively (Fig. 3.1).

The binary segmentations were generated by applying a threshold to the probability map from the linear regression technique. Different values of threshold reflect different levels of sensitivity/specificity in the segmentations. Fig. 3.6 shows the SI between the binary

segmentation and gold standard manual segmentations for different values of threshold for the three datasets. Confidence intervals indicate the standard deviation of mean SI across 10 folds in the cross validation. From Fig. 3.6, we can see that the optimal threshold for generating binary segmentations is different for each case, since the number of available modalities is different, and consequently the number of features in the model are different for the 3 datasets.

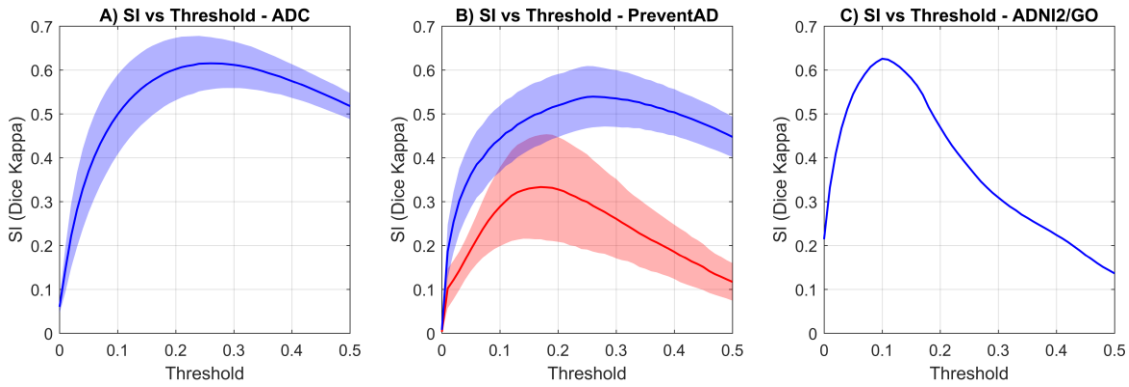


Fig. 3.6. SI (Dice Kappa) vs threshold for A) ADC B) PREVENT-AD C) ADNI2/GO datasets. Blue and red curves in B represent the results for the periventricular and deep WMHs, respectively.

SI, ICC, sensitivity, FPR, and PPV were calculated for all subjects with the optimal thresholds calculated as hyper-parameters through cross validation. Since the purpose of using ADNI2/GO dataset was to show the performance of the method on independent data from different scanners (not previously used in training or in parameter optimization), only information from ADC dataset was used to determine the optimal threshold for classification on ADNI2/GO dataset. (Note that additional investigation showed that using ADNI2/GO data to determine the optimal threshold would not lead to a significant improvement over using the ADC-derived threshold, p=0.31). The results are summarized in Table 3.2. For PREVENT-AD, the binary segmentation is a union of the periventricular and deep segmentations. Fig. 3.7 shows a boxplot diagram of SI values for the 3 categories of lesion loads in each dataset.

Table 3.2. Similarity measures between the manual and automatic segmentations for ADC, PREVENT-AD, and ADNI2/GO datasets.

Dataset	SI	ICC	Sensitivity	FPR	PPV
ADC	0.62±0.16	0.96±0.09	0.63±0.18	0.0002±0.0001	0.69±0.17
PREVENT-AD	0.51±0.16	0.78±0.21	0.52±0.20	0.0002±0.0002	0.59±0.15
ADNI2/GO	0.64±0.17	0.93	0.71±0.23	0.0014±0.0014	0.60±0.09

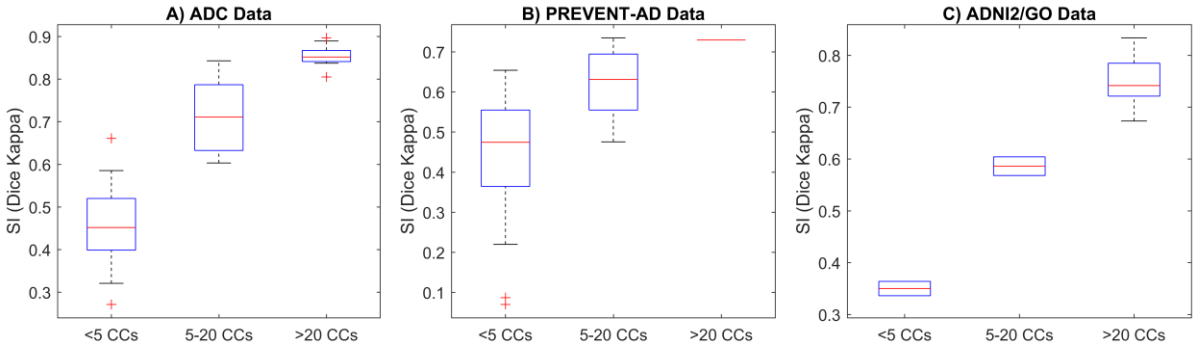


Fig. 3.7. Boxplot diagrams of SI (Dice Kappa) (<5 CCs, 5-20 CCs, and >20 CCs) for A) ADC B) PREVENT-AD C) ADNI2/GO datasets.

An SI value of 0.7 or higher indicates an excellent agreement (Bartko, 1991). The SI values suggest excellent agreement for medium and large lesion loads for ADC and ADNI2/GO datasets, and very good agreement for medium and excellent agreement for large lesion loads for PREVENT-AD dataset. To investigate this further, SI values were plotted against total lesion loads obtained from manual segmentations (Fig. 3.8). All of the small SI values occur in subjects that have smaller total lesion loads.

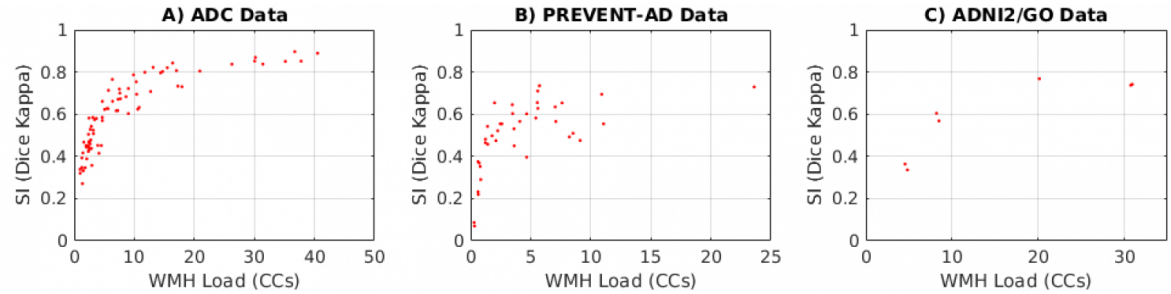


Fig. 3.8. SI (Dice Kappa) vs manually segmented WMH loads (CCs) for A) ADC B) PREVENT-AD C) ADNI2/GO datasets.

3.3.3. Contribution of the features

In order to show how much each of the proposed feature sets contributes to the performance of the classifier, the classifier was trained without each feature set. Table 3 shows the percentage of drop in SI (Dice Kappa) after removing each set of features for each dataset.

Table 3.3. Percentage of drop in SI (Dice Kappa) by removing feature sets for ADC, PREVENT-AD (periventricular-deep), and ADNI2/GO.

Dataset	Voxel Intensity	Spatial Prior	Average Intensity	P_{WMH}	P_H	$\frac{P_{WMH}}{P_H}$
ADC	5.5	8.6	5.3	6.6	5.5	1.6
PREVENT-AD	3.3-2.9	75.7-78.1	1.9-3.1	2.6-2.3	2.2-1.9	3.8-4.2
ADNI2/GO	8.0	9.3	8.4	19.9	7.7	9.7

3.3.4. Comparison between classifiers

Linear discriminant analysis (LDA), LogitBoost, and random forest classifiers were also trained and validated on the same features (Breiman, 2001; Friedman et al., 2000). For these classifiers, MATLAB toolbox implementations were used. Table 3.4 summarizes the results.

Table 3.4. Performance (SI) of LDA, LogitBoost and Random Forests classifiers. LDA= Linear Discriminant Analysis

	ADC	PREVENT-AD (PV - Deep)	ADNI2/GO
LDA	0.58±0.24	0.17±0.21 - 0.11±0.12	0.41±0.25
LogitBoost	0.70±0.14	0.62±0.15 - 0.52±0.21	0.31±0.24
Random Forest	0.68±0.15	0.61±0.15 - 0.51±0.22	0.32±0.23

3.3.5. Impact of Size of the Training Set

One of the important concerns for any supervised classification method that is dependent on training samples is the number of previously labeled samples that are required

to reach desirable performance on new unobserved data. To evaluate this dependence, we trained and validated the performance of the method using different sizes of training sets for the ADC dataset. The results of our investigations as shown in Fig. 3.9. suggest that the method shows acceptable performance (SI~0.6 and ICC~0.9) and can be used with as few as 40 labeled training subjects.

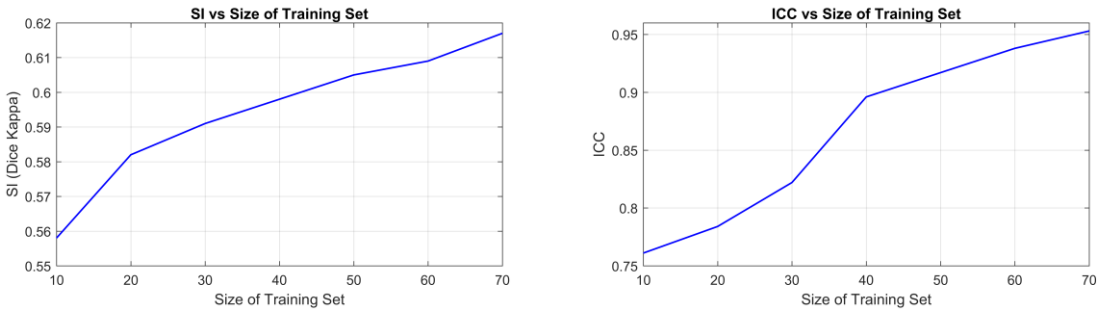


Fig. 3.9. Impact of the number of training subjects on SI (Dice Kappa) and intra-class correlation (ICC). Plotted SI and ICC values between the manually labeled gold standard WMHs and the WMH labels estimated by our automated method for the ADC dataset for different sizes of training sets.

3.4. Discussion

In this paper, we proposed and validated a new method for fully automated segmentation of WMHs from MR images. The proposed method uses a variety of location and intensity based features and a linear regression technique to create a continuous output that can be considered as a subject specific probability map of lesions, which can then be thresholded to create binary WMH segmentations. The advantage of creating these subject specific continuous WMH maps over binary segmentations is that they can be thresholded with different values, balancing the desired level of sensitivity/specificity depending on the purpose of segmentation. Furthermore, such lesion probability maps can provide more information about the voxel tissue than a simple binary valued segmentation; e.g., lesion probabilities may be useful to identify dirty white matter compared to healthy white matter

tissue (Beggs et al., 2016). These continuous values may also reflect the level of damage to the tissue, since higher WMH intensities can indicate more extensive cognitive deficits (Lindemer et al., 2015). Finally, we demonstrated that the thresholded determined on one dataset (ADC) was applicable to a previously unseen dataset (ADNI2/GO), underlining the robustness and generalizability of the proposed method.

A linear regression classifier was selected over other classification techniques for two reasons. First, because it provides a smooth continuous output that can be used as a subject specific probability map at low computational cost. But more importantly, our experiments showed that choosing more complex nonlinear classifiers may reduce the generalizability and applicability of the technique to new previously unseen data. For example, Random Forests and LogitBoost classifiers had a higher performance on ADC, but a much poorer performance on the independent ADNI2/GO datasets, as opposed to the simpler and more generalizable linear LDA and linear regression classifiers (Table 3.4).

The automated WMH segmentation method was evaluated on three different datasets (n1=80, n2=40, and n3=10) with the gold standard labels obtained from manual segmentations and measures such as SI (Dice Kappa), intra-class correlation (ICC), sensitivity and specificity. The automated labels showed high agreement with manual labels across all the datasets. The good performance of the algorithm on the ADNI2/GO subjects, which were not used in training the classifier, suggests that the method is robust in dealing with inter-site variability and enables us to apply the classifier to other datasets.

One of the major complications for automated segmentation of WMHs is caused by resampling. Since most automated tools use multiple contrasts of images to increase segmentation accuracy, it is necessary to co-register all the modalities to a common space. However, in most studies, the FLAIR scans (i.e. the modality with the optimal contrast for

lesion detection) as well as T2w and PD scans are obtained with thick slices (usually 3-5 mms) in clinical studies due to acquisition timing constraints. This results in blurring effects after resampling. To avoid resampling as much as possible, we transformed all data (i.e. the spatial priors, brain masks, etc.) to the native FLAIR space for the ADC and ADNI2/GO datasets and performed segmentation in the native FLAIR space. This improved the segmentation performance significantly for the ADC (SI=0.62 native vs SI=0.53 resampled) and ADNI2/GO (SI=0.64 native vs. SI=0.55 resampled) datasets while it did not have any effect on the PREVENT-AD dataset due to its inherent high spatial resolution ($1mm^3$ isotropic voxels). In the PREVENT-AD dataset, separating the WMHs into periventricular and deep classes yielded an improvement of 10.87% in SI. This was expected since deep WMHs have a different contrast and were more likely to be missed if the same threshold as periventricular WMHs was used.

The SI was used to validate the performance of the method as well as to determine the optimal threshold for creating binary segmentations from probabilistic lesion maps. However, as can be deduced from Fig. 3.8. the algorithm yields smaller SI values for small lesion loads and larger values for relatively larger WMH loads. This is not specific to the proposed method and is in fact due to the nature of the definition of SI, which causes the same amount of difference to yield lower SI values if the total volume is smaller. This prevents SI from being considered as the ideal similarity measure for lesion segmentation applications, since the reported results will then depend on the average lesion load across the population under study. However, since metrics such as ICC depend only on the total load rather than the actual segmentations, SI still remains the most informative metric, if its values are reported along with the average lesion loads across the population.

Another possible set of metrics that are commonly used to study the performance of lesion segmentation techniques (especially those applied to MS lesion segmentation due to the clinical relevance of lesion count when evaluating treatment strategies) are per lesion metrics. However, since most of the WMHs in AD and aging populations are relatively large and confluent, such measures are not as informative in these studies. In fact, most of the per lesion metrics that were calculated for the ADC dataset showed nearly excellent performance.

The average Dice Kappa was lower for the PREVENT-AD dataset in comparison with ADC and ADNI2/GO due to several reasons. First, the PREVENT-AD subjects are much younger and drawn from a healthy population without any cognitive complaints, and as a result have significantly lower WMH loads and smaller lesions when compared to the ADNI2/GO and ADC subjects ($p<0.0001$, $p=0.0044$). Second, different techniques were used for manual detection of the WMHs in each of the three datasets. Specifically, for the PREVENT-AD dataset where sensitive detection was desired, the union of two raters was used as the gold standard. This would naturally lead to more generous segmentations as opposed to using the intersection between the two labels, which would have the opposite effect. Third, the contrast between the healthy tissue and WMHs in the FLAIR scans was lower in the PREVENT-AD FLAIRs, leading to a significant overlap in the intensity histograms, and thus making the classification task more prone to errors, both for the manual raters and automated tools. On the other hand, the PREVENT-AD FLAIR scans had a much better spatial resolution (i.e. 1 mm slice thickness) enabling the method to identify smaller lesions. In the future, it would be interesting to study the dependence of lesion contrast on lesion age and level of tissue damage.

FLAIR is the optimal modality to detect WMHs due to the high contrast between WMHs and surrounding tissue. However, many studies forego FLAIR acquisition in favor of

other modalities. As a result, segmentation techniques that can detect WMHs without using FLAIR are highly desirable. The proposed technique was able to detect WMHs in the ADC dataset using only T1w, T2w, PD data with SI=0.45±0.18. While this is not as high as when using FLAIR, it shows that it is possible to segment some WMHs without using FLAIR information.

The training time for the proposed method using an Intel Core i3-2120 processor at 3.30 GHz was approximately 19 minutes for 40 subjects and the segmentation time for each subject after training was approximately 1.6 seconds. The low computational expense enables us to use this technique on large MRI databases without being concerned with computation burden.

The proposed technique was also compared with FAST toolbox (FMRIB's Automated Segmentation Tool) (Zhang et al., 2001) from FSL (Jenkinson et al., 2012; Smith et al., 2004; Woolrich et al., 2009), LGA (Lesion Growth Algorithm; Schmidt et al., 2012) as implemented in the LST toolbox version 2.0.15 (www.statistical-modelling.de/lst.html) for SPM (Penny et al., 2011), as well as W2MHS from Ithapu et al. (Ithapu et al., 2014) as three well-known freely available segmentation techniques in the literature on the same 3 datasets that were used for our validations. The results showed that the proposed technique outperforms all three methods in terms of Dice Kappa (SI) in segmenting all categories (small, medium, and large) of lesion loads across all three datasets. FAST from FSL oversegmented artifacts and bright regions near the cortex and was only able to segment large lesions with high contrast yielding SI=0.11±0.15 for the ADC dataset and SI=0.23±0.34 for the ADNI2/GO dataset. LGA from SPM tended to under-segment the lesions, especially deep WMHs across all three datasets (SI=0.09±0.12 for ADC and SI=0.20±0.24 for ADNI2/GO). W2MHS had a better performance for both ADC and ADNI2/GO datasets (SI=0.20±0.18 for

ADC and SI=0.39±0.29 for ADNI2/GO datasets). All techniques had poor performance in cases with small lesions or low contrast between the healthy tissue and WMHs: in the PREVENT-AD dataset, neither technique was able to detect the WMHs (SI=0.003±0.003 for FAST, SI=0.01±0.02 for LGA, and SI=0.01±0.02 for W2MHS). All results were tested for statistical significance (using paired t-tests) in comparison with SI values obtained from the proposed method for the same datasets ($p<0.0001$).

It is difficult to compare our technique to previously published results. The difficulty lies in the differences between populations, MR image contrasts, anatomical definition of WMHs, and quality of manual segmentations. All three datasets have a much higher number of subjects with small or even no WMHs (in case of deep WMHs in PREVENT-AD) for which disagreement in a few voxels would lead to a very small SI (or even zero for cases with no WMHs). In addition, there are other factors that might lead into differences between the reported performances of the methods, which do not necessarily reflect the superiority of the WMH segmentation technique, such as: masking out difficult/prone to artifact regions, using WM masks, using a rule for minimum number of neighboring voxels for manually or automatically labelling a voxel as WMH (See Table 3.5). Still, our results are comparable to those published in literature, yielding the best results for patients with large lesion loads, and among the best for medium lesion loads (See Table 3.5). Future work will focus on improving the technique for small lesion loads to facilitate application of this technique to datasets of cognitively normal individuals and at-risk populations.

Table 3.5. Comparison of SI (Dice Kappa) for different lesion loads in various studies. (S: small load, M: medium load, L: large load). Notes: 1- No exclusion mask. 2- No post processing. 3- Subjects with vascular disease. 4- Excluded areas between lateral ventricles. 5- Excluded small lesions. 6- Population did not have subjects with small WMH loads. 7- Used tissue segmentation. 8- Validation on 20 slices per subject on average, selected based on presence of lesions with clear borders. 9- Removed periventricular flow artifacts. 10- Excluded areas outside WM mask. 11- Post processing to remove noisy detections. 12- Subjects with small vessel disease (based on appearance of WMHs and/or lacunas). 13- Aging/AD and vascular disease patients with minor strokes. Used exclusion mask containing dilated CSF and subcortical structures (basal ganglia) and entorhinal cortex.

			Dice (SI)			
Method	Notes	Number (S-M-L%)	S	M	L	Total
Proposed Method	1,2	80 (58-31-11)	0.49	0.74	0.87	0.62
Admiraal (Admiraal-Behloul et al., 2005)	3,4,5	100 (40-35-25)	0.70	0.75	0.82	0.75
Anbeek (Anbeek et al., 2004)	3	20 (40-35-25)	0.50	0.75	0.85	0.61
Beare (Beare et al., 2009)	6	30		0.50	0.65	0.58
Boer (de Boer et al., 2009)	5,7	20		0.72		0.72
Steenwijk (Steenwijk et al., 2013)	5,7	20 (15-45-40)	0.78	0.85	0.91	0.84
Khayati (Khayati et al., 2008)	5,6,8	20 (35-50-15)	0.72	0.75	0.80	0.75
Sajja (Sajja et al., 2006)	5,7	23 (35-65)	0.67		0.84	0.78
Schmidt (Schmidt et al., 2012)	7	53	0.66	0.79	0.85	0.75
Ong (Ong et al., 2012)	9,10	38	0.36	0.56	0.71	0.47
Ithapu (Ithapu et al., 2014)	9,11	38				0.67
Herskovits (Herskovits et al., 2008)	2,7	42				0.60
Dyrby (Dyrby et al., 2008)	10	362	0.45	0.62	0.65	0.56
Erus (Erus et al., 2014)	6	33		0.54		0.54
Ghafoorian (Ghafoorian et al., 2016b)	12	46				0.79
Simões (Simões et al., 2013)	7,10	28 (14-9-5)	0.51	0.70	0.84	0.62
Yoo (Yoo et al., 2014)	5,6,10	32 (7-10-15)	0.59	0.73	0.86	0.76
Griffanti (Griffanti et al., 2016)	13	21	0.70	0.69	0.80	0.76

Quantification of WMH volumes is critical for evaluation of the vascular burden of AD. As well, this will prove especially useful in vascular cognitive impairment where cerebrovascular disease is believed to be the primary cause of the disease and the lesion load is thought to reflect the severity of disease (Gorelick et al., 2011). There is growing evidence that controlling vascular risk factors which are the primary cause of WMHs is associated with decline in dementia (Langa et al., 2016). Here, quantification of WMH will be essential for assessing severity, for monitoring progression and response to treatment. The proposed

method has several advantages including robustness, not requiring any manual intervention, and fast computation time. Our results suggest that the proposed automated tool can provide fast, robust, and accurate segmentations for WMHs and holds good potential for clinical studies. Hence, it is particularly useful given the emergence of large MRI databases such as ADNI (http://www.loni.ucla.edu/ADNI/).

Chapter 4. Performance Comparison of 10 Different Classification Techniques in Segmenting White Matter Hyperintensities in Aging

Preface

In this chapter, we extend the previous work to build an automated pipeline for segmenting WMHs in large multi-cite and multi-scanner datasets. To achieve this, we obtained manual segmentations from subjects scanned on different scanner models and with different protocols from the multi-center ADNI1, ADNI2, and NACC databases. We trained and validated 10 different linear and nonlinear classifiers on the manual segmentations and compared their performance in detecting WMHs using various combinations of input sequences, including T1w+T2w+PD+FLAIR, T1w+T2+PD, T1w+FLAIR, and only T1w. Further, we made the WMH segmentation pipeline along with the pretrained classifiers publicly available[1].

Our results showed that Random Forests classifier has the best performance in detecting WMHs in all of the experiments.

This work has been published as:

Dadar, M., J. Maranzano, K. Misquitta, C. J. Anor, V. S. Fonov, M. C. Tartaglia, O. T. Carmichael, C. Decarli, D. L. Collins, and Alzheimer's Disease Neuroimaging Initiative. "Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging." *NeuroImage* 157 (2017): 233.

[1] http://nist.mni.mcgill.ca/?p=221

Performance Comparison of 10 Different Classification Techniques in Segmenting White Matter Hyperintensities in Aging

Mahsa Dadar, Josefina Maranzano, Karen Misquitta, Cassandra J. Anor, Vladimir S. Fonov, M. Carmela Tartaglia, Owen T. Carmichael, Charles Decarli, D. Louis Collins, Alzheimer's Disease Neuroimaging Initiative[2]

Abstract

Introduction: White matter hyperintensities (WMHs) are areas of abnormal signal on magnetic resonance images (MRIs) that characterize various types of histopathological lesions. The load and location of WMHs are important clinical measures that may indicate the presence of small-vessel disease in aging and Alzheimer's disease (AD) patients. Manually segmenting WMHs is time consuming and prone to inter-rater and intra-rater variabilities. Automated tools that can accurately and robustly detect these lesions can be used to measure the vascular burden in individuals with AD or the elderly population in general. Many WMH segmentation techniques use a classifier in combination with a set of intensity and location features to segment WMHs, however, the optimal choice of classifier is unknown.

Methods: We compare 10 different linear and nonlinear classification techniques to identify

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

WMHs from MRI data. Each classifier is trained and optimized based on a set of features obtained from co-registered MR images containing spatial location and intensity information. We further assessed the performance of the classifiers using different combinations of MRI contrast information. The performances of the different classifiers were compared on three heterogeneous multi-site datasets, including images acquired with different scanners and different scan-parameters. These included data from the ADC study from University of California Davis, the NACC database and the ADNI study. The classifiers (naïve Bayes, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, bagging, and boosting) were evaluated using a variety of voxel-wise and volumetric similarity measures such as Dice Kappa similarity index (SI), Intra-Class Correlation (ICC), and sensitivity as well as computational burden and processing times. These investigations enable meaningful comparisons between the performances of different classifiers to determine the most suitable classifiers for segmentation of WMHs. In the spirit of open-source science, we also make available a fully automated tool for segmentation of WMHs with pre-trained classifiers for all these techniques.

Results: Random Forests yielded the best performance among all classifiers with mean Dice Kappa (SI) of 0.66±0.17 and ICC=0.99 for the ADC dataset (using T1w, T2w, PD, and FLAIR scans), SI=0.72±0.10, ICC=0.93 for the NACC dataset (using T1w and FLAIR scans), SI=0.66±0.23, ICC=0.94 for ADNI1 dataset (using T1w, T2w, and PD scans) and SI=0.72±0.19, ICC=0.96 for ADNI2/GO dataset (using T1w and FLAIR scans). Not using the T2w/PD information did not change the performance of the Random Forest classifier (SI=0.66±0.17, ICC=0.99). However, not using FLAIR information in the ADC dataset significantly decreased

the Dice Kappa, but the volumetric correlation did not drastically change (SI=0.47±0.21, ICC=0.95).

Conclusion: Our investigations showed that with appropriate features, most off-the-shelf classifiers are able to accurately detect WMHs in presence of FLAIR scan information, while Random Forests had the best performance across all datasets. However, we observed that the performances of most linear classifiers and some nonlinear classifiers drastically decline in absence of FLAIR information, with Random Forest still retaining the best performance.

Keywords: White matter hyperintensities, Segmentation, Classification, Alzheimer's Disease

4.1. Introduction

White matter hyperintensities (WMHs), commonly identified as areas of increased signal in relation with the surrounding white matter regions on T2w, PD and FLAIR MRIs, are one of the non-specific yet typical and constant MRI expressions of cerebral small vessel disease (CSVD), along with lacunar infarcts and microhemorrhages (Conklin et al., 2014; Gouw et al., 2010). They have been shown to be more extensive in patients with Alzheimer's disease compared to age-matched healthy normal populations (Yoshita et al., 2005). WMHs reflect ischemic injury in the elderly and AD populations and the existence and severity of WMHs can lead to or accelerate decline in cognitive as well as executive functions (Dubois et al., 2014). As a result, the location and load of WMHs are important clinical measures, raising substantial need for their accurate quantifications. WMHs are generally detected using fluid attenuated inversion recovery (FLAIR) or T2w/PD scans. Manually labeling WMHs is challenging due to time constraints as well as inter-rater and intra-rater variabilities (Grimaud et al., 1996). As a result, automated tools that can segment WMHs robustly and with high accuracy are extremely useful,

particularly in large scale studies such the Alzheimer's Disease Neuroimaging Initiative (http://www.loni.ucla.edu/ADNI/), the National Alzheimer's Coordinating Center (NACC) database (https://www.alz.washington.edu/) and others where it is desired to estimate the contribution of neurovascular disease to cognitive decline.

The heterogeneity in the distribution and patterns of WMHs makes the segmentation task intrinsically complex (Caligiuri et al., 2015). Automated segmentation tools usually integrate information from multiple complementary MRI contrasts including T1w, T2w, PD and FLAIR to reduce uncertainty and improve segmentation accuracy. Most successful fully automated WMH segmentation techniques extract a combination of location and intensity features from these images and use them as inputs to a linear or nonlinear classifier. Here we review the most commonly used linear and nonlinear classifiers in general as well as their application to the task of segmenting lesions in general or WMHs of vascular etiology specifically.

While there have been many studies attempting to segment WMHs using these classification techniques, drawing meaningful comparisons between their performance is not possible since they have been applied to different datasets and results are highly variable across different populations and imaging protocols (García-Lorenzo et al., 2013; Caligiuri et al., 2015). To our knowledge, no studies have compared the performance of these classification techniques for detecting WMHs against one another on the same datasets, especially for cases where classification is attempted without using the optimal FLAIR information. In this paper, we have extensively compared the performance of these different classification techniques in detecting WMHs with and without FLAIR information using 3 different large publicly available datasets with different scanners and acquisition protocols. This enables us to draw more generalizable conclusions regarding the performance of the classifiers. Our contributions include an extensive

comparison of 10 widely used classification techniques in detecting WMHs across 4 different datasets, three of which are from multi-site and multi-scanner studies and across different combinations of imaging modalities. In addition, we make publicly available an implementation of the segmentation tool along with all the pre-trained classifiers (http://nist.mni.mcgill.ca/?p=221). The proposed tool is generalizable to data from different scanners since it has been trained on data from multiple scanners.

4.2. Materials and methods

4.2.1. Subjects

The performances of the different classifiers were assessed based on four datasets of subjects with different ranges of WMH loads. Table 4.1 shows the demographic information for each dataset.

(i) ADC**:** This dataset consists of 70 individuals (70-90 years old) with normal cognition, mild cognitive impairment (MCI), and AD dementia from University of California, Davis Alzheimer's Disease Center (ADC) who were scanned using T1w, double-echo T2w/PD, and FLAIR MRI modalities.

(ii) NACC**:** This dataset consists of a patient sample of 32 MCI and AD subjects obtained from the National Alzheimer's Coordinating Center (NACC) database which is a database of subjects with a range of cognitive status, i.e. normal cognition, MCI, and demented who received T1w, and FLAIR MRI scans (https://www.alz.washington.edu/). Data consisted of variables from a Uniform Data Set collected from more than 30 Alzheimer's disease centers (ADC) throughout the

United States and cataloged at the NACC. ADCs are National Institute on Aging–funded centers that enroll patients using different participation recruiting practices. A full description of the NACC data set has been previously provided (Beekly et al., 2004; Morris et al., 2006). NACC data used here has been acquired at six different ADCs using eight different scanner models of three different manufacturers. Subjects were selected to have low, medium, and large WMH loads.

ADNI: Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

(iii) ADNI1**:** This dataset consists of T1w, T2w, and PD scans of 53 subjects from ADNI1 study. Despite the fact that all subjects had to have Hachinski Ischemic Score of less than or equal to 4 as part of the inclusion criteria (Petersen et al., 2010), we found many subjects that had high WMH loads. Subjects were selected from different sites and scanners and a preliminary assessment was performed to evaluate their WMH load with the goal of acquiring subjects with different scanner information as well as different loads of WMHs. For each scanner model, we selected datasets that had low, medium and high lesion loads. Approximately equal number of male and female subjects were selected. The age of the subjects was also considered for the selection, with the aim of achieving a normal distribution.

(iv) ADNI2/GO: This dataset consists of T1w and FLAIR scans of 46 subjects from ADNI2/GO studies. Subject selection criteria were the same as ADNI1.

Table 4.1. Demographic information for ADC, NACC, ADNI1 and ADNI2/GO datasets.

Dataset	ADC	NACC	ADNI1	ADNI2/GO
N	70	32	53	46
Sex	35 M	15 M	27 M	25 M
Age	78.0±7.3	74.9±8.0	75.7±6.6	74.1±6.5

4.2.2. MR imaging

Table 4.2 summarizes the scanner information as well as the MR imaging parameters for each of the datasets.

Table 4.2. Scanner information and MRI acquisition parameters for ADC, NACC, ADNI1, and ADNI2/GO datasets.

Modality	Dataset	ADC	NACC	ADNI1	ADNI2/GO
	Scanner Manufacturer	GE MS Philips MS	GE MS	GE MS Philips MS SIEMENS	Philips MS SIEMENS
T1w	Slice thickness (mm)	1.5	1.5	1.2	1.2
	No. of slices	128	124	160	196
	Field of view (cm^2)	250×250	256×256	192×192	256×256
	Scan Matrix (cm^2)	256×256	256×256	192×192	256×256
	TR: Repetition time (ms)	9	9	3000	7.2
	TE: Echo time (ms)	2.9	1.8	3.55	3.0
	Pulse Sequence	FSPGR	FSPGR	MPRAGE	GR
Other	Contrast	FLAIR	FLAIR	T2w/PD	FLAIR
	Slice thickness (mm)	3	3	3	5
	No. of slices	48	48	56	42
	Field of view (cm^2)	220×220	256×256	256×256	256×256
	Scan Matrix (cm^2)	256×192	256×256	256×256	256×256
	TR: Repetition time (ms)	11000	11002	3000/3000	11000
	TE: Echo time (ms)	144	147	95.2/10.5	150
	Pulse Sequence	FSE	Obl	FSE	SE/IR

4.2.3. Manual segmentation

In ADC, NACC, and ADNI2/GO datasets, the WMHs were manually segmented by experts with FLAIR used as the primary contrast and the other image contrasts used to aid in the decision process to include or exclude a voxel from the lesion mask. For the ADNI1 dataset, T2w was used as the primary contrast. All WMH masks were created fully manually, without using any thresholding technique. ADC, ADNI1 and ADNI2/GO datasets were scored by JM, an MD with training in general radiology, and specialized in MRI imaging methods of quantifying WM pathologies in MS and AD. JM has more than 12 years of experience in reading MRI and developing standardized MRI guidelines to detect WM lesions using different image modalities (Maranzano et al., 2016). The lesions were fully manually traced using the interactive software package Display, part of the MINC Tool Kit (https://github.com/BIC-MNI) developed at the McConnell Brain Imaging Center of the Montreal Neurological Institute. The program allows simultaneous viewing and segmentation in the coronal, sagittal and axial planes, and cycling between each image volume. The image volumes were co-registered so that, when assessing a given voxel or region and switching from one contrast to another (e.g. T1w to FLAIR), the rater can assess the intensity signal of the same region of the brain on each contrast. In the NACC dataset, images were similarly segmented by two raters that had previously received training to segment WMHs, and ascertained by an expert neurologist. The between rater agreement was verified (Dice Kappa=0.70). All the manual raters were also asked to segment 3 scans with low (<5cm^3), medium (5-20 cm^3), and high (>20 cm^3) WMH loads a second time without consulting the initial segmentations. Table 4.3 shows the intra-rater Dice Kappa obtained from these segmentations as well as WMH volume information for each dataset. Figure 4.1 shows examples

of the available contrasts as well as the manual labels for each dataset.

Table 4.3. Intra-rater mean Dice Kappa, range of WMHLs, and number (N) of subjects with low (<5cm^3), medium (5-20cm^3), and high (>20cm^3) WMHLs for manual segmentations of WMHs in different datasets. WMHL= White Matter Hyperintensity Load.

Dataset	ADC	NACC	ADNI1	ADNI2/GO
Dice Kappa	0.72	0.78	0.80	0.86
WMHL Range	0.2-148.6	0.2-109.0	0.0-119.3	0.2-63.0
$N_{WMHL<5cm^3}$	36	6	14	16
$N_{5cm^3<WMHL<20cm^3}$	23	11	10	11
$N_{WMHL>20cm^3}$	11	12	27	18

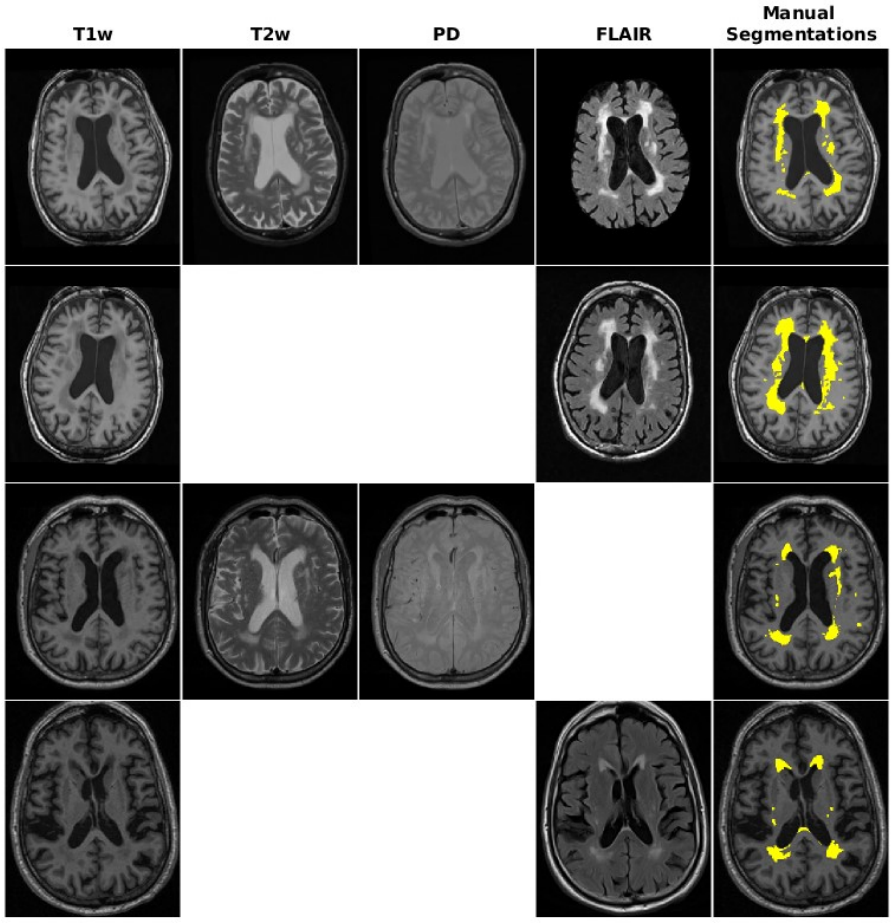


Fig. 4.1. Axial slices comparing manual segmentations and T1w, T2w, PD, and FLAIR information for subjects from ADC, NACC, ADNI1, and ADNI2 datasets. Yellow color indicates regions labeled as WMH in manual segmentations.

4.2.4. Pre-processing

All the images were preprocessed using our standard pipeline from MINC toolkit, publicly available at https://github.com/BIC-MNI/minc-tools (Aubert-Broche et al., 2013) through three steps: I) Image noise reduction using mincnlm tool (Coupe et al., 2008), II) Correction of image intensity non-uniformity using nu_estimate tool (Sled et al., 1998) and III) Normalization of image intensity into range (0-100) using an intensity histogram matching algorithm (volume_pol tool). The T1w, T2w, PD, and FLAIR images were linearly co-registered using a 6 parameter rigid registration (Collins et al., 1994). The T1w images were linearly and then nonlinearly registered to an average template (Collins and Evans, 1997) created based on data from the ADNI1 study (Fonov et al., 2011b; Fonov et al., 2011a), enabling the use of anatomical priors in the segmentation process. Brain extraction was performed on the linearly registered T1w images as part of the standard pipeline (Aubert-Broche et al., 2013).

4.2.5. Features

The classical features that are most commonly used in lesion segmentation tasks are the intensity of the voxel in each MRI contrast (García-Lorenzo et al., 2013). Here, these classical features as well as a variety of intensity and spatial features were used to train the classifiers. These features have been previously validated and verified to be informative in detecting WMHs. The rationale behind the selection of the suggested feature set as well as the contribution of each of the features has been described in more detail in an earlier work (Dadar et al., 2017a).

(i) Voxel intensity from T1w, T2w, PD, and FLAIR images.

(ii) Average voxel intensity of non-WMH tissue from T1w, T2w, PD, and FLAIR images for the specific voxel location obtained from averaging non-WMH voxels of the training subjects in stereotaxic space. Since datasets were selected to include subjects with very small WMH loads, there were at least several subjects in each training set that had no WMHs in each specific voxel location. The average intensity of non-WMH tissue feature was calculated using data from these subjects.

(iii) Probability of voxel being a lesion (P_{WMH}) obtained by creating a probability distribution function (PDF) based on the intensity histogram of the WMH labels from manually segmented training data across all WMH voxels.

(iv) Probability of voxel being healthy tissue (P_H) obtained by creating a PDF of Non-WMH voxels from manually segmented training data across all non-WMH voxels.

(v) Ratio of P_H / P_{WMH}

(vi) Spatial WMH probability map created by averaging the WMH maps from the training dataset.

(vii) Ratio of T2w/T1w, PD/T1w, FLAIR/T1w.

The WMH segmentations were performed in the native space of the primary image contrast, i.e. T2w for ADNI1 and FLAIR for ADC, NACC, and ADNI2/GO datasets to avoid the blurring caused by resampling of the primary image contrast. To achieve this, all images were non-linearly transformed to the ADNI template space, and all the priors and averages were calculated in this stereotaxic space and then registered back and resampled in the native space using the inverse nonlinear transformations. The final segmentations were performed using the

features in the native space of the image with optimal contrast. Therefore, the image with optimal contrast is not resampled, and only a 6-parameter rigid transformation is applied to the other co-registered contrasts (as opposed to other techniques where the nonlinearly registered images are used for segmentation). Figure 4.2 illustrates a flow-chart of the preprocessing, registration, and feature selection steps of the pipeline.

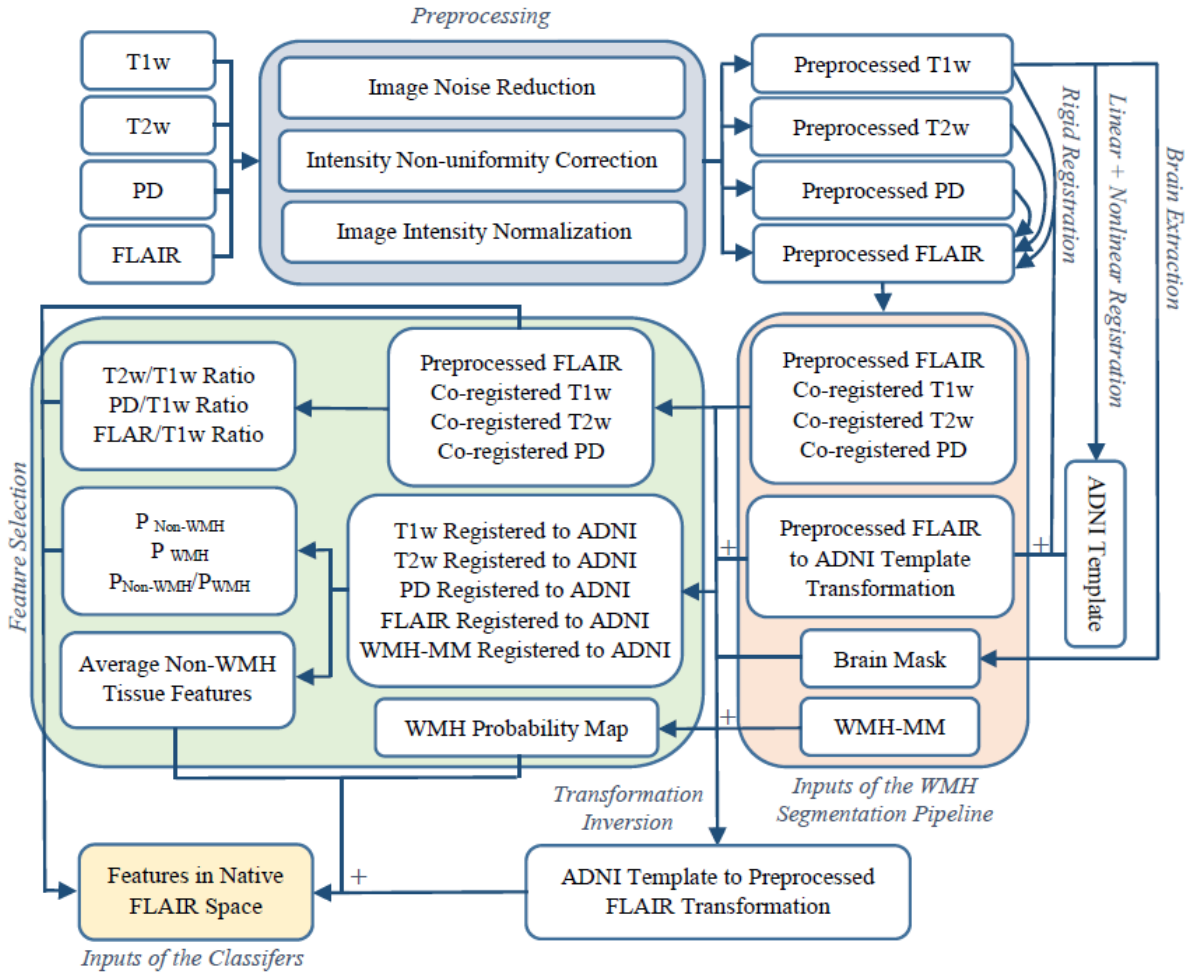


Fig. 4.2. Flow-chart of the preprocessing, registration, and feature selections steps. WMH-MM= White Matter Hyperintensity Manual Mask.

4.2.6. Classification Methods

In a binary classification setting, a classifier is a function that maps a set of input feature vectors $x = (x_1, x_2, \dots, x_n)^T$ from feature space X to an output class label set y in $Y = \{0,1\}$. Here, we select and compare supervised methods as unsupervised techniques have been shown to be less robust, dependent on initialization, and do not necessarily arrive at meaningful segmentations (Clarke et al., 1995). Specifically for the task of WMH segmentation, supervised methods generally outperform unsupervised techniques (Anbeek et al., 2004; Caligiuri et al., 2015).

Naive Bayes

Naïve Bayes classifiers are a family of probabilistic classifiers that have been used for many simple classification tasks (Lewis, 1998). Naïve Bayes is a probabilistic classifier that returns the label that maximizes the posterior probability $p(y|x)$ as the output, with the underlying assumption that given the class label, all the features are conditionally independent

$$arg\ max_y\ p(y|x) = arg\ max_y \frac{p(y)\prod_{i=1}^{n} p(x_i|y)}{p(x)} = arg\ max_y\ p(y)\prod_{i=1}^{n} p(x_i|y)$$

Naïve Bayes classifiers have previously been used to segment diabetic retinopathy lesions (Köse et al., 2012).

Discriminant Analysis

Linear and Quadratic Discriminant Analysis methods (LDA and QDA) are generalizations of Fisher's linear discriminant method that can be used for performing

classification (Fisher, 1936; McLachlan, 2004). Using the assumption that the conditional probability density functions of the classes are normally distributed with identical covariance, i.e.

$$p(x|y=k) = \frac{1}{(2\pi)^n|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^t\Sigma^{-1}(x-\mu_k)\right) \qquad k\epsilon\{0,1\}$$

LDA predicts input vector x as belonging to a class y based on the log likelihood ratio $ln\frac{p(y=1|x)}{p(y=0|x)}$. QDA is similar to LDA, without the identical covariance assumption.

$$p(x|y=k) = \frac{1}{(2\pi)^n|\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)\right) \qquad k\epsilon\{0,1\}$$

Amato et al. proposed a non-parametric discriminant analysis technique for segmenting MS lesions (Amato et al., 2003). Akselrod-Ballin et al. have used LDA technique along with Random Forests to segment MS lesions (Akselrod-Ballin et al., 2009).

Logistic Regression

The idea of logistic regression was introduced by Cox with the purpose of estimating a binary response based on a set of independent features (Cox, 1958). The Logistic regression classifier models $p(y|x)$ as a logistic function $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ and estimates the error using a cumulative logistic distribution function.

$$E(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(-y^i\log(h_\theta(x^i)) - (1-y^i)\log(1-h_\theta(x^i))\right)$$

Sánchez et al. used a logistic regression classifier for automatic detection of micro-aneurysms in retinal images (Sánchez et al., 2009).

Decision Trees

The idea of performing induction using decision trees was first proposed by Hunt et al. (Hunt et al., 1966) and later developed by Quinlan for classification tasks (Quinlan, 1986). Decision tree classifiers map the feature vector x to conclusions about the target value y using a tree structure in which the leaves represent class labels y and the nodes represent partitionings of feature x that lead to these class labels. The decision tree is generally constructed in 2 phases: 1) A recursive, top-down procedure "grows" a tree to fit the training data. 2) A "pruning" phase to avoid overfitting. Decision tree classifiers have since been used for tissue classification (Chao et al., 2009) and lesion segmentation in Multiple Sclerosis (MS) (Kamber et al., 1992) (Kamber et al., 1995).

Random forests

Initially introduced by Breiman (Breiman, 2001), Random decision forests perform classification and regression by constructing a multitude of independent decision trees and using the mode or mean of their predictions as the final output for classification or regression tasks, respectively. They have since been widely used for lesion segmentation in MS (Geremia et al., 2011; Maier et al., 2015; Mitra et al., 2014; Akselrod-Ballin et al., 2009) as well as for WMH segmentation in aging and AD populations (Ithapu et al., 2014).

K-nearest neighbors

The K-nearest neighbours (KNN) is a non-parametric instance based algorithm developed by Altman for classification and regression (Altman, 1992)**.** The KNN classifier uses majority voting between the labels for the K closest data points in the feature space in the training data to

assign a label to the new unseen test data. The distance metric used for determining the closest data points is generally the Euclidian distance for continuous variables or Hamming distance for discrete variables. Due to its simplicity, it has been popular for various applications including segmentation of MS lesions (Wu et al., 2006b) and WMHs (Anbeek et al., 2004).

Support Vector Machines

The idea of performing nonlinear classification using support vector machines (SVMs) was introduced by Boser et al. (Boser et al., 1992)**.** SVMs perform classification by finding a maximum-margin hyperplane that separates the two classes while maximizing the distance between the nearest point from either class. SVMs have been widely used for lesion segmentation tasks in MS populations (Ferrari et al., 2003; Abdullah et al., 2011) as well as for WMH segmentation in aging and AD populations (Ithapu et al., 2014; Quddus et al., 2005)**.**

Bagging

Bootstrap aggregating, also called bagging, is a model averaging technique initially introduced by Brieman et al. with the purpose of improving stability and reducing variance (Breiman, 1996). Bagging is an ensemble method that builds multiple classifiers such as decision trees by uniformly sampling the training data with replacement, and voting, to output a consensus prediction. Madabhushi used bagging for detecting prostatic adenocarcinoma from high resolution MR images (Madabhushi et al., 2006).

AdaBoost

Adaptive Boosting or AdaBoost was developed by Freund and Schapire (Freund et al.,

1999). AdaBoost performs classification by aggregating the outputs of other learning algorithms (also called weak learners) into a weighted sum that represents the final output of the boosted classifier. The subsequent weak learners are tweaked in favor of the instances that were misclassified by previous classifiers to improve classification accuracy. It has been used for MS lesion segmentation (Wels et al., 2008), interactive lesions segmentation (Li et al., 2007), as well as segmentation of WMHs (Quddus et al., 2005; Ghafoorian et al., 2016a).

For all classification tasks, the Scikit-learn Python library implementations were used (Pedregosa et al., 2011). For Naïve Bayes, LDA, QDA, SVM, and Decision Tree classifiers, the default settings were used. For KNN, 10 neighbours were used. For Bagging, KNN classifiers were used with the default parameters. For AdaBoost and Random Forests classifiers, 100 estimators were used. Ten-fold cross validation across subjects was used to train and validate the performance of the classifiers; i.e. no voxels from subjects used for validation were used in training and feature selection stages. It is worthwhile noting that the spatial WMH probability maps, average intensities, and P_{WMH} and P_H were also calculated through the cross-validation to avoid any overfitting (no data used in testing was used to generate the priors). All the segmentations were performed in the native space for the optimal primary modality to avoid resampling and further blurring of the lesion borders. To achieve this, all the priors and averages were first calculated in the stereotaxic template space and then registered back and recalculated in the native space using the inverse nonlinear transformations.

4.2.7. Evaluation metrics

There is no single similarity measure that can perfectly reflect the level of agreement between WMH segmentation maps. While Dice Kappa similarity measure (Dice, 1945) is the

most commonly used, the Kappa values are highly dependent on the WMH loads and lesion sizes. To address this, the mean Dice Kappa values are generally reported for different ranges of WMH loads, i.e. small (<5 cm^3), medium (5-20 cm^3), and large (>20 cm^3) separately (Admiraal-Behloul et al., 2005; Griffanti et al., 2016; Schmidt et al., 2012; Simões et al., 2013; Steenwijk et al., 2013; Dadar et al., 2017a). In this study, while Dice Kappa was used as the primary similarity measure for validation of the classifiers, other similarity measures such as the two-way mixed single measures with absolute agreement intra-class correlation coefficient (ICC) for the total WMH loads to assess the volumetric correspondence between the manual and automatic segmentations (Koch, 1982), true positive rate (TPR), positive prediction value (PPV), outline error rate (OER) measuring agreement of the raters in outlining of the same lesion (Wack et al., 2012), and detection error rate (DER) measuring agreement in detecting the same regions (Wack et al., 2012) are reported to facilitate comparison with previously published papers. Table 4 shows the list of these metrics along with their definitions.

Table 4.4. List of similarity measures and their definitions. The metrics are listed in the table below using the following abbreviations: true positive (TP), true negative (TN), false positive (FP), false negative (FN), true positive rate (TPR), Mean Square Within samples based upon the anova (MSW), Mean Square F Statistic Regression Slope (MSR). CR_1, CR_2, and C_{12} represent region from only rater 1, region from only rater 2, and the combination of both raters, respectively. $|cr|$ represents area of the connected region, $cr \in CR_1 \; or \; CR_2$ represents the set of connected regions that can be labeled either as CR_1 or CR_2. $|R_1(cr)|$, $|R_2(cr)|$ represent the areas of rater 1 and rater 2 regions within cr, respectively (Wack et al., 2012).

Name	Dice Kappa	Intra-class	Sensitivity	Outline Error Rate	Detection Error
Abbreviation	SI	ICC	TPR	OER	DER
Equation	$\frac{2 \times TP}{FP + FN + 2 \times TP}$	$\frac{MSR - MSW}{MSR + MSW}$	$\frac{TP}{TP + FN}$	$\sum_{cr \in C_{12}} \lvert cr \rvert - \lvert R_1(cr) \cap R_2(cr) \rvert$	$\sum_{cr \in CR_1 \; or \; CR_2} \lvert cr \rvert$

4.3. Results

4.3.1. Segmentation using T1w, T2w, PD, and FLAIR

The performance of each classifier was validated through 10-fold cross validation using T1w, T2w, PD, and FLAIR images for the ADC dataset. All voxels within a brain mask that contained the cerebrum, cerebellum and brain stem were classified. Table 4.5 shows the average Dice Kappa, detection/outline error rates (DER/OER), ICC, TPR, and PPV values for different classifiers. Figure 3 shows boxplot diagrams for the same results separately for subjects with small, medium and large WMH loads. Figure 4.4 shows the manual and automatic segmentation results of different classifiers on axial slices of one subject. To assess the statistical significance of the results, paired t-tests were performed on the Dice Kappa values of all pairs of classifier comparisons, and p-values were corrected for multiple comparisons using false discovery rate (FDR). Figure 4.5 shows the negative logarithm of the FDR corrected p-values.

Table 4.5. Comparison between mean Dice Kappa, detection/outline error rate (DER/OER), intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs using T1w, T2w, PD and FLAIR data in the ADC dataset. Blue color indicates the best performance in terms of SI.

Dataset	SI	DER	OER	ICC	TPR	PPV
Naïve Bayes	0.32±0.27	0.53±0.34	0.82±0.21	0.27	0.23	0.96
Logistic	0.57±0.22	0.32±0.36	0.54±0.14	0.97	0.65	0.57
LDA	0.56±0.23	0.41±0.38	0.46±0.20	0.88	0.48	0.83
QDA	0.36±0.26	0.55±0.36	0.74±0.17	0.44	0.26	0.96
KNN	0.66±0.17	0.18±0.18	0.52±0.18	0.99	0.73	0.65
Decision Trees	0.57±0.18	0.27±0.28	0.58±0.18	0.96	0.58	0.62
Random Forests	0.66±0.17	0.16±0.15	0.53±0.19	0.99	0.73	0.64
Bagging	0.63±0.19	0.21±0.26	0.57±0.03	0.99	0.75	0.58
SVM	0.57±0.24	0.32±0.42	0.54±0.11	0.98	0.66	0.60
AdaBoost	0.63±0.20	0.21±0.24	0.53±0.10	0.98	0.70	0.65

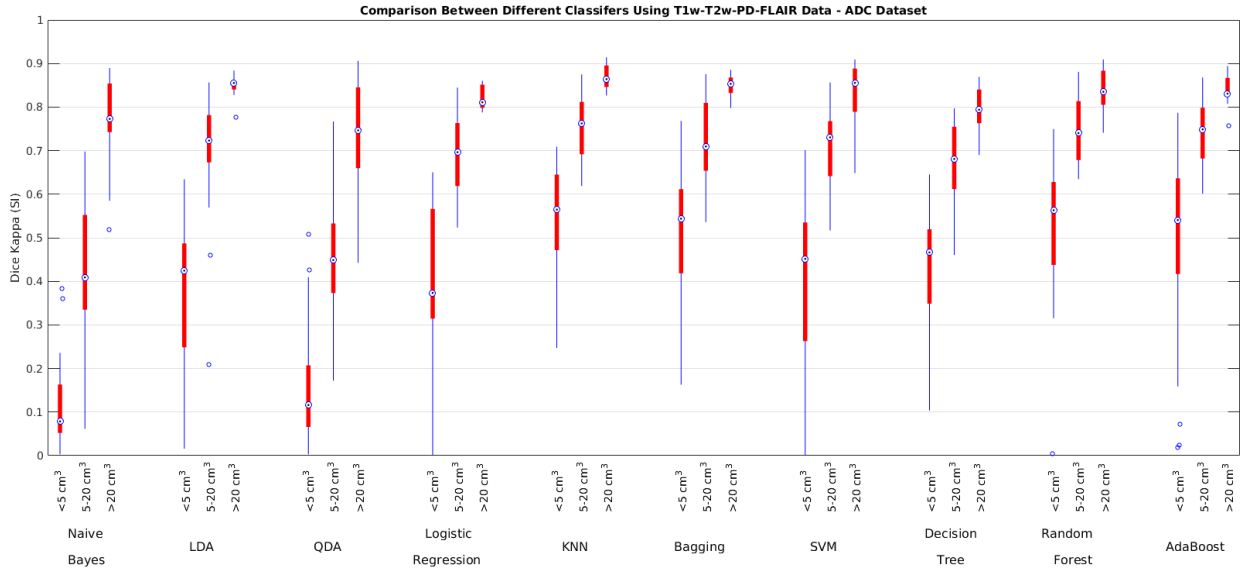


Fig. 4.3. Dice Kappa (SI) for different classification methods for (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w, T2w, PD, and FLAIR information for the ADC dataset.

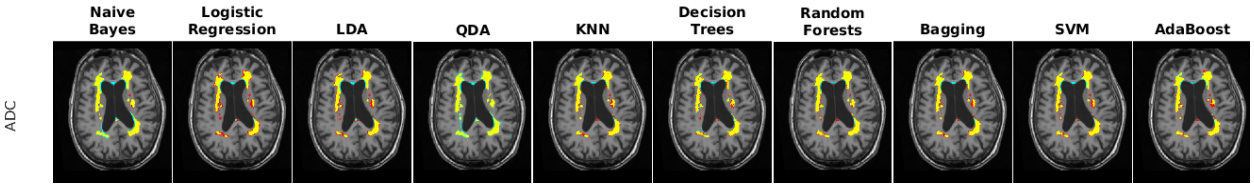


Fig. 4.4. Axial slices comparing manual and automatic segmentations using T1w, T2w, PD, and FLAIR information for a subject from ADC dataset. Yellow color indicates regions labeled as WMH in both manual and automatic segmentations, blue color indicates regions only segmented by the automatic technique, and red color indicates regions only segmented by the manual rater.

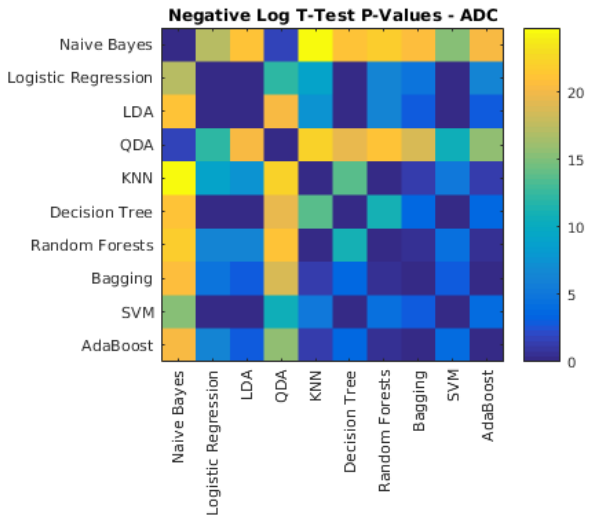


Fig. 4.5. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. Values higher than 1.3 are statistically significant.

4.3.2. Segmentation using T1w and FLAIR data

The performance of each classifier was validated through 10-fold cross validation using T1w and FLAIR images for the ADC, NACC, and ADNI2/GO datasets (recall that ADNI1 does not have FLAIR data). Table 4.6 shows the average Dice Kappa and detection/outline error rate (DER/OER) values for different classifiers. Table 4.7 shows corresponding ICC, TPR, and PPV values. Figure 4.6 shows boxplot diagrams for the same results separately for subjects with small, medium and large WMH loads. Figure 4.7 shows the manual and automatic segmentation results of different classifiers on axial slices of one subject. Figure 4.8 shows the negative logarithm of the FDR corrected p-values of t-tests on Dice Kappa values of different classifier pairs. To assess the contribution of T2w+PD features in the performance of different classifiers, paired t-tests were performed between the Dice Kappa values of the segmentations based on T1w+T2w+PD+FLAIR and T1w+FLAIR in the ADC dataset. The "*" in Table 4.7 indicates the significant differences between the two segmentations, after correction for multiple comparisons using FDR. The performance of Naïve Bayes, QDA, and Bagging has significantly dropped

without using FLAIR information.

Table 4.6. Comparison between mean Dice Kappa and detection/outline error rate (DER/OER) values of different classifiers for segmentation of WMHs using T1w and FLAIR data in the ADC, NACC, and ADNI2/GO datasets. Blue color indicates the best performance in terms of SI.

Dataset	ADC			NACC			ADNI2/GO		
Measure	SI	DER	OER	SI	DER	OER	SI	DER	OER
Naïve Bayes	0.42±0.25*	0.34±0.27	0.82±0.30	0.50±0.21	0.34±0.20	0.65±0.25	0.50±0.29	0.35±0.26	0.65±0.39
Logistic	0.56±0.18	0.27±0.24	0.61±0.19	0.65±0.13	0.13±0.13	0.58±0.18	0.64±0.20	0.19±0.25	0.52±0.22
LDA	0.58±0.19	0.35±0.33	0.49±0.17	0.69±0.13	0.15±0.15	0.50±0.19	0.60±0.23	0.15±0.17	0.66±0.37
QDA	0.42±0.23*	0.44±0.32	0.73±0.22	0.54±0.21	0.39±0.25	0.54±0.20	0.51±0.29	0.40±0.29	0.57±0.34
KNN	0.65±0.16	0.18±0.18	0.51±0.18	0.71±0.13	0.09±0.09	0.49±0.21	0.72±0.18	0.14±0.21	0.42±0.20
Decision Trees	0.58±0.16	0.25±0.25	0.58±0.14	0.65±0.12	0.16±0.16	0.54±0.14	0.65±0.22	0.21±0.28	0.49±0.21
Random Forests	0.66±0.14	0.18±0.18	0.50±0.16	0.72±0.10	0.09±0.10	0.46±0.16	0.72±0.19	0.14±0.21	0.42±0.22
Bagging	0.14±0.16*	0.27±0.28	0.63±0.27	0.69±0.13	0.10±0.11	0.51±0.21	0.69±0.17	0.14±0.22	0.46±0.21
SVM	0.56±0.24	0.31±0.37	0.56±0.26	0.67±0.13	0.09±0.08	0.56±0.22	0.68±0.22	0.19±0.28	0.46±0.28
AdaBoost	0.65±0.15	0.18±0.18	0.50±0.17	0.72±0.11	0.09±0.11	0.46±0.16	0.71±0.20	0.14±0.21	0.43±0.23

Table 4.7. Comparison between intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs in different datasets using T1w and FLAIR data for ADC, NACC, and ADNI2/GO datasets. Blue color indicates the best performance in terms of SI.

Dataset	ADC			NACC			ADNI2/GO		
Measure	ICC	TP	PP	ICC	TP	PP	ICC	TP	PP
Naïve Bayes	0.81	0.31	0.93	0.45	0.38	0.89	0.54	0.41	0.91
Logistic	0.98	0.73	0.48	0.85	0.78	0.59	0.86	0.70	0.70
LDA	0.98	0.56	0.65	0.92	0.78	0.63	0.80	0.63	0.71
QDA	0.77	0.30	0.94	0.53	0.42	0.91	0.57	0.41	0.95
KNN	0.99	0.76	0.60	0.94	0.80	0.69	0.96	0.74	0.78
Decision Trees	0.99	0.62	0.58	0.94	0.67	0.69	0.96	0.65	0.76
Random Forests	0.99	0.62	0.58	0.93	0.79	0.71	0.96	0.72	0.80
Bagging	0.16	0.63	0.09	0.89	0.83	0.63	0.91	0.76	0.70
SVM	0.95	0.70	0.56	0.90	0.83	0.60	0.95	0.67	0.79
AdaBoost	0.99	0.73	0.63	0.94	0.78	0.72	0.96	0.71	0.81

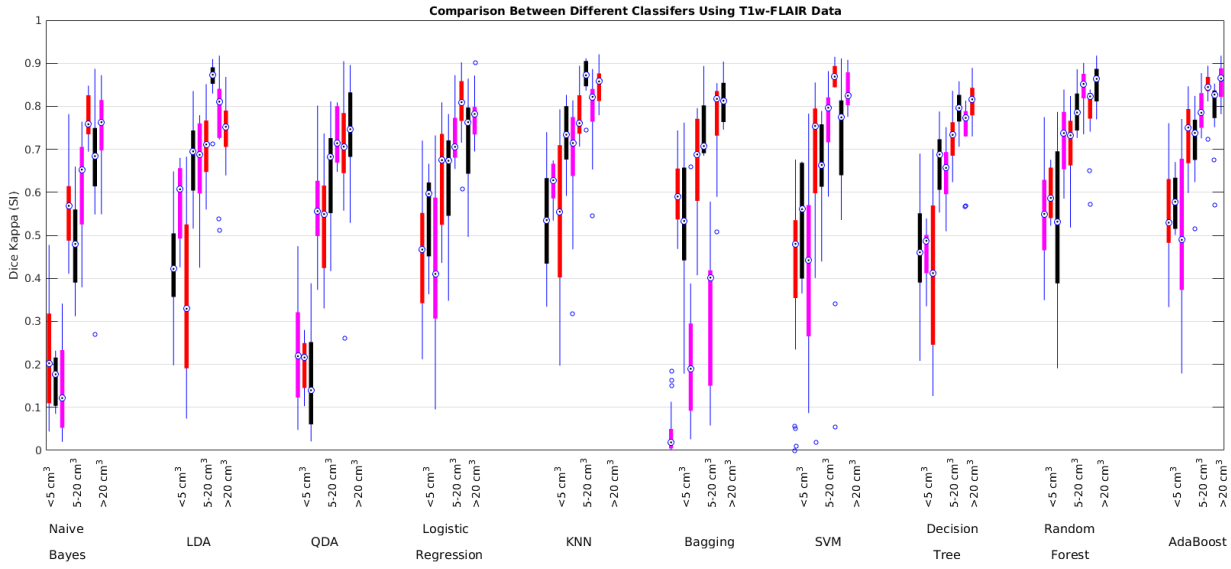


Fig. 4.6. Dice Kappa (SI) for different classification methods for low (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w and FLAIR information for ADC (red), NACC (black), and ADNI2/GO (magenta) datasets.

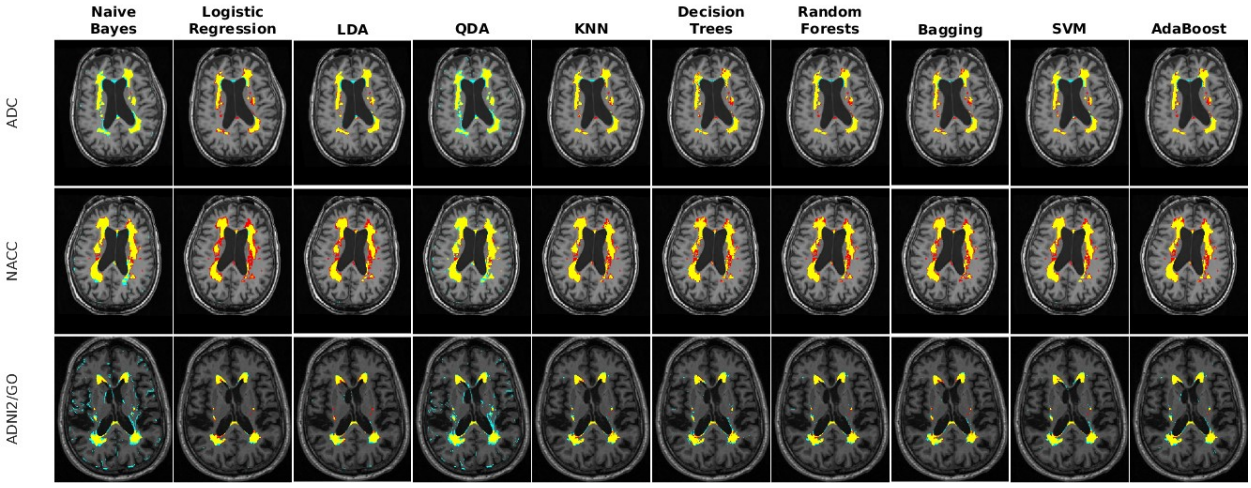


Fig. 4.7. Axial slices comparing manual and automatic segmentations using T1w and FLAIR information in one subject from each of ADC, NACC, and ADNI2/GO datasets. Yellow color indicates regions labeled as WMH in both segmentations, blue color indicates regions only segmented by the automatic technique, and red color indicates regions only segmented by the manual rater.

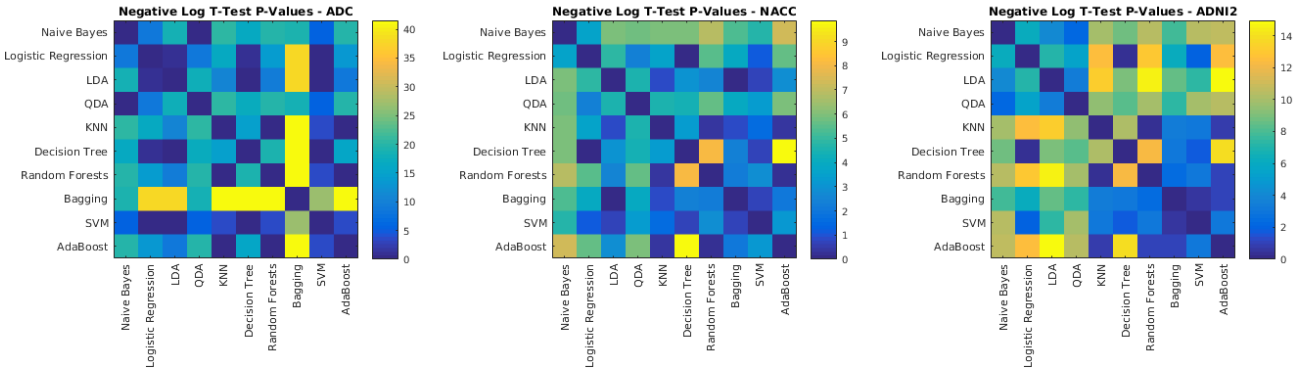


Fig. 4.8. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. Values higher than 1.3 are statistically significant.

4.3.3. Segmentation using T1w, T2w, and PD data

While FLAIR scans have the optimal contrast for differentiating WMHs from normal appearing white matter (Barkhof and Scheltens, 2002; Alexander et al., 1996; Bakshi et al., 2001), many studies forgo acquisition of FLAIR images in favour of other modalities. In order to take advantage of large studies such as ADNI1 that do not have FLAIR, segmentation methods that can provide accurate segmentation results without using the optimal FLAIR contrast are highly advantageous. A relatively easier task (in comparison to using FLAIR) is to segment WMHs from T1w, T2w, and PD or T1w, and T2w images. While segmenting WMHs solely from T1w images with high accuracy proves to be extremely difficult, being able to obtain an estimate of the WMH load that is significantly correlated with the actual loads can still be useful.

To address the first challenge, we trained and validated the performance of the classifiers using the features obtained from T1w, T2w, and PD images from the ADC and ADNI1 datasets. Table 8 shows the mean Dice Kappa, detection/outline error rates (DER/OER), and the corresponding ICC, TPR, and PPV values for each classifier and dataset, respectively. Figure 4.9 shows the corresponding boxplot diagrams for these results separately for subjects with small,

medium and large WMH loads. Figure 4.10 shows the segmentation results on the axial slices for different classifiers and datasets. Figure 4.11 shows the negative logarithm of the FDR corrected p-values of t-tests on Dice Kappa values of different classifier pairs. To assess the contribution of FLAIR features in the performance of different classifiers, paired t-tests were performed between the Dice Kappa values of the segmentations based on T1w+T2w+PD+FLAIR and T1w+T2w+PD in the ADC dataset. The "*" symbol in Table 4.8 indicates the significant differences between the two segmentations, after correction for multiple comparisons using FDR. The performance of all classifiers has significantly dropped without using FLAIR information.

Table 4.8. Comparison between mean Dice Kappa (SI), detection/outline error rate (DER/OER), intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs using T1w, T2w, and PD data, ADC and ADNI1 datasets. Blue color indicates the best performance in terms of SI.

Dataset	ADC						ADNI1					
Measure	SI	DER	OER	ICC	TPR	PPV	SI	DER	OER	ICC	TPR	PPV
Naïve Bayes	0.17±0.17*	0.88±0.39	0.77±0.32	0.10	0.11	0.84	0.34±0.22	0.73±0.33	0.59±0.32	0.09	0.24	0.89
Logistic	0.09±0.14*	1.26±0.73	0.55±0.55	0.46	0.24	0.06	0.44±0.23	0.38±0.48	0.73±0.33	0.68	0.65	0.39
LDA	0.28±0.21*	0.08±0.05	1.35±0.42	0.45	0.24	0.66	0.48±0.28	0.08±0.27	0.96±0.55	0.62	0.46	0.73
QDA	0.13±0.13*	0.63±0.38	1.11±0.37	0.08	0.08	0.90	0.31±0.21	0.64±0.34	0.74±0.31	0.12	0.20	0.93
KNN	0.28±0.24*	0.77±0.68	0.66±0.34	0.69	0.44	0.22	0.59±0.23	0.26±0.37	0.57±0.28	0.74	0.67	0.58
Decision Trees	0.38±0.20*	0.55±0.47	0.69±0.19	0.93	0.38	0.44	0.57±0.25	0.30±0.41	0.56±0.25	0.94	0.57	0.67
Random Forests	0.47±0.21*	0.36±0.34	0.69±0.21	0.95	0.60	0.42	0.66±0.23	0.18±0.33	0.50±0.27	0.94	0.67	0.71
Bagging	0.17±0.18*	0.88±0.70	0.77±0.46	0.37	0.51	0.11	0.54±0.22	0.22±0.38	0.70±0.33	0.59	0.75	0.47
SVM	0.31±0.21*	0.62±0.54	0.76±0.31	0.65	0.48	0.31	0.61±0.24	0.18±0.33	0.61±0.37	0.83	0.62	0.70
AdaBoost	0.44±0.21*	0.43±0.47	0.69±0.21	0.93	0.53	0.42	0.64±0.24	0.18±0.33	0.55±0.34	0.94	0.66	0.73

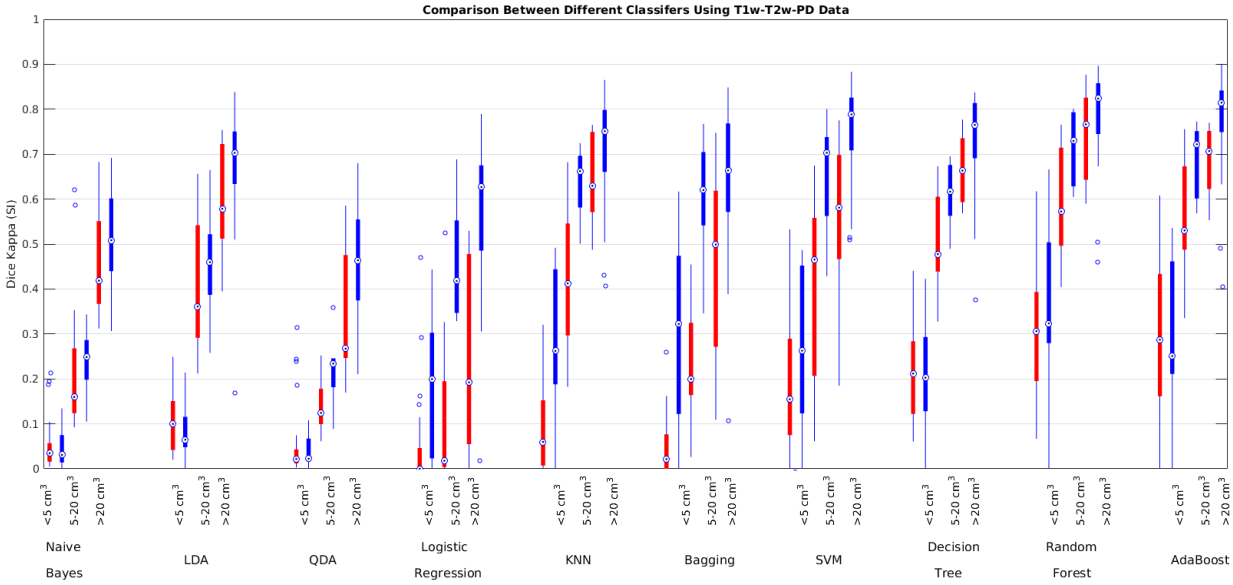


Fig. 4.9. Dice Kappa (SI) for different classification methods for low (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w, T2w, and PD information for ADC (red) and ADNI1 (blue) datasets.

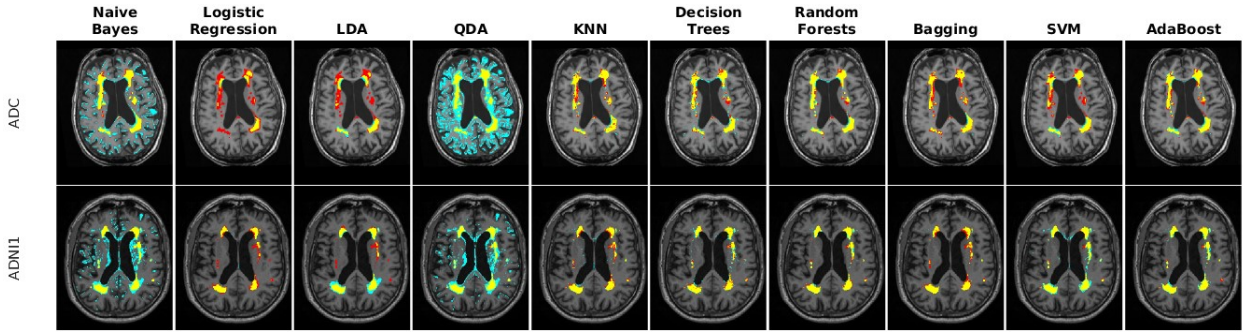


Fig. 4.10. Axial slice comparing manual and automatic segmentations using T1w, T2w, and PD information for a subject from each of ADC and ADNI1 datasets. Yellow color indicates regions labeled as WMH in both segmentations, blue color indicates regions only segmented by the automatic technique, and red color indicates regions only segmented by the manual rater.

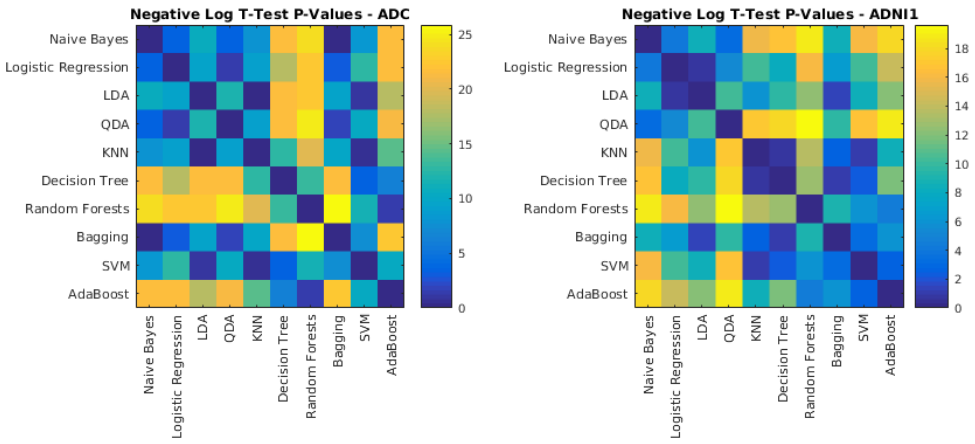


Fig. 4.11. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. Values higher than 1.3 are statistically significant.

4.3.4. Segmentation using T1w, and T2w data

Many studies forgo acquisition of PD images in favour of acquiring a higher resolution T2w image. Here we assess the performance of the classifiers without using PD images. Table 4.9 shows the mean Dice Kappa and detection/outline error rates (DER/OER), and the corresponding ICC, TPR, and PPV values for each classifier and dataset, respectively. Figure 4.12 shows the corresponding boxplot diagrams for these results separately for subjects with small, medium and large WMH loads. Figure 4.13 shows the segmentation results on the axial slices for different classifiers and datasets. Figure 4.14 shows the negative logarithm of the FDR corrected p-values of t-tests on Dice Kappa values of different classifier pairs. To assess the contribution of PD feature in the performance of different classifiers, paired t-tests were performed between the Dice Kappa values of the segmentations based on T1w+T2w+PD and T1w+T2w in ADC, and ADNI1 datasets. The "*" symbols in Table 4.9 indicate the significant differences between the two segmentations, after correction for multiple comparisons using FDR. No classifier has performed significantly worse after removing PD features for either dataset.

Table 4.9. Comparison between mean Dice Kappa (SI), detection/outline error rate (DER/OER), intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) values of different classifiers for segmentation of WMHs using T1w, and T2w data – ADC and ADNI1 datasets. Blue color indicates the best performance in terms of SI.

Dataset	ADC						ADNI1					
Measure	SI	DER	OER	ICC	TPR	PPV	SI	DER	OER	ICC	TPR	PPV
Naïve Bayes	0.25±0.22	0.59±0.37	0.68±0.41	0.37	0.18	0.79	0.43±0.26	0.51±0.30	0.62±0.38	0.51	0.33	0.87
Logistic Regression	0.16±0.16	0.62±0.70	0.71±0.47	0.43	0.42	0.12	0.42±0.25	0.38±0.51	0.78±0.40	0.79	0.66	0.35
LDA	0.28±0.21	0.08±0.23	1.08±0.55	0.46	0.24	0.66	0.48±0.28	0.08±0.28	0.96±0.55	0.61	0.46	0.72
QDA	0.20±0.18	0.61±0.32	0.80±0.38	0.23	0.13	0.86	0.36±0.23	0.61±0.33	0.68±0.33	0.28	0.25	0.92
KNN	0.27±0.24	0.46±0.59	0.54±0.25	0.83	0.44	0.21	0.58±0.23	0.30±0.42	0.54±0.21	0.90	0.67	0.55
Decision Trees	0.37±0.21	0.38±0.45	0.59±0.22	0.93	0.38	0.43	0.57±0.25	0.30±0.42	0.56±0.24	0.94	0.57	0.66
Random Forests	0.45±0.22	0.25±0.37	0.55±0.27	0.93	0.56	0.41	0.65±0.23	0.19±0.34	0.51±0.27	0.95	0.67	0.70
Bagging	0.24±0.22	0.49±0.65	0.60±0.34	0.68	0.55	0.17	0.57±0.24	0.26±0.43	0.59±0.26	0.86	0.73	0.52
SVM	0.37±0.18	0.47±0.51	0.71±0.35	0.58	0.48	0.48	0.46±0.23	0.36±0.42	0.73±0.37	0.23	0.54	0.62
AdaBoost	0.44±0.21	0.29±0.47	0.57±0.31	0.92	0.53	0.42	0.64±0.25	0.19±0.34	0.54±0.32	0.95	0.66	0.72

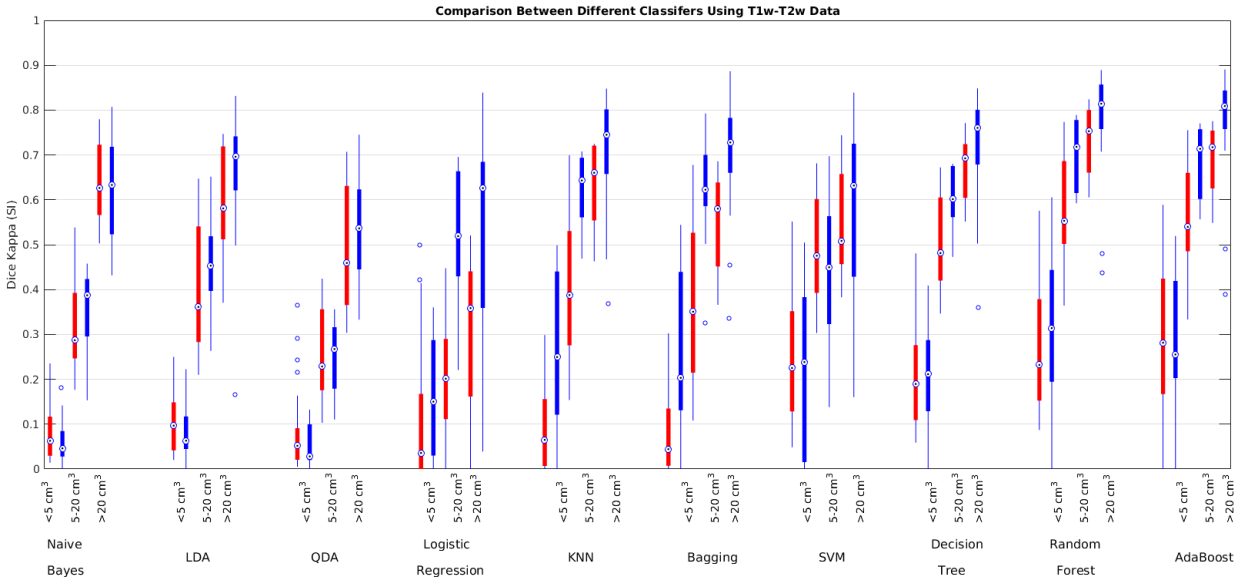


Fig. 4.12. Dice Kappa (SI) for different classification methods for low (<5 cm^3, left), medium (5-20 cm^3, middle), and high (>20 cm^3, right) WMH load using T1w and T2w information for ADC (red) and ADNI1 (blue) datasets.

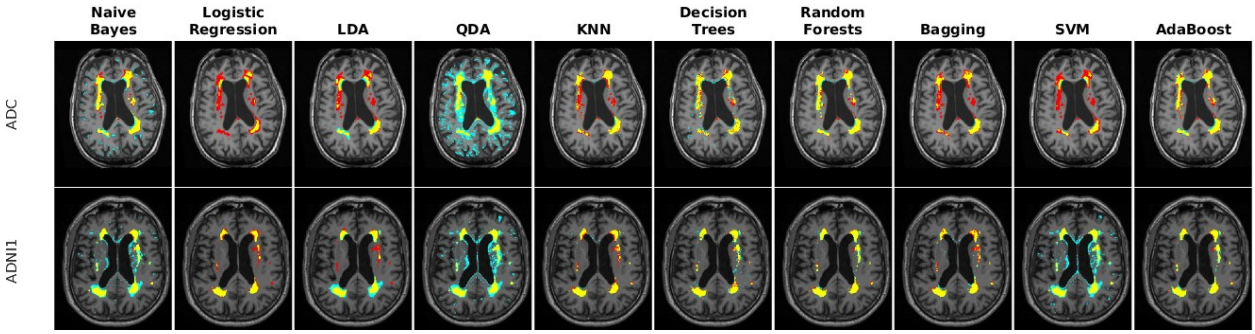


Fig. 4.13. Axial slice comparing manual and automatic segmentations using T1w and T2w information for ADC and ADNI1 datasets. Yellow color indicates regions labeled as WMH in both segmentations, blue color indicates regions only segmented by the automatic technique, and red color indicates regions only segmented by the manual rater.

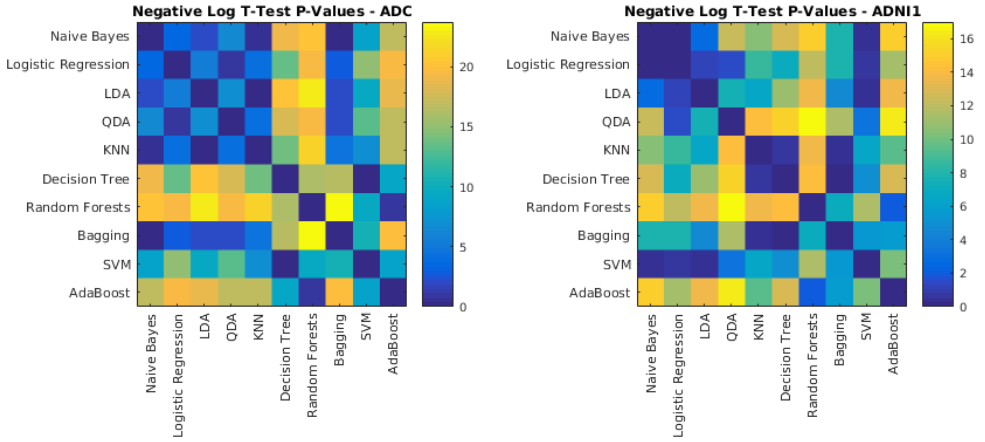


Fig. 4.14. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. Values higher than 1.3 are statistically significant.

4.3.5. Segmentation using only T1w data

To address the second challenge, we trained and validated the performance of the classifiers with features only from T1w images from ADC, NACC, ADNI1, and ADNI2/GO datasets. Table 4.10 shows the mean Dice Kappa and detection/outline error rates (DER/OER), for each classifier and dataset. Table 4.11 shows the corresponding ICC, TPR, and PPV values. Figure 4.15 shows boxplot diagrams for these results separately for subjects with small, medium and large WMH loads. Figure 4.16 shows the segmentation results on the axial slices for

different classifiers from each study. Figure 4.17 shows the negative logarithm of the FDR corrected p-values of t-tests on Dice Kappa values of different classifier pairs.

Table 4.10. Comparison between mean Dice Kappa and detection/outline error rate (DER/OER) values of different classifiers for segmentation of WMHs using T1 data for ADC, NACC, ADNI1, and ADNI2/GO datasets. Blue color indicates the best performance in terms of SI.

Dataset	ADC			NACC		
Measures	SI	DER	OER	SI	DER	OER
Naïve Bayes	0.24±0.21	0.08±0.03	1.44±0.21	0.32±0.15	0.07±0.04	1.27±0.30
Logistic	0.11±0.14	1.26±0.80	0.52±0.14	0.08±0.13	1.32±0.85	0.51±0.64
LDA	0.25±0.20	0.09±0.05	1.40±0.20	0.34±0.14	0.08±0.05	1.25±0.29
QDA	0.20±0.17	0.28±0.16	1.30±0.17	0.32±0.14	0.18±0.09	1.17±0.30
KNN	0.28±0.18	0.57±0.48	0.86±0.18	0.33±0.13	0.26±0.20	1.06±0.21
Decision Trees	0.24±0.18	0.76±0.49	0.75±0.18	0.30±0.11	0.34±0.15	1.05±0.13
Random	0.34±0.19	0.51±0.44	0.82±0.19	0.40±0.12	0.22±0.14	0.97±0.20
Bagging	0.03±0.03	0.98±0.72	0.86±0.03	0.08±0.12	0.91±0.77	0.96±0.68
SVM	0.16±0.11	0.65±0.41	1.02±0.11	0.28±0.10	0.24±0.17	1.20±0.18
AdaBoost	0.26±0.10	0.48±0.36	0.99±0.10	0.36±0.11	0.20±0.12	1.08±0.20
	ADNI1			ADNI2/GO		
Measures	SI	DER	OER	SI	DER	OER
Naïve Bayes	0.42±0.27	0.07±0.28	1.10±0.55	0.38±0.25	0.05±0.04	1.19±0.50
Logistic	0.37±0.19	0.19±0.34	1.07±0.37	0.31±0.15	0.29±0.34	1.10±0.29
LDA	0.44±0.26	0.09±0.27	1.04±0.52	0.41±0.24	0.08±0.05	1.10±0.47
QDA	0.44±0.27	0.07±0.28	1.05±0.55	0.41±0.25	0.06±0.04	1.12±0.49
KNN	0.51±0.22	0.27±0.42	0.72±0.30	0.46±0.19	0.31±0.42	0.77±0.25
Decision Trees	0.41±0.23	0.34±0.35	0.84±0.28	0.39±0.21	0.34±0.27	0.89±0.21
Random	0.51±0.24	0.25±0.40	0.73±0.31	0.48±0.21	0.26±0.34	0.77±0.23
Bagging	0.30±0.21	0.35±0.51	1.12±0.45	0.20±0.17	0.47±0.61	1.12±0.51
SVM	0.36±0.24	0.15±0.32	1.13±0.48	0.39±0.18	0.15±0.15	1.07±0.36
AdaBoost	0.50±0.23	0.18±0.35	0.81±0.37	0.48±0.19	0.20±0.27	0.84±0.27

Table 4.11. Comparison between intra-class correlation (ICC), true positive rate (TPR), and positive prediction value (PPV) of different classifiers for segmentation of WMHs in different datasets using T1w data for ADC, NACC, ADNI1, and ADNI2/GO datasets. Blue color indicates the best performances in terms of SI.

Dataset	ADC			NACC			ADNI1			ADNI2/GO		
Measure	ICC	TP	PP	ICC	TP	PP	ICC	TP	PP	ICC	TP	PP
Naïve Bayes	0.24	0.20	0.67	0.01	0.28	0.56	0.30	0.37	0.74	0.06	0.33	0.74
Logistic Regression	0.08	0.27	0.0	0.00	0.20	0.07	0.22	0.61	0.34	0.17	0.65	0.25
LDA	0.32	0.22	0.64	0.07	0.29	0.54	0.45	0.42	0.68	0.19	0.39	0.68
QDA	0.40	0.15	0.67	0.52	0.25	0.56	0.38	0.40	0.74	0.17	0.37	0.73
KNN	0.36	0.49	0.22	0.14	0.61	0.25	0.55	0.63	0.50	0.54	0.61	0.43
Decision Trees	0.55	0.24	0.37	0.36	0.30	0.38	0.63	0.41	0.54	0.62	0.40	0.51
Random Forests	0.56	0.45	0.31	0.54	0.55	0.36	0.60	0.59	0.56	0.65	0.57	0.50
Bagging	0.01	0.59	0.02	0.23	0.58	0.05	0.23	0.73	0.21	0.16	0.71	0.15
SVM	0.03	0.49	0.13	0.10	0.54	0.22	0.20	0.59	0.41	0.14	0.56	0.43
AdaBoost	0.25	0.49	0.21	0.10	0.61	0.30	0.52	0.60	0.56	0.51	0.59	0.51

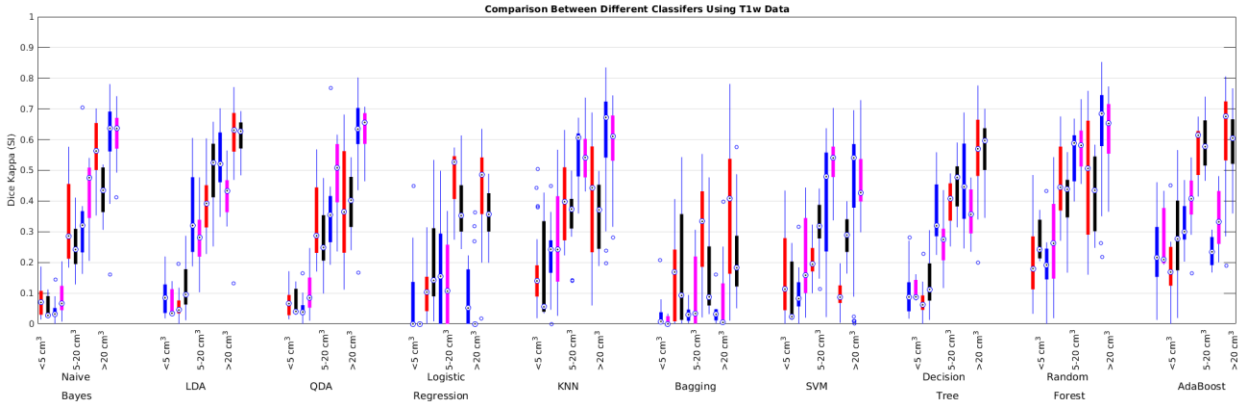


Fig. 4.15. Dice Kappa (SI) for different classification methods for low (<5 cm^3), medium (5-20 cm^3), and high (>20 cm^3) WMH load using only T1w information for ADC (red), NACC (black), ADNI1 (blue), and ADNI2/GO (magenta) datasets.

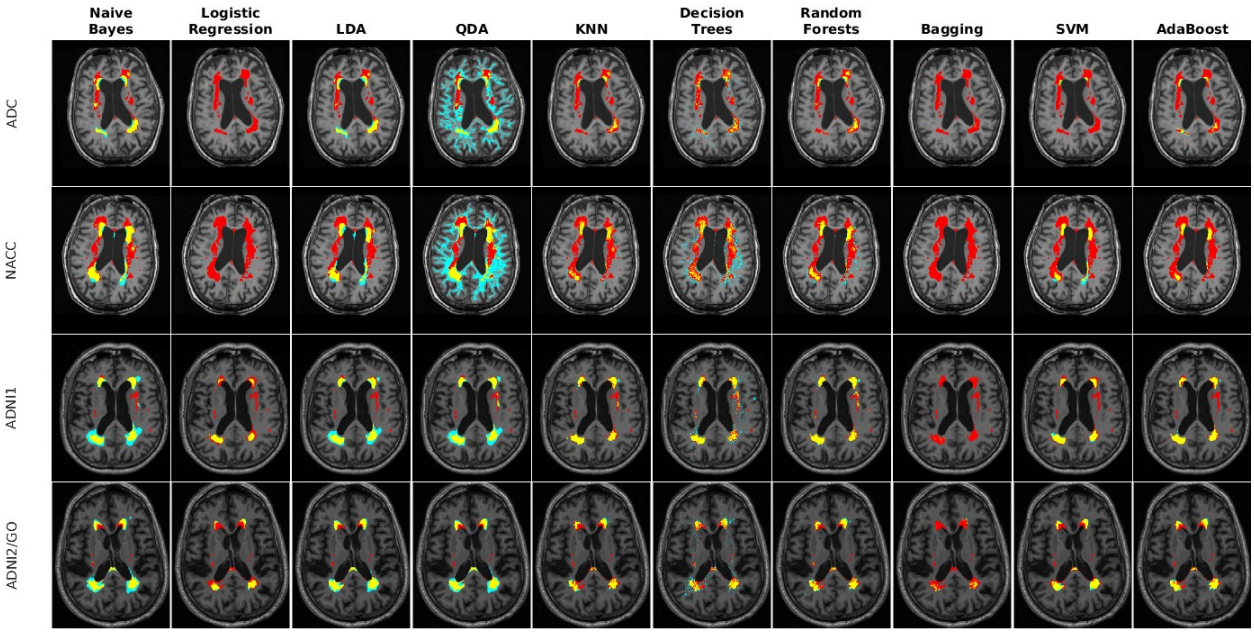


Fig. 4.16. Axial slice comparing manual and automatic segmentations using T1w information for ADC, NACC, ADNI1, and ADNI2/GO datasets. Yellow color indicates regions labeled as WMH in both segmentations, cyan color indicates regions only segmented by the automatic technique, and red color indicates regions only segmented by the manual rater.

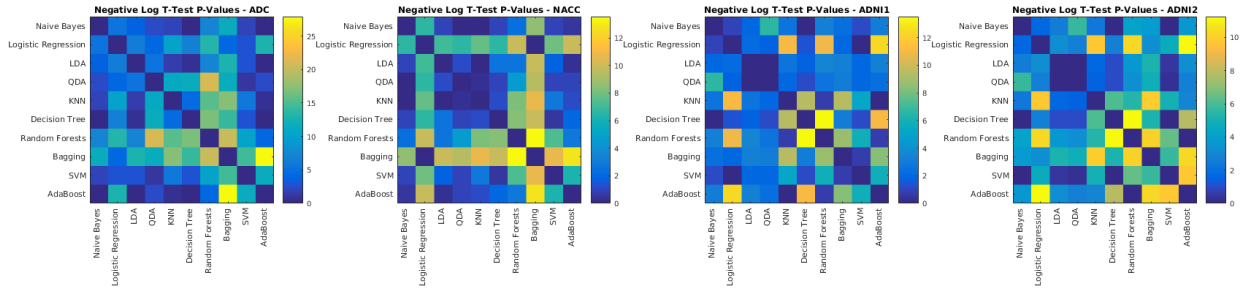


Fig. 4.17. Negative logarithm of FDR corrected p-values of paired t-tests between Dice Kappa values of classifier pairs. Values higher than 1.3 are statistically significant.

4.3.6. Over-segmentation/Under-segmentation

To provide information regarding over-segmentation/under-segmentation of WMHs, paired t-tests were performed between total WMH loads in small, medium, and large groups on T1w+FLAIR (n=147) and T1w+T2w+PD (n=123) experiments. Table 4.12 shows the mean and standard deviation of the volumes as well as statistical significance of the differences after

correcting for multiple comparisons using FDR correction. From the results, we can see that Naïve Bayes and QDA significantly oversegment WMHs in all three groups. Logistic regression and Bagging significantly undersegment medium and large WMHs. LDA and Decision Trees seem to work well with T1w+FLAIR images, but they tend to significantly oversegment when dealing with T1w+T2w+PD sequences. AdaBoost, KNN, SVM and Random Forest seem to work very well for medium and large WMHs, but slightly oversegment small lesions. However, KNN and SVM seem to show a lot of variability (high standard deviations) for small lesions using T1w+T2w+PD sequences.

Table 4.12. Mean ± standard deviation of WMH loads in small (<5cm^3), medium (5-20cm^3), and large (>20cm^3) groups. Statistically significant differences from manual segmentations after corrections for multiple comparisons using false discovery rate (FDR) correction are indicated with *.

Sequences	T1w-FLAIR			T1w-T2w-PD		
Method	Small	Medium	Large	Small	Medium	Large
Manual	1.85±1.39	12.07±4.58	40.02±23.33	1.47±1.14	10.81±4.27	47.62±25.74
Naïve Bayes	18.65±12.52*	33.45±16.79*	68.31±37.88*	91.29±98.89*	89.85±33.56*	148.31±48.49*
Logistic Regression	2.53±2.59*	9.83±4.98*	34.03±25.35*	2.18±5.68	4.44±8.32*	32.80±27.80*
LDA	3.97±3.09*	12.48±5.76	39.49±30.83	22.44±8.16*	28.72±12.14*	53.47±19.22
QDA	17.26±11.21*	32.53±16.05*	73.04±42.40*	121.10±89.30*	129.28±43.98*	205.15±69.82*
KNN	2.42±2.46*	11.00±5.81	41.53±30.49	4.74±21.28	8.74±8.53	45.50±25.54
Decision Trees	3.44±3.24*	12.64±6.35	41.63±25.35	5.10±4.41*	14.03±9.86*	54.30±26.60*
Random Forests	2.62±2.61*	11.47±5.87	40.79±25.37	2.58±3.29	11.51±9.72	51.57±27.38
Bagging	1.46±2.41	7.13±6.23*	28.13±21.69*	3.12±14.95	5.16±7.17*	32.86±20.56*
SVM	2.76±3.23*	11.05±5.90	41.17±30.55	5.49±8.53*	11.98±14.87	49.28±29.49
AdaBoost	2.85±2.78*	11.93±5.97	41.08±25.60	3.42±4.05*	11.94±10.25	52.47±29.93

4.3.7. Computational burden

In order for a segmentation technique to be applicable to large-scale datasets, reasonable computation time and memory demands are crucial. To assess this, all classifiers were trained on the same dataset consisting of 50 subjects and used to segment 20 subjects on an Intel(R)

Core(TM) i7-5600 CPU @ 2.60 GHz machine with 20.0 GBs RAM. Table 4.13 shows the training as well as segmentation time per subject in seconds for each classifier.

Table 4.13. Comparison between training and segmentation times (s) between different classifiers

Classifier	Training Time (s)	Segmentation Time
Naïve Bayes	12.45	0.38
Logistic Regression	333.38	0.11
LDA	66.31	0.52
QDA	100.56	0.45
KNN	7718.98	3021.88
Decision Trees	1225.63	0.53
Random Forests	22620.11	7.29
Bagging	8992.54	981.55
SVM	14581.04	0.26
AdaBoost	100766.02	71.16

4.4. Discussion

In the recent years, there have been many different studies in the literature that address the challenge of automatically segmenting WMHs (Caligiuri et al., 2015; Admiraal-Behloul et al., 2005; Anbeek et al., 2004; Beare et al., 2009; De Boer et al., 2009; Dyrby et al., 2008; Ghafoorian et al., 2016a; Griffanti et al., 2016; Ithapu et al., 2014; Lao et al., 2008; Ong et al., 2012; Schmidt et al., 2012; Simões et al., 2013; Steenwijk et al., 2013; Wu et al., 2006a, 2006b; Yoo et al., 2014; García-Lorenzo et al., 2013; Shiee et al., 2010). However, drawing meaningful comparisons between these segmentation techniques proves to be practically impossible since the results are greatly influenced by the MRI acquisition characteristics and resolution as well as the quality of the manually segmented labels that are used for training and validation. Here we have validated and compared the performance of a variety of different supervised linear and nonlinear classifiers in segmenting WMHs using 4 relatively large datasets. We also provide our fully automated tool for segmentation of WMHs from multiple contrasts of MR images along

with the pre-trained classifiers.

Several commonly used linear and nonlinear classifiers with different levels of computational complexity were employed for segmentation of WMHs from multiple contrasts of MR images. In presence of FLAIR information, most methods performed relatively well and can be employed for WMH segmentation. However, the performance of the classifiers declined significantly in absence of the optimal FLAIR modality information, with Random forests and AdaBoost classifiers still retaining the best performance. Using only T1w images, the performance of all classifiers declined drastically with random forest and AdaBoost classifiers still providing the best results. These segmentations tend to detect only the brightest of the WMHs. However, their high volumetric correlation with the gold standard values shows that while not perfectly accurate, they still might be used as surrogate measures to reflect WMH burden if they are also associated with risk factors and clinical measures. This can prove extremely valuable in studies that only have T1w scans and need to take into account the WMH burden.

One of the major issues when using automated techniques for segmenting WMHs is the variability caused by differences in the scanner and acquisition sequences which would in turn lead to differences in contrast and borders of WMHs. As a result, classifiers that are trained on data from a single scanner with a specific acquisition sequence tend to perform poorly on data from different scanners and/or sequences. To increase the generalizability of our tools, we have trained and validated our classifiers using data from different scanners/sites.

It would be worthwhile to note that all of the voxels inside the brain were input to the classifiers and no white matter mask or any mask excluding either ventricles or cerebrospinal

fluids were used. This makes the classification task more challenging, but on the other hand, makes the performance of the classifiers more easily comparable with other methods since the results will not be dependent on the quality of the tissue segmentation algorithm or whether specific regions such as brainstem or cerebellum which are generally more challenging to segment are masked out. Another valid concern in using tissue segmentation results is that most tissue classification techniques use only T1w images, on which some of the WMHs appear hypointense. This makes the tissue classification results prone to error since they will be likely to classify WMHs as grey matter while most WMHs occur in the white matter. This misclassification in the initial tissue segmentation will add an extra level of noise to the data that can significantly affect WMH segmentation results. One limitation of our technique is that it has not been validated on patients with stroke; the intensity profile in such subjects is likely very different from the subjects evaluated here.

In detecting WMHs, FLAIR is of the highest importance since it provides the best lesion to WM contrast when compared with T1w, T2w and PD sequences. PD provides the most variable contrast difference between tissue types directly related to the parameters used in its acquisition. The more T2 weighted the PD sequence, the less supplemental contrast information it provides (since the information is already provided by the T2w sequence). Hence, the PD sequence is most meaningful if the parameters allow the CSF to be of the lowest possible signal. The T1w sequence on its own should only be considered in cases where other modalities are not available or their poor image quality prevents their use. The lower information given by T1w images resides in a poorer contrast between the signal of lesions and surrounding WM. Lesion intensity spans from iso-intense to WM to deep hypointense, causing the difficulty in detecting lesions using only T1w images. Another factor that can significantly affect the quality of both

manual and automated segmentations is the signal to noise ratio (SNR). A lower SNR will impact the image quality and number of artifacts, which would then translate into poorer performance of either software or manual rater. The ADC, NACC, and ADNI2 FLAIRs had an average SNR value of 17.25±2.37, 20.11±5.52, and 35.11±7.26 as estimated by our denoising tool (Coupe et al., 2008), respectively. This may partially explain the poorer results for ADC data. As a general rule, the highest possible SNR should be attained in each modality employed. In addition to SNR, ringing or ghosting caused by movement and inter-package motion also contribute to the deterioration of image quality.

Manually segmenting WMHs is a challenging task. Lesion edges always exhibit a degree of hyperintense signal that decreases gradually towards the healthy surrounding WM. In other words, no lesion edge goes from one pathologic hyperintense voxel, to a contiguous healthy hypointense WM voxel, and the edges may shift from scoring to scoring by one or two voxels. Additionally, when cases have multiple lesions, the surface to volume ratio of the lesions increases. Even when the rater identifies exactly the same lesions, one extra voxel around the edge of a small lesion may have a large impact on the Dice Kappa value. The small DER values for the manual segmentations further confirm that most of the disagreement between the manual segmentations occurs around the edges (0.03±0.04, 0.05±0.04, 0.03±0.04, and 0.04±0.04 for ADC, NACC, ADNI1, and ADNI2, respectively). Also, the poorer image quality, in terms of SNR, of the ADC dataset, could partially account for the worse intra-rater performance for that dataset.

Segmenting WMHs without the optimal FLAIR modality is a challenging task. Additional errors might arise from comparing segmentations obtained without FLAIR with manual labels that are based on FLAIR images. The extent and borders of WMHs generally do

not look the same on the different MRI sequences (Filippi et al., 1996). It has been shown that FLAIR sequence is less sensitive in detecting thalamic lesions in vascular disease populations (Leite et al., 2004). Furthermore, FLAIR may present hyperintense artifacts that can lead to an increase in false positives such as the hyperintensities often observed in insula (Hirai et al., 2000). As a result, a certain degree of disagreement between segmentations obtained with and without FLAIR information is expected. This explains the higher SI values for the ADNI1 dataset where the manual segmentations are based on T2w/PD scans compared with automatic segmentations with the same contrasts in ADC dataset (T1w, T2w, and PD) where FLAIR information was used for the manual segmentations. Additionally, the difference in tissue contrast between the PD sequence of the ADC dataset and ADNI1 may also partially account for the higher SI value for ADNI1. The PD scans in ADNI1 dataset had a higher white-to-grey matter contrast, higher white matter-to-lesion contrast, and better delineation of CSF as a different tissue type, given its low signal. All these characteristics were absent in the ADC dataset, where PD was heavily T2 weighted. These differential characteristics are critical in the WMHs segmentation process either by a rater or an automatic tool, improving the accuracy of the segmentation in the ADNI1 cases.

Using classifiers such as KNN and Bagging with KNN has the additional drawback of longer computation time for segmenting new data. The fact that they do not require rigorous training is generally outweighed by their longer classification times, especially when one needs to segment 100s or 1000s of MRI volumes in larger datasets. In addition, these methods are generally more susceptible to skewedness in class distributions, which is the case in lesion segmentation tasks, since most voxels in the brain are non-WMHs. As a result, the examples of the more frequent non-WMH class tend to dominate the new predictions, simply owing to the

fact that they are more common.

Accurate quantification and localization of WMHs is critical since they are important clinical measures in the elderly and AD populations. A Dice Kappa value of 0.7 is considered as a good segmentation in the literature (Caligiuri et al., 2015). Random forest was able to obtain average Dice Kappa values higher than 0.7 for the medium lesion load and 0.8 for large lesion load groups, which is considered as excellent agreement. Their average Dice Kappa for the small lesion group was higher than 0.5, which is still considered as a very good agreement, especially considering the fact that Dice Kappa values are smaller for objects with a high surface to volume ratio, as is the case for subjects with small lesion loads.

The Random Forests technique consistently had the best results across all the experiments when using Dice Kappa (SI) as the primary measure of comparison. Considering the fact that it also had a shorter computational time than the second-best classifier (AdaBoost), Random Forests was the best classifier amongst the nonlinear classification techniques tested. The Linear Discriminant Analysis method was the best linear classifier considering the Dice Kappa results and computation times.

Random Forests and AdaBoost classifiers are both highly non-linear (locally) compared to other methods such as SVM, QDA, and LDA which estimate a model to explain the variability in the data and perform classification. In addition, due to its nature, Random Forest works well with a mixture of categorical variables and numerical variables with various scales, while classifiers that rely on a notion of distance such as SVM and KNN have difficulty in such problems.

In cases where different classifiers have different strengths and weaknesses, using an ensemble of all the classifiers can improve the overall classification accuracy. Here, performing a voting between the outputs of all 10 classifiers achieved Dice Kappa values of 0.68±0.17, 0.74±0.10, 0.66±0.22, and 0.72±0.19 (versus 0.66±0.17, 0.72±0.10, 0.66±0.23, and 0.72±0.19 for the Random Forest classifier) for ADC, NACC, ADNI1, and ADNI2 datasets, respectively, suggesting a slight improvement for ADC ($p=0.001$), and NACC ($p=0.004$), and no difference for ADNI1 and ADNI2 ($p>0.05$).

As mentioned previously, drawing meaningful comparisons between techniques that have been applied to different datasets, using different brain masks, and with different definitions of WMHs should be done with care. Taking these considerations into account, our Random Forests classifier performs very well in comparison with other methods in the field (Table 4.14).

Table 4.14. Comparison of SI (Dice Kappa) for different lesion loads in various studies. (S: small load, M: medium load, L: large load).

Method	Technique	Number (S-M-L%)	Dice (SI)			
			S	M	L	Total
Proposed pipeline	Random Forests	70 (36-23-11)	0.55	0.75	0.84	0.66
		32 (6-11-12)	0.57	0.73	0.84	0.72
Dadar (Dadar et al., 2017a)	Linear regression + thresholding	80 (58-31-11)	0.49	0.74	0.87	0.62
		40 (25-14-1)	0.48	0.64	0.74	0.51
Admiraal (Admiraal-Behloul et al., 2005)	Fuzzy inference	100 (40-35-25)	0.70	0.75	0.82	0.75
Anbeek (Anbeek et al., 2004)	K-nearest neighbors	20 (40-35-25)	0.50	0.75	0.85	0.61
Beare (Beare et al., 2009)	AdaBoost	30		0.50	0.65	0.58
Boer (De Boer et al., 2009)	K-nearest neighbors	20		0.72		0.72
Steenwijk (Steenwijk et al., 2013)	K-nearest neighbors	20 (15-45-40)	0.78	0.85	0.91	0.84
		18 (40-33-17)	0.65	0.72	0.81	0.75
Khayati (Khayati et al., 2008)	Adaptive Mixture Model	20 (35-50-15)	0.72	0.75	0.80	0.75
Sajja (Sajja et al., 2006)	Parzen Window	23 (35-65)	0.67		0.84	0.78
Schmidt (Schmidt et al., 2012)	Markov random field	53	0.66	0.79	0.85	0.75
Sheei (Shiee et al., 2010)	Fuzzy segmentation	10		0.63		0.63
Ong (Ong et al., 2012)	Adaptive trimmed mean	38	0.36	0.56	0.71	0.47

Ithapu (Ithapu et al., 2014)	Random Forests	38		0.67		0.67
	Support Vector Machine			0.54		0.54
Herskovits (Herskovits et al., 2008)	Bayesian classification	42		0.60		0.60
Dyrby (Dyrby et al., 2008)	Neural networks	362	0.45	0.62	0.65	0.56
Erus (Erus et al., 2014)	Abnormality detection +	33		0.54		0.54
	principal component analysis	47		0.66		0.66
Ghafoorian (Ghafoorian et al., 2016b)	Convolutional neural networks	46		0.79		0.79
Simões (Simões et al., 2013)	Gaussian Mixture Model	28 (14-9-5)	0.51	0.70	0.84	0.62
Yoo (Yoo et al., 2014)	Variable thresholding	32 (7-10-15)	0.59	0.73	0.86	0.76
Griffanti (Griffanti et al., 2016)	K-nearest neighbors	21	0.70	0.69	0.80	0.76
		109	0.41	0.58	0.68	0.52

Accurate quantification of WMHs is critical for evaluating the vascular burden contributing to cognitive deficits in the vascular dementia and AD patients as well as the aging population in general. Due to the high variability across different populations, image acquisition parameters and manual segmentation protocols, comparing different techniques in a meaningful way is practically impossible. Here we have extensively compared 10 most widely used off-the-shelf classifiers in segmenting WMHs with and without FLAIR information in terms of accuracy and computational burden. These experiments have enabled us to draw meaningful and generalizable comparisons between different methods and determine which classifiers are best suited to the task of segmenting WMHs.

Chapter 5. Validation of T1w-based Segmentations of White Matter Hyperintensity Volumes in Large Scale Datasets of Aging

Preface

In this chapter, we assess whether T2w/PD and FLAIR sequences are necessary for studying WMHs in aging and AD populations. WMHs in ADNI1 and ADNI2/GO datasets were segmented using the pipeline proposed in the previous chapter and a Random Forest classifier, once using all available modalities and once using only T1w images. The WMH volumes were extracted for each lobe and hemisphere separately using these two sets of segmentations and correlated with i) manually segmented volumes, ii) each other, iii) cognitive and clinical measures.

The results showed that even though the T1w segmentations underestimated the WMH volumes, they still were able to hold strong correlations with the actual WMH loads as well as cognitive and clinical measures, and that there was no statistically significant difference between the correlations based on these volumes and the more accurate volumes based on all available modalities.

This work has been published as:

Dadar, M., Maranzano, J., Ducharme, S., Carmichael, O. T., Decarli, C., & Collins, D. L. (2018). Validation of T1w-based segmentations of white matter hyperintensity volumes in large-scale datasets of aging. *Human brain mapping*, *39*(3), 1093-1107.

Validation of T1w-based Segmentations of White Matter Hyperintensity Volumes in Large Scale Datasets of Aging

Mahsa Dadar, Josefina Maranzano, Simon Ducharme, Owen T. Carmichael, Charles Decarli, D. Louis Collins, Alzheimer's Disease Neuroimaging Initiative[3]

Abstract:

Introduction: Fluid-attenuated Inversion Recovery (FLAIR) and dual T2w and Proton Density (PD) magnetic resonance images (MRIs) are considered to be the optimum sequences for detecting white matter hyperintensities (WMHs) in aging and Alzheimer's disease populations. However, many existing large multi-site studies forgo their acquisition in favour of other MRI sequences due to economic and time constraints.

Methods: In this paper, we have investigated whether FLAIR and T2w/PD sequences are necessary to detect WMHs in Alzheimer's and aging studies, compared to using only T1w images. Using a previously validated automated tool based on a Random Forests classifier, WMHs were segmented for the baseline visits of subjects from ADC, ADNI1, and ADNI2/GO studies with and without T2w/PD and FLAIR information. The obtained WMH loads (WMHLs) in different lobes were then correlated with manually segmented WMHLs, each other, age,

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

cognitive and clinical measures to assess the strength of the correlations with and without using T2w/PD and FLAIR information.

Results: The WMHLs obtained from T1w-Only segmentations correlated with the manual WMHLs (ADNI1: r=0.743, p<0.001, ADNI2/GO r=0.904, p<0.001), segmentations obtained from T1w+T2w+PD for ADNI1 (r=0.888, p<0.001) and T1w+FLAIR for ADNI2/GO (r=0.969, p<0.001), age (ADNI1: r=0.391, p<0.001, ADNI2/GO: r=0.466, p<0.001), and ADAS13 (ADNI1: r=0.227, p<0.001, ADNI2/GO: r=0.190, p<0.001), and NPI (ADNI1: r=0.290, p<0.001, ADNI2/GO: r=0.144, p<0.001), controlling for age.

Conclusion: Our results suggest that while T2w/PD and FLAIR provide more accurate estimates of the true WMHLs, T1w-Only segmentations can still provide estimates that hold strong correlations with the actual WMHLs, age, and performance on various cognitive/clinical scales, giving added value to datasets where T2w/PD or FLAIR are not available.

Keywords: White Matter Hyperintensities, Aging, Alzheimer's Disease

5.1. Introduction

White matter hyperintensities (WMHs), defined as regions of higher signal than the surrounding normal appearing white matter (NAWM) on T2w or FLAIR MR images, are one of the most common findings in structural MR imaging in older adults, reflecting demyelination and axonal loss (Prins and Scheltens, 2015). While sensitive as an expression of abnormality in the white matter (WM) tissue, the etiology of WMHs is quite varied, with ischemia due to cerebral small vessel disease playing an important role in the majority of older subjects (Gouw et al., 2010; Yoshita et al., 2005). This age-related ischemic small vessel disease is also referred to

as arteriolosclerosis, hypertension-related or vascular-risk-factor-related small vessel disease (Pantoni, 2010). However, the term small vessel disease is also related to other pathologies that affect small arteries, arterioles, venules and capillaries, such as cerebral amyloid angiopathy, genetic small vessel disease distinct from amyloid angiopathy (e.g. cerebral autosomal dominant/recessive arteriopathy with subcortical infarcts and leukoencephalopathy or CADASIL/CARASIL), inflammatory mediated small vessel disease (e.g. primary angiitis of central nervous system CNS, Wegener's granulomatosis), and venous collagenosis (Pantoni, 2010).

Some of the consequences of small vessel disease include lacunar infarcts, WMHs, micro and macro bleeding. The first two phenomena are easily detected on MR images. In contrast, small vessels cannot be seen using MRI, so the term small vessel disease on MRI has been used for (and become equivalent to) WMHs and lacunar infarcts (Pantoni, 2010). Unfortunately, there is great heterogeneity across neuropathological centers regarding the definition of MRI small vessel disease, with overall agreement lower than 50% (Pantoni et al., 2006). Since there is no conclusive data showing the levels of specificity and sensitivity of WMHs on MRI as a reflection of a specific etiology of small vessel disease, we consider, in our cases, that the two major groups (arteriolosclerosis and amyloid angiopathy) are probably the main substrates of the WMHs. These two etiologies on their own have a crucial role in three major clinical areas: stroke, neurocognitive disorders (dementia), and aging related cognitive decline (Pantoni, 2010).

The location and load of WMHs have been shown to correlate with age, a history of hypertension, hyperinsulinemia (Hawkins et al., 2017), as well as cognitive deficits (Biesbroek et al., 2017; DeCarli et al., 1995a; Dubois et al., 2014). Therefore, WMHs constitute a clinically meaningful biomarker of cognitive decline related to general aging and pathological vascular

processes, which are known contributors to multi-factorial neurodegenerative diseases (Iturria-Medina et al., 2016). They are a particularly important clinical measure in the elderly populations (Carmichael et al., 2010; De Groot et al., 2002; DeCarli et al., 2001; Dubois et al., 2014; Longstreth et al., 1996; van Straaten et al., 2008).

WMHs are generally assessed using FLAIR or T2w/PD scans which have optimum contrast for detecting such lesions (Caligiuri et al., 2015). T2w/PD and FLAIR WMHs have been shown to correspond to myelin stain lesions in post-mortem histology studies (Fernando et al., 2004; Takao et al., 1999). However, WMHs can also be detected on T1w scans to some extent. The characteristic bright WMH signal of FLAIR, T2w and PD, manifests in the T1w sequence as a hypointense area, heterogeneous in the value of the lower signal, ranging from iso-intense to hypo-intense in relation to the surrounding NAWM. In other words, a FLAIR, T2w/PD homogeneous hyperintense area would correspond, in the T1w modality, to a similar area of heterogeneous hypointense signal, ranging from values close to fluid, to isointense in relation to the surrounding NAWM. This phenomenon is probably determined by the different types and degrees of change occurring in the WM at the same time (e.g. more/less intense demyelination and axonal loss). This range of T1w hypointensities in a region of WM tissue is more homogeneously represented by the bright signal of T2w and FLAIR sequences. Since the quantification of lesion volumes in a given MRI modality depends on the contrast between the lesional area and the surrounding NAWM, these volumes will always be larger if the detection considers the bright signal of FLAIR or T2w scans, as opposed to using only T1w images. The hypointensity seen on T1w images presumably reflects the most severe spectrum of WM injury.

Although FLAIR or T2w/PD scans are the optimal sequences to detect WMHs, many especially large-scale studies forgo acquisition of either one or all of the optimal modalities

because of time and financial constraints. There can also be differences in WMH volume levels when comparing T2w/PD with FLAIR scans, with FLAIR scans tending to give higher overall levels. Consequently, it would be extremely useful if one can get an estimate of the load and location of WMHs without requiring these optimal modalities. While there have been other studies that define and use T1w white matter signal abnormality (WMSA) detected by Freesurfer (Fischl, 2012) as a measure of WMH in aging and AD populations (Jacobs et al., 2013; Leritz et al., 2014; Salat et al., 2010), to our knowledge, no studies have investigated and validated the relationship between these T1w hypointensities and FLAIR or T2w/PD based WMH segmentations and whether there is a significant difference in their relationships with clinical measures.

In this paper, we aimed to compare the ability of T1w, T2w/PD and FLAIR scans in differentiating between healthy tissue and WMHs, both in terms of (i) detection in comparison with manually segmented labels and (ii) correlation with a variety of clinical measures. Our goal is to determine if WMHs can be partially but accurately segmented based only on T1w images, and how reliable T1w-based assessments are in comparison with the more accurate estimates obtained based on FLAIR or T2w/PD sequences.

5.2. Materials and Methods

5.2.1. Subjects

The WMHs were segmented both manually and automatically in three different datasets to ensure generalizability of the results. Table 5.1 summarizes the information for each dataset.

(i) The first dataset (ADC) consists of 70 elderly individuals who received a full clinical workup and structural MR scans including T1w, double-echo PD/T2w, and FLAIR scans at their enrollment into the University of California, Davis Alzheimer's Disease Center (ADC) (Hinton et al., 2010). Subjects were 70-90 years old with either normal cognition, mild cognitive impairment, or AD. All subjects were manually segmented by an expert rater.

(ii) The second dataset included subjects selected from ADNI study. This data was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). ADNI was carried out with the goal of recruiting 800 adults aged from 55 to 90, and consists of approximately 200 cognitively normal, 400 MCI, and 200 AD subjects. ADNIGO is a later study that followed ADNI participants that were in cognitively normal or early MCI stages (http://www.adcs.org/studies/imagineadni.aspx). ADNI2 study followed patients in the same categories as well as recruiting 550 new subjects (http://www.adcs.org/studies/ImagineADNI2.aspx). Baseline visit data from ADNI1 and ADNI2/GO subjects were used in this study (Table 5.1). 46 subjects with T1w and FLAIR scans and different loads of WMHs were selected from ADNI2/GO study for manual segmentation. To ensure that the datasets used

for training and validation of the method have a wide range of WMHs, segmentation techniques generally make sure to include subjects with small, medium and large WMH loads (Dadar et al., 2017b; Griffanti et al., 2016; Schmidt et al., 2012; Simões et al., 2013). Here, subjects were selected from different sites and scanners and a preliminary assessment was performed to evaluate their WMH load with the goal of acquiring subjects with different scanner information as well as different loads of WMHs. For each scanner model, we selected datasets that had low (<5 CCs), medium (5-20 CCs) and high lesion loads (> 20 CCs). Equal numbers of male and female subjects were selected. The age of the subjects was also considered for the selection, with the aim of achieving a normal distribution. Using a similar strategy, 53 subjects with T1w, T2w and PD scans and different WMH loads were selected from the ADNI1 study for manual segmentation.

5.2.2. Clinical evaluations

We did not have the clinical evaluations available for the ADC study. The clinical assessment and cognitive testing of ADNI study followed a standardized protocol that has been described previously (Petersen et al., 2010). At each visit, the participants underwent a standardized clinical evaluation and cognitive tests including Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), Functional Assessment Questionnaire (FAQ), Neuropsychiatric Inventory (NPI), a composite score for executive function (Gibbons et al., 2012), and Immediate, Forgetting and Learning sub-scores from Rey Auditory Verbal Learning Task (RAVLT). Each individual had a self-reported history

of hypertension, and cardiovascular risk factors available. Table 5.1 summarizes this information for the subjects that were used in this study. For details on the administration and scoring see http://www.adni-info.org/Scientists/ADNIData.html.

Table 5.1. Descriptive statistics for the ADNI subjects enrolled in this study. Data are number or mean± standard deviation. ADNI=Alzheimer's Disease Neuroimaging Initiative. MMSE= Mini-Mental State Examination. ADAS= Alzheimer's Disease Assessment Scale. FAQ= Functional Assessment Questionnaire. RAVLT= Rey Auditory Verbal Learning Task (I=Immediate, F=Forgetting, L=Learning). Executive= Executive Function. NPI= Total Neuropsychiatric Inventory Score.

Test	ADNI1	ADNI2/GO
Number (Male)	669 (393)	481 (298)
Age	75.26±6.84	72.62±7.54
MMSE	26.71±2.70	27.58±2.76
ADAS11	11.63±6.24	9.89±6.80
ADAS13	18.45±9.08	15.29±9.66
FAQ	4.88±6.47	3.44±5.70
RAVLTI	32.09±11.17	37.72±13.08
RAVLTF	4.31±2.38	4.29±2.62
RAVLTL	3.61±2.62	4.65±2.80
Executive	0.71±0.52	0.91±0.60
NPI	10.30±0.51	7.39±7.77

5.2.3. MR imaging

This section describes the scanner information and image acquisition parameters for the abovementioned datasets. Table 5.2 shows the summary of this information for each sequence in each dataset.

(i) ADC dataset: MRI data was acquired on two 1.5T MRI scanners: a GE MEDICAL SYSTEMS Signa scanner located at UCD Medical Center (Sacramento, CA), and a Philips Eclipse scanner located at the Veterans Administration Northern California Health Care System (Martinez, CA). T1w

scans were acquired with a FSPGR pulse sequence with 1.5 mm slice thickness, 128 slices covering the entire brain, a 250×250 mm field of view, a 256×256 scan matrix, voxel size of 0.9765×0.9765×1.5 mm, repetition time (TR) of 9 ms and echo time (TE) of 2.9 ms. FLAIR scans were acquired with a fast spin echo (FSE) sequence with 3 mm slice thickness, a 220×220 mm field of view, and a 256×192 scan matrix, voxel size of 0.9765×0.9765×3 mm, TR = 11000 ms and TE = 144 ms. Analogous sequences were installed on both the GE and Philips scanners.

(ii) ADNI1 dataset: The MRI data used was acquired on scanners from three different manufacturers, GE, Philips, and SIEMENS. All patients had similar MRI protocols for T1w and T2w/PD scans. T1w scans were acquired in 3D with a gradient recalled sequence with 1.2 mm slice thickness, 160 sagittal slices, covering the entire brain, a 192×192 mm field of view, and a 192×192 scan matrix, voxel size of 1.2×0.9375 ×0.9375 mm, TR = 3000 ms and TE = 3.55 ms. T2w/PD scans were acquired in 2D with a FSE sequence with 3.0 mm slice thickness, 56 axial slices covering the entire brain, a 256×256 mm field of view, and a 256×256 scan matrix, a voxel size of 0.8594×0.8594×5 mm, with TR = 3000 ms, TE = 95.2 ms for T2w and TE=10.5 ms for PD images.

(iii) ADNI2/GO dataset: The MRI data used was acquired on scanners from three different manufacturers, Philips, GE and SIEMENS. All patients had similar MRI protocols for T1w and FLAIR scans. T1w scans were acquired in 3D with a gradient recalled sequence with 1.2 mm slice thickness, 196 sagittal slices, covering the entire brain, a 256×256 mm field of view, and a 256×256 scan matrix, voxel size of 1×1×1.2 mm, TR = 7.2 ms and TE = 3.0 ms. FLAIR scans

were acquired in 2D with a spin echo inversion recovery sequence with 5.0 mm slice thickness, 42 axial slices covering the entire brain, a 256×256 mm field of view, and a 256×256 scan matrix, voxel size of 0.8594×0.8594×5 mm, with TR = 11000 ms, TE = 150 ms.

Table 5.2. Scanner information and MRI acquisition parameters for ADC, ADNI1, and ADNI2/GO datasets.

Modality	Dataset	ADC	ADNI1	ADNI2/GO
T1w	Scanner Manufacturer	GE/Philips	GE/Philips/SIEMEN	Philips/SIEMENS
	Slice thickness (mm)	1.5	1.2	1.2
	No. of slices	128	160	196
	Field of view (cm^2)	250×250	192×192	256×256
	Scan Matrix (cm^2)	256×256	192×192	256×256
	TR: Repetition time (ms)	9	3000	7.2
	TE: Echo time (ms)	2.9	3.55	3.0
	Pulse Sequence	FSPGR	MPRAGE	MPRAGE
Other	Contrast	T2w/PD/FLAIR	T2w/PD	FLAIR
	Scanner Manufacturer	3/3/3	3	5
	Slice thickness (mm)	42/42/48	56	42
	No. of slices	240×240/240×240/220×220	256×256	256×256
	Field of view (cm^2)	256×256/256×256/256×192	256×256	256×256
	Scan Matrix (cm^2)	2420/2420/11000	3000/3000	11000
	TR: Repetition time (ms)	90/20/144	95.2/10.5	150
	TE: Echo time (ms)	DSE/DSE/FSE	FSE	SE/IR

5.2.4. Pre-processing

All MRI scans were pre-processed using our standardized pipeline. Images were denoised (Manjón et al., 2010), corrected for image intensity inhomogeneity (Sled et al., 1998) and intensity scaled (Fonov et al., 2011a). The T2w, PD, and FLAIR scans were then co-registered to the structural T1w scan of the same subject using a six-parameter rigid body registration (Collins et al., 1994). The T1w scans were nonlinearly registered to the ADNI template based on intensity correlation coefficient (Collins and Evans, 1997). The quality of the nonlinear registrations was visually assessed and the results that did not pass this quality control were

discarded. Using the T1w-to-template transformations (i.e., linear + nonlinear), the other modalities (e.g., FLAIR, T2w, PD) were registered to the ADNI template as well. The manually segmented lesion maps were also registered to the ADNI template using the transformations of their corresponding FLAIR images.

5.2.5. Manual segmentation

In ADC and ADNI2/GO datasets, the WMHs were manually segmented by an expert with more than 12 years of experience in reading MRI and developing standardized MRI guidelines to detect WM lesions using different image modalities (Maranzano et al., 2016) with FLAIR used as the primary contrast and with T1w used to aid in the decision process to include or exclude a voxel from the lesion mask. In ADNI1, T2w images were used as the primary contrast with T1w and PD used to aid in the decision process. The rater defined the voxel of interest according to anatomical location and intensity information in all given MRI modalities. To be considered a WMH, a given voxel had to be hyperintense in relation to the surrounding NAWM on T2w, PD or FLAIR. The same voxel had to be iso to hypointense in relation to the NAWM on T1w images. Previous work (Dadar et al., 2017b) showed that intra-rater Dice Kappa was 0.72, 0.80 and 0.86 for ADC, ADNI1 and ADNI2/GO, respectively.

5.2.6. Automatic segmentation

A previously validated fully automatic WMH segmentation technique was used to automatically segment the WMHs in all three datasets using a set of intensity and spatial features and a Random Forest classifier (Dadar et al., 2017b, 2017a). The intensity features include voxel intensity for all available modalities, the probability of a specific intensity value being a WMH

(P_{WMH}) or non-WMH ($P_{non\text{-}WMH}$) for each available modality, and the ratio of these two probabilities for each available modality. The spatial features include a spatial WMH probability map indicating the probability of a voxel at a specific location being a WMH and the average intensity of a non-WMH voxel at that specific voxel location for each available modality. After preprocessing and co-registration of all available modalities, these spatial and intensity features are calculated for each modality. The Random Forest classifier is then trained using these features to segment the WMHs (Dadar et al., 2017b).

For each dataset, the automatic technique was first trained and validated based on the manual segmentations. Two sets of automatic segmentations were completed, first with all available modalities (referred to as All-Contrasts segmentations) and second without using the T2w, PD, and FLAIR information (referred to as T1w-Only segmentations). The trained classifiers were then used to segment the entire ADC, ADNI1 and ADNI2/GO datasets. The quality of the segmentations was then assessed and verified by a human expert. The volumes of the WMHs for the left and right frontal, parietal, temporal, and occipital lobes as well as the entire brain were calculated by nonlinearly warping the Hammers atlas (Hammers et al., 2003) to the T1w scans of individual subjects and normalized for head size to make population comparisons possible.

The WMH volumes obtained from the T1w-only segmentations were then correlated with All-Contrasts volumes as well as clinical scores. False discovery rate (FDR) correction was performed to correct for all multiple comparisons separately for each dataset (significance threshold = 0.05). All the correlations with clinical scores were performed with log transformed WMHs (to achieve normal distribution), controlling for age.

The key concern when using T1w-only segmentations is whether the WMH portions that are not captured by T1w-only segmentations are clinically significant. To assess whether the difference between the WMH volumes provided by T1w-Only and All-Contrast segmentations is statistically significant, a general linear model was used to regress the contrast ($WMHL_{All\text{-}Contrast}$-$WMHL_{T1\text{-}Only}$) and each measure.

Figure 5.1 shows axials slices of T1w, T2w, PD, and FLAIR images for a subject from ADC, T1w, T2w, and PD images for a subject from ADNI1 and T1w and FLAIR images for a subject from ADNI2/GO, along with the manual segmentations, as well as T1-Only and All-Contrasts automated segmentations. While All-Contrasts segmentations conform very well with the manual labels, the T1-Only segmentations seem to mostly capture the brightest of the WMHs.

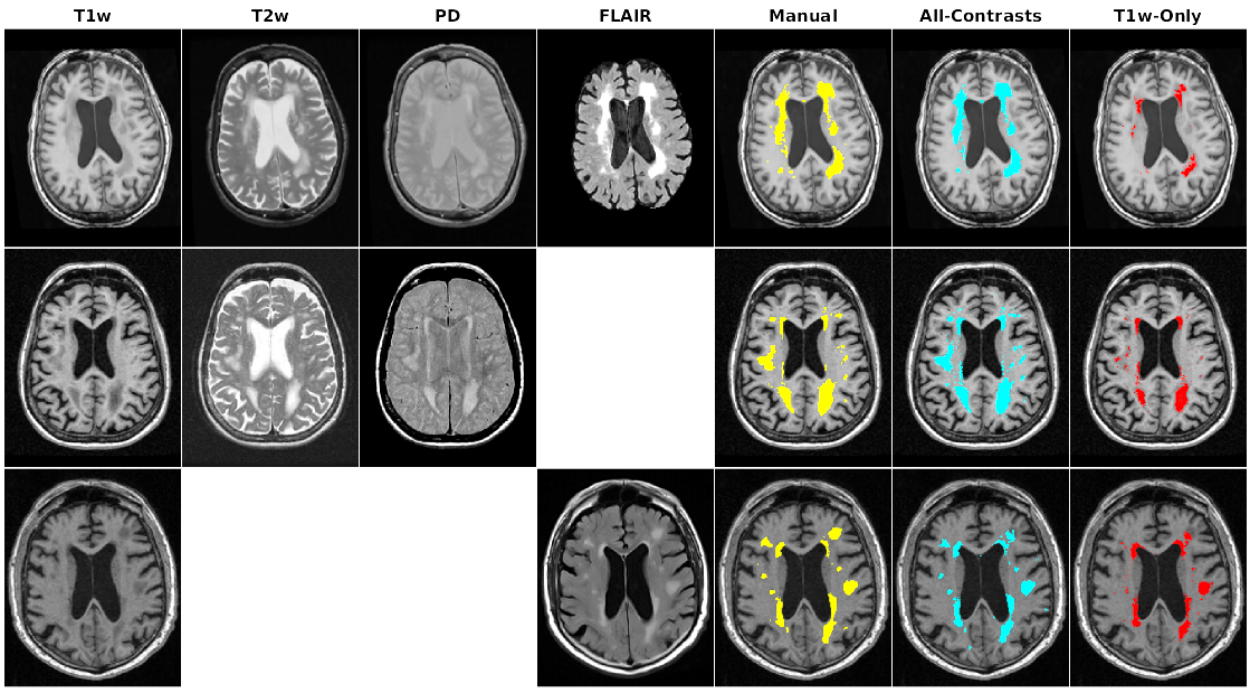


Fig. 5.1. Axial slices showing T1w, T2w, PD, and FLAIR images as well as manual (yellow), All-Contrasts (cyan), and T1w-Only (red) segmentations of subjects from ADC (top), ADNI1 (middle), and ADNI2/GO (bottom) datasets.

5.3. Results

5.3.1. Comparison of Tissue Histograms

Here we compare the histograms of WMHs with white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) density histograms using the manually segmented labels for each dataset (Fig. 5.2). Table 3 shows the percentage of overlap between the density histograms of WMHs and WM, GM, and CSF. The tissue histograms show the greatest separation of WMH with GM, WM, and CSF in FLAIR contrast images, followed by T2w, PD, and T1w.

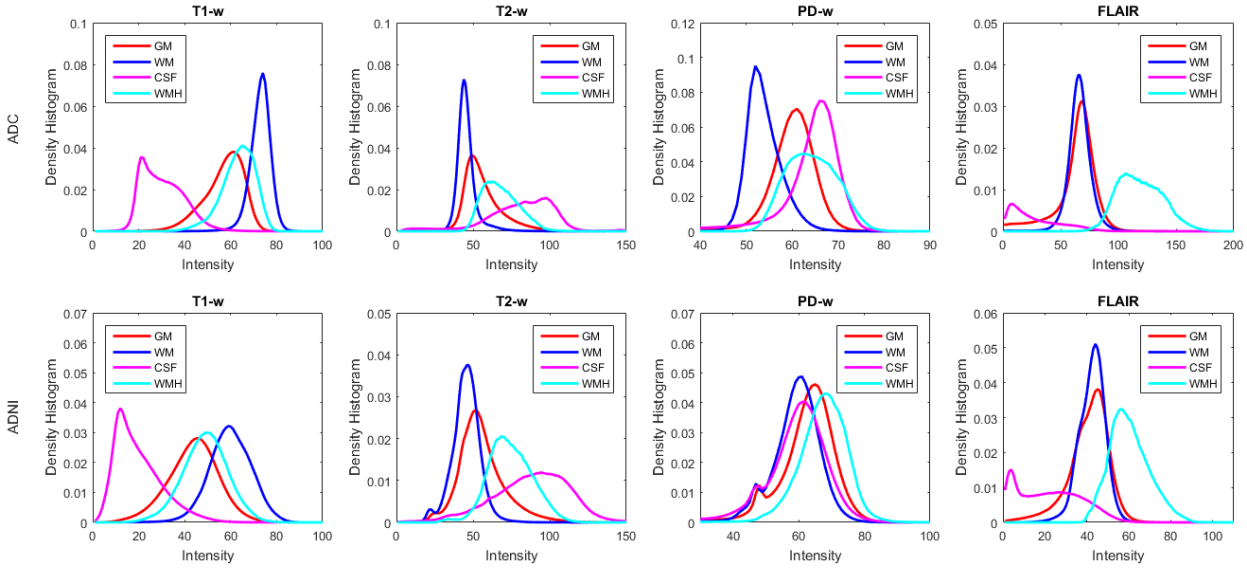


Fig. 5.2. Intensity histograms of white matter (WM), grey matter (GM), cerebrospinal fluid (CSF), and white matter hyperintensities (WMHs) for a. ADC and b. ADNI datasets.

Table 5.3. Percentage of overlap between density histograms of WMHs and GM, WM, and CSF. WMH= White Matter Hyperintensity. GM= Gray Matter. WM= White Matter.

Dataset	Dataset	WM	GM	CSF
ADC	T1w	22.33	37.71	5.33
	T2w	9.02	29.80	30.75
	PD	15.77	36.47	56.91
	FLAIR	2.73	3.76	1.02
ADNI	T1w	29.69	42.20	9.45
	T2w	9.32	23.95	32.95
	PD	28.96	39.90	32.37
	FLAIR	12.41	14.53	5.00

5.3.2. Comparisons with Manual Segmentations

In order to assess the importance of using the T2w, PD, and FLAIR information, the WMH loads obtained from segmentations with and without the information of the optimal modality (i.e. T2w/PD and FLAIR sequences for ADNI1 and ADNI2/GO datasets, respectively) were correlated with the equivalent volumes obtained from the manual segmentations. Correlations were computed for total brain WMH, and for lobar WMH loads. Fig. 3 shows this information for the three datasets. While the T1-Only volumes systematically underestimate the gold standard volumes (obtained from the manual segmentations), they are still able to retain high correlations in all regions and datasets ($r=0.963$, $p<0.001$ for ADC, $r=0.743$, $p<0.001$ for ADNI1, and $r=0.904$, $p<0.001$ for ADNI2/GO for whole brain T1w-Only correlations), the dynamic WMH range (i.e. the range of the WMH volumes obtained from T1w-Only and All-Contrasts segmentations) is also substantially reduced (e.g. 0-42.5 CCs vs. 0-124 CCs for ADC, 0-74 CCs vs. 0-116 CCs for ADNI1, and 0-67 CCs vs. 0-104 CCs for ADC, for T1w-Only and All-Contrasts segmentations, respectively).

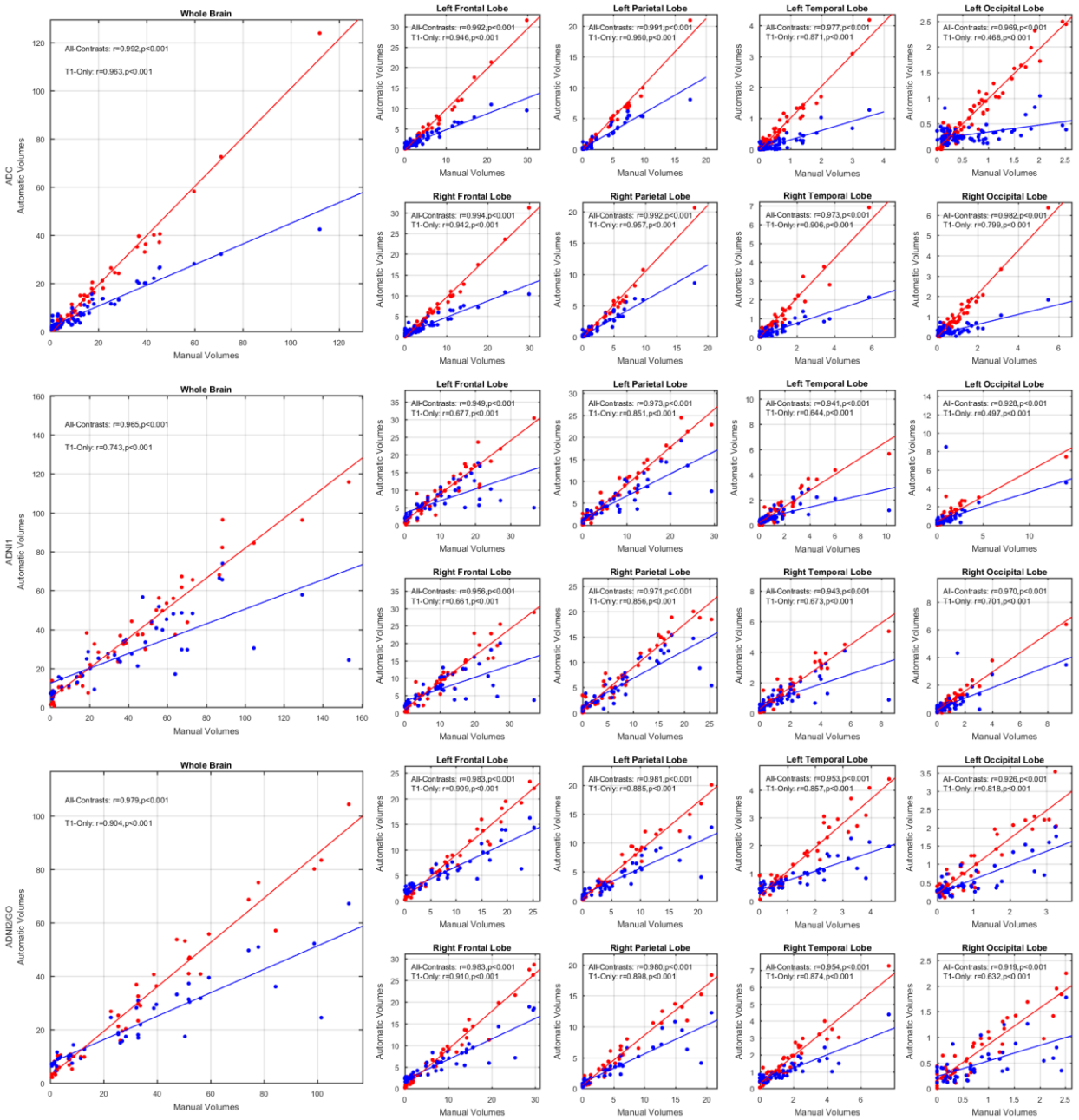


Fig. 5.3. Total brain and per lobe correlations for automatic versus manually segmented WMH volumes (CCs), using all contrasts available (red) and using only T1w contrast (blue) for a. ADC, b. ADNI1, c. ADNI2/GO datasets.

5.3.3. Large Scale Correlations

A correlation analysis of the WMH loads for the whole brain as well as different lobes

was performed between the automatic segmentations obtained with and without T2w/PD and FLAIR information (All-Contrasts and T1w-Only, respectively) for the subjects from ADNI1 and ADNI2/GO datasets. Figure 5.4 shows the results of these comparisons for each dataset. The correlations were significant for every lobe in both datasets (r=0.888, p<0.001 for ADNI1, and r=0.969, p<0.001 for ADNI2/GO for whole brain correlations).

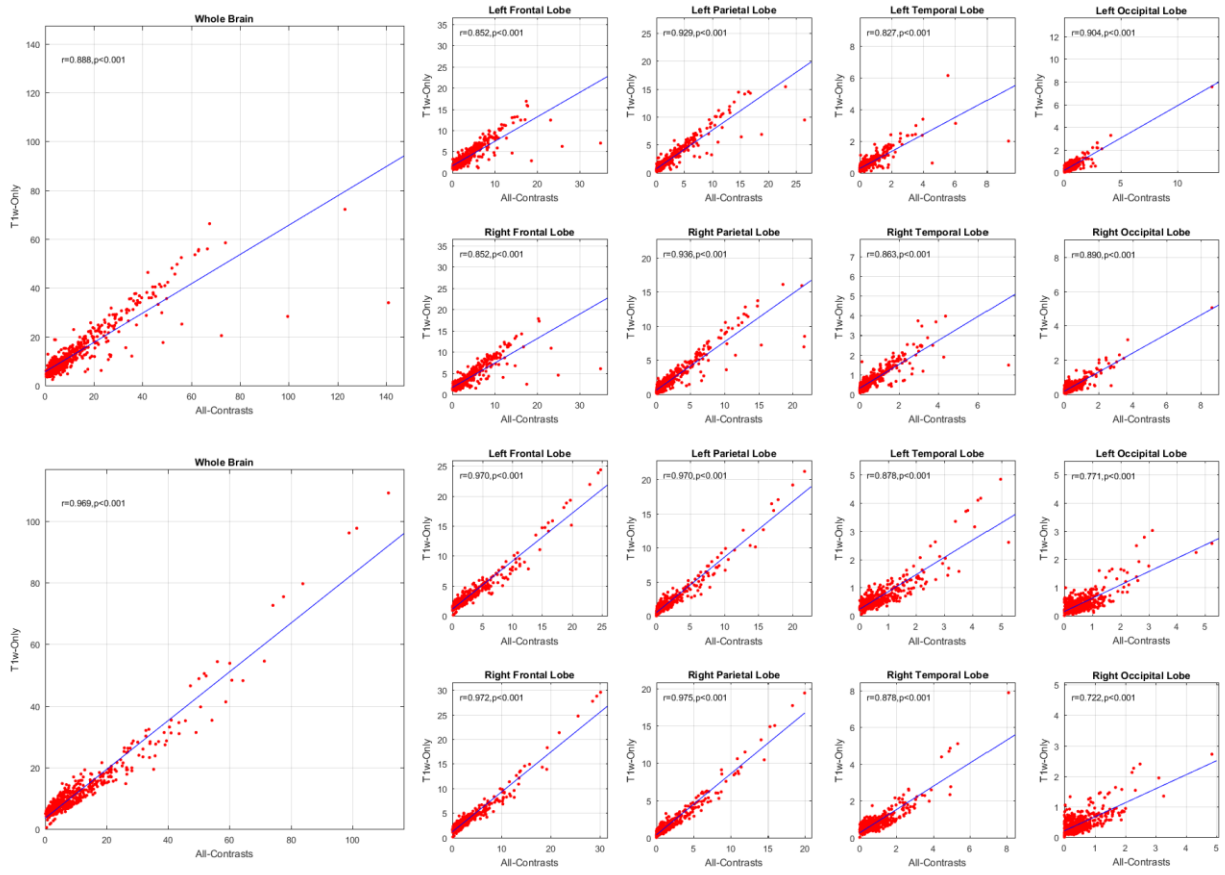


Fig. 5.4. Total brain and per lobe correlations for automatically segmented WMHs volumes (CCs) based on all available contrast images versus only T1w contrast image for ADNI1 (left) and ADNI2/GO (right) datasets.

The results show a trend of under-segmentation that remains consistent with the change in the WMH load. The amount of this underestimation is also highly correlated with the total WMH load for both datasets (r= 0.790, p<0.00001 for ADNI1 and r= 0.717, p<0.00001 for ADNI2/GO).

The log transformed WMH loads were also correlated with age. Figure 5.5 shows the results of the correlations with age for ADNI1 and ADNI2/GO datasets.

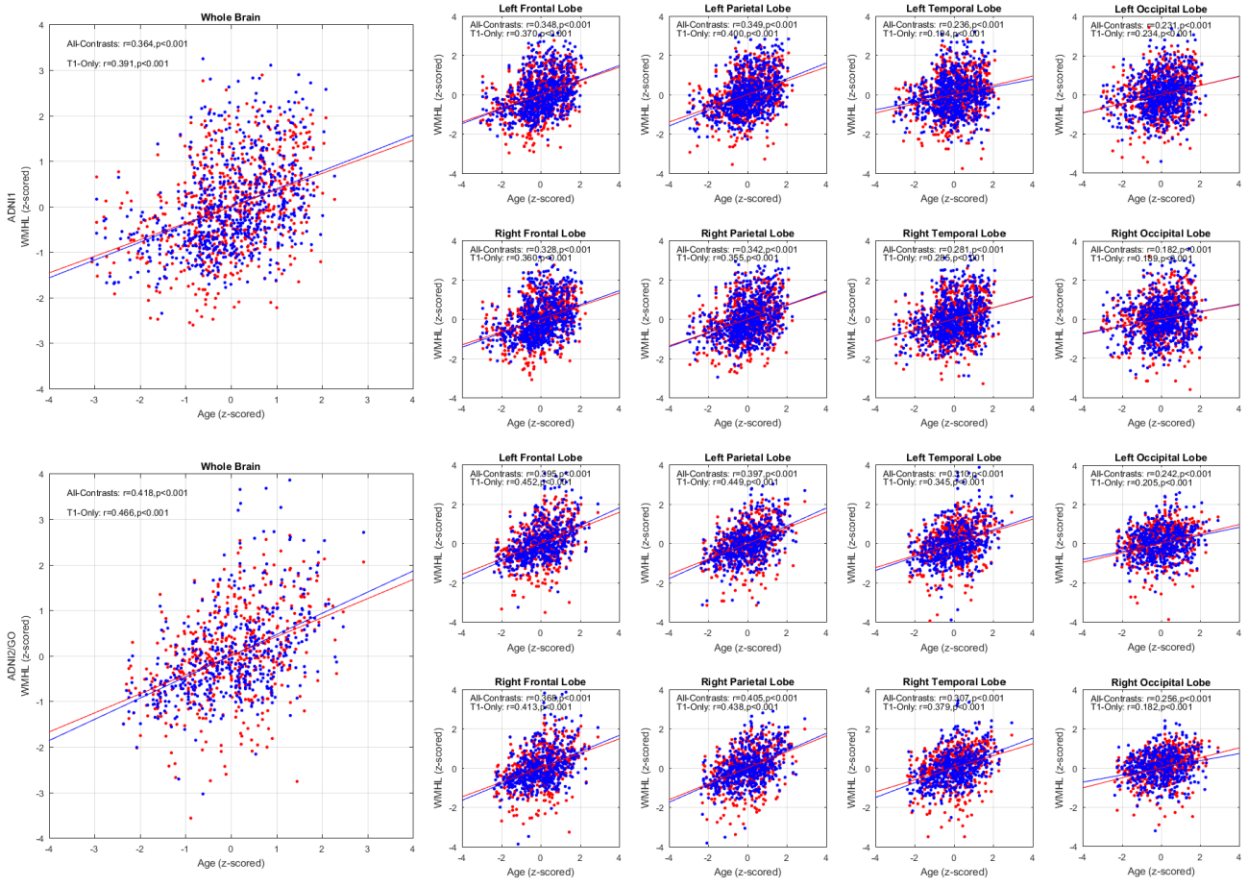


Fig. 5.5. Total brain and per lobe correlations for z-scored log transformed automatically segmented WMH volumes correlated with age (z-scored) based on all available contrasts (red) versus only T1w images (blue) for ADNI1 (left) and ADNI2/GO (right) datasets. WMHL= White Matter Hyperintensity Load.

If one considers the T1w intensity profile of the tissue that is segmented as WMH based on FLAIR or T2w images, the intensities range from hypo-intense to iso-intense in relation to the neighboring tissue. To investigate whether the T1w-Only segmentations have different sensitivity levels for detecting WMHs (i.e. whether the T1w-Only segmentations label the hypo-intense portion of the WMHs and not the iso-intense areas), the manual WMH masks created by the expert were thresholded based on T1w intensity of NAWM at different values, reflecting

different levels of sensitivity. The resulting WMH masks were then compared with the T1-Only and All-Contrasts WMH segmentations for different threshold values using Dice Kappa similarity measure and volumetric correlation (Fig. 6). The results show higher Dice Kappa values as well as stronger correlations with T1-Only segmentations at lower threshold values, confirming that the T1w-Only segmentations correspond to the more hypointense regions of the manual WMH masks, with the optimal intensity threshold at approximately 63 (note that the intensity range of all the images has been normalized to the range of 0-100).

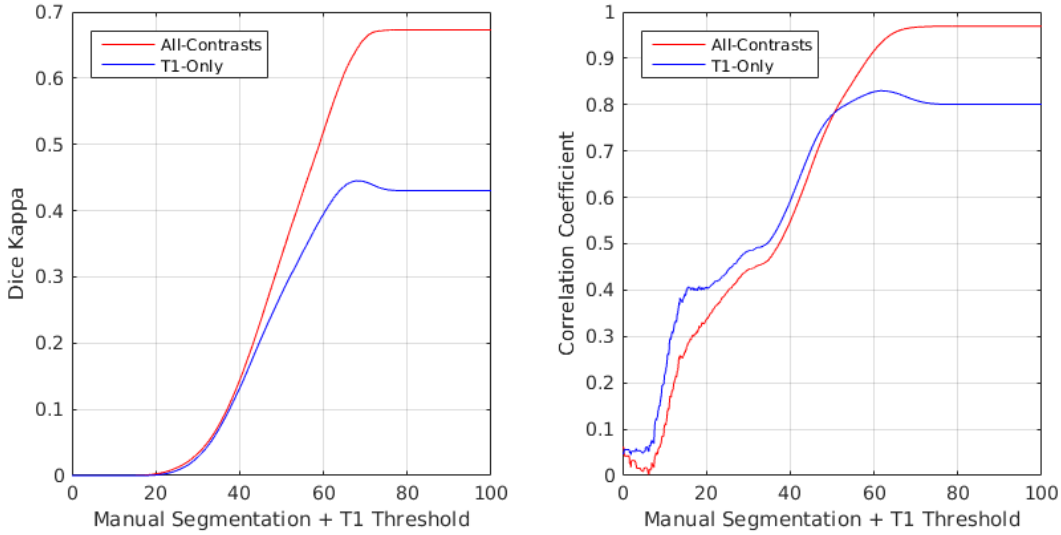


Fig 5.6. Dice Kappa and volumetric correlation between T1-Only and All-Contrasts segmentations and thresholded manual labels based on T1w image intensity.

5.3.4. WMHs and clinical measures

Tables 5.4 and 5.5 summarize the results of correlating different cognitive measures and the log transformed WMH loads (to achieve normal distribution) in different lobes obtained from All-Contrasts and T1w-Only segmentations, controlling for age for ADNI1 and ADNI2/Go datasets, respectively. The "*" indicates significant correlations, after multiple comparisons correction using false discovery rate (FDR), controlling for age for ADNI1 and ADNI2/GO

datasets. Figure 5.7 shows an example of these correlations for ADAS13 score.

Table 5.4. Coefficients of correlation between WMH loads in different lobes and cognitive measures for ADNI1 subjects, controlling for age. WMH= White Matter Hyperintensity. LFL/RFL=Left/Right Frontal Lobe. LPL/RPL= Left/Right Parietal Lobe. LTL/RTL=Left/Right Temporal Lobe. LOL/ROL= Left/Right Occipital Lobe. WB= Whole Brain. MMSE= Mini-Mental State Examination. ADAS= Alzheimer's Disease Assessment Scale. FAQ= Functional Assessment Questionnaire. RAVLT= Rey Auditory Verbal Learning Task (I=Immediate, F=Forgetting, L=Learning). Executive= Executive Function (Gibbons et al., 2012). NPI= Total Neuropsychiatric Inventory Score.

	Test	LFL	RFL	LPL	RPL	LTL	RTL	LOL	ROL	WB
T1+T2+PD	MMSE	-0.165*	-0.171*	-0.139*	-0.145*	-0.081*	-0.090*	-0.125*	-0.134*	-0.174*
	ADAS11	0.203*	0.199*	0.183*	0.192*	0.108*	0.131*	0.129*	0.172*	0.217*
	ADAS13	0.219*	0.217*	0.201*	0.199*	0.127*	0.151*	0.141*	0.190*	0.236*
	FAQ	0.209*	0.215*	0.171*	0.192*	0.109*	0.128*	0.139*	0.208*	0.216*
	RAVLTI	-0.174*	-0.179*	-0.161*	-0.163*	-0.119*	-0.168*	-0.130*	-0.157*	-0.197*
	RAVLTF	0.074	0.051	0.062	-0.028	-0.005	-0.005	-0.023	-0.020	-0.061
	RAVLTL	-0.084*	-0.100*	-0.079	-0.079	-0.070	-0.085*	-0.084*	-0.095*	-0.103*
	Executive	-0.111*	-0.110*	-0.121*	-0.107*	-0.112*	-0.113*	-0.182*	-0.139*	-0.130*
	NPI	0.210*	0.199*	0.279*	0.273*	0.270*	0.239*	0.242*	0.221*	0.277*
Only T1	MMSE	-0.203*	-0.211*	-0.127*	-0.147*	-0.031	-0.057	-0.144*	-0.162*	-0.187*
	ADAS11	0.231*	0.230*	0.149*	0.180*	0.054	0.083*	0.132*	0.146*	0.217*
	ADAS13	0.251*	0.251*	0.168*	0.186*	0.069	0.092*	0.148*	0.156*	0.236*
	FAQ	0.226*	0.238*	0.150*	0.182*	0.065	0.089*	0.160*	0.224*	0.218*
	RAVLTI	-0.220*	-0.217*	-0.166*	-0.179*	-0.100*	-0.130*	-0.133*	-0.141*	-0.217*
	RAVLTF	0.074	0.067	0.054	0.043	0.053	0.033	0.036	0.044	0.073
	RAVLTL	-0.109*	-0.113*	-0.059	-0.070	-0.020	-0.037	-0.070	-0.084*	-0.097*
	Executive	-0.128*	-0.135*	-0.117*	-0.113*	-0.018	-0.077	-0.124*	-0.092*	-0.130*
	NPI	0.228*	0.237*	0.308*	0.300*	0.211*	0.200*	0.200*	0.228*	0.290*

Table 5.5. Coefficients of correlation between WMH loads in different lobes and cognitive measures for ADNI2/GO subjects, controlling for age. WMH= White Matter Hyperintensity. LFL/RFL=Left/Right Frontal Lobe. LPL/RPL= Left/Right Parietal Lobe. LTL/RTL=Left/Right Temporal Lobe. LOL/ROL= Left/Right Occipital Lobe. WB= Whole Brain. MMSE= Mini-Mental State Examination. ADAS= Alzheimer's Disease Assessment Scale. FAQ= Functional Assessment Questionnaire. RAVLT= Rey Auditory Verbal Learning Task (I=Immediate, F=Forgetting, L=Learning). Executive= Executive Function. NPI= Total Neuropsychiatric Inventory Score.

	Test	LFL	RFL	LPL	RPL	LTL	RTL	LOL	ROL	WB
T1+FLAIR	MMSE	-0.153*	-0.124*	-0.145*	-0.137*	-0.095	-0.086	-0.147*	-0.083	-0.152*
	ADAS11	0.210*	0.191*	0.199*	0.189*	0.166*	0.130*	0.231*	0.135*	0.219*
	ADAS13	0.197*	0.184*	0.184*	0.180*	0.161*	0.119*	0.227*	0.138*	0.208*
	FAQ	0.160*	0.148*	0.155*	0.148*	0.132*	0.104*	0.124*	0.099*	0.167*
	RAVLTI	-0.108*	-0.115*	-0.096	-0.104*	-0.064	-0.024	-0.180*	-0.070	-0.121*
	RAVLTF	-0.069	-0.048	-0.077	-0.046	-0.036	-0.087	-0.022	-0.061	-0.072
	RAVLTL	-0.124*	-0.143*	-0.126*	-0.138*	-0.096	-0.102*	-0.088	-0.011	-0.137*
	Executive	-0.057	-0.078	-0.056	-0.074	-0.036	-0.013	-0.104	-0.100	-0.080
	NPI	0.179*	0.204*	0.175*	0.213*	0.200*	0.247*	0.175*	0.195*	0.215*
Only T1	MMSE	-0.126*	-0.094	-0.109*	-0.106*	-0.055	-0.067	-0.097	-0.041	-0.118*
	ADAS11	0.209*	0.183*	0.186*	0.158*	0.136*	0.102*	0.140*	0.097	0.199*
	ADAS13	0.192*	0.165*	0.162*	0.145*	0.120*	0.084	0.126*	0.083	0.179*
	FAQ	0.151*	0.118*	0.113*	0.096	0.071	0.037	0.055	0.051	0.127*
	RAVLTI	-0.115*	-0.105*	-0.101*	-0.112*	-0.057	-0.033	-0.072	-0.012	-0.111*
	RAVLTF	-0.057	-0.056	-0.062	-0.024	-0.037	-0.042	-0.050	-0.073	-0.061
	RAVLTL	-0.158*	-0.156*	-0.094	-0.105*	-0.134*	-0.108*	-0.067	-0.021	-0.142*
	Executive	-0.075	-0.077	-0.069	-0.095	-0.016	-0.055	-0.046	-0.031	-0.083
	NPI	0.125*	0.130*	0.113*	0.152*	0.145*	0.153*	0.034	0.108*	0.144*

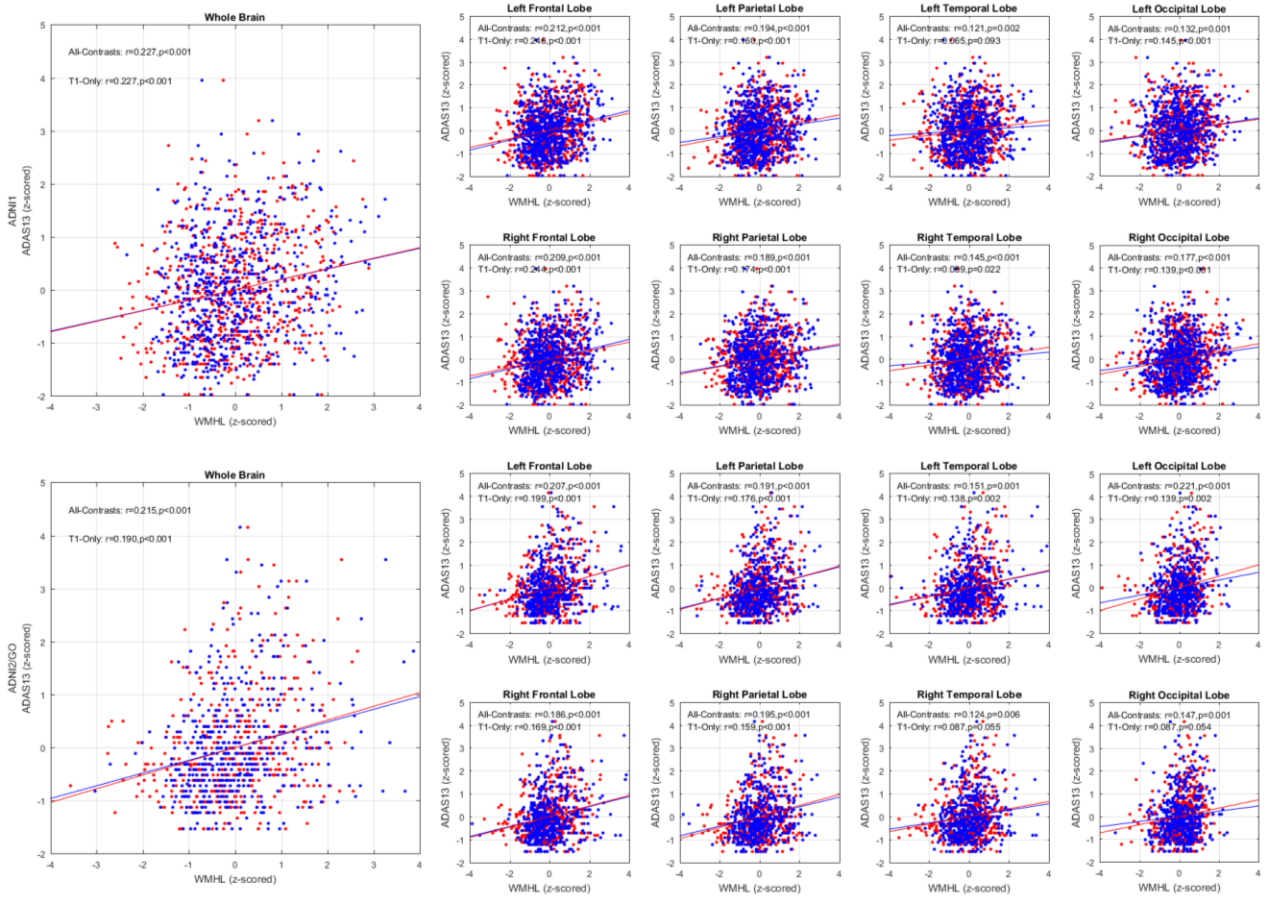


Fig. 5.7. Total brain and per lobe correlations for log transformed automatically segmented WMH volumes (CCs) versus ADAS13 while controlling for age, based on all available contrasts (red) and based only on T1w images (blue) for ADNI1 (left) and ADNI2/GO (right) datasets.

The key concern when using T1w-only segmentations is whether the WMH portions that are not captured by T1-only segmentations are clinically significant. To assess whether the difference between the WMH volumes provided by T1-Only and All-Contrast segmentations is statistically significant, a general linear model was used to regress the contrast ($WMHL_{All\text{-}Contrast}$-$WMHL_{T1\text{-}Only}$) and each measure. None of the contrasts was significant after correcting for multiple comparisons using FDR.

5.3.5. WMHs and other measures

The log transformed total WMH loads were significantly different between subjects with and without cardiovascular risk factors and history of hypertension, for both T1w-Only and All-Contrasts segmentations ($p<0.0001$). Figure 8 shows boxplots of the log transformed WMH loads in Normal Control (NC), Mild Cognitive Impairment (MCI), and Dementia groups, separately for T1w-Only and All-Contrasts loads in ADNI1 and ADNI2/GO datasets. In all cases, the Dementia cohort has significantly larger WMH loads than the other two groups (Note that the log transformed values are plotted here, and that a 0.4 difference in the log transformed values is equivalent to approximately 3 CCs of WMHs).

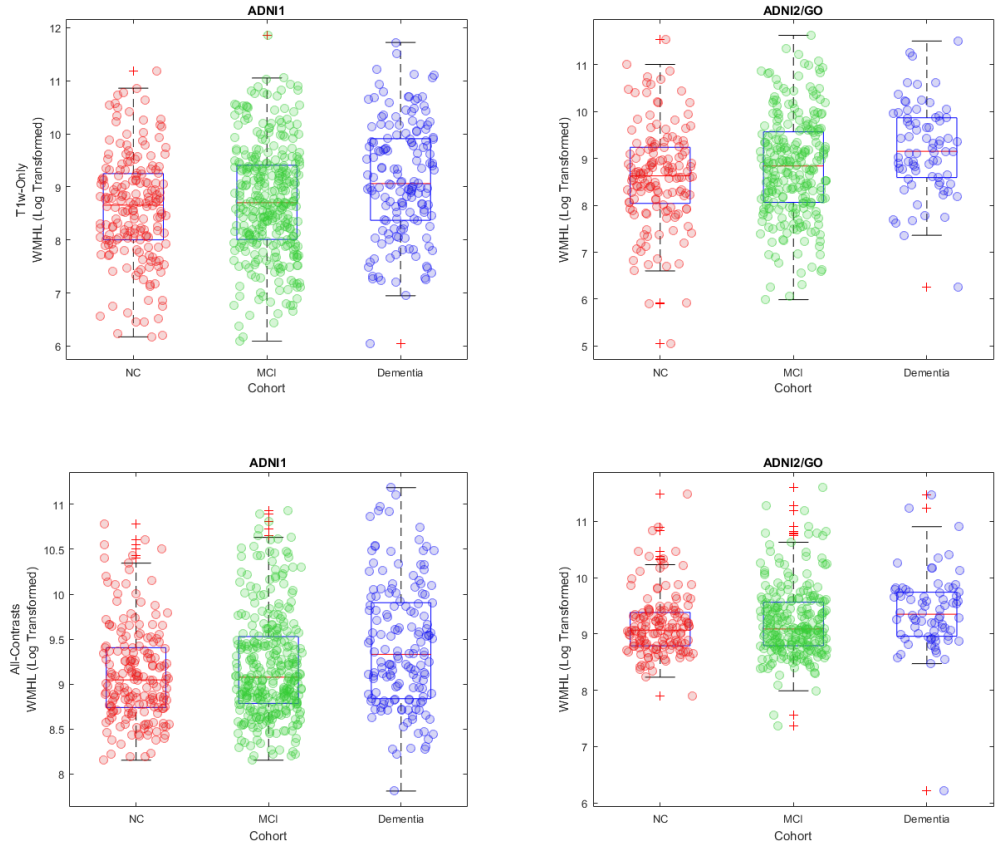


Fig. 5.8. Boxplots of log transformed WMHLs for NC, MCI, and Dementia Cohorts. In all cases, the Dementia cohort has significantly larger WMH loads than the other two groups. WMHL= White Matter Hyperintensity Load.

NC= Normal Control. MCI= Mild Cognitive Impairment.

Tables 5.6 summarizes the results of correlating different risk factors with the log transformed total WMH loads from T1w-Only images and All-Contrasts segmentations, for ADNI1 and ADNI2/Go datasets, respectively. The "*" indicates significant correlations, after multiple comparisons correction using false discovery rate (FDR). None of the contrasts ($WMHL_{All\text{-}Contrast}$-$WMHL_{T1\text{-}Only}$) was significant after correcting for multiple comparisons.

Table 5.6. Coefficients of correlation between WMH loads in different lobes and measures for ADNI1 and ADNI2/GO subjects. WMH= White Matter Hyperintensity. FDG= Fluorodeoxyglucose PET. AV45= Florbetapir. PET= Positron Emission Tomography. CSF= Cerebrospinal Fluid.

Dataset	ADNI1			ADNI2/GO		
Measure	N	T1+T2+PD	T1	N	T1+FLAIR	T1
Systolic blood pressure	148	0.002	-0.014	409	0.162*	0.199*
Diastolic blood pressure	148	-0.052	-0.027	409	0.103	0.089
Hyperhomocysteinaemia	667	0.124*	0.146*	47	0.016	0.048
FDG	342	0.183*	0.135*	448	0.284*	0.261*
AV45	0	-	-	443	0.172*	0.162
Serum Glucose	626	0.119*	0.112*	417	0.048	0.043
Cholesterol	557	0.065	0.080	382	0.153*	0.105
CSF Protein	333	0.209*	0.125*	73	0.232*	0.247*

5.4. Discussion

White matter hyperintensities are an important clinical marker of small vessel disease in aging, and patients suffering from stroke and dementia (Carmichael et al., 2010; DeCarli et al., 1995a; Pantoni et al., 2006; van Straaten et al., 2008). In recent years, there has been an increasing interest in using WMHs as an outcome in clinical trials investigating cerebral small vessel disease in the context of stroke and dementia (Debette and Markus, 2010). They reflect the burden of the disease in relation to a small-vessel component (Pantoni, 2010), and are associated with decline in different cognitive domains. Specifically in AD, WMHs are emerging

as a potential biomarker of the preclinical vulnerability risk for the disease (Brickman et al., 2012; Deoni et al., 2013; Provenzano et al., 2013).

The optimal MRI sequences for detecting WMHs are FLAIR and T2w/PD scans. However, many previous large-scale datasets have forgone the acquisition of these sequences in favor of other imaging modalities or due to time and financial concerns. WMHs are also visible as hypointense regions on T1w images, but their range of hypointensity is more spread out when compared to the bright signal obtained in FLAIR and T2w/PD, with large isointense areas, or almost isointense in relation with the surrounding NAWM (Fig. 9). This lower contrast between the NAWM tissue and WMHs on T1w sequence makes accurate detection of the full extent of WMHs more challenging in this MRI modality. However, by assessing, even to some extent, the load and location of WMHs using only T1w data, it is possible to study them in other retrospective datasets, where T2w/PD, or FLAIR information are not available. The key concern when using T1w-Only segmentations would be whether the WMH portions that are not captured by T1-Only segmentations are clinically significant.

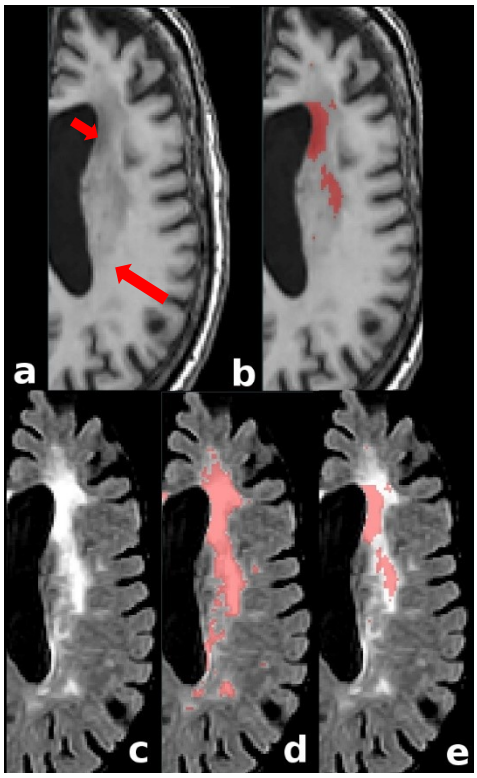


Fig. 5.9. Comparing T1w hypointensities and FLAIR hyperintensities. a) T1w image showing different hypointense values: short and long arrows indicate lower and higher signals, respectively. b) T1w manual mask corresponding to a given level of hypointensity in T1w image. Note that only the lower signal areas are labeled as WMHs. c) FLAIR image d) FLAIR and manual mask corresponding to WMHs detected based on FLAIR image d) FLAIR and T1w-based mask. Note how the hypointense information of the T1w sequence colocalizes only with a portion of the WMHs on FLAIR even though the signal on the FLAIR sequence is mostly homogeneous.

In our study, a previously validated automated technique was used to segment the WMHs with and without the optimal FLAIR and T2w/PD information. In a previous study, we have shown that this automated tool was able to detect WMHs using different combinations of MRI sequences (Dadar et al., 2017b). Here, a random forest classifier was chosen to report our results based on our previous experiments and validations since it had the best performance in detecting WMHs among the studied linear and nonlinear classifiers. However, similar experiments were also performed using the other available classifiers such as AdaBoost which showed similar results regardless of the choice of classifier.

Our results suggest that the measures obtained from using only T1w images underestimate the actual WMH burden, since they only capture the portion of the lesions that co-localizes with the lower intensity signal of the overall T1w hypointense area (Fig. 5.9). We speculate that these deeper hypointense portions are probably due to more severely affected tissue (i.e. more extensive demyelination, and/or more axonal loss) which are likely to be more clinically relevant (hence the maintained correlations). Future studies correlating histological specimens and MR T1w information would be necessary to clarify the specific histological substrate of the full range of hypointense signal captured by T1w sequence and the more homogeneous bright signal on FLAIR/T2w/PD. Nevertheless, despite the absence of a histopathology gold standard, T1w-Only WMHs volumes are still able to hold strong correlations with both manual and automatic segmentations obtained using the optimal modalities. Additionally, the assessment of colocalization shows higher Dice Kappa values for the T1w-Only classifications when they are compared with the manual mask thresholded to a percentage value of the T1w NAWM, confirming that the partially detected WMH area colocalizes with the lower T1w signal voxels (Fig. 5.6). Likewise, the volumetric correlation of the T1w-Only volumes with the threholded manual expert volumes reinforces the concept of specific partial detection of hypointense voxels on T1w images.

Regarding the assessment of clinical outcomes, the T1w-Only WMH volumes mostly correlate with age, cognitive and clinical measures as strongly as the WMH volumes determined using the optimal modalities of FLAIR or T2/PD. This is true for WMHs in whole brain and in different single brain lobes. Likewise, vascular risk factors show a significant correlation with T1-Only WMHs volumes, similar to those obtained with all the optimal modalities. This suggest that, although the WMH burden might be underestimated in T1w-Only segmentations, the

identified lesions can be used for clinical correlations in datasets where the optimal sequences are not available. This will enable us to generate and use the WMH loads obtained from the T1w images as an estimate of the actual WMH loads in datasets where the FLAIR or T2w/PD information is unavailable, and also in T1w scans that do not have full FLAIR or T2w/PD brain coverage.

Finally, many studies acquire T2w/PD or FLAIR information in their baseline visit, or at longer intervals compared with the T1w scans, which are acquired at every MRI visit. Even though the T1w-Only segmentations systematically underestimate the volume of WMHs, having an estimate of the accurate WMH load from the baseline T2w/PD or FLAIR images, one may correct for this bias to obtain more accurate estimates of WMH loads, for the visits that only have T1w acquisitions. In addition, T1w images are generally acquired at higher resolutions (i.e. 1 mm thick slices) than the T2w/PD or FLAIR scans which are generally acquired at 3-5 mm slice thicknesses. The higher spatial resolution of T1w images can also be used to obtain more spatially accurate segmentations, where such information may not available due to the higher slice thickness of the T2w/PD and FLAIR images.

Our study indicates that datasets that lack T2w/PD and FLAIR modalities may still benefit from the estimation of WMHs using our T1w-Only segmentation approach, in order to correlate this MRI data with other clinical variables available for the subjects.

Chapter 6. White Matter Hyperintensities Predict Cognitive Decline in de Novo Parkinson's Disease Patients

Preface

In this chapter, we use the proposed WMH segmentation pipeline to study the effect of WMHs in Parkinson's disease. We use data from the Parkinson's Progression Markers Initiative (PPMI), a multi-center and multi-scanner database of unmedicated early stage Parkinson's disease patients and age matched healthy controls. The WMHs were segmented using either T1w+T2w or T1w+FLAIR sequences from baseline visits of the subjects. Using mixed effects modeling and survival analysis, the relationship between baseline WMHs and future cognitive decline was assessed.

Our results showed that Parkinson's patients that had higher loads of WMHs, have higher rates of decline in different cognitive domains than can be explained by normative aging.

This work has been revised and re-submitted to Jama Neurology as:

Dadar, M., Y. Zeighami, Y. Yau, S.M. Fereshtehnejad, J. Maranzano, R. Postuma, A. Dagher, D. L. Collins. " White Matter Hyperintensities and Cognitive Decline in de Novo Parkinson's Disease Patients."

White Matter Hyperintensities and Cognitive Decline in de Novo Parkinson's Disease Patients

Mahsa Dadar, Yashar Zeighami, Yvonne Yau, Seyed-Mohammad Fereshtehnejad, Josefina Maranzano, Ronald Postuma, Alain Dagher, D. Louis Collins

Key points:

Question: Does existence of white matter hyperintensities (WMHs) in *de Novo* Parkinson's disease (PD) patients affect their future cognition differently from non-affected controls subjects?

Findings: In a cohort of early stage drug naïve PD patients and age matched healthy controls, we measured baseline WMH loads. Using longitudinal cognitive scores (follow-up: 4.09±1.14 years), we found that PD subjects with high WMH loads had significantly higher decline in cognition, compared with i) PD subjects with low WMH loads and ii) HC subjects with high WMH loads.

Meaning: WMHs in PD patients are associated with more pronounced cognitive decline than healthy controls.

Abstract:

Importance: White Matter Hyperintensities (WMHs) are associated with cognitive decline in normative aging and Alzheimer's disease. However, the pathogenesis of cognitive decline in Parkinson's disease (PD) is not as clearly related to vascular causes, and therefore the role of WMHs as a marker of small-vessel disease (SVD) in PD remains unclear. Currently, SVD in PD is assessed and treated independently of PD. However, if WMH as the major MRI sign of SVD has a higher impact on cognitive decline in PD patients than in healthy controls, vascular pathology needs to be assessed and treated with a higher priority in this population.

Objective: To assess at the earliest stages of PD, whether WMHs are associated with faster cognitive decline, and if these effects relate to cortical thickness alterations.

Design, setting, and participants: Cohort study of the role of WMHs in PD in recently diagnosed and non-treated *(de novo)* PD patients. *De novo* PD patients (N_{PD} =365) and age-matched controls (N_{HC} =174) with FLAIR or T2w MRI scans at baseline were selected from the multi-center Parkinson's Progression Markers Initiative (PPMI) database. Baseline WMHs and longitudinal cortical thickness were measured to analyse the relationship between baseline WMHs and future cognitive decline (follow-up: 4.09±1.14 years) and cortical thinning (follow-up: 1.05±0.10 years).

Main Outcomes and Measures: WMH loads, cortical thickness, and cognitive scores.

Result: Within *de novo* PD patients, high WMH burden at baseline was associated with increased cognitive decline, significantly more than i) PD patients with low WMH loads (χ^2=31.5, p<0.00001) and ii) controls with high WMH loads (χ^2=6.2, p=0.012). Furthermore, PD patients with higher baseline WMH loads showed more cortical thinning in right frontal lobe

than subjects with low WMH loads. Cortical thinning of this region also predicted poorer performance on a cognitive test as well as the change in performance between baseline and follow-up.

Conclusion and relevance: Presence of WMHs in *de novo* PD patients predicts greater future cognitive decline and cortical thinning than can be accounted for by normal aging. Recognizing WMHs as a potential predictor of cognitive deficit in PD provides an opportunity for timely interventions.

Key words: Parkinson's disease, white matter hyperintensities, magnetic resonance imaging, cognitive decline, *de Novo* Parkinson's disease patients

6.1. Introduction

While Parkinson's disease (PD) is typically characterized by motor symptoms, cognitive deficits occur in approximately 15% of patients in early drug-naïve stages(Poletti et al., 2012). Two decades after disease onset, this prevalence increases to over 80% (Hely et al., 2008). Early mild cognitive impairment (MCI) is a strong predictor of later development of dementia (Anang et al., 2014; Pedersen et al., 2017), which is a key determinant of mortality and poorer quality of life in PD (de Lau et al., 2005). Cognitive impairment in PD is related to subcortical dysfunction in early stages, followed by cortical α-synuclein pathology and loss of neurotransmitters. However, it remains unclear to what degree white matter changes, historically described as leukoaraiosis (Hachinski et al., 1987) which are major signs of small-vessel disease (SVD) (Halliday et al., 2014; Merino and Hachinski, 2000) may contribute to cognitive dysfunction in PD.

White matter hyperintensities (WMHs) or leukoaraiosis are areas of increased signal in

T2w and FLAIR structural MRI. The neuropathologic correlates of WMHs are varied: loss of axons and glial cells, myelin rarefaction, spongiosis, perivascular demyelination, gliosis, subependymal glial accumulation and loss of the ependymal lining (Merino and Hachinski, 2000). Despite the various findings, consensus exists regarding the association of WMHs and SVD (Pantoni and Garcia, 1997). The term small-vessel disease is mainly related to two etiologies: 1) age-related vascular disease, also referred as arteriolosclerosis, or vascular-risk-factor related small-vessel disease (Debette and Markus, 2010; de Leeuw et al., 2002), and 2) cerebral amyloid angiopathy (Pantoni, 2010). These two play a crucial role in stroke, dementia and aging, and could also be relevant in PD. Therefore, early detection of WMHs and treatment of cardiovascular risk factors could have a positive impact on cognitive decline in PD (Biesbroek et al., 2017; Dufouil et al., 2005; Hawkins et al., 2017; Veselỳ and Rektor, 2016). In AD, WMHs have been extensively studied and strongly predict rapid cognitive decline in individuals with MCI (Dubois et al., 2014; Tosto et al., 2014). In PD, the pathogenic role of vascular risk factors is less clear (de Lau et al., 2005) and results have been contradictory (Veselỳ and Rektor, 2016). The WMHs might cause cognitive decline independent of PD, or the synergy between the two mechanisms may accelerate cognitive impairment (Veselỳ and Rektor, 2016). Alternatively, the WMHs might aggravate the pathologic spread of misfolded α-synuclein or amyloid-β proteins. Of the few studies that have investigated WMHs and cognitive decline in PD, most are cross-sectional, include patients that are on dopaminergic medication, and are typically from cohorts that are at later stages of disease (Auning et al., 2014; Jones et al., 2017; Mak et al., 2015). Additionally, different groups implement different tests to assess cognition and many do not perform a comprehensive neuropsychological battery.

Capitalizing on the longitudinal assessment of cognitive abilities and imaging biomarkers

in the multi-centre cohort of *de novo* PD patients from the Parkinson's Progression Markers Initiative (Marek et al., 2011), we investigated the relationship between WMH burden and: 1) cognitive decline over time and 2) cortical grey matter changes over time (as indexed by cortical thinning) in early stages of PD.

6.2. Methods

6.2.1. Patients

The Parkinson's Progression Markers Initiative (PPMI) is a longitudinal multi-site clinical study of *de novo* PD patients and age-matched healthy controls (HC) (Marek et al., 2011) (http://www.ppmi-info.org). The study was approved by the institutional review board of all participating sites and written informed consent was obtained from all participants before inclusion in the study. In the present study, we included all subjects that had either FLAIR or T2w MR images at their baseline visit and had follow-up visits for at least one year after the baseline scan (N_{PD}=365, N_{HC}=174). All subjects were regularly assessed (yearly follow-ups, mean total follow-up period of 4.09±1.14 years) for clinical characteristics (motor, non-motor and neuropsychological performance) by site investigators, including Montreal Cognitive Assessment (MoCA), Hopkins Verbal Learning Test–Revised (HVLT), Benton judgement of line orientation test for visuospatial skills, Letter-Number Sequencing test for verbal working memory, and semantic fluency test to detect cognitive decline (Table 6.1). The executive function score is calculated as the sum of letter number sequencing and semantic fluency scores (Chan et al., 2008). To validate the correlation between these two components, we verified their relationship in the PD population ($r=0.56$, $p<0.0001$). For more information on clinical measurements, see Appendix, section Cognitive Testing.

Table 6.1. Descriptive statistics for the PPMI subjects enrolled in this study. Data are number of participants in each category (N), percentage of the total population (%), and mean (SD) of key variables. PPMI=Parkinson's Progression Marker Initiative. FLAIR= Fluid Attenuated Inversion Recovery. MoCA= Montreal Cognitive Assessment Score. HVLT= Hopkins Verbal Learning Test Revised Total Score. Benton= Benton Judgement of Line Orientation Score. WMH= White Matter Hyperintensity.

	Control	*De novo* PD
Participants (N_{Total})	174	365
Female (N)	57 (33%)	114 (32%)
T1w and FLAIR Scans ($N_{Baseline}$)	79 (45%)	167 (46%)
T1w and T2w Scans ($N_{Baseline}$)	95 (55%)	198 (54%)
Follow-up 3T T1w scans ($N_{Follow-up}$)	55 (32%)	100 (27%)
Age at Baseline (years)	60.07 (±11.34)	60.51 (±9.86)
MoCA at Baseline	28.25 (±1.12)	27.24 (±2.22)
HVLT at Baseline	35.05 (±6.78)	32.01 (±7.95)
Benton at Baseline	26.13 (±4.12)	25.60 (±4.07)
Executive Function at Baseline	20.94 (±4.73)	22.29 (±4.58)
WMH Load at Baseline (cm^3)	7.66 (±10.38)	6.93 (±8.03)

6.2.2. Procedures

All MR images were preprocessed using our standard pipeline (Aubert-Broche et al., 2013) in three steps: noise reduction, intensity non-uniformity correction, and intensity normalization. T2w and FLAIR images were linearly co-registered to the T1w images using a 6-parameter rigid registration. The T1w images were first linearly and then nonlinearly registered to the standard template (MNI-ICBM-152). The WMHs were segmented using a previously validated automatic multi-modality segmentation technique in the native space of FLAIR or T2w scans to avoid further blurring caused by resampling of the images (Dadar et al., 2017a, 2017b). This technique uses a set of location and intensity features obtained from a library of manually segmented scans in combination with a random forest classifier to detect the WMHs in new images. The libraries used in this study were obtained from *Alzheimer's Disease Neuroimaging* Initiative (ADNI) dataset since the T2w and FLAIR sequences for the PPMI images follow the same acquisition protocol as ADNI. The quality of the registrations and segmentations was

visually assessed and cases that did not pass this quality control were discarded (n=43). WMH load was defined as the volume (in cm^3) of all segmented WMH voxels in the standard space, i.e. the WMH volumes were corrected for total intracranial volume (ICV). All MRI processing and segmentation steps were blinded to clinical outcomes.

For voxel-wise analysis of WMHs, the WMH probability maps generated by the segmentation tool were nonlinearly transformed to the template space at 2×2×2 mm^3 resolution and blurred with a 3D Gaussian kernel with full width at half maximum of 5 mm to compensate for the variability caused by differences in voxel sizes in the native FLAIR and T2w images. Rates of cognitive decline were calculated for subjects that had at least one-year follow-up information as the change of the score per year (N_{PD}=365, N_{HC}=174), using a linear regression between time and the score values at different time points along with an intercept term.

Only subjects with T1w 3T MRI data at both initial/baseline visit and at a one-year follow-up MRI were included for cortical thickness analysis (N_{Total}=155, see Table 6.1). Cortical models were generated using the CIVET 2.1 preprocessing pipeline (Ad-Dab'bagh et al., 2006), registered to MNI-ICBM-152 template, and analyzed using the SurfStat software package (http://www.math.mcgill.ca/keith/surfstat/). Distances between inner and outer cortical surfaces were evaluated to provide a measure of cortical thickness at each vertex. Changes in cortical thickness were calculated by subtracting the values ($\Delta t = t_1 - t_2$) at the one-year follow-up (t_2) from the baseline (t_1). The average time between the baseline and follow-up visits was 1.05±0.11 and 1.05±0.09 years for the PD and control subjects, respectively.

6.2.3. Statistical Analysis

We tested two major hypotheses: (1) greater WMH load will lead to more extensive and

faster decline in cognition of the PD patients (2) patients with a higher WMH load (WMHL) will show more cortical thinning in their follow-up visit after one year.

Survival analysis was used to investigate the relationship between WMH burden and decline in cognition. It has been previously shown that a threshold of WMHs should be present before cognitive deficits are observed (Boone et al., 1992; Price et al., 2012). The question of interest was whether there is a significant difference between the *cognitive* survival curves of subjects (normal controls and PD patients) with low versus high WMHL. The threshold for differentiating between high and low WMHL was set at 5 cm^3 (median value, 0.7% of WM volume, 0.27% of brain volume). Similar to previous studies (Baracchini et al., 2012; Suzuki et al., 2015; Joana et al., 2013), a stable 2-point drop in MoCA (a drop that persists over the follow-up visits) was considered as the terminal event in the survival analysis and the time from baseline MoCA measurement to the visit where the 2-points drop was detected was considered as survival time. This was consistent with recommendations from our in-house clinical consultation. Drop in MoCA was selected as the main terminal event since MoCA has been previously validated as a sensitive measure for detecting and monitoring cognitive change over time (Costa et al., 2014) in general and MCI or dementia in PD specifically (Hoops et al., 2009). Robustness of the results was verified for a WMHL threshold of 10 cm^3 and 1 to 4 point drops in MoCA. For survival analysis, the survdiff function from R package *survival* was used (ftp://centos.ustc.edu.cn/CRAN/web/packages/survival/survival.pdf). The function implements the two-sample G^{ρ} statistics family of Harrington and Fleming, with weights on each event (2-point drop in MoCA) of $S(t)^{\rho}$, where $S(t)$ is the Kaplan-Meier estimate of survival, i.e. the probability that a subject survives longer than time t (Harrington and Fleming, 1982).

Furthermore, Longitudinal mixed-effects models were used to assess the association of

WMHs with changes in cognition. MoCA, Benton, HVLT, and executive function scores were used as measures of cognition (dependent variables). The log-transformed WMH loads and age were used as continuous predictors, and cohort (HC versus PD) was used as a categorical predictor. All continuous variables were z-scored prior to the analysis. All models contained first order interactions with age. Subject and contrast used for segmentation (T2w versus FLAIR) were considered as categorical random effects in all the models. Models were fitted using fitlme in MATLAB version R2015b.

Differences in cortical thickness between high and low WMHL classes [(high$WMHL_{t1}$-high$WMHL_{t2}$)-(low$WMHL_{t1}$-low$WMHL_{t2}$)] were analyzed statistically based on Gaussian random field theory with a threshold of $p<0.05$ (Worsley et al., 1996). Similar to the survival analysis, the threshold for differentiating between high and low WMHL was 5 cm^3. Observed differences in cortical thickness were then correlated to cognitive measures using Pearson partial correlations correcting for age.

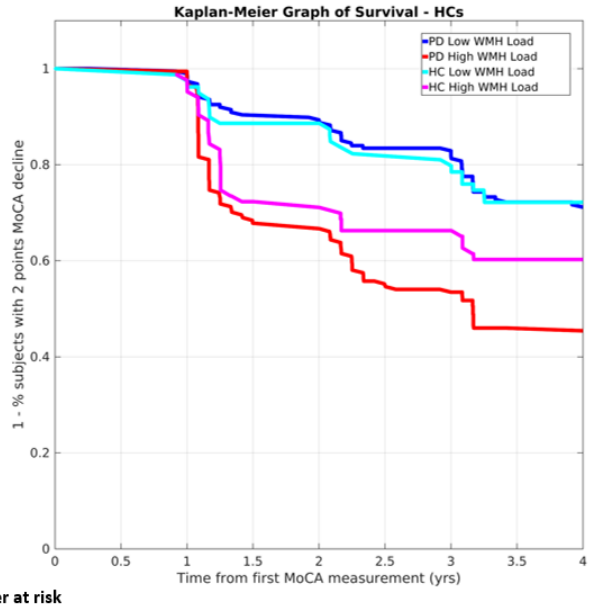
6.3. Results

6.3.1. Baseline WMH Load as a Predictor of Longitudinal Cognition

Survival Analysis:

Baseline WMH loads were not significantly different in control and PD populations ($p>0.05$) (see Appendix for further information on baseline measures). Controlling for age, the rate of decline in MoCA score was significantly correlated with baseline WMH load ($r=-0.145$, $p=0.007$) in the PD cohort, but not in controls ($r=0.045$, $p=0.577$). Figure 1 shows the Kaplan-Meier plot for the survival analysis for progression decline in MoCA. The 4-year survival rate

(i.e. rate of patients maintaining MoCA stability) for the low and high WMHL groups were estimated as 63% (95 CI=0.55-0.70) and 37% (95 CI=0.29-0.45) in PDs and 65% (95 CI=0.52-0.75) and 56% (95 CI=0.45-0.67) in controls, respectively ($N_{PD\text{-}Low}$=186, $N_{PD\text{-}High}$=174, $N_{HC\text{-}Low}$=79, $N_{HC\text{-}High}$=83). In PD, the high WMHL cohort experienced a significantly lower survival rate than the low WMHL cohort (χ^2=30.9, p<0.00001, hazard ratio= 2.42). There was no high vs low difference in controls (χ^2=2.5, p=0.11, hazard ratio= 1.52). Furthermore, PD patients showed significantly lower survival rate compared to controls in the high WMHL group (χ^2=6.7, p=0.009, hazard ratio=1.58) while the survival rate was not significantly different between two groups in low WMHL group (χ^2=0.1, p = 0.8, hazard ratio=1.0). Similar results were obtained with a threshold of 10 cm^3 and 1-4 point drops in MoCA, suggesting that WMH load-based dichotomization is sensitive to a range in the cognitive decline as measured by MoCA.



Number at risk

PD Low WMH Load	186	176	153	134
PD High WMH Load	174	163	103	72
HC Low WMH Load	79	70	60	52
HC High WMH Load	83	82	59	53

Fig. 6.1. Kaplan-Meier graph of survival showing the survival curves of control and PD patients with low versus high WMH loads demonstrating the compounded affect of PD and WMH load. A 2-point drop in MoCA was considered as the survival event and the time from baseline MoCA measurement to the visit where the 2-point drop occurred was considered as survival time. HC=Healthy Control. PD=Parkinson's Disease. MoCA= Montreal Cognitive Assessment Score.

Mixed-Effects Modelling:

The mixed-effects modelling results based on age, baseline WMH, and their interaction (Table 6.2, Fig. 6.2) showed a significant negative relationship between MoCA, Benton, HVLT, and Executive function scores and age in both PD and HC cohorts. More importantly, in the PD cohort, there was a significant interaction between Age and baseline WMH load for MoCA, Benton, and HVLT which was not observed in the HC cohort.

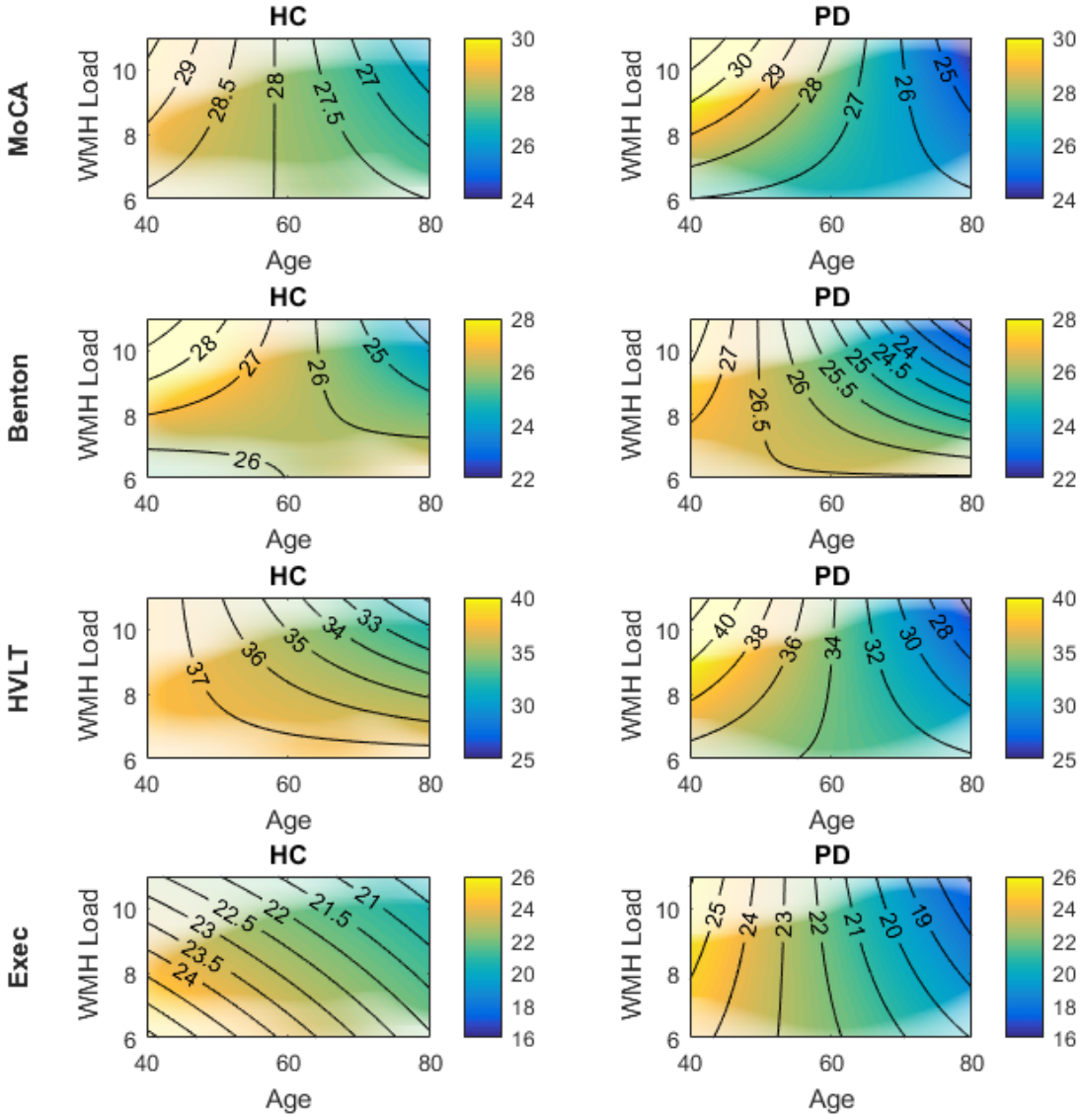


Fig. 6.2. Density plots of longitudinal cognitive changes versus age and log transformed baseline WMH load. The colors indicate predicted cognitive scores by the mixed effects models, with warmer colors representing higher scores, and cooler colors representing lower scores. The transparency in the figures indicates the density of the data, i.e. areas of low transparency indicate regions where there are no subjects and the model is extrapolating (e.g. young subjects with high WMH loads, or old subjects with low WMH loads). The contour lines imply the direction of changes (i.e. horizontal orientation indicates predominance of age effects and vertical orientation indicates predominance of WHM load effects). WMH=White Matter Hyperintensities. HC= Healthy Control. PD= Parkinson's Disease. MoCA= Montreal Cognitive Assessment Score. HVLTRT= Hopkins Verbal Learning Test Revised Total Score. Benton= Benton Judgement of Line Orientation Score. Exec= Executive Function Score.

Table 6.2. Summary of the mixed effects models of association between baseline WMH Load and change in cognition. Entries show the regression coefficients for the listed fixed effect followed by the associated p values. Baseline WMH load was log transformed and z-scored along with age, MoCA, HVLTRT, and Benton scores prior to analysis. WMHL=White Matter Hyperintensity Load. HC= Healthy Control. ":" indicates the interaction between two variables. Global Cognition= Montreal Cognitive Assessment Score (MoCA). Memory= Hopkins Verbal Learning Test Revised Total Score (HVLT). Visuospatial= Benton Judgement of Line Orientation Score. Executive= Executive Function Score (Letter Number Sequencing + Semantic Fluency). HC= Healthy Control. PD= Parkinson's Disease.

	Cognitive Score	Global Cognition		Memory		Visuospatial		Executive	
	Variable	ß	p-value	ß	p-value	ß	p-value	ß	p-value
PD	Intercept	-0.063	0.180	-0.098	0.029	0.013	0.737	-0.086	0.059
	Age	**-0.413**	**<0.001**	**-0.341**	**<0.001**	**-0.164**	**<0.001**	**-0.374**	**<0.001**
	WMHL	0.035	0.428	-0.029	0.485	**-0.093**	**0.021**	-0.049	0.236
	Age:WMHL	**-0.122**	**<0.001**	**-0.091**	**0.006**	**-0.062**	**0.059**	-0.048	0.139
HC	Intercept	0.251	<0.001	0.263	<0.001	0.116	0.067	0.186	0.005
	Age	**-0.215**	**<0.001**	**-0.113**	**0.030**	**-0.131**	**0.019**	**-0.167**	**0.002**
	WMHL	-0.031	0.495	-0.093	0.083	-0.017	0.777	-0.088	0.113
	Age:WMHL	-0.047	0.180	-0.043	0.330	-0.087	0.072	0.011	0.816

Cortical Thickness:

Mean whole-brain cortical thickness decreased significantly among PD patients with both low (t_1 = 3.3177mm ± 0.0993; t_2 = 3.3087mm ± 0.1082) and high (t_1 = 3.2932mm ± 0.0996; t_2 = 3.2786mm ± 0.0966) WMH at baseline. Among PD patients, baseline WMH load did not correlate with whole-brain cortical thickness at baseline (r=-0.09, p>0.05) or at one-year follow-up (r=-0.19, p>0.05), but did correlate with cortical thickness change across the one-year period (r=0.26, p=0.01). When comparing high and low WMH groups in PD, cortical thinning was greater in the high WMH group with a significant cluster observed in the right frontal lobe ($N_{Vertices}$=1523, resels=7.99, p<0.001) which covers the lateral precentral, superior frontal, and middle frontal gyri (Fig. 3). Cortical thinning of this cluster at baseline was not significantly correlated with poorer performance on the HVLT at baseline (r=-0.169, p>0.05), but was at one-

year follow-up (r=-0.335, p<0.001) and with declining performance over the one-year period (*r*=0.196, p<0.05). No significant correlation or vertex/cluster-wise difference was observed in the HC cohort. No significant correlation was observed between MoCA, Benton, and executive function and cortical thickness in PD cohort.

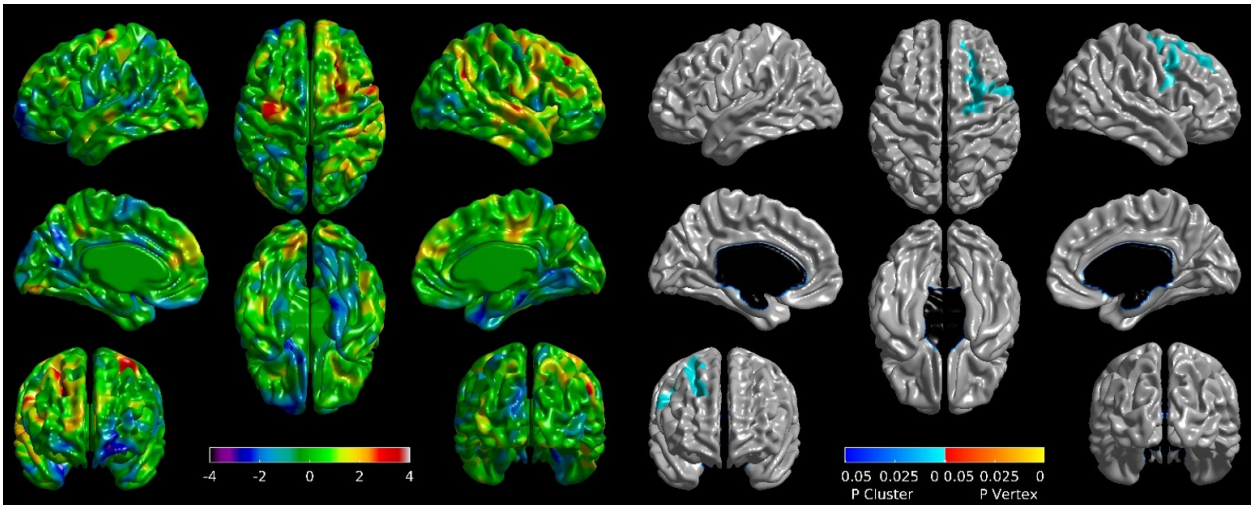


Fig. 6.3. Differences in cortical thickness changes between high and low WMHL cohorts in PD subjects. T-maps (left) and areas of significant cortical thickness decreases (right) covering the precentral, superior frontal, and middle frontal gyri. WMHL= White Matter Hyperintensity Load. PD= Parkinson's Disease.

Voxel-wise Analysis:

Within the PD cohort, significant voxel-wise correlations were observed between WMH localization maps and the slope of MoCA and Benton scores, corrected for multiple comparisons using false discovery rate (FDR) adjustment and controlled for age and modality (Fig. 4). The significant regions include voxels in all lobes: frontal, temporal, parietal, occipital, and also insular subcortical WM bilaterally. No significant associations were found for the HC cohort. No significant associations were found for HVLT and Executive Function scores in the PD cohort. No significant differences were observed between the baseline voxel-wise WMH maps of PD and HC cohorts after FDR correction.

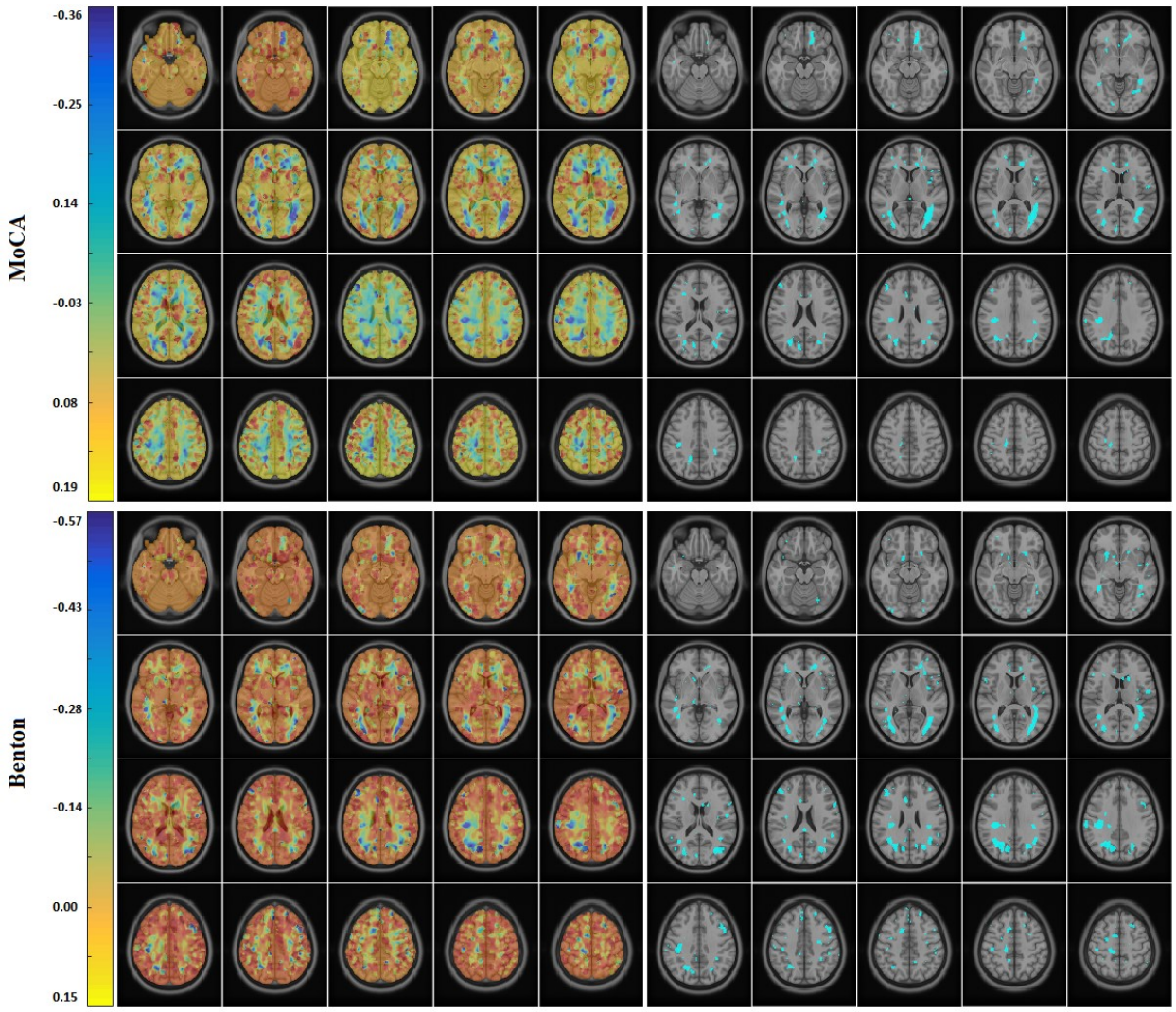


Fig. 6.4. Correlation between WMH location and slope of MoCA (top) and Benton (bottom) score in the PD cohort, controlled for age and modality. Correlation coefficients (left) and thresholded areas of significant correlations after FDR correction. WMH=White Matter Hyperintensity. MoCA= Montreal Cognitive Assessment. PD=Parkinson's Disease. FDR= False Discovery Rate.

6.4. Discussion

High WMHL PD patients experienced significantly higher decline than i) low WMHL PD patients and ii) high WMHL control subjects. Additionally, WMHL was significantly associated with whole-brain cortical thinning after only one-year follow-up in PD patients, but not in controls. Moreover, PD patients with a high WMHL at baseline showed significant

cortical thinning of a frontal cluster compared to those with low WMHL. Taken together, these findings suggest that measures of WMHL in *de novo* PD patients can predict later cognitive decline, even in patients exhibiting no cognitive symptoms at baseline.

As with previous studies (Dalaker et al., 2009), cross-sectional WMHL at baseline in early PD was not significantly associated with baseline cognitive performance. Rather, WMHL at baseline was associated with future cognitive deterioration across multiple cognitive domains including visuospatial, memory, and global cognition corrected for age. This suggests that we can extend previous work on later stages of PD, where WMH burden was significantly associated with conversion to dementia in patients with MCI (Kandiah et al., 2014; Sunwoo et al., 2014), to the earliest stages of the disease. In line with these findings, post-mortem studies have shown that vascular lesions are common in idiopathic PD (Lewy body disease of the brainstem type) (Jellinger, 2003).

MoCA has been validated as a sensitive measure for detecting and monitoring cognitive change over time (Costa et al., 2014). Controlling for age, MoCA decline was significantly correlated with baseline WMHL in the PD cohort, but not in controls. Additionally, PD subjects with high WMHLs were more likely to experience a 2-point drop in MoCA than (i) the low WMHL PD and (ii) the high WMHL HC subjects, as evaluated by the survival analysis. The driver for cognitive decline in controls and PD appear to differ in that the former is largely driven by age, while the latter is affected by both advancing age and greater baseline WMH load.

While the literature on PD and WMH is scarce, there has been substantial progress in understanding the relationship between WMHs and cognitive impairment/dementia in AD, especially in the context of amyloid pathology. WMHs associated with vascular risk factors (e.g.,

hypoperfusion and inflammation) are thought to precede Aβ aggregation. Previous work found significant associations between baseline WMHs and later progression of amyloid load (Grimmer et al., 2012). This further supports the hypothesis of a chain of events; namely WMH impairs clearance of amyloid, which builds up and contributes to cognitive impairment and AD symptoms. While amyloid deposition strongly predicts progression to AD, WMH burden can provide additional independent information to this prediction (Provenzano et al., 2013), suggesting that WMH is not solely related to amyloid pathology, but can directly impact cognitive impairment. Whether a similar interaction between vascular lesions and α-synuclein formation or deposition occurs in PD remains unclear.

WMH burden can also precede irreversible neurological damage as indexed by cortical atrophy. Previous studies have found higher WMHL to be correlated with lower cortical thickness in frontotemporal regions which in turn are related to cognitive decline (Tuladhar et al., 2015). Cortical thinning caused by direct or indirect effects of WMHs (tract-specific damage) might lead to cognitive decline and eventually dementia. Cortical thickness might be a sensitive measurement to detect regional grey matter micro-changes that are missed by conventional voxel-based techniques at the earlier stages of the neurodegeneration due to partial volume effect (Hutton et al., 2009; Seo et al., 2012). While we observed whole-brain cortical thinning among all PD patients, those with high WMH load showed greater cortical thinning of a frontal cluster, mostly encompassing the right dorsolateral prefrontal cortex (rDLPFC) which was further associated with decline in memory performance in HVLT over the one-year period. This is consistent with previous studies that have found significant associations between rDLPFC and HVLT scores (Qiao et al., 2016; Ries et al., 2012). Our results suggest that cortical changes in early PD are potentially moderated by WMH load, and might in turn presage cognitive decline.

Regardless of etiology, prevention and treatment of vascular risk factors associated with WMHs is a promising avenue to slow down cognitive deterioration, especially in *de novo* PD patients who are largely cognitively asymptomatic. The classical and most explored strategy regarding reduction of vascular disease risk and WMHs has been to control hypertension, which subsequently reduces the risk of cognitive deterioration (Debette and Markus, 2010; Dufouil et al., 2001; de Leeuw et al., 2002). In a randomized trial, active lowering of blood pressure was shown to stop or lower the progression of WMHs in patients with cerebrovascular disease over 3 years of follow-up (Dufouil et al., 2005). In the present cohort, we observed an association between WMH load and (systolic-diastolic) blood pressure for both PDs and controls ($p<0.001$). However, there is also evidence linking WMHs and dementia in PD to orthostatic hypotension, a common occurrence in PD which can be aggravated with anti-hypertensive medication, especially as the disease progresses (Oh et al., 2013). This further indicates the need for a tailored blood pressure management in PD patients, while extreme care should be taken to avoid overtreating hypertension. Finally, other small-vessel disease risk factors (some of which have been explored in the context of other pathologies, mainly AD, showing significant correlations with WMHs (Biesbroek et al., 2017; Veselỳ and Rektor, 2016)) should be further explored to assess their relevance in WMHs severity and cognitive decline in PD. More importantly, most of these factors are potentially modifiable: percentage of small dense LDL cholesterol, triglycerides level, body mass index, tobacco consumption, type II diabetes, and insulin levels. More studies should focus on assessment of these risk factors in the context of PD and WMHL.

From a practical standpoint, WMHs can be quantified reliably and non-invasively on large samples and can be measured as a continuous trait, thus providing increased statistical power to detect potential associations (Debette and Markus, 2010). The image processing and

WMH segmentation pipelines used in this study have been designed to process data from multi-center studies, are able to control biases due to multi-site MRI scanning (i.e. differences in acquisition parameters), and have been previously applied successfully to a number of multi-site projects (Boucetta et al., 2016; Zeighami et al., 2015). The WMH segmentation pipeline has been trained and extensively validated on data from multiple scanners and different acquisition parameters to ensure inter-site and inter-scanner generalizability (Dadar et al., 2017b).

We acknowledge there are limitations to the present study. First, though their differences were accounted for in our analysis, segmentations were based on either T2w or FLAIR images, of which the latter has the better contrast for detecting WMHs. Second, subjects had these scans only at their baseline visit; therefore, we were not able to study the longitudinal changes of WMHs. Future studies investigating WMHs in PD during prodromal and pre-clinical stages are warranted, though there are inherent constraints in recruiting such a cohort. Also, the population under study included relatively cognitively intact individuals (none of the subjects met criteria for dementia), limiting the ability to detect important contributors. Longer follow-ups might further increase the observed differences. One potential confounding factor could be PD medication. However, previous studies have found no significant difference between PD patients on PD medications and PD patients off medications in MoCA and several other cognitive tasks (Cools et al., 2006). Similarly, we found no relationship between MoCA and medication in PD patients (see Appendix, Medication Information). Another limitation is that we cannot identify the underlying mechanism. The WMHs might cause cognitive decline independent of PD, however the synergy between the two mechanisms may accelerate the cognitive decline. Alternatively, the WMHs might aggravate the pathologic spread of misfolded α-synuclein proteins in PD. Another possibility is that WMHs in PD may promote amyloid propagation,

similar to AD.

In conclusion, our findings suggest that WMH burden is an important predictor of subsequent acceleration in cortical thinning and cognitive decline in early-stage *de novo* PD. Recognizing WMHs as early indicators of cognitive deficit, prior to onset of MCI or dementia, provides an opportunity for timely interventions (Marek et al., 2011; Zeighami et al., 2015).

Chapter 7. Conclusions and Future Work

7.1. Discussions

This chapter provides a summary of the rationale as well as the main findings of each of the four manuscripts included in this thesis and the prospective future work that can be pursued based on these contributions. The overall goal of this thesis is to create and validate a fully automated tool that can be used to detect and study WMHs from various MRI sequences in different large multi-site and multi-center studies.

The main challenge in this task lies in the fact that the appearance of WMHs can vary greatly as the scanner model or MRI sequence parameters change, which is often the case in multi-center and longitudinal studies. In order to be able to effectively study the effect of WMHs in such studies, one needs an automated tool that can segment these lesions robustly using MRI scans obtained from different scanners and sequences (Caligiuri et al., 2015). Tools that have been developed and validated on data from a single scanner and sequence generally perform poorly in detecting WMHs using data from other scanners or different protocols. To assess the generalizability of segmentation tools, it is necessary to validate the performance of the techniques on an independent dataset, ideally collected from multiple MRI scanners and acquisition sequences that are not used for training the model.

7.1.1. Feature Selection and Linear Classifier

As an initial attempt to address this problem, we proposed a linear regression technique to segment WMHs from a combination of multiple contrasts of MR images as described in Chapter 3 (Dadar et al., 2017a). The proposed technique uses a combination of intensity and

location features from different input sequences along with a spatial prior and a linear regression model to provide continuous subject-specific WMH maps that can reflect different levels of damage to the tissue. These continuous maps can later be thresholded to obtain binary WMH labels. The threshold value can determine the desired level of sensitivity in detecting the WMHs.

Using data from three different datasets (ADC, PREVENT-AD, and ADNI2/GO), the performance of the classifier was extensively validated and the contribution of each of the proposed intensity and location features in improving its performance was assessed. We further showed that the linear regression classifier outperforms several publicly available WMH segmentation tools as well as more complicated nonlinear classifiers when it comes to detecting WMHs in data from other scanners, not used in their training, confirming the generalizability of its results. The results showed that our proposed linear regression classifier can provide a fast and computationally efficient tool to detect WMHs robustly and accurately.

7.1.2. Nonlinear Classifiers and Automated Pipeline

The proposed technique in Chapter 3 was later expanded in Chapter 4 to a WMH segmentation pipeline that can use any combination of input sequences (e.g. T1w, T1w+FLAIR, T1w+T2w+PD, T1w+T2w+PD+FLAIR, etc.) and any of 10 linear and nonlinear classifier options from Scikit-learn library implementations in Python (naïve Bayes, linear and quadratic discriminant analysis, logistic regression, support vector machines, k nearest neighbors, decision trees, random forests, bagging, and AdaBoost) to detect WMHs (Dadar et al., 2017b). To achieve generalizability while taking advantage of the higher complexity of nonlinear classification techniques, we obtained manual segmentations on data selected from different scanner and MR

sequences, taking care to choose subjects with low, medium, and high WMH loads from each scanner.

We extensively validated the performance of these ten different linear and nonlinear classifiers using different combinations of input sequences in 4 different multi-center and multi-scanner datasets (ADC, NACC, ADNI1, and ADNI2/GO). We found that the random forests classifier had the best performance among all the classifiers evaluated in all the experiments with different input sequences and datasets. The results of our experiments with different classifiers and datasets also showed that a combination of T1w (3D isotropic) and axial FLAIR sequences (similar acquisition protocols to ADNI sequences rather than T2 space isotropic FLAIRs) would lead to the best automatic segmentations. We further made the WMH segmentation pipeline publicly available along with the pre-trained classifiers, for different combinations of input sequences. This enables verification of our work by other groups and facilitates comparisons with new techniques developed in other laboratories.

We have since successfully applied the Random Forest technique (using the pre-trained classifiers) to detect WMHs in a number of different populations acquired with different acquisition sequences and scanners, including:

(i) The entire NACC database which includes approximately 2000 subjects with different pathologies (e.g. AD, PD, FTD, vascular dementia, depression, autism, etc.) that have been scanned at 30 different centers using T1w and FLAIR images.

(ii) A multi-scanner dataset of subjects with FTD and age matched controls using T1w and FLAIR images.

(iii) A single-scanner dataset of HIV patients and age matched controls using T1w and T2w images.

(iv) Individuals with the fat mass and obesity-associated protein (FTO) gene using T1w and FLAIR images.

The successful application of the WMH segmentation pipeline to these datasets indicates that the proposed method is able to deal with the variations in the imaging protocols and scanners. Additionally, our experiments showed that the acquisition protocols used for FLAIR sequences from the ADNI dataset are the best in differentiating WMHs and normal tissue. They have since been used by several other studies such as the PPMI.

7.1.3. Detecting WMHs from T1w Images

Many large multi-center studies forgo FLAIR and T2w/PD sequence acquisition in favor of other modalities due to time and/or financial constraints. Being able to obtain reliable estimates of WMH loads in such datasets would be highly advantageous, as it would make it possible to study small-vessel disease and its interactions with other diseases without being limited to datasets with T2w/PD or FLAIR sequence acquisitions. In Chapter 5, we assessed whether the developed WMH segmentation pipeline can be used to detect and study WMHs in datasets that only have T1w images available (Dadar et al. 2018). To achieve this, we used baseline data from 1150 subjects from ADNI study and extensively validated the relationship between the WMH loads obtained from only T1w images and (i) manual labels, (ii) the WMHs loads obtained from automatic segmentations using the optimal FLAIR and T2w/PD sequences and (iii) several cognitive and clinical measures and WMH related risk factors.

Our results showed that while the T1w segmentations generally underestimate the true WMH loads, they still hold strong correlations with clinical and cognitive measures and can be

used as estimates of the WMHs to study the vascular burden in datasets where the optimal sequences are not available. This enables us to take advantage of many multi-center databases that were previously not analyzable due to the lack of T2w/PD or FLAIR acquisitions. In addition, segmenting these lesions from T1w-Only data provides more opportunities to study WMHs and their relationship and interactions with other neurodegenerative diseases.

7.1.4. WMHs in Parkinson's Disease

Finally, we used the developed pipeline to segment and study WMHs in a multi-center database of early stage, drug naïve *de Novo* Parkinson's patients (the PPMI database) as well as age-matched controls (Dadar et al. 2017c). Using longitudinal cognitive measures, survival analysis, and mixed-effects modeling, we were able to show that the *de Novo* Parkinson's patients with higher WMH loads at baseline were more likely to cognitively decline in the follow-up visits than (i) the Parkinson's patients with lower WMH loads and (ii) the control subjects with higher WMH loads. Further, using longitudinal cortical thickness measures, we observed that the Parkinson's patients with higher WMH loads experienced significantly more cortical thinning in the following year in a cluster in the frontal lobe. The cortical thinning in this cluster was also found to be associated with the 1-year decline in a cognitive measure.

Our findings in the PPMI population suggest that the early on co-occurrence of WMHs and Parkinson's disease can increase the rate of future cognitive decline, more than that expected from normative aging, in as early as the *de Novo* stage. In other words, WMHs have a higher impact on the cognitive decline in Parkinson's patients than non-diseased individuals with the same amount of WMH burden, suggesting that vascular pathology needs to be assessed and treated with a higher priority in the PD population.

7.2. Future work

The potential future works in the context of this thesis is discussed below.

7.2.1. WMH Segmentation

In the WMH segmentation front, there are several areas worth pursuing to further improve the accuracy and robustness of the classification techniques. To be specific,

(i) Longitudinal segmentation of WMHs: In studies that assess the longitudinal progression of WMHs, reproducibility is essential to ensure that the differences that are observed between segmentations obtained from different timepoints result from changes in the pathology and not from the variabilities in the automated segmentation tool (García-Lorenzo et al., 2013). All the proposed techniques in this thesis so far are cross-sectional; i.e. multiple timepoints from the same subject are treated as independent. However, combining the information from multiple timepoints of the same subject can increase accuracy by enabling detection of smaller, less hyperintense lesions at earlier visits, as well as by reducing false positive detections.

(ii) Adding a post-processing step: certain brain regions are known to be more susceptible to imaging artifacts and false positives, such as the hyperintensities observed in the insular regions or in the choroid plexus. In addition, in many studies, detections smaller than a certain size are not considered WMHs and consequently are not of interest. Using a second classifier, such false positives can be removed in a post-processing step (Caligiuri et al., 2015).

(iii) Using deep neural networks for segmenting WMHs: more recently, convolutional neural networks (CNNs) have shown promising results specifically in image

classification tasks. So far, the main hindrance in using such structures was the small number of available manually segmented labels which was not sufficient for training such networks. Through our collaborations on different projects, we have now acquired manual segmentations for more than 320 subjects. This will enable us to train complex CNN architectures to improve our WMH segmentations.

(iv) Detecting dirty white matter: as was mentioned before, WMHs do not necessarily have sharp borders. In many cases, the fuzzy borders gradually dissolve into the normal appearing white matter. The most hyperintense regions are detected as WMHs in either manual or automatic segmentations. However, these less hyperintense regions (generally referred to as dirty white matter) can also be detected as continuous maps reflecting the level of damage to the tissue and studied separately, especially considering the fact that they might turn into WMHs later on (Beggs et al., 2016).

(v) Detecting lacunar infarcts and microhemorrhages: WMHs are the major signs of cerebral small-vessel disease, along with lacunar infarcts and microhemorrhages, both of which have also been shown to significantly affect the cognition in the elderly population (Conklin et al., 2014; Gouw et al., 2010; Sam et al., 2016). Segmenting these two addition aspects of small-vessel disease can provide a more thorough representation of the vascular disease burden in these individuals.

(vi) Detecting WMHs in presence of other pathology: WMHs can be accompanied by other pathologies such as infarcts, stroke and tumors. However, the datasets that were used for training and validation of the proposed techniques generally exclude individuals with these pathologies. It would be important to assess the impact of these

pathological changes on WMH segmentations in datasets that have subjects with mixed pathologies. Our preliminary investigations have shown that since stroke lesions have a different intensity and location profile than WMHs (i.e. they are generally more hyperintense and in the surface of the brain), they are not classified as WMHs. However, a more thorough investigation to assess the accuracy of the segmentations with different combinations of input sequences in populations with mixed pathologies is necessary.

7.2.2. Studying WMHs in Aging and Diseases

The purpose of developing accurate WMH segmentation tools is to enable us to study them in different populations, to see how they correlate with disease and how they affect cognition. Using the accurate segmentations obtained from the developed WMH segmentation pipeline, we can attempt to study their causes and effects more thoroughly.

(i) So far, we have only used the volume of WMHs in different lobes as WMH related features. While WMH loads are informative, other features might be able to provide further information on the vascular burden (Lindemer et al., 2015). Using WMH segmentations, various WMH related features such as first order statistics, shape and size based features, intensity and textural features and wavelet features can be extracted through a radiomics approach (Lambin et al., 2012; Gillies et al., 2015). Based on their association with cognitive measures or disease progression, a subset of descriptive WMH features can be selected and used as WMH related biomarkers. A machine learning diagnostic model can be generated based on a combination of these informative WMH related features and other relevant features obtained from

structural MRI (e.g. measures of grey matter atrophy in different structures), and clinical data (e.g. age, gender, level of education) to predict the cognitive state and the potential future decline of the patients in different neurodegenerative diseases.

(ii) WMHs have been studied mostly in patients with small-vessel disease and to some extent in Alzheimer's disease populations. However, as we were able to show in Chapter 6 with Parkinson's disease, the co-occurrence of WMHs along with other neurodegenerative diseases such as fronto-temporal dementia or Huntington's disease can lead to a more extensive decline in cognition than caused by the (pure) disease related changes. Studying and comparing the prevalence and distribution pattern of WMHs along with their effects on cognition and course of the disease in these populations can provide further insights into the mechanisms through which WMHs interact with neurodegenerative processes and affect cognition.

(iii) Specifically in PD, we have shown that the baseline WMHs are linked to the cognitive decline of the patients in as early as the *de novo* stage. Further studies to assess this relationship longitudinally and in the later stages of the disease are necessary. Additionally, adding information from other imaging metrics such as DTI and fMRI, one can investigate the interactions between WMHs and other disease components that lead to cognitive decline.

7.3. Conclusions

In conclusion, the proposed WMH segmentation pipeline has been found to provide robust and accurate segmentations in multi-center and multi-scanner databases. The pipeline can be used to detect WMHs from any combination of input sequences, and further validations

showed that even the segmentations that are produced using only T1w images correlate significantly with the manual labels and clinical and cognitive measures.

From studying the longitudinal effects of WMHs in *de Novo* Parkinson's disease patients, we were able to show that the cognitive performance of the subjects with higher WMH burdens declines significantly faster in the following years compared with the age matched controls with similar WMH loads, suggesting an interaction between the disease and vascular burden from the early drug naïve stages.

Continuing studies on the effect of WMHs in neurodegenerative disease is of great importance since multiple studies are showing that these vascular risk factors might be one of the initial stages in the chain of reactions that lead to neurodegeneration. Our WMH segmentation pipeline provides the opportunity to accurately detect and study both the causes and the effects of these lesions in large multi-center and multi-scanner aging and neurodegenerative disease databases. That can enhance our understanding of the processes that lead to neurodegeneration and their associated influencing risk factors, and potentially lead to intervention of their progression.

Appendix

Validation of the segmentations

31 subjects from the PPMI study had both T2w and FLAIR scans available. In order to assess the consistency between the segmentations based on these two sequences, the WMHs were segmented once based on T1w and T2w, and once based on T1w and FLAIR scans. The log transformed total WMH loads obtained from these segmentations were significantly correlated ($r=0.94$, $p<0.0001$). Figure 1 shows the WMH segmentations on axial slices for a subject that had both T2w and FLAIR scans.

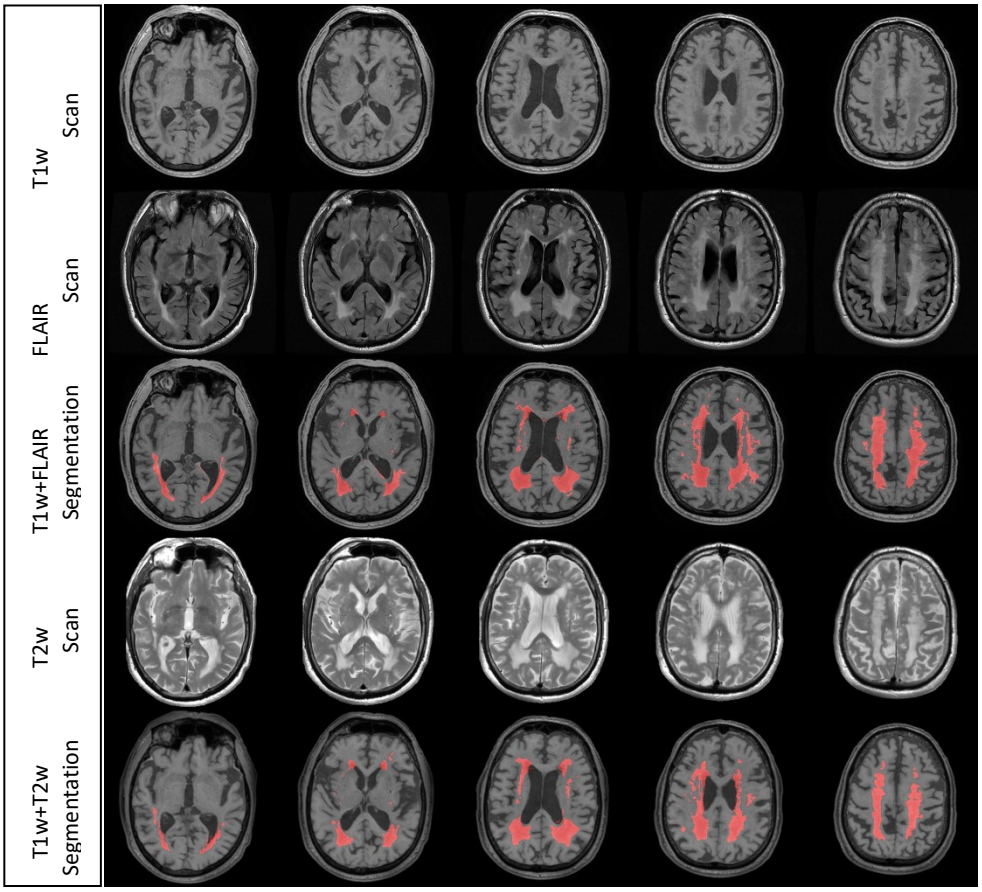


Fig. A.5. Axial slices of a subject with T1w, T2w, and FLAIR scans along with the automatic WMH segmentations based on T1w + FLAIR and T1w + T2w images. WMH= White Matter Hyperintensities.

Cognitive Testing:

Montreal Cognitive Assessment (MoCA): MoCA is a 10-minute 30-point cognitive screening tool for detection and assessment of mild cognitive impairment (MCI) (Nasreddine et al., 2005). The test involves short-term memory recall (5 points), visuospatial ability (4 points), executive function (4 points), attention and working memory (6 points), language (5 points), and orientation to time and place (6 points). A cut-off threshold of 26 (out of 30) is generally used for detecting MCI.

Hopkins Verbal Learning Test–Revised (HVLT): HVLT is a brief verbal learning and memory task ideal for repeated neuropsychological examinations consisting of a 12-item word list read to subjects on 3 successive trials (Benedict et al., 1998). Free recall scores are recorded for each trial, followed by a yes/no recognition task. Delayed recall trial follows a 20-25-min interval filled with unrelated tasks with no cues, similar to free recall trial. The score is calculated based on the total number of items recalled per trial.

Benton judgement of line orientation: Benton is a 30-item task that assesses the ability for discriminating the direction of lines (Benton et al., 1978). The response-choice display consists of an array of 11 lines each separated by an angle of 18 degrees. Each stimulus consists of two lines that represent either the proximal, middle, or distal half of a response-choice line. The performance is scored based on the number of correct responses.

Letter-Number Sequencing (LNS): LNS is used to assess working memory (Crowe, 2000). It consists of repeating a sequence of letters and numbers. Subjects have to repeat the numbers in ascending order, followed by the letter in alphabetical order. Total scores range from 0 to 21.

Semantic fluency: Semantic fluency is often used to study semantic memory (Capitani et

al., 1999). The subject is asked to name as many items as possible belonging to a given category, in the fixed time of one minute per category. The score is the number of names produced, excluding repetitions and circumlocutions. The scores generally range from 0 to 20.

Baseline Information

Table A.1. summarizes the results of correlating WMH loads with different cognitive measures at baseline.

Table A.6. Correlation of baseline cognitive measures with WMH loads.

Cognitive Score	Global Cognition		Memory		Visuospatial		Executive	
Cohort	r	p-value	r	p-value	r	p-value	r	p-value
HC	-0.070	0.39	-0.206	<0.001	-0.026	0.74	-0.101	0.217
PD	-0.004	0.93	-0.198	<0.001	-0.195	<0.001	-0.228	<0.001

Table A.2 summarizes the cognitive scores for HC and PD subjects with low and high WMH loads separately.

Table A.2. Descriptive statistics at baseline for the PPMI low and high WMHL subjects. Data are mean (SD) of key variables. PPMI=Parkinson's Progression Marker Initiative. WMHL= White Matter Hyperintensity Load. MoCA= Montreal Cognitive Assessment Score. HVLT= Hopkins Verbal Learning Test Revised Total Score. Benton= Benton Judgement of Line Orientation Score.

Cohort	Control		*De novo* PD	
Class	Low WMHL	High WMHL	Low WMHL	High WMHL
Age at Baseline (years)	60.48 (±12.00)	59.72 (±9.58)	60.54 (±11.02)	61.25 (±9.41)
MoCA at Baseline	27.21 (±2.00)	27.32 (±1.95)	27.30 (±2.32)	27.19 (±2.35)
HVLT at Baseline	34.10 (±7.10)	34.26 (±6.27)	33.28 (±6.89)	32.71 (±7.15)
Benton at Baseline	26.05 (±4.05)	26.38 (±4.30)	25.54 (±4.37)	25.17 (±4.30)
Executive Function at Baseline	21.59 (±4.65)	22.10 (±3.62)	21.19 (±4.50)	20.58 (±4.88)

Mixed-effects modeling

The following mixed-effects model with random and fixed effects was used for fitting

different cognitive scores:

$Score_{ij} = \beta_0 + \beta_1$ Age $+ \beta_2$ WMH Load $+ \beta_3$ Age:WMH Load $+ \gamma_{0i} + E_{ij}$

Where $Score_{ij}$ is one of the four cognitive scores (i.e. MoCA, Benton, HVLT, and Executive function) for timepoint j of subject i, β_0 is the intercept, β_1 is the linear term for Age, β_2 is the linear term for WMH Load, β_3 is the interaction term between Age and WMH Load. γ_{0i} is the random effects coefficient specific for subject i and E_{ij} is the error for time point j in subject i.

Figure A.2 shows the WMH probability distribution map for the PD group. The highest lesion probabilities are in the periventricular, occipital WM and centrum semi-ovale areas.

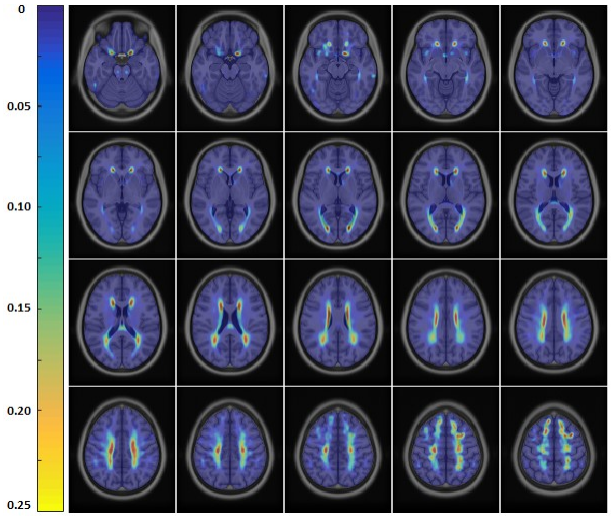


Fig. A.2. WMH distribution map for the PD cohort. WMH=White Matter Hyperintensity. PD= Parkinson's Disease.

Medication Information

None of the patients were on medication for the baseline visit for WMH and clinical measurements. 85.32 % of the PD subjects used in this study were put on PD medications at some point during the study. The Levodopa Equivalent Dose was calculated for the subjects that

received medications (mean= 285.01, std= 192.88). Controlling for age, we assessed the relationship between MoCA and LED using a mixed-effects model and found no significant correlation (r=0.01, p=0.47). The following is a summary of the drugs that the some of the patients received during the study: AMANDATIN, APOMORFIN, AZILECT, BENSERAZIDA, CABIDOPA/LEVODOPA, CLARIUM, CO-CARLEDOPA, CR SINEMET, DOPA MUCONA, DOPICAR, ELDEPRYL, ENTACAPONA, LEGANTO, LEVOCOMP RETARD, LEVODOPA BENSERAZIDE, LEVODOPA/CARBIDOPA, LEVODOPA/BENSERACID, LEVODOPA/ENTACAPONE, LEVOPAR, MADOPAR, MELEVODOPA, MIRAPEX, NACOM, NEUPRO, NEURPRO, PIRIBEDIL, PK MERZ, PRAMIPEXOL, RASAGALIN, REQUIP, RIGOTINE, ROPINEROL, ROTIGOTIN, RYTARI, SELEGELIN, SIFROL, SINEMET, STALEVO, SYMMETREL.

REFERENCES

Abdullah, B.A., Younis, A.A., Pattany, P.M., and Saraf-Lavi, E. (2011). Textural Based SVM for MS Lesion Segmentation in FLAIR MRIs. Open J. Med. Imaging *01*, 26–42.

Ad-Dab'bagh, Y., Lyttelton, O., Muehlboeck, J.S., Lepage, C., Einarson, D., Mok, K., Ivanov, O., Vincent, R.D., Lerch, J., Fombonne, E., et al. (2006). The CIVET image-processing environment: a fully automated comprehensive pipeline for anatomical neuroimaging research. In Proceedings of the 12th Annual Meeting of the Organization for Human Brain Mapping, (Florence, Italy), p. 2266.

Admiraal-Behloul, F., Van Den Heuvel, D.M.J., Olofsen, H., Van Osch, M.J.P., Van Der Grond, J., Van Buchem, M.A., and Reiber, J.H.C. (2005). Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage *28*, 607–617.

Akselrod-Ballin, A., Galun, M., Gomori, J.M., Filippi, M., Valsasina, P., Basri, R., and Brandt, A. (2009). Automatic segmentation and classification of multiple sclerosis in multichannel MRI. Biomed. Eng. IEEE Trans. On *56*, 2461–2469.

Alexander, J.A., Sheppard, S., Davis, P.C., and Salverda, P. (1996). Adult cerebrovascular disease: role of modified rapid fluid-attenuated inversion-recovery sequences. Am. J. Neuroradiol. *17*, 1507–1513.

Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. *46*, 175–185.

Amato, U., Larobina, M., Antoniadis, A., and Alfano, B. (2003). Segmentation of magnetic resonance brain images through discriminant analysis. J. Neurosci. Methods *131*, 65–74.

Anang, J.B., Gagnon, J.-F., Bertrand, J.-A., Romenets, S.R., Latreille, V., Panisset, M., Montplaisir, J., and Postuma, R.B. (2014). Predictors of dementia in Parkinson disease A prospective cohort study. Neurology *83*, 1253–1260.

Anbeek, P., Vincken, K., van Osch, M., Bisschops, B., Viergever, M., and van der Grond, J. (2003). Automated white matter lesion segmentation by voxel probability estimation. In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003, (Springer), pp. 610–617.

Anbeek, P., Vincken, K.L., van Osch, M.J., Bisschops, R.H., and van der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage *21*, 1037–1044.

Au, R., Massaro, J.M., Wolf, P.A., Young, M.E., Beiser, A., Seshadri, S., D'Agostino, R.B., and DeCarli, C. (2006). Association of white matter hyperintensity volume with decreased cognitive functioning: the Framingham Heart Study. Arch. Neurol. *63*, 246–250.

Aubert-Broche, B., Fonov, V.S., García-Lorenzo, D., Mouiha, A., Guizard, N., Coupé, P., Eskildsen, S.F., and Collins, D.L. (2013). A new method for structural volume analysis of longitudinal brain MRI data and its application in studying the growth trajectories of anatomical brain structures in childhood. NeuroImage *82*, 393–402.

Auning, E., Kjærvik, V.K., Selnes, P., Aarsland, D., Haram, A., Bjørnerud, A., Hessen, E., Esnaashari, A., and Fladby, T. (2014). White matter integrity and cognition in Parkinson's disease: a cross-sectional study. BMJ Open *4*, e003976.

Bakshi, R., Ariyaratana, S., Benedict, R.H.B., and Jacobs, L. (2001). Fluid-Attenuated Inversion Recovery Magnetic Resonance Imaging Detects Cortical and Juxtacortical Multiple Sclerosis Lesions. Arch. Neurol. *58*, 742–748.

Bao, P., and Zhang, L. (2003). Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. IEEE Trans. Med. Imaging *22*, 1089–1099.

Baracchini C, Mazzalai F, Gruppo M, Lorenzetti R, Ermani M, Ballotta E. Carotid endarterectomy protects elderly patients from cognitive decline: A prospective study. Surgery. 2012;151:99–106.

Barkhof, F., and Scheltens, P. (2002). Imaging of white matter lesions. Cerebrovasc. Dis. 1*3*, 21–30.

Bartko, J.J. (1991). Measurement and reliability: statistical thinking considerations. Schizophr. Bull. *17*, 483–489.

Beare, R., Srikanth, V., Chen, J., Phan, T.G., Stapleton, J., Lipshut, R., and Reutens, D. (2009). Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. Neuroimage *47*, 199–203.

Beekly, D.L., Ramos, E.M., Van Belle, G., Deitrich, W., Clark, A.D., Jacka, M.E., Kukull, W.A., and others (2004). The National Alzheimer's Coordinating Center (NACC) Database: an Alzheimer disease database. Alzheimer Dis. Assoc. Disord. *18*, 270–277.

Beggs, C.B., Magnano, C., Shepherd, S.J., Belov, P., Ramasamy, D.P., Hagemeier, J., and Zivadinov, R. (2016). Dirty-Appearing White Matter in the Brain is Associated with Altered

Cerebrospinal Fluid Pulsatility and Hypertension in Individuals without Neurologic Disease. J. Neuroimaging *26*, 136–143.

Benedict, R.H.B., Schretlen, D., Groninger, L., and Brandt, J. (1998). Hopkins Verbal Learning Test – Revised: Normative Data and Analysis of Inter-Form and Test-Retest Reliability. Clin. Neuropsychol. *12*, 43–55.

Benedictus, M.R., Binnewijzend, M.A.A., Kuijer, J.P.A., Steenwijk, M.D., Versteeg, A., Vrenken, H., Scheltens, P., Barkhof, F., van der Flier, W.M., and Prins, N.D. (2014). Brain volume and white matter hyperintensities as determinants of cerebral blood flow in Alzheimer's disease. Neurobiol. Aging *35*, 2665–2670.

Benton, A.L., Varney, N.R., and deS Hamsher, K. (1978). Visuospatial judgment: A clinical test. Arch. Neurol. *35*, 364–367.

Biesbroek, J.M., Weaver, N.A., and Biessels, G.J. (2017). Lesion location and cognitive impact of cerebral small vessel disease. Clin. Sci. *131*, 715–728.

Boone, K.B., Miller, B.L., Lesser, I.M., Mehringer, C.M., Hill-Gutierrez, E., Goldberg, M.A., and Berman, N.G. (1992). Neuropsychological correlates of white-matter lesions in healthy elderly subjects: a threshold effect. Arch. Neurol. *49*, 549–554.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, (ACM), pp. 144–152.

Boucetta, S., Salimi, A., Dadar, M., Jones, B.E., Collins, D.L., and Dang-Vu, T.T. (2016).

Structural brain alterations associated with rapid eye movement sleep behavior disorder in Parkinson's disease. Sci. Rep. *6*.

Boyle, P.A., Yu, L., Fleischman, D.A., Leurgans, S., Yang, J., Wilson, R.S., Schneider, J.A., Arvanitakis, Z., Arfanakis, K., and Bennett, D.A. (2016). White matter hyperintensities, incident mild cognitive impairment, and cognitive decline in old age. Ann. Clin. Transl. Neurol. *3*, 791–800.

Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140.

Breiman, L. (2001a). Random Forests. Mach. Learn. *45*, 5–32.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). Classification and regression trees (CRC press).

Brickman, A.M., Provenzano, F.A., Muraskin, J., Manly, J.J., Blum, S., Apa, Z., Stern, Y., Brown, T.R., Luchsinger, J.A., and Mayeux, R. (2012). Regional white matter hyperintensity volume, not hippocampal atrophy, predicts incident Alzheimer disease in the community. Arch. Neurol. *69*, 1621–1627.

Brickman, A.M., Schupf, N., Manly, J.J., Stern, Y., Luchsinger, J.A., Provenzano, F.A., Narkhede, A., Razlighi, Q., Collins-Praino, L., Artero, S., et al. (2014). APOE ε4 and risk for Alzheimer's disease: Do regionally distributed white matter hyperintensities play a role? Alzheimers Dement. *10*, 619–629.

Brickman, A.M., Zahodne, L.B., Guzman, V.A., Narkhede, A., Meier, I.B., Griffith, E.Y., Provenzano, F.A., Schupf, N., Manly, J.J., Stern, Y., et al. (2015). Reconsidering harbingers of

dementia: progression of parietal lobe white matter hyperintensities predicts Alzheimer's disease incidence. Neurobiol. Aging *36*, 27–32.

Brinkmann, B.H., Manduca, A., and Robb, R.A. (1998). Optimized homomorphic unsharp masking for MR grayscale inhomogeneity correction. IEEE Trans. Med. Imaging *17*, 161–171.

Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. Neuroinformatics *13*, 261–276.

Callahan, B.L., Ramirez, J., Berezuk, C., Duchesne, S., and Black, S.E. (2015). Predicting Alzheimer's disease development: a comparison of cognitive criteria and associated neuroimaging biomarkers. Alzheimers Res. Ther. *7*, 68.

Capitani, E., Laiacona, M., and Barbarotto, R. (1999). Gender Affects Word Retrieval of Certain Categories in Semantic Fluency Tasks. Cortex *35*, 273–278.

Carmichael, O., Schwarz, C., Drucker, D., Fletcher, E., Harvey, D., Beckett, L., Jack, C.R., Weiner, M., and DeCarli, C. (2010). Longitudinal Changes in White Matter Disease and Cognition in the First Year of the Alzheimer Disease Neuroimaging Initiative. Arch. Neurol. *67*, 1370–1378.

Chan, R.C., Shum, D., Toulopoulou, T., and Chen, E.Y. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. Arch. Clin. Neuropsychol. *23*, 201–216.

Chao, W.-H., Chen, Y.-Y., Lin, S.-H., Shih, Y.-Y.I., and Tsang, S. (2009). Automatic

segmentation of magnetic resonance images using a decision tree with spatial information. Comput. Med. Imaging Graph. *33*, 111–121.

Ciccarelli, O., Brex, P.A., Thompson, A.J., and Miller, D.H. (2002). Disability and lesion load in MS: a reassessment with MS functional composite score and 3D fast FLAIR. J. Neurol. *249*, 18–24.

Clarke, L.P., Velthuizen, R.P., Camacho, M.A., Heine, J.J., Vaidyanathan, M., Hall, L.O., Thatcher, R.W., and Silbiger, M.L. (1995). MRI segmentation: methods and applications. Magn. Reson. Imaging *13*, 343–368.

Collins, D.L., and Evans, A.C. (1997a). Animal: validation and applications of nonlinear registration-based segmentation. Int. J. Pattern Recognit. Artif. Intell. *11*, 1271–1294.

Collins, D.L., Neelin, P., Peters, T.M., and Evans, A.C. (1994a). Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. *18*, 192–205.

Conklin, J., Silver, F.L., Mikulis, D.J., and Mandell, D.M. (2014). Are acute infarcts the cause of leukoaraiosis? Brain mapping for 16 consecutive weeks. Ann. Neurol. *76*, 899–904.

Cools, R., Altamirano, L., and D'Esposito, M. (2006). Reversal learning in Parkinson's disease depends on medication status and outcome valence. Neuropsychologia *44*, 1663–1673.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

Costa, A.S., Reich, A., Fimm, B., Ketteler, S.T., Schulz, J.B., and Reetz, K. (2014). Evidence of the sensitivity of the MoCA alternate forms in monitoring cognitive change in early Alzheimer's

disease. Dement. Geriatr. Cogn. Disord. *37*, 95–103.

Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., and Barillot, C. (2008). An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images. IEEE Trans. Med. Imaging *27*, 425–441.

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. IEEE Trans. Inf. Theory *13*, 21–27.

Cox, D.R. (1958). The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B Methodol. 215–242.

Crowe, S.F. (2000). Does the letter number sequencing task measure anything more than digit span? Assessment *7*, 113–117.

Dadar, M., Pascoal, T., Manitsirikul, S., Misquitta, K., Tartaglia, C., Brietner, J., Rosa-Neto, P., Carmichael, O., DeCarli, C., and Collins, D.L. (2017a). Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease. IEEE Trans. Med. Imaging.

Dadar, M., Maranzano, J., Misquitta, K., Anor, C.J., Fonov, V.S., Tartaglia, M.C., Carmichael, O.T., Decarli, C., Collins, D.L., and Alzheimer's Disease Neuroimaging Initiative (2017b). Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. NeuroImage *157*, 233–249.

Dadar, M., Zeighami, Y., Yau, Y., Fereshtehnejad, S. M., Maranzano, J., Postuma, R. B., ... & Collins, D. L. (2017c). White Matter Hyperintensities and Cognitive Decline in de Novo Parkinson's Disease Patients. *bioRxiv*, 230896.

Dadar, M., Maranzano, J., Ducharme, S., Carmichael, O. T., Decarli, C., & Collins, D. L. (2018). Validation of T1w-based segmentations of white matter hyperintensity volumes in large-scale datasets of aging. *Human brain mapping*, *39*(3), 1093-1107.

Dalaker, T.O., Larsen, J.P., Dwyer, M.G., Aarsland, D., Beyer, M.K., Alves, G., Bronnick, K., Tysnes, O.-B., and Zivadinov, R. (2009). White matter hyperintensities do not impact cognitive function in patients with newly diagnosed Parkinson's disease. Neuroimage *47*, 2083–2089.

De Boer, R., Vrooman, H.A., Van Der Lijn, F., Vernooij, M.W., Ikram, M.A., Van Der Lugt, A., Breteler, M.M., and Niessen, W.J. (2009). White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage *45*, 1151–1161.

De Groot, J.C., De Leeuw, F.-E., Oudkerk, M., Van Gijn, J., Hofman, A., Jolles, J., and Breteler, M. (2002). Periventricular cerebral white matter lesions predict rate of cognitive decline. Ann. Neurol. *52*, 335–341.

De la Torre, J.C. (2000). Critically attained threshold of cerebral hypoperfusion: the CATCH hypothesis of Alzheimer's pathogenesis. Neurobiol. Aging *21*, 331–342.

Debette, S., and Markus, H.S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. Bmj *341*, c3666.

DeCarli, C., Murphy, D.G.M., Tranh, M., Grady, C.L., Haxby, J.V., Gillette, J.A., Salerno, J.A.,

Gonzales-Aviles, A., Honvitz, B., Rapoport, S.I., et al. (1995a). The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults. Neurology *45*, 2077–2084.

DeCarli, C., Murphy, D.G.M., Tranh, M., Grady, C.L., Haxby, J.V., Gillette, J.A., Salerno, J.A., Gonzales-Aviles, A., Honvitz, B., Rapoport, S.I., et al. (1995b). The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults. Neurology *45*, 2077–2084.

DeCarli, C., Miller, B.L., Swan, G.E., Reed, T., Wolf, P.A., and Carmelli, D. (2001). Cerebrovascular and brain morphologic correlates of mild cognitive impairment in the National Heart, Lung, and Blood Institute Twin Study. Arch. Neurol. *58*, 643–647.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 1–38.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. Syst. Man Cybern. IEEE Trans. On *25*, 804–813.

Deoni, S.C., Matthews, L., and Kolind, S.H. (2013). One component? Two components? Three? The effect of including a nonexchanging “free” water component in multicomponent driven equilibrium single pulse observation of T1 and T2. Magn. Reson. Med. *70*, 147–154.

Dice, L.R. (1945). Measures of the amount of ecologic association between species. Ecology *26*, 297–302.

Doi, T., Makizako, H., Shimada, H., Tsutsumimoto, K., Hotta, R., Nakakubo, S., Park, H., and

Suzuki, T. (2015). Objectively measured physical activity, brain atrophy, and white matter lesions in older adults with mild cognitive impairment. Exp. Gerontol. *62*, 1–6.

Dubois, B., Feldman, H.H., Jacova, C., Hampel, H., Molinuevo, J.L., Blennow, K., DeKosky, S.T., Gauthier, S., Selkoe, D., Bateman, R., et al. (2014a). Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol. *13*, 614–629.

Dufouil, C., de Kersaint–Gilly, A., Besancon, V., Levy, C., Auffray, E., Brunnereau, L., Alperovitch, A., and Tzourio, C. (2001). Longitudinal study of blood pressure and white matter hyperintensities The EVA MRI Cohort. Neurology *56*, 921–926.

Dufouil, C., Chalmers, J., Coskun, O., Besançon, V., Bousser, M.-G., Guillon, P., MacMahon, S., Mazoyer, B., Neal, B., Woodward, M., et al. (2005). Effects of Blood Pressure Lowering on Cerebral White Matter Hyperintensities in Patients With Stroke. Circulation *112*, 1644–1650.

Dyrby, T.B., Rostrup, E., Baaré, W.F.C., van Straaten, E.C.W., Barkhof, F., Vrenken, H., Ropele, S., Schmidt, R., Erkinjuntti, T., Wahlund, L.-O., et al. (2008). Segmentation of age-related white matter changes in a clinical multi-center study. NeuroImage *41*, 335–345.

Edelman, R.R., Hesselink, J.R., and Zlatkin, M.B. (1996). Clinical magnetic resonance imaging Vol 1 (Philadelphia: Saunders, 1996).

Erus, G., Zacharaki, E.I., and Davatzikos, C. (2014). Individualized statistical learning from medical image databases: Application to identification of brain lesions. Med. Image Anal. *18*, 542–554.

Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Wassef, S.N.,

Østergaard, L.R., Collins, D.L., Initiative, A.D.N., et al. (2012). BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage *59*, 2362–2373.

Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I., and Zimmerman, R.A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. Am. J. Roentgenol. *149*, 351–356.

Fazekas, F., Kleinert, R., Offenbacher, H., Schmidt, R., Kleinert, G., Payer, F., Radner, H., and Lechner, H. (1993). Pathologic correlates of incidental MRI white matter signal hyperintensities. Neurology *43*, 1683–1683.

Fazekas, F., Barkhof, F., Wahlund, L.O., Pantoni, L., Erkinjuntti, T., Scheltens, P., and Schmidt, R. (2002). CT and MRI Rating of White Matter Lesions. Cerebrovasc. Dis. *13*, 31–36.

Fernando, M.S., O'brien, J.T., Perry, R.H., English, P., Forster, G., McMeekin, W., Slade, J.Y., Golkhar, A., Matthews, F.E., Barber, R., et al. (2004). Comparison of the pathology of cerebral white matter with post-mortem magnetic resonance imaging (MRI) in the elderly brain. Neuropathol. Appl. Neurobiol. *30*, 385–395.

Ferrari, R.J., Wei, X., Zhang, Y., Scott, J.N., and Mitchell, J.R. (2003). Segmentation of multiple sclerosis lesions using support vector machines. In Medical Imaging 2003, (International Society for Optics and Photonics), pp. 16–26.

Filippi, M., Baratti, C., Yousry, T., Horsfield, M.A., Mammi, S., Becker, C., Voltz, R., Spuler, S., Campi, A., Reiser, M.F., et al. (1996). Quantitative assessment of MRI lesion load in multiple sclerosis. Brain *119*, 1349–1355.

Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen. *7*, 179–188.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., and Collins, D.L. (2011a). Unbiased average age-appropriate atlases for pediatric studies. NeuroImage *54*, 313–327.

Fonov, V., Coupé, P., Eskildsen, S.F., and Collins, L.D. (2011b). Atrophy specific MRI brain template for Alzheimer's disease and Mild Cognitive Impairment. In Alzheimer's Association International Conference, (France), p. S58.

Freifeld, O., Greenspan, H., and Goldberger, J. (2009). Multiple Sclerosis Lesion Detection Using Constrained GMM and Curve Evolution. J. Biomed. Imaging *2009*, 14:1–14:13.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. J.-Jpn. Soc. Artif. Intell. *14*, 1612.

Friedman, J., Hastie, T., Tibshirani, R., and others (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. *28*, 337–407.

Fujishima, M., Maikusa, N., Nakamura, K., Nakatsuka, M., Matsuda, H., and Meguro, K. (2014). Mild cognitive impairment, poor episodic memory, and late-life depression are associated with cerebral cortical thinning and increased white matter hyperintensities. Front. Aging Neurosci. *6*.

Gao, F., Swartz, R.H., Scheltens, P., Leibovitch, F.S., Kiss, A., Honjo, K., and Black, S.E. (2011). Complexity of MRI white matter hyperintensity assessments in relation to cognition in

aging and dementia from the Sunnybrook Dementia Study. J. Alzheimers Dis. *26*, 379–388.

García-Lorenzo, D., Lecoeur, J., Arnold, D.L., Collins, D.L., and Barillot, C. (2009). Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009, (Springer), pp. 584–591.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., and Collins, D.L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. *17*, 1–18.

Gauthier, S., Albert, M., Fox, N., Goedert, M., Kivipelto, M., Mestre-Ferrandiz, J., and Middleton, L.T. (2016). Why has therapy development for dementia failed in the last two decades? Alzheimers Dement. *12*, 60–64.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 721–741.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., and Ayache, N. (2011). Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage *57*, 378–390.

Gerig, G., Kubler, O., Kikinis, R., and Jolesz, F.A. (1992). Nonlinear anisotropic filtering of MRI data. IEEE Trans. Med. Imaging *11*, 221–232.

Ghafoorian, M., Karssemeijer, N., van Uden, I.W.M., de Leeuw, F.-E., Heskes, T., Marchiori, E., and Platel, B. (2016a). Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. Med. Phys. *43*, 6246–6258.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I., Sanchez, C., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., and Platel, B. (2016b). Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities. ArXiv Prepr. ArXiv161004834.

Gibbons, L.E., Carle, A.C., Mackin, R.S., Harvey, D., Mukherjee, S., Insel, P., Curtis, S.M., Mungas, D., Crane, P.K., Initiative, A.D.N., et al. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. Brain Imaging Behav. *6*, 517–527.

Gillies, R.J., Kinahan, P.E., and Hricak, H. (2015). Radiomics: images are more than pictures, they are data. Radiology *278*, 563–577.

Gordon, B.A., Najmi, S., Hsu, P., Roe, C.M., Morris, J.C., and Benzinger, T.L. (2015). The effects of white matter hyperintensities and amyloid deposition on Alzheimer dementia. NeuroImage Clin. *8*, 246–252.

Gorelick, P.B., Scuteri, A., Black, S.E., DeCarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D., et al. (2011). Vascular contributions to cognitive impairment and dementia a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke *42*, 2672–2713.

Gouw, A.A., Flier, W.M. van der, Straaten, E.C.W. van, Pantoni, L., Bastos-Leite, A.J., Inzitari, D., Erkinjuntti, T., Wahlund, L.O., Ryberg, C., Schmidt, R., et al. (2008). Reliability and Sensitivity of Visual Scales versus Volumetry for Evaluating White Matter Hyperintensity Progression. Cerebrovasc. Dis. *25*, 247–253.

Gouw, A.A., Seewann, A., Van Der Flier, W.M., Barkhof, F., Rozemuller, A.M., Scheltens, P., and Geurts, J.J. (2010). Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations. J. Neurol. Neurosurg. Psychiatry jnnp–2009.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., et al. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. NeuroImage *141*, 191–205.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G.J., Plummer, D.L., Tofts, P.S., McDonald, W.I., and Miller, D.H. (1996). Quantification of MRI lesion load in multiple sclerosis: A comparison of three computer-assisted techniques. Magn. Reson. Imaging *14*, 495–505.

Grimmer, T., Faust, M., Auer, F., Alexopoulos, P., Förstl, H., Henriksen, G., Perneczky, R., Sorg, C., Yousefi, B.H., Drzezga, A., et al. (2012). White matter hyperintensities predict amyloid increase in Alzheimer's disease. Neurobiol. Aging *33*, 2766–2773.

Gunning-Dixon, F.M., and Raz, N. (2000). The cognitive correlates of white matter abnormalities in normal aging: a quantitative review. Neuropsychology *14*, 224.

Hachinski, V.C., Potter, P., and Merskey, H. (1987). Leuko-Araiosis. Arch. Neurol. *44*, 21–23.

Halliday, G.M., Leverenz, J.B., Schneider, J.S., and Adler, C.H. (2014). The neurobiological basis of cognitive impairment in Parkinson's disease. Mov. Disord. *29*, 634–650.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N.,

Brooks, D.J., and Duncan, J.S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. *19*, 224–247.

Harrington, D.P., and Fleming, T.R. (1982). A class of rank test procedures for censored survival data. Biometrika 553–566.

Hawkins, K.A., Emadi, N., Pearlson, G.D., Winkler, A.M., Taylor, B., Dulipsingh, L., King, D., Pittman, B., and Blank, K. (2017). Hyperinsulinemia and elevated systolic blood pressure independently predict white matter hyperintensities with associated cognitive decrement in the middle-aged offspring of dementia patients. Metab. Brain Dis. 1–9.

Hely, M.A., Reid, W.G., Adena, M.A., Halliday, G.M., and Morris, J.G. (2008). The Sydney multicenter study of Parkinson's disease: the inevitability of dementia at 20 years. Mov. Disord. *23*, 837–844.

Herskovits, E., Bryan, R., and Yang, F. (2008). Automated Bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study. Adv. Med. Sci. *53*, 182.

Hinton, L., Carter, K., Reed, B.R., Beckett, L., Lara, E., DeCarli, C., and Mungas, D. (2010). Recruitment of a community-based cohort for research on diversity and risk of dementia. Alzheimer Dis. Assoc. Disord. *24*, 234.

Hirai, T., Korogi, Y., Yoshizumi, K., Shigematsu, Y., Sugahara, T., and Takahashi, M. (2000). Limbic Lobe of the Human Brain: Evaluation with Turbo Fluid-attenuated Inversion-Recovery MR Imaging. Radiology *215*, 470–475.

Honjo, K., Black, S.E., and Verhoeff, N.P. (2012). Alzheimer's disease, cerebrovascular disease,

and the β-amyloid cascade. Can. J. Neurol. Sci. *39*, 712–728.

Hoops, S., Nazem, S., Siderowf, A.D., Duda, J.E., Xie, S.X., Stern, M.B., and Weintraub, D. (2009). Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. Neurology *73*, 1738–1745.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Netw. *2*, 359–366.

Huang, Y.H., Zhang, W.W., Lin, L., Feng, J., Zhao, X.X., Guo, W.H., and Wei, W. (2010). Could changes in arterioles impede the perivascular drainage of interstitial fluid from the cerebral white matter in leukoaraiosis? Neuropathol. Appl. Neurobiol. *36*, 237–247.

Hunt, E.B., Marin, J., and Stone, P.J. (1966). Experiments in induction.

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

Ithapu, V., Singh, V., Lindner, C., Austin, B.P., Hinrichs, C., Carlsson, C.M., Bendlin, B.B., and Johnson, S.C. (2014). Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. Hum. Brain Mapp. *35*, 4219–4235.

Iturria-Medina, Y., Sotero, R.C., Toussaint, P.J., Mateos-Pérez, J.M., and Evans, A.C. (2016). Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. Nat. Commun. *7*, 11934.

Iturria-Medina, Y., Carbonell, F.M., Sotero, R.C., Chouinard-Decorte, F., and Evans, A.C. (2017). Multifactorial causal model of brain (dis)organization and therapeutic intervention: Application to Alzheimer's disease. NeuroImage *152*, 60–77.

Jack, C.R., and Holtzman, D.M. (2013). Biomarker modeling of Alzheimer's disease. Neuron *80*, 1347–1358.

Jack, C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., and Erickson, B.J. (2001). FLAIR histogram segmentation for measurement of leukoaraiosis volume. J. Magn. Reson. Imaging *14*, 668–676.

Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., and Trojanowski, J.Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. *9*, 119–128.

Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. *12*, 207–216.

Jacobs, H.I.L., Leritz, E.C., Williams, V.J., Van Boxtel, M.P.J., Elst, W. van der, Jolles, J., Verhey, F.R.J., McGlinchey, R.E., Milberg, W.P., and Salat, D.H. (2013). Association between white matter microstructure, executive functions, and processing speed in older adults: The impact of vascular health. Hum. Brain Mapp. *34*, 77–95.

Jellinger, K.A. (2003). Prevalence of cerebrovascular lesions in Parkinson's disease. A postmortem study. Acta Neuropathol. (Berl.) *105*, 415–419.

Jellinger, K.A., Schmidt, R., and Windisch, M. (2013). Advances in Dementia Research (Springer Science & Business Media).

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., and Smith, S.M. (2012). Fsl. Neuroimage *62*, 782–790.

Joana B, Fonseca MJ, Moreira A, Abelha F. Quality of life in patients with cognitive decline after major surgery: 17AP2-9. Eur J Anaesthesiol EJA. 2013;30:239–239.

Jones, J.D., Tanner, J.J., Okun, M., Price, C.C., and Bowers, D. (2017). Are Parkinson's Patients More Vulnerable to the Effects of Cardiovascular Risk: A Neuroimaging and Neuropsychological Study. J. Int. Neuropsychol. Soc. *23*, 322–331.

Joutel, A., and Faraci, F.M. (2014). Cerebral Small Vessel Disease. Stroke *45*, 1215–1221.

Kamber, M., Collins, D.L., Shinghal, R., Francis, G.S., and Evans, A.C. (1992). Model-based 3-D segmentation of multiple sclerosis lesions in dual-echo MRI data. In Visualization in Biomedical Computing, (International Society for Optics and Photonics), pp. 590–600.

Kamber, M., Shinghal, R., Collins, D.L., Francis, G.S., and Evans, A.C. (1995). Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. Med. Imaging IEEE Trans. On *14*, 442–453.

Kandiah, N., Zainal, N.H., Narasimhalu, K., Chander, R.J., Ng, A., Mak, E., Au, W.L., Sitoh, Y.Y., Nadkarni, N., and Tan, L.C. (2014). Hippocampal volume and white matter disease in the prediction of dementia in Parkinson's disease. Parkinsonism Relat. Disord. *20*, 1203–1208.

Karimaghaloo, Z., Shah, M., Francis, S.J., Arnold, D.L., Collins, D.L., and Arbel, T. (2012).

Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields. Med. Imaging IEEE Trans. On *31*, 1181–1194.

Kertesz, A., Black, S.E., Tokar, G., Benke, T., Carr, T., and Nicholson, L. (1988). Periventricular and subcortical hyperintensities on magnetic résonance imaging:"rims, caps, and unidentified bright objects". Arch. Neurol. *45*, 404–408.

Khayati, R., Vafadust, M., Towhidkhah, F., and 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.

Kindermann, R., and Snell, L. (1980). Markov random fields and their applications.

Klunk, W.E., Engler, H., Nordberg, A., Wang, Y., Blomqvist, G., Holt, D.P., Bergström, M., Savitcheva, I., Huang, G.-F., Estrada, S., et al. (2004). Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Ann. Neurol. *55*, 306–319.

Koch, G.G. (1982). Intraclass correlation coefficient. Encycl. Stat. Sci.

Köse, C., Şevik, U., İkibaş, C., and Erdöl, H. (2012). Simple methods for segmentation and measurement of diabetic retinopathy lesions in retinal fundus images. Comput. Methods Programs Biomed. *107*, 274–293.

Lam, C.L.M., Yiend, J., and Lee, T.M.C. (2017). Imaging and neuropsychological correlates of white matter lesions in different subtypes of Mild Cognitive Impairment: A systematic review. NeuroRehabilitation *Preprint*, 1–16.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R.G., Granton, P.,

Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer *48*, 441–446.

Langa, K.M., Larson, E.B., Crimmins, E.M., Faul, J.D., Levine, D.A., Kabeto, M.U., and Weir, D.R. (2016). A comparison of the prevalence of dementia in the United States in 2000 and 2012. JAMA Intern. Med.

Lange, C., Suppa, P., Mäurer, A., Ritter, K., Pietrzyk, U., Steinhagen-Thiessen, E., Fiebach, J.B., Spies, L., and Buchert, R. (2016). Mental speed is associated with the shape irregularity of white matter MRI hyperintensity load. Brain Imaging Behav. 1–11.

Lao, Z., Shen, D., Liu, D., Jawad, A.F., Melhem, E.R., Launer, L.J., Bryan, R.N., and Davatzikos, C. (2008). Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. Acad. Radiol. *15*, 300–313.

de Lau, L.M., Schipper, C.M.A., Hofman, A., Koudstaal, P.J., and Breteler, M.M. (2005). Prognosis of Parkinson disease: risk of dementia and mortality: the Rotterdam Study. Arch. Neurol. *62*, 1265–1269.

Lee, S., Viqar, F., Zimmerman, M.E., Narkhede, A., Tosto, G., Benzinger, T.L.S., Marcus, D.S., Fagan, A.M., Goate, A., Fox, N.C., et al. (2016). White matter hyperintensities are a core feature of Alzheimer's disease: Evidence from the dominantly inherited Alzheimer network. Ann. Neurol.

Leemput, K.V., Maes, F., Vandermeulen, D., Colchester, A., and Suetens, P. (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging *20*, 677–688.

Leeuw, F.-E. de, Groot, J.C. de, Achten, E., Oudkerk, M., Ramos, L.M.P., Heijboer, R., Hofman, A., Jolles, J., Gijn, J. van, and Breteler, M.M.B. (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.

Leeuw, F.-E., de Groot, J.C., Oudkerk, M., Witteman, J.C.M., Hofman, A., Van Gijn, J., and Breteler, M.M.B. (2002). Hypertension and cerebral white matter lesions in a prospective cohort study. Brain *125*, 765–772.

Leite, A.J.B., Straaten, E.C.W. van, Scheltens, P., Lycklama, G., and Barkhof, F. (2004). Thalamic Lesions in Vascular Dementia Low Sensitivity of Fluid-Attenuated Inversion Recovery (FLAIR) Imaging. Stroke *35*, 415–419.

Leritz, E.C., Shepel, J., Williams, V.J., Lipsitz, L.A., McGlinchey, R.E., Milberg, W.P., and Salat, D.H. (2014). Associations between T1 white matter lesion volume and regional white matter microstructure in aging. Hum. Brain Mapp. *35*, 1085–1100.

Lewis, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In Machine Learning: ECML-98, (Springer), pp. 4–15.

Li, J.-Q., Tan, L., Wang, H.-F., Tan, M.-S., Tan, L., Xu, W., Zhao, Q.-F., Wang, J., Jiang, T., and Yu, J.-T. (2016). Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies. J Neurol Neurosurg Psychiatry *87*, 476–484.

Li, Y., Hara, S., Ito, W., and Shimura, K. (2007). A machine learning approach for interactive lesion segmentation. pp. 651246-651246–651248.

Lindemer, E.R., Salat, D.H., Smith, E.E., Nguyen, K., Fischl, B., Greve, D.N., Initiative, A.D.N., and others (2015). White matter signal abnormality quality differentiates mild cognitive impairment that converts to Alzheimer's disease from nonconverters. Neurobiol. Aging *36*, 2447–2457.

Longstreth, W.T., Manolio, T.A., Arnold, A., Burke, G.L., Bryan, N., Jungreis, C.A., Enright, P.L., O'leary, D., Fried, L., and others (1996). Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people. Stroke *27*, 1274–1282.

Lopez, O.L., Jagust, W.J., DeKosky, S.T., Becker, J.T., Fitzpatrick, A., Dulberg, C., Breitner, J., Lyketsos, C., Jones, B., Kawas, C., et al. (2003). Prevalence and classification of mild cognitive impairment in the Cardiovascular Health Study Cognition Study: part 1. Arch. Neurol. *60*, 1385–1389.

Madabhushi, A., Shi, J., Feldman, M., Rosen, M., and Tomaszewski, J. (2006). Comparing Ensembles of Learners: Detecting Prostate Cancer from High Resolution MRI. In Computer Vision Approaches to Medical Image Analysis, R.R. Beichel, and M. Sonka, eds. (Springer Berlin Heidelberg), pp. 25–36.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997). Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging *16*, 187–198.

Maier, O., Wilms, M., von der Gablentz, J., Krämer, U.M., Münte, T.F., and Handels, H. (2015). Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. J. Neurosci. Methods *240*, 89–100.

Mak, E., Dwyer, M.G., Ramasamy, D.P., Au, W.L., Tan, L., Zivadinov, R., and Kandiah, N. (2015). White matter hyperintensities and mild cognitive impairment in Parkinson's disease. J. Neuroimaging *25*, 754–760.

Mamdani, E.H., and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud. *7*, 1–13.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., and Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Magn. Reson. Imaging *31*, 192–203.

Mäntylä, R., Erkinjuntti, T., Salonen, O., Aronen, H.J., Peltonen, T., Pohjasvaara, T., and Standertskjöld-Nordenstam, C.-G. (1997). Variable Agreement Between Visual Rating Scales for White Matter Hyperintensities on MRI: Comparison of 13 Rating Scales in a Poststroke Cohort. Stroke *28*, 1614–1623.

Maranzano, J., Rudko, D.A., Arnold, D.L., and Narayanan, S. (2016). Manual Segmentation of MS Cortical Lesions Using MRI: A Comparison of 3 MRI Reading Protocols. Am. J. Neuroradiol. *37*, 1623–1628.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al. (2011). The Parkinson Progression Marker Initiative (PPMI). Prog. Neurobiol. *95*, 629–635.

Masdeu, J.C., and Pascual, B. (2008). Neuroimaging of disorders leading to dementia. Curr. Neurol. Neurosci. Rep. *8*, 443.

Matsusue, E., Sugihara, S., Fujii, S., Ohama, E., Kinoshita, T., And Ogawa, T. (2006). White matter changes in elderly people: MR-pathologic correlations. Magn. Reson. Med. Sci. *5*, 99–104.

McLachlan, G. (2004). Discriminant analysis and statistical pattern recognition (John Wiley & Sons).

Mechrez, R., Goldberger, J., and Greenspan, H. (2016). Patch-based Segmentation with Spatial Consistency: Application to MS Lesions in Brain MRI. J. Biomed. Imaging *2016*, 3:3–3:3.

Merino, J.G., and Hachinski, V. (2000). Leukoaraiosis: reifying rarefaction. Arch. Neurol. *57*, 925–926.

Mitra, J., Bourgeat, P., Fripp, J., Ghose, S., Rose, S., Salvado, O., Connelly, A., Campbell, B., Palmer, S., Sharma, G., et al. (2014). Lesion segmentation from multimodal MRI using random forest following ischemic stroke. NeuroImage *98*, 324–335.

Moore, G.R.W., Laule, C., MacKay, A., Leung, E., Li, D.K.B., Zhao, G., Traboulsee, A.L., and Paty, D.W. (2008). Dirty-appearing white matter in multiple sclerosis. J. Neurol. *255*, 1802.

Morgen, K., Schneider, M., Frölich, L., Tost, H., Plichta, M.M., Kölsch, H., Rakebrandt, F., Rienhoff, O., Jessen, F., Peters, O., et al. (2015). Apolipoprotein E-dependent load of white matter hyperintensities in Alzheimer's disease: a voxel-based lesion mapping study. Alzheimers Res. Ther. *7*, 27.

Morris, J.C., Weintraub, S., Chui, H.C., Cummings, J., DeCarli, C., Ferris, S., Foster, N.L., Galasko, D., Graff-Radford, N., Peskind, E.R., et al. (2006). The Uniform Data Set (UDS):

clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. Alzheimer Dis. Assoc. Disord. *20*, 210–216.

Mortazavi, D., Kouzani, A.Z., and Soltanian-Zadeh, H. (2012). Segmentation of multiple sclerosis lesions in MR images: a review. Neuroradiology *54*, 299–320.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., and Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. J. Am. Geriatr. Soc. *53*, 695–699.

Nordahl, C.W., Ranganath, C., Yonelinas, A.P., DeCarli, C., Reed, B.R., and Jagust, W.J. (2005). Different mechanisms of episodic memory failure in mild cognitive impairment. Neuropsychologia *43*, 1688–1697.

Nyúl, L.G., Udupa, J.K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging *19*, 143–150.

Ogama, N., Sakurai, T., Shimizu, A., and Toba, K. (2014). Regional white matter lesions predict falls in patients with amnestic mild cognitive impairment and Alzheimer's disease. J. Am. Med. Dir. Assoc. *15*, 36–41.

Oh, Y.-S., Kim, J.-S., and Lee, K.-S. (2013). Orthostatic and supine blood pressures are associated with white matter hyperintensities in Parkinson disease. J. Mov. Disord. *6*, 23–27.

Ong, K.H., Ramachandram, D., Mandava, R., and Shuaib, I.L. (2012). Automatic white matter lesion segmentation using an adaptive outlier detection method. Magn. Reson. Imaging *30*, 807–823.

Pantoni, L. (2010). Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol. *9*, 689–701.

Pantoni, L., and Garcia, J.H. (1997). Pathogenesis of leukoaraiosis. Stroke *28*, 652–659.

Pantoni, L., Sarti, C., Alafuzoff, I., Jellinger, K., Munoz, D.G., Ogata, J., and Palumbo, V. (2006). Postmortem Examination of Vascular Lesions in Cognitive Impairment. Stroke *37*, 1005–1009.

Pedersen, K.F., Larsen, J.P., Tysnes, O.-B., and Alves, G. (2017). Natural course of mild cognitive impairment in Parkinson disease A 5-year population-based study. Neurology 10–1212.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., and Nichols, T.E. (2011). Statistical parametric mapping: the analysis of functional brain images (Academic press).

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., et al. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI) Clinical characterization. Neurology *74*, 201–209.

Poletti, M., Frosini, D., Pagni, C., Baldacci, F., Nicoletti, V., Tognoni, G., Lucetti, C., Del Dotto, P., Ceravolo, R., and Bonuccelli, U. (2012). Mild cognitive impairment and cognitive-motor relationships in newly diagnosed drug-naive patients with Parkinson's disease. J. Neurol.

Neurosurg. Psychiatry jnnp–2011.

Price, C.C., Mitchell, S.M., Brumback, B., Tanner, J.J., Schmalfuss, I., Lamar, M., Giovannetti, T., Heilman, K.M., and Libon, D.J. (2012). MRI-leukoaraiosis thresholds and the phenotypic expression of dementia. Neurology *79*, 734–740.

Prince, M., Albanese, E., and Guerchet, M. (2014). World Alzheimer Report 2014.

Prins, N.D., and Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: an update. Nat. Rev. Neurol. *11*, 157–165.

Prins, N.D., Van Straaten, E.C.W., Van Dijk, E.J., Simoni, M., Van Schijndel, R.A., Vrooman, H.A., Koudstaal, P.J., Scheltens, P., Breteler, M.M.B., and Barkhof, F. (2004a). Measuring progression of cerebral white matter lesions on MRI Visual rating and volumetrics. Neurology *62*, 1533–1539.

Prins, N.D., van Dijk, E.J., den Heijer, T., Vermeer, S.E., Koudstaal, P.J., Oudkerk, M., Hofman, A., and Breteler, M.M. (2004b). Cerebral white matter lesions and the risk of dementia. Arch. Neurol. *61*, 1531–1534.

Provenzano, F.A., Muraskin, J., Tosto, G., Narkhede, A., Wasserman, B.T., Griffith, E.Y., Guzman, V.A., Meier, I.B., Zimmerman, M.E., Brickman, A.M., et al. (2013). White matter hyperintensities and cerebral amyloidosis: necessary and sufficient for clinical expression of Alzheimer disease? JAMA Neurol. *70*, 455–461.

Qiao, J., Jin, G., Lei, L., Wang, L., Du, Y., and Wang, X. (2016). The positive effects of high-frequency right dorsolateral prefrontal cortex repetitive transcranial magnetic stimulation on

memory, correlated with increases in brain metabolites detected by proton magnetic resonance spectroscopy in recently detoxified alcohol-dependent patients. Neuropsychiatr. Dis. Treat. *12*, 2273–2278.

Quddus, A., Fieguth, P., and Basir, O. (2005). Adaboost and Support Vector Machines for White Matter Lesion Segmentation in MR Images. In Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of The, pp. 463–466.

Quinlan, J.R. (1986). Induction of Decision Trees. Mach. Learn. *1*, 81–106.

Ries, M.L., McLaren, D.G., Bendlin, B.B., GuofanXu, Rowley, H.A., Birn, R., Kastman, E.K., Sager, M.A., Asthana, S., and Johnson, S.C. (2012). Medial prefrontal functional connectivity—Relation to memory self-appraisal accuracy in older adults with and without memory disorders. Neuropsychologia *50*, 603–611.

Rydberg, J.N., Hammond, C.A., Grimm, R.C., Erickson, B.J., Jack Jr, C.R., Huston 3rd, J., and Riederer, S.J. (1994). Initial clinical experience in MR imaging of the brain with a fast fluid-attenuated inversion-recovery pulse sequence. Radiology *193*, 173–180.

Rydberg, J.N., Riederer, S.J., Rydberg, C.H., and Jack, C.R. (1995). Contrast optimization of fluid-attenuated inversion recovery (flair) imaging. Magn. Reson. Med. *34*, 868–877.

Sachdev, P., Kalaria, R., O'Brien, J., Skoog, I., Alladi, S., Black, S.E., Blacker, D., Blazer, D., Chen, C., Chui, H., et al. (2014). Diagnostic criteria for vascular cognitive disorders: a VASCOG statement. Alzheimer Dis. Assoc. Disord. *28*, 206.

Sajja, B.R., Datta, S., He, R., Mehta, M., Gupta, R.K., Wolinsky, J.S., and Narayana, P.A.

(2006). Unified approach for multiple sclerosis lesion segmentation on brain MRI. Ann. Biomed. Eng. *34*, 142–151.

Salat, D.H., Tuch, D.S., van der Kouwe, A.J.W., Greve, D.N., Pappu, V., Lee, S.Y., Hevelone, N.D., Zaleta, A.K., Growdon, J.H., Corkin, S., et al. (2010). White matter pathology isolates the hippocampal formation in Alzheimer's disease. Neurobiol. Aging *31*, 244–256.

Sam, K., Crawley, A.P., Conklin, J., Poublanc, J., Sobczyk, O., Mandell, D.M., Venkatraghavan, L., Duffin, J., Fisher, J.A., Black, S.E., et al. (2016). Development of White Matter Hyperintensity Is Preceded by Reduced Cerebrovascular Reactivity. Ann. Neurol. *80*, 277–285.

Sánchez, C.I., Hornero, R., Mayo, A., and García, M. (2009). Mixture model-based clustering and logistic regression for automatic detection of microaneurysms in retinal images. p. 72601M–72601M–8.

Scheltens, P.H., Barkhof, F., Leys, D., Pruvo, J.P., Nauta, J.J.P., Vermersch, P., Steinling, M., and Valk, J. (1993). A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. J. Neurol. Sci. *114*, 7–12.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., et al. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. Neuroimage *59*, 3774–3783.

Schmidt, R., Fazekas, F., Kapeller, P., Schmidt, H., and Hartung, H.-P. (1999). MRI white matter hyperintensities Three-year follow-up of the Austrian Stroke Prevention Study. Neurology *53*, 132–132.

Schrag, A., Siddiqui, U.F., Anastasiou, Z., Weintraub, D., and Schott, J.M. (2017). Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. Lancet Neurol. *16*, 66–75.

Schwarz, C., Fletcher, E., DeCarli, C., and Carmichael, O. (2009). Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR. In Information Processing in Medical Imaging, (Springer), pp. 239–251.

Scott, J.A., Braskie, M.N., Tosun, D., Thompson, P.M., Weiner, M., DeCarli, C., and Carmichael, O.T. (2014). Cerebral Amyloid and Hypertension are Independently Associated with White Matter Lesions in Elderly. Front. Aging Neurosci. *7*, 221–221.

Selkoe, D.J. (2005). Defining molecular targets to prevent Alzheimer disease. Arch. Neurol. *62*, 192–195.

Seo, S.W., Lee, J.-M., Im, K., Park, J.-S., Kim, S.-H., Kim, S.T., Ahn, H.-J., Chin, J., Cheong, H.-K., Weiner, M.W., et al. (2012). Cortical thinning related to periventricular and deep white matter hyperintensities. Neurobiol. Aging *33*, 1156–1167.e1.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., and Pham, D.L. (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage *49*, 1524–1535.

Silbert, L.C., Nelson, C., Howieson, D.B., Moore, M.M., and Kaye, J.A. (2008). Impact of white matter hyperintensity volume progression on rate of cognitive and motor decline. Neurology *71*, 108–113.

Simões, R., Mönninghoff, C., Dlugaj, M., Weimar, C., Wanke, I., van Walsum, A.-M. van C., and Slump, C. (2013). Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. Magn. Reson. Imaging *31*, 1182–1189.

Sled, J.G., Zijdenbos, A.P., and Evans, A.C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. Med. Imaging IEEE Trans. On *17*, 87–97.

Smart, S.D., Firbank, M.J., and O'Brien, J.T. (2011). Validation of Automated White Matter Hyperintensity Segmentation. J. Aging Res. *2011*, e391783.

Smith, S.M. (2002). Fast robust automated brain extraction. Hum. Brain Mapp. *17*, 143–155.

Smith, E.E., Egorova, S., Blacker, D., Killiany, R.J., Muzikansky, A., Dickerson, B.C., Tanzi, R.E., Albert, M.S., Greenberg, S.M., and Guttmann, C.R.G. (2008). Magnetic Resonance Imaging White Matter Hyperintensities and Brain Volume in the Prediction of Mild Cognitive Impairment and Dementia. Arch. Neurol. *65*, 94–100.

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., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage *23*, S208–S219.

Snyder, H.M., Corriveau, R.A., Craft, S., Faber, J.E., Greenberg, S.M., Knopman, D., Lamb, B.T., Montine, T.J., Nedergaard, M., Schaffer, C.B., et al. (2015). Vascular contributions to cognitive impairment and dementia including Alzheimer's disease. Alzheimers Dement. *11*, 710–717.

Sperling, R.A., Jack, C.R., and Aisen, P.S. (2011a). Testing the right target and right drug at the right stage. Sci. Transl. Med. *3*, 111cm33–111cm33.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., et al. (2011b). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. *7*, 280–292.

Steenwijk, M.D., Pouwels, P.J., Daams, M., van Dalen, J.W., Caan, M.W., Richard, E., Barkhof, F., and Vrenken, H. (2013). Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage Clin. *3*, 462–469.

Storandt, M., Mintun, M.A., Head, D., and Morris, J.C. (2009). Cognitive decline and brain volume loss as signatures of cerebral amyloid-β peptide deposition identified with Pittsburgh compound B: cognitive decline associated with Aβ deposition. Arch. Neurol. *66*, 1476–1481.

van Straaten, E.C., Harvey, D., Scheltens, P., Barkhof, F., Petersen, R.C., Thal, L.J., Jack, C.R., DeCarli, C., and others (2008). Periventricular white matter hyperintensities increase the likelihood of progression from amnestic mild cognitive impairment to dementia. J. Neurol. *255*, 1302.

Sunwoo, M.K., Jeon, S., Ham, J.H., Hong, J.Y., Lee, J.E., Lee, J.-M., Sohn, Y.H., and Lee, P.H. (2014). The burden of white matter hyperintensities is a predictor of progressive mild cognitive impairment in patients with Parkinson's disease. Eur. J. Neurol. *21*, 922–e50.

Suzuki H, Kawai H, Hirano H, et al. One-Year Change in the Japanese Version of the Montreal

Cognitive Assessment Performance and Related Predictors in Community-Dwelling Older Adults. J Am Geriatr Soc. 2015;63:1874–1879.

Takagi, T., and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. 116–132.

Takao, M., Koto, A., Tanahashi, N., Fukuuchi, Y., Takagi, M., and Morinaga, S. (1999). Pathologic findings of silent hyperintense white matter lesions on MRI. J. Neurol. Sci. *167*, 127–131.

Tosto, G., Zimmerman, M.E., Carmichael, O.T., and Brickman, A.M. (2014). Predicting aggressive decline in mild cognitive impairment: the importance of white matter hyperintensities. JAMA Neurol. *71*, 872–877.

Tremblay-Mercier, J., Madjar, C., Etienne, P., Poirier, J., and Breitner, J. (2014). A Program of Pre-Symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease (PREVENT-AD): Design, Methods, And Perspectives. Alzheimers Dement. J. Alzheimers Assoc. *10*, P808.

Tuladhar, A.M., Reid, A.T., Shumskaya, E., Laat, K.F. de, Norden, A.G.W. van, Dijk, E.J. van, Norris, D.G., and Leeuw, F.-E. de (2015). Relationship Between White Matter Hyperintensities, Cortical Thickness, and Cognition. Stroke *46*, 425–432.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging *18*, 885–896.

van der Kouwe, A., and Fischl. B. (2015). Anatomical MRI for Human Brain Morphometry. (2015): *3-28.*

Veselỳ, B., and Rektor, I. (2016). The contribution of white matter lesions (WML) to Parkinson's disease cognitive impairment symptoms: A critical review of the literature. Parkinsonism Relat. Disord. *22*, S166–S170.

Vovk, U., Pernus, F., and Likar, B. (2007). A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Med. Imaging *26*, 405–421.

Wack, D.S., Dwyer, M.G., Bergsland, N., Di Perri, C., Ranza, L., Hussein, S., Ramasamy, D., Poloni, G., and Zivadinov, R. (2012). Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates. BMC Med. Imaging *12*, 17.

Wahlund, L.O., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjögren, M., Wallin, A., Ader, H., Leys, D., Pantoni, L., et al. (2001). A New Rating Scale for Age-Related White Matter Changes Applicable to MRI and CT. Stroke *32*, 1318–1322.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., T O'Brien, J., Barkhof, F., Benavente, O.R., et al. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. *12*, 822–838.

Wardlaw, J.M., Hernández, M.C.V., and Muñoz-Maniega, S. (2015). What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. J. Am. Heart Assoc. *4*, e001140.

Wels, M., Huber, M., and Hornegger, J. (2008). Fully automated segmentation of multiple sclerosis lesions in multispectral MRI. Pattern Recognit. Image Anal. *18*, 347–350.

Woolrich, M.W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S.M. (2009). Bayesian analysis of neuroimaging data in FSL. Neuroimage *45*, S173–S186.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., and others (1996). A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. *4*, 58–73.

Wu, M., Rosano, C., Butters, M., Whyte, E., Nable, M., Crooks, R., Meltzer, C.C., Reynolds, C.F., and Aizenstein, H.J. (2006a). A fully automated method for quantifying and localizing white matter hyperintensities on MR images. Psychiatry Res. Neuroimaging *148*, 133–142.

Wu, Y., Warfield, S.K., Tan, I.L., Wells, W.M., Meier, D.S., van Schijndel, R.A., Barkhof, F., and Guttmann, C.R. (2006b). Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. NeuroImage *32*, 1205–1215.

Yamada, M. (2012). Predicting cerebral amyloid angiopathy-related intracerebral hemorrhages and other cerebrovascular disorders in Alzheimer's disease. Front. Neurol. *3*, 64.

Yoo, B.I., Lee, J.J., Han, J.W., Oh, S.Y.W., Lee, E.Y., MacFall, J.R., Payne, M.E., Kim, T.H., Kim, J.H., and Kim, K.W. (2014). Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. Neuroradiology *56*, 265–281.

Yoshita, M., Fletcher, E., and DeCarli, C. (2005). Current concepts of analysis of cerebral white matter hyperintensities on magnetic resonance imaging. Top. Magn. Reson. Imaging TMRI *16*, 399.

Zeighami, Y., Ulla, M., Iturria-Medina, Y., Dadar, M., Zhang, Y., Larcher, K.M.-H., Fonov, V., Evans, A.C., Collins, D.L., and Dagher, A. (2015). Network structure of brain atrophy in de novo Parkinson's disease. ELife *4*, e08440.

Zhang, Y., Brady, M., and 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.P., Forghani, R., and Evans, A.C. (2002). Automatic pipeline analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. Med. Imaging IEEE Trans. On *21*, 1280–1291.